Pontificia Universidad
JAVERIANA
Cali

MÉTODOS BAYESIANOS Y FRECUENTISTAS FLEXIBLES PARA LA CARACTERIZACIÓN Y ANÁLISIS PREDICTIVO Y PRESCRIPTIVO DE CADENAS DE VALOR

Jesús Andrés Burbano Gómez

Proyecto Aplicado para optar al título de

Magister en Ciencia de Datos

Directora

Isabel Cristina García Arboleda

FACULTAD DE INGENIERÍA Y CIENCIAS

MAESTRÍA EN CIENCIA DE DATOS

SANTIAGO DE CALI, JUNIO 8 DE 2025

FICHA RESUMEN

TÍTULO DEL PROYECTO: MÉTODOS BAYESIANOS Y FRECUENTISTAS FLEXIBLES PARA LA CARACTERIZACIÓN Y ANÁLISIS PREDICTIVO Y PRESCRIPTIVO DE CADENAS DE VALOR

1. ÁREA DE TRABAJO: Economía organizacional, análisis de mercado laboral.
2. TIPO DE PROYECTO: Aplicado.
3. ESTUDIANTE(S): Jesús Andrés Burbano Gómez
4. CORREO ELECTRÓNICO: jesus.burbano@javerianacali.edu.co
5. DIRECCIÓN Y TELEFONO: Calle 18N #4-40 Urbanización Pomona, Popayán, Cauca.
6. DIRECTOR: PhD. Isabel Cristina García Arboleda.
7. VINCULACIÓN DEL DIRECTOR: Docente del Departamento de Ciencias Naturales y Matemáticas – Pontificia Universidad Javeriana Sede Cali
8. CORREO ELECTRÓNICO DEL DIRECTOR: isabel.garcia@javerianacali.edu.co
9. GRUPO O EMPRESA QUE LO AVALA: Caja de Compensación Familiar del Cauca – Comfacauca.
10. PALABRAS CLAVE (al menos 5): Cadenas de valor, aprendizaje supervisado, mercado laboral, estadística bayesiana, estadística frecuentista.
11. FECHA DE INICIO: 1 de mayo de 2024.
12. FECHA DE FINALIZACIÓN: 30 de junio de 2025.
13. RESUMEN:

La implementación de metodologías de Ciencia de Datos para la caracterización y análisis predictivo y/o prescriptivo de las actividades organizacionales toma fuerza como un instrumento predilecto para la minimización de la incertidumbre y el incremento de la productividad en cada vez más sectores de las economías y la sociedad en general. De tal forma que la ausencia de este tipo de herramientas analíticas puede implicar una limitación importante para el desarrollo de un rango de acción eficiente y planeación efectiva de las organizaciones. Tal es el caso de la Caja de Compensación Familiar del Cauca – Comfacauca, la cual, como figura de administración de recursos parafiscales en el marco del Sistema de Seguridad Social colombiano para el departamento del Cauca, posee un aparato operativo y de caracterización socioeconómica de alcance poblacional que le ha permitido consolidarse como una institución ejemplar a nivel nacional en su medio normativo. Sin embargo, la ausencia de instrumentos de caracterización estadística y optimización para la toma de decisiones al interior de la organización, representa una debilidad que puede comprometer el cumplimiento de sus objetivos estratégicos y metas de cobertura en el

mediano y largo plazo. Para resolver este problema, este proyecto aplicado propone tomar como referencia la conceptualización organizacional de cadenas de valor y el contraste entre las metodologías de modelación bayesiana y frecuentista, con el fin de desarrollar un protocolo para la caracterización y análisis de operaciones de la organización en el marco de su Agencia de Empleo y Mecanismo de Protección al Cesante (MPC), el cual pueda servir como insumo para la toma de decisiones óptimas fundamentadas en criterios cuantitativos rigurosos, de cara al incremento de la eficiencia multidimensional y, por ende, del bienestar socioeconómico de la población caucana. En general, esta propuesta sugiere aplicaciones de Ciencia de Datos desde el planteamiento de un problema de clasificación en el contexto de los mecanismos de matching entre oferta y demanda del mercado laboral captado por la organización desde su Agencia. Además, por su esencia metodológica, también propone aplicaciones en las áreas de economía organizacional, construcción de protocolos integrados para la evaluación y monitoreo de vacantes laborales, y el estudio teórico y empírico de cadenas de valor para servicios y productos intangibles.

TABLA DE CONTENIDO

LISTA DE ANEXOS

1. Licencia de autorización para la publicación de la obra en Vitela, repositorio institucional de la Pontificia Universidad Javeriana Cali.
2. Carta de presentación para evaluación del proyecto aplicado.
3. Acuerdo de colaboración para la autorización de uso de datos por parte de la Caja de Compensación Familiar del Cauca.
4. Informe final de resultados para la organización Comfacauca.
5. Protocolo analítico implementado para la ejecución del proyecto (Excel).
6. Archivo máster de candidatos ordenados por probabilidades estimadas de colocación (Excel).

Los anexos 5 y 6 en esta lista se almacenan en: https://osf.io/87xcg/files/osfstorage.

LISTA DE TABLAS

LISTA DE FIGURAS

INTRODUCCIÓN

La Caja de Compensación Familiar del Cauca – Comfacauca es la organización encargada de la administración de fondos de aportes parafiscales para el departamento del Cauca en el marco del Sistema de Seguridad Social Colombiano, a través de la prestación de servicios sociales y el otorgamiento de subsidios de distinta índole a trabajadores afiliados y sus grupos familiares, además de la asignación de recursos específicos entre la población caucana para la protección social en vivienda, atención integral a la primera infancia, jornada escolar complementaria, fomento al empleo y protección al cesante [1].

A pesar de poseer una estructura organizacional sólida, una infraestructura operacional de alcance departamental para el cumplimiento de su propósito misional, y, consecuentemente, una amplia base de datos de caracterización socioeconómica, Comfacauca carece de herramientas para la caracterización estadística de sus actividades de cara al empleo de dinámicas de análisis predictivo y/o prescriptivo para la toma de decisiones.

Si bien se reconoce la consolidación de Comfacauca desde una gestión ejemplar de su fuero normativo como Caja de Compensación Familiar, la ausencia de herramientas para la caracterización y optimización estadística de sus operaciones significa una limitación importante para el incremento de su eficiencia técnica, económica y financiera en el mediano y largo plazo, lo cual puede incluso comprometer el cumplimiento de metas específicas enmarcadas en sus objetivos estratégicos y Visión para 2025. En este sentido, se justifica la pertinencia del desarrollo de protocolos específicos para la medición estadística y análisis predictivo y/o prescriptivo en el marco de las operaciones de la organización.

Este proyecto aplicado aborda esta labor desde la práctica de la Ciencia de Datos, a través de la conceptualización organizacional de cadenas de valor y el contraste entre las metodologías de modelación bayesiana y frecuentista, con el fin de establecer una rutina analítica sobre las operaciones asociadas a la cadena de valor de la Agencia de Empleo y Mecanismo de Protección al Cesante (MPC) de la organización, y proveer como resultado un informe puntual de caracterización y recomendaciones para el desarrollo de escenarios óptimos, las cuales sirvan como insumo de toma de decisiones en busca de mejoras significativas en la cobertura y eficiencia de la Agencia, y, por ende, en una proporción significativa del mercado laboral del departamento del Cauca.

El capítulo inicial de este documento provee la definición del problema a abordar desde su planteamiento y formulación, desglosando la solución planteada desde el objetivo general y objetivos específicos de la investigación, su alcance, justificación y marco de referencia. El capítulo 2 aborda la contextualización del caso de estudio a la luz del concepto de cadena de valor de la Agencia de Empleo, y la estructura de datos a emplear. El capítulo 3 aborda

los pormenores de las herramientas de modelación contrastadas, para así proveer una justificación concreta de los modelos de clasificación más eficientes para los propósitos de este trabajo en el capítulo 4. Los resultados y su discusión en términos de una caracterización del mercado laboral captado por la Agencia se presentan en el capítulo 5, finalizando con conclusiones y el relacionamiento de trabajos futuros potenciales en el capítulo 6.

1. CONTEXTUALIZACIÓN DEL PROYECTO

1.1 DEFINICIÓN DEL PROBLEMA

1.1.1 PLANTEAMIENTO DEL PROBLEMA

En el contexto de Comfacauca como caso de referencia, el problema de investigación consiste en la ausencia de un protocolo flexible para la caracterización estadística de la relación entre las estructuras de datos asociadas a las operaciones de la organización, y las cadenas de valor definidas para la provisión de servicios en el marco de sus funciones como Caja de Compensación en el departamento del Cauca, descritas por los manuales de procesos y procedimientos existentes para el manejo de los Fondos a su cargo.

Comfacauca ha venido desempeñando su labor de manera exitosa, siendo encargada de la administración de fondos asociados a la prestación de servicios sociales y el otorgamiento de subsidios a trabajadores afiliados y sus grupos familiares, además de la asignación de recursos específicos entre la población caucana para la protección social en vivienda, atención integral a la primera infancia, jornada escolar complementaria, fomento al empleo y protección al cesante. Siendo una organización consolidada en su funcionamiento y alcance a nivel departamental durante 57 años, Comfacauca cuenta con una estructura organizacional amplia e infraestructura adecuada para el desarrollo y registro de sus operaciones, lo cual le ha permitido aumentar progresivamente su cobertura en el marco de su misión como Caja de Compensación del Sistema Colombiano de Seguridad Social, existiendo un incremento del 9% en la provisión de subsidios con respecto a 2022 [1].

Sin embargo, es de resaltarse que, a pesar de poseer una estructura operativa sólida y bases de datos de alcance poblacional en el Departamento del Cauca, la organización no cuenta con un protocolo analítico para la medición de sus operaciones y modelación estadística de escenarios para la toma de decisiones en función de éstas. Esto representa una limitación importante al potencial de acción de la organización en el departamento, en tanto no posee una herramienta cuantitativa de análisis para la planeación de operaciones que permita reconocer patrones y tendencias de cara al desarrollo de estrategias bajo un riesgo estadístico fiable.

La causa principal de esta situación puede hallarse en la naturaleza monopólica de la organización, pues es la única figura jurídica autorizada para la provisión de subsidios y servicios provenientes de recursos parafiscales en el departamento. La ausencia de "competencia" en su entorno, así como la predisposición normativa a obtener un alcance poblacional en sus operaciones, implican la ausencia previa de incentivos para el desarrollo

de tecnologías y protocolos analíticos que incrementen orgánicamente la demanda por sus servicios o la eficiencia organizacional.

Sin embargo, el hecho de no poseer competidores en su misión no es sinónimo de poseer una cobertura plena en el Departamento o de alcanzar la máxima eficiencia en su estructura organizacional, aspectos clave a mejorar que se reconocen en los objetivos estratégicos de la organización, y, particularmente, en su visión para 2025, la cual propone alcanzar una cobertura con al menos un servicio en el 80% de los afiliados y el 40% de la población del Cauca [1].

Dado el contexto de dispersión del tejido social en el departamento del Cauca y la consecuente vulnerabilidad de una proporción significativa de su población, es notorio el hecho de que el carácter monopólico de la organización no es suficiente para la satisfacción de la meta planteada para el 2025, haciéndose necesario el empleo de recursos adicionales asociados al uso estructurado de información, con el fin de segmentar la distribución de servicios y subsidios de manera más eficiente y con un mayor alcance.

De esta manera, es posible incluso afirmar que la ausencia de protocolos de análisis predictivo y/o prescriptivo en el marco de las dinámicas operacionales de la organización puede representar un impedimento para la progresión de su eficiencia en el futuro, y para el cumplimiento de sus objetivos en el mediano y largo plazo.

1.1.2 FORMULACIÓN DEL PROBLEMA

La necesidad de un protocolo para la medición y modelación estadística en función de las operaciones de la organización se expresó con claridad en el planteamiento del problema que evoca este proyecto aplicado. Ahora, en cuanto a la estructura del problema y posterior proceder con respecto a éste, la siguiente pregunta de investigación retoma el concepto transversal de cadena de valor [2] para determinar su formulación, restringiéndose al caso de la Ruta de Atención de la Agencia de Empleo de Comfacauca:

¿Cómo establecer un protocolo para la optimización estadística del desempeño de la cadena de valor asociada a la Ruta de Atención de la Agencia de Empleo de la Caja de Compensación Familiar del Cauca?

La pregunta fundamental de investigación formulada determina la orientación principal del proyecto aplicado a realizar. Además, el problema formulado implica una especificación en función de preguntas de sistematización.

Inicialmente, es necesario considerar en un sentido descriptivo a la cadena de valor seleccionada para un análisis de la dimensión de este proyecto aplicado, así como su

comportamiento. Para este propósito, la pregunta de sistematización formulada toma una perspectiva técnica del concepto de cadenas de valor analizada en [2]:

¿Cómo identificar y describir el comportamiento de la cadena de valor de la Ruta de Atención de la Agencia de Empleo, por sub actividades sucesivas interconectadas hacia la provisión de bienes o servicios finales?

La siguiente pregunta de sistematización indaga sobre la selección de un modelo(s) estadístico(s) idóneo(s) para los propósitos de análisis predictivo/prescriptivo de este proyecto aplicado:

¿Cuáles son los mejores modelos a emplear para el protocolo de análisis predictivo/prescriptivo de la cadena de valor evaluada?

Además, es necesario considerar el conjunto de criterios técnicos, económicos y contextuales para la determinación del modelo(s) idóneo(s) de análisis predictivo de la cadena de valor:

¿Qué modelo presenta los mejores resultados considerando los criterios de eficiencia técnica, estadística y económica planteados bajo el contexto de la organización?

Finalmente, los resultados del ejercicio práctico a realizar en el marco del proyecto aplicado implican una comunicación efectiva hacia las partes interesadas en este problema de negocio, lo cual evoca la última pregunta de sistematización de este problema:

¿Cómo comunicar efectivamente los resultados del proyecto y sus implicaciones potenciales a las partes interesadas de la organización?

1.2 OBJETIVOS DEL PROYECTO

Como solución al problema planteado, la implementación de un protocolo analítico que fundamente la toma de decisiones en hechos estadísticos, puede significar una herramienta clave para la proliferación de mejoras en la eficiencia multidimensional de la Agencia de Empleo, y de Comfacauca en general, convirtiéndose en un instrumento indispensable para su funcionamiento en el largo plazo. El desarrollo de tal protocolo puede proveer un marco estandarizado para la medición de desempeño de cadenas de valor, y posterior mejoramiento de la cobertura, eficiencia (técnica, financiera y organizacional) y planeación de las actividades de la organización, a través del análisis de escenarios óptimos provisto por métodos formales en el marco de la Ciencia de Datos.

En términos puntuales, el proyecto aplicado propuesto abordará el problema formulado a través de la caracterización de la cadena de valor de la Agencia de empleo desde la identificación de las estructuras de datos clave en las actividades encadenadas de la Ruta de Atención, y el contraste de métodos bayesianos y frecuentistas de aprendizaje automático, en función de criterios de eficiencia técnica, computacional y restricciones normativas de la organización. Nótese que las respuestas concretas a la pregunta central de investigación y sus preguntas auxiliares de investigación, se plasman respectivamente en el objetivo general del proyecto y sus objetivos específicos.

1.2.1 OBJETIVO GENERAL

Diseñar y aplicar un protocolo para la optimización estadística del desempeño de la cadena de valor de la Ruta de Atención de la Agencia de Empleo de la Caja de Compensación Familiar del Cauca - Comfacauca.

1.2.2 OBJETIVOS ESPECÍFICOS

- Caracterizar la cadena de valor asociada a las actividades de la Ruta de Atención de la Agencia de Empleo de Comfacauca, desde el registro de los candidatos al mercado laboral, hasta la evaluación de sus resultados con respecto a las vacantes emitidas por la demanda laboral del Departamento del Cauca.
- Modelar el comportamiento estadístico de la cadena de valor de la Ruta de Atención por medio del contraste de eficiencia entre métodos bayesianos y frecuentistas de aprendizaje supervisado.

- Evaluar la eficiencia de los modelos seleccionados bajo criterios técnicos, económicos y asociados al contexto de la organización.
- Proveer un informe de resultados con una descripción directa y comprensible de sus impactos potenciales en materia socioeconómica y de eficiencia (técnica, financiera y organizacional).

1.2.3 ALCANCE

El proyecto aplicado a desarrollar resume su alcance de manera sucesiva en función del siguiente conjunto de etapas de la investigación:

- Planteamiento y contextualización de la cadena de valor analizada y sus problemas de negocio.
- Caracterización y análisis exploratorio de variables relevantes en la cadena de valor de la Ruta de Atención de la Agencia de Empleo.
- Selección e implementación de modelos estadísticos.
- Contraste y validación de modelos ejecutados.
- Determinación de resultados finales y desarrollo de informe entregable con conclusiones y recomendaciones estratégicas para la organización.

En este sentido, el alcance del proyecto aplicado propuesto va desde el planteamiento del problema de negocio alrededor de la cadena de valor de la Ruta de Atención de la Agencia de Empleo de Comfacauca, hasta la entrega de recomendaciones en función del análisis cuantitativo realizado desde las estructuras de modelación estadística empleadas.

Es necesario reconocer dos restricciones clave del caso analizado:

- El alcance de los patrones a analizar se limita al departamento del Cauca y sus subdivisiones, y a la información relevante disponible en las bases de datos de la Agencia.
- El grado de acceso a información sobre los afiliados será regido por las políticas de protección de datos de la organización, y se hará de forma completamente anonimizada.

Es pertinente enfatizar que la implementación práctica por parte de la organización de las recomendaciones generadas en función de esta investigación no se halla dentro del alcance de este proyecto aplicado. Así, una evaluación de impactos post-proyecto se presenta como una sofisticación potencial al trabajo propuesto, tanto para la organización y su dinámica institucional en términos prácticos y de eficiencia, como para el abordaje de implicaciones interesantes desde la academia.

1.2.4 JUSTIFICACIÓN

La justificación esencial para el desarrollo de este proyecto aplicado se basa en la propiedad básica que los métodos de Ciencia de Datos otorgan a las organizaciones en la industria, y, en general, en la sociedad. Ésta propiedad es la capacidad de formular instrumentos específicos para la cuantificación y monitoreo de las actividades y operaciones de manera estructurada, permitiendo el incremento de la eficiencia multidimensional desde la reducción de la incertidumbre y del riesgo en la ejecución de recursos valiosos.

La idea anterior se encuentra perfectamente alineada con el planteamiento y formulación del problema desarrollados anteriormente, pues se reconoce plenamente que la ausencia de un protocolo analítico próximo a la Ciencia de Datos como disciplina en Comfacauca implica la pérdida de los beneficios que puede traer un instrumento para la reducción de la incertidumbre de cara a la mejora en el desempeño de sus actividades misionales, particularmente considerando las metas cuantitativas enmarcadas en la visión de la organización.

La idoneidad del proyecto no sólo se evidencia desde su utilidad potencial como herramienta de la organización, también se justifica desde la disponibilidad de la información necesaria para llevarlo a cabo correctamente, pues el alcance departamental de Comfacauca implica que tanto sus operaciones como sus bases de datos de afiliados se hallan inmersas en la actividad socioeconómica, productiva y laboral de una porción mayoritaria de la población del Cauca. Lo anterior hace del proyecto aplicado una propuesta viable, y, además, sujeta a sofisticaciones y continuaciones que pueden arrojar conclusiones y otros instrumentos que maximicen el beneficio de tal base de datos de corte socioeconómico en el departamento.

Finalmente, la implementación de un protocolo en el sentido del proyecto aplicado propuesto no sólo puede significar una mejora global en la eficiencia de las operaciones de la Agencia de Empleo, sino que, dada la misión y naturaleza jurídica de la organización, tal mejoramiento generalizado en la eficiencia de Comfacauca se hallaría directamente traducido en impactos socioeconómicos positivos para la región, implicando beneficios potenciales en rubros clave de la economía departamental, específicamente, en el mercado laboral.

1.3 MARCO DE REFERENCIA

1.3.1 MARCO CONTEXTUAL

Con sede principal ubicada en Popayán, Cauca, Comfacauca es la organización asignada como Caja de Compensación Familiar para el departamento del Cauca, regida por la Ley 21 de 1982, la cual dicta las disposiciones para la administración de fondos de subsidio familiar y demás rubros correspondientes a los recursos provenientes de aportes parafiscales en el marco del Sistema de Seguridad Social Colombiano.

Con poco más de 57 años de operaciones, la organización viene desempeñando su rol de administradora de fondos de aportes parafiscales a través de la prestación de servicios sociales y el otorgamiento de subsidios a trabajadores afiliados y sus grupos familiares, además de la asignación de recursos específicos entre la población caucana para la protección social en vivienda, atención integral a la primera infancia, jornada escolar complementaria, fomento al empleo y protección al cesante; reportando para el año 2023 un total de ingresos por aportes parafiscales de $147.646 millones por parte de 8.936 empleadores afiliados, los cuales fueron administrados para la provisión de servicios y subsidios entre 313.665 afiliados entre trabajadores, cónyuges y sus personas a cargo.

A continuación, se presenta un resumen de las operaciones de los rubros principales de la organización sobre los fondos a su cargo para el año 2023 según [1], con el fin de ilustrar la pluralidad de servicios, y, por ende, de cadenas de valor estructuradas por Comfacauca:

- En general, la organización realizó entrega de 968 mil cuotas monetarias, equivalentes a $20.483 millones para subsidiar la prestación de servicios en educación, recreación, capacitación y crédito, así como $10.948 millones de pesos destinados a la entrega de subsidios en especie.
- En el marco del Fondo de Vivienda de Interés social – FOVIS la organización asignó 247 subsidios de vivienda, equivalentes a un valor de $8.328 millones al 92% de los trabajadores con salarios por debajo de los 2 SMMLV.
- En el marco de su Mecanismo de Protección al Cesante – MPC, se registraron desembolsos de subsidios de desempleo por $14.069 millones a 3.738 beneficiarios, además, se registró la capacitación de 33.372 trabajadores activos de manera completamente subsidiada. Adicionalmente, la Agencia de Empleo de la organización registró 1.117 empresas que inscribieron 6.032 vacantes, ocupadas por 4.824 oferentes.
- En cuanto a servicios complementarios en educación y cultura, la organización financia o administra educación de educación media y superior en Popayán y el norte del Cauca.

- En materia de sus programas de Atención Integral a la Primera Infancia y Jornada Escolar Complementaria, se beneficiaron 16.846 jóvenes de 34 municipios del departamento del Cauca, siendo esto un referente regional y nacional en materia de atención institucional a población vulnerable.
- Finalmente, en materia de recreación, deportes y turismo, la organización registró 1.283.968 usos en sus 6 centros recreativos y deportivos, ubicados en Popayán, Puerto Tejada y Santander de Quilichao.

Como puede apreciarse, el grueso de operaciones y cadenas de valor para la provisión de bienes y servicios específicos por parte de la organización se ve descrito por una ramificación de múltiples ramas y componentes, decantando en un complejo de servicios de alcance poblacional en el departamento, el cual se halla estrechamente vinculado con información de base del mercado laboral y sus características socioeconómicas.

Dado el problema planteado en secciones anteriores, el escenario que propone Comfacauca se muestra como propicio para el uso de métodos de Ciencia de Datos con orientaciones predictivas y/o prescriptivas, este proyecto provee un ejercicio piloto para la implementación de estos métodos en la cadena de valor correspondiente a la Ruta de Atención de la Agencia de Empleo de la organización.

Como subdivisión encargada de la administración del Mecanismo de Protección al Cesante, la Agencia de Empleo de Comfacauca basa sus funciones en el despliegue de un conjunto de subprocesos secuenciales establecidos con el fin de subsidiar a la población desempleada del departamento desde un conjunto de actividades que, principalmente, decantan en el facilitar la consecución de nuevos empleos en función de vacantes disponibles en el mercado laboral captado por la organización, o en la gestión y rastreo de las etapas de desempleo de los candidatos registrados a través de capacitaciones y/o subsidios monetarios. Este conjunto de subprocesos se estructura en una Ruta de Atención, la cual determina la cadena de valor puntual a analizar en este proyecto aplicado.

Dado lo anterior, y como se expondrá más adelante en la descripción de la cadena de valor que fundamenta esta investigación, es crítico resaltar que el eje de la Ruta de Atención de la Agencia de Empleo gira en torno a la determinación de los candidatos al mercado laboral en el departamento como colocados, es decir, como receptores de nuevos empleos vacantes, o como no colocados, es decir, como población que aún permanece desempleada.

De esta manera, se tiene que los resultados esperados de esta investigación y sus implicaciones se basan específicamente en la aplicación de métodos de Ciencia de Datos para obtener un marco estadístico sobre el cuál predecir de manera confiable la posibilidad de los candidatos de obtener un empleo dada la estructura de la demanda laboral del departamento a partir de los atributos inmersos en las bases de datos de la Agencia, y, dada la construcción de este marco estadístico, caracterizar cuantitativamente el panorama del

mercado laboral Caucano para generar recomendaciones, tanto para el proceso establecido por la organización para la Ruta de Atención de la Agencia en el sentido de una cadena de valor, como para el análisis normativo del mercado laboral del departamento y sus implicaciones socioeconómicas. Así, es pertinente resaltar que el núcleo de esta investigación, dado su marco contextual, consiste en un problema de clasificación de los candidatos al mercado laboral en función de la información disponible con respecto a los mismos en las bases de datos de la organización.

1.3.2 MARCO TEÓRICO

La conceptualización fundamental para el desarrollo de este proyecto parte de la definición más básica de cadena de valor. Enunciada inicialmente por Michael Porter, se entiende como cadena de valor a aquel proceso secuencial estructurado por las organizaciones para la consecución de bienes y/o servicios en función de un conjunto de insumos de entrada, de tal forma que el proceso de producción puede comprenderse como una especie de sistema compuesto por subsistemas encadenados de acuerdo a la naturaleza del sector en que se desenvuelven [2]. Nótese además que esta definición acapara la idea de cadenas de valor de forma independiente al tamaño del mercado analizado o el sector en que se desenvuelve, por lo cual se trata de un espectro teórico de amplitud transversal.

Como concepto práctico de amplio uso, las cadenas de valor se dirigen a la localización teórica de este trabajo a través de los postulados y líneas de investigación base de la economía organizacional como sub disciplina de la ciencia económica, en la cual se estudian las relaciones entre los agentes, factores productivos y tecnologías que conforman las cadenas de valor a través de ideas como sueldos, incentivos, costos de transacción, restricciones normativas y diseños institucionales de los entornos organizacionales [3].

Como puede inferirse a priori, los conceptos mencionados anteriormente se asemejan de manera más precisa al contexto descrito por el caso de Comfacauca a nivel organizacional y operacional como un componente clave del sistema de seguridad social del departamento del Cauca, sin embargo, para fundamentar teóricamente las externalidades o efectos del rol de una organización como la analizada en la sociedad, es necesario traer a colación la sub disciplina de evaluación de impactos socioeconómicos, la cual desarrolla un conjunto de protocolos e instrumentos econométricos con el fin de otorgar medidas confiables de los impactos (positivos y negativos) de políticas y organizaciones en las variables socioeconómicas de sectores y contextos de distintas escalas [4] [5].

Ahora, el contexto descrito y los objetivos planteados anteriormente otorgan un sentido específico a la orientación metodológica de los instrumentos de modelación tentativos para

el desarrollo de este proyecto, específicamente, la consecución de los propósitos de esta investigación se basa en el planteamiento y resolución de un problema de clasificación de candidatos en el mercado laboral caucano captado por la Agencia de Empleo de Comfacauca, en función de la exploración de los dos paradigmas fundamentales para la modelación y el manejo de la incertidumbre en Ciencia de Datos: la orientación frecuentista y la orientación bayesiana. A continuación, se presenta una breve descripción de la estructura teórica de cada uno de los instrumentos de modelación a evaluar para ambos paradigmas.

CLASIFICADORES FRECUENTISTAS

La perspectiva frecuentista, escuela del análisis estadístico de mayor uso dentro de las disciplinas asociadas al uso de datos, comprende la incertidumbre como un producto de la medida de la frecuencia de eventos en función de la relación entre muestras y poblaciones, por lo cual trae a colación conceptos clave que resultan familiares para el practicante de la Ciencia de Datos, especialmente la conceptualización de construcción de estimadores estadísticos y su inferencia en función de paradigmas distribucionales, por lo cual a esta escuela también se le conoce como "perspectiva objetiva" de la estadística.

El modelo de mayor uso y generalidad en el framework frecuentista es sin duda el desarrollo de estimadores de máxima verosimilitud [4]. En este modelo se considera una muestra de datos de dimensión arbitraria $y = (y_1, \dots, y_n)$, regida por la definición de un vector de parámetros de dimensión arbitraria $\theta = [\theta_1, \dots, \theta_n]$, tal que su distribución de probabilidad pertenece a una familia paramétrica $\{f(.;\theta) | \, \theta \in \Theta\}$ con Θ siendo su espacio de parámetros. La densidad conjunta de la distribución relacionada implica la existencia de la función de verosimilitud del conjunto de datos en función del espacio de parámetros:

$$L_n(\theta) = f_n(y;\theta) \qquad (1).$$

La función de verosimilitud en (1) determina la probabilidad conjunta de aparición del conjunto de datos y, sujeta a θ como elemento del conjunto de parámetros en función de la familia paramétrica escogida para la modelación. De esta manera, el estimador de máxima verosimilitud consiste en la determinación del parámetro que maximice (1), emergiendo la siguiente estructura fundamental de optimización:

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmax}} \; L_n(\theta; y) \qquad (2).$$

$\hat{\theta}$ se conoce como el estimador de máxima verosimilitud, en tanto el proceso de optimización del cual se obtiene es análogo a la selección del parámetro de modelación que maximice el ajuste de los datos provistos a las predicciones de su modelo relacionado. Nótese que la estimación de parámetros por máxima verosimilitud es también una generalización del conocido método por Mínimos Cuadrados Ordinarios (OLS),

ampliamente utilizado en modelos de Ciencia de Datos, lo cual refuerza su calidad de modelo estándar en el paradigma frecuentista.

Así, en general, se tiene que los modelos de clasificación de mayor prominencia en Ciencia de Datos se hallan en un espectro entre el cálculo de parámetros basados en frecuencias muestrales, y la inferencia sobre paradigmas distribucionales derivados de o relacionados con la estimación por máxima verosimilitud anteriormente relacionada. A continuación, se describen los modelos frecuentistas evaluados en esta investigación.

Regresión Logística

El algoritmo de clasificación por regresión logística se perfila como el modelo introductorio por excelencia para el practicante de la Ciencia de Datos, en tanto es capaz de otorgar un alto grado de interpretabilidad en sus resultados, y tiene un amplio rango de aplicación preservando precisión en sus predicciones.

En su versión esencial, la clasificación por regresión logística consiste en la determinación óptima de la probabilidad de pertenencia de las observaciones de una muestra a una clase objetivo en función de los atributos presentes en el dataset (para el caso de este proyecto aplicado, la clase de colocados en el mercado laboral), de tal forma que el modelo genera predicciones de clasificación en función de las probabilidades estimadas, a la vez que permite observar la contribución media de cada atributo del dataset a esta probabilidad [6].

Este proceso se realiza por medio de la definición de un modelo lineal de la forma $z = w_1x_1 + w_2x_2 + \cdots + w_nx_n + b$, donde los ponderadores w_i corresponden a los coeficientes de contribución de cada atributo x_i, y b es un término ajustable de sesgo. La expresión para z se transforma por medio de la función sigmoide del modelo logit $\sigma(z)$, de tal forma que la probabilidad de pertenencia de una observación a la instancia positiva de la etiqueta de clase binaria ($y = 1$) está dada por:

$$P(y = 1|X) = \sigma(z) = \frac{1}{1 + e^{-(w^T X + b)}} \qquad (3).$$

Donde $z = w^T X$ es la expresión de la forma lineal z como un producto interior entre el vector de coeficientes w y el vector de atributos X. En general, el modelo optimiza sus salidas de predicción de probabilidad por medio de la minimización de la entropía cruzada sobre la muestra (Log-loss), a través del afinamiento iterativo de los coeficientes de los atributos y el componente de sesgo por medio de descenso de gradiente, permitiendo también la inserción de expresiones de regularización para prevenir el sobreajuste del modelo a los datos de entrenamiento. Finalmente, la decisión de clasificación del algoritmo se basa en un umbral sobre la probabilidad estimada, generalmente del 50%.

Métodos por árboles: Random Forest.

Definido como una extensión por ensamble del algoritmo de Árbol de Decisión, Random Forest se perfila como un método robusto para problemas de predicción y clasificación donde abundan atributos categóricos.

Inicialmente, retoma la metodología de Árbol de Decisión, sobre la cual se establece un proceso sucesivo de evaluación condicional sobre los atributos de un dataset, la cual divide condicionalmente la muestra analizada a partir de un criterio de optimización informacional como el coeficiente de Gini o el cálculo de entropía, hasta llegar a un máximo de profundidad donde cada observación se clasifica en función de la moda entre las observaciones de las 'hojas' del árbol construido.

Random Forest mejora las propiedades de Árbol de Decision al entrenar un conjunto de árboles seleccionando aleatoriamente subconjuntos de la muestra de entrenamiento, y limitando aleatoriamente los atributos sobre los cuáles se dividen las muestras y se obtienen las predicciones de cada árbol, de tal manera que la predicción final de la etiqueta de clase de cada observación, para un problema de clasificación, está dada por la moda de las predicciones del total de árboles entrenados.

Esta extensión por ensamble del método de Árbol de Decisión se destaca por abordar eficientemente problemas de entrenamiento donde las relaciones entre la etiqueta de clase y los atributos del dataset superan la complejidad de modelos lineales como la Regresión Logística, mientras que previenen el sobreajuste a los datos de entrenamiento debido al criterio de clasificación sobre la moda entre distintas muestras aleatorias correspondientes a cada árbol [6].

Si bien el afinamiento de hiperparámetros de Random Forest puede implicar un análisis más extendido que en el caso de otros modelos, en tanto involucra propiedades como la cantidad de árboles a entrenar, la cantidad de atributos por división del árbol, la profundidad de los árboles a entrenar, el mínimo de muestras requeridas para dividir los nodos, y el mínimo de muestras para conformar las hojas; este modelo otorga alta eficiencia para casos de datasets de altas dimensiones y variables categóricas con bastantes instancias. Además, a pesar de que su mecanismo de entrenamiento no es de fácil interpretación para el público objetivo, Random Forest calcula valores de importancia de los atributos en el modelo de clasificación final, equiparando la propiedad análoga existente en Regresión Logística.

Métodos por árboles: Gradient Boosting.

Esta extensión por ensamble del método de Árbol de Decisión consiste en el entrenamiento secuencial a partir de un árbol inicial, cuyas predicciones se corrigen desde sus árboles sucesivos hasta obtener una predicción óptima por medio de descenso de gradiente. Es importante resaltar que lo anterior se da en contraposición al caso de Random Forest, donde cada árbol se entrena de forma paralela e independiente [6].

Para el caso de problemas de clasificación, en general, Gradient Boosting inicia con un conjunto de predicciones desde un árbol de decisión inicial, estructurando el cálculo del logaritmo del odds ratio de las probabilidades de pertenencia a las etiquetas binarias de clase como referencia. Seguido de esto, se calculan valores residuales que miden la diferencia entre las probabildades estimadas por el árbol y las etiquetas de clase reales en el conjunto de entrenamiento, esta versión del descenso de gradiente se emplea para actualizar las predicciones del árbol inicial como una corrección, considerando una tasa de aprendizaje como hiperparámetro del modelo. Tras la conversión de las predicciones a valores puntuales de probabilidad estimada, este proceso se realiza de forma iterativa hasta alcanzar un resultado óptimo.

Gradient boosting suele presentar resultados con mayor eficiencia que Random Forest, sin embargo, es de resaltarse que su estructura presenta mayor sensibilidad a la configuración de hiperparámetros debido a su flexibilidad con respecto a las funciones de pérdida a emplear y su dependencia de una tasa de aprendizaje (en adición a los hiperparámetros anteriormente relacionados para Random Forest). Además, la esencia secuencial del entrenamiento de los árboles implica un mayor costo computacional, y una propensión al sobreajuste. Para propósitos de interpretabilidad, cabe resaltar que este algoritmo también ofrece un cálculo de la importancia de los atributos para la determinación del proceso de clasificación.

CLASIFICADORES BAYESIANOS

En contraposición a la esencia “objetiva” del paradigma frecuentista, la perspectiva bayesiana surge de una comprensión “subjetiva” con respecto al concepto de probabilidad, en tanto ésta se entiende como una medida en función del conocimiento previo del practicante, conocimiento que depende del modelo empleado en el ejercicio, y que puede actualizarse en función de la obtención y estructuración de nueva información del entorno analizado.

En el contexto de los problemas de clasificación, la piedra angular de los modelos bayesianos es el conocido Teorema de Bayes, el cual expresa el cálculo de la probabilidad

posterior de un evento de forma condicionada a la información previa que se posee del mismo, de la siguiente forma:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (4).$$

Donde $P(A|B)$ representa la probabilidad posterior de la clase A dado el atributo (o conjunto de atributos) B, $P(B|A)$ representa la verosimilitud (likelihood) del atributo B, es decir, la probabilidad de ocurrencia de las instancias del (los) atributo(s) B dada la clase A; $P(A)$ representa la probabilidad previa (prior) de ocurrencia de la clase A, y $P(B)$ representa la evidencia con respecto al atributo(s) B.

De esta forma, las derivaciones de modelación del Teorema de Bayes decantan en un conjunto de construcciones que apelan a la noción de actualización en función de evidencia para establecer predicciones para los problemas de clasificación [7]. A continuación, se describen brevemente los modelos de este paradigma para el contraste a realizar.

Naive Bayes

El algoritmo de clasificación Naive Bayes (Bayes ingenuo) toma su nombre de la 'ingenuidad' del supuesto que los fundamenta. En este modelo, se asume que los atributos en el dataset son independientes dada la clase a predecir. Éste supuesto es muy poco realista, en tanto difícilmente se puede hallar un caso donde no se hallen grados de dependencia entre los atributos de un dataset, sin embargo, asumir esta propiedad facilita el cálculo de probabilidades posteriores para la predicción de una clase, en tanto el componente de verosimilitud en la expresión (4) se convierte en un producto, de la siguiente forma:

$$P(X_1, X_2, \dots, X_n|Y) = P(X_1|Y)P(X_2|Y) \dots P(X_n|Y) \quad (5).$$

Donde $X_1, X_2, \dots, X_n$ es el conjunto de atributos y Y es la clase a predecir. Lo anterior decanta en una regla de clasificación simple y de computación conveniente, ya que, apelando al Teorema de Bayes en (4) y al supuesto insertado por (5), la regla de clasificación para cada observación del dataset puede expresarse de la siguiente forma:

$$Y = arg \max_{c_k} P(Y = c_k) \prod_{i=1}^{n} P(X_i | Y = c_k) \quad (6).$$

Así, cada observación se asigna a la clase con la probabilidad posterior más alta, a esta regla se le conoce como la regla de decisión del máximo a posteriori (MAP) [7]. Notoriamente, el algoritmo Naive Bayes otorga resultados de alta precisión a pesar de que el supuesto en el que se basa no se cumple casi nunca en realidad. Además, posee las ventajas que tener un muy bajo costo computacional, adaptarse a datasets de alta dimensionalidad, no implicar un proceso de afinamiento de hiperparámetros, y no requerir volúmenes altos de observaciones para los datos de entrenamiento.

Entre sus principales desventajas destacan un bajo rendimiento cuando se aplica sobre atributos con correlaciones muy altas, y el hecho de que este algoritmo no puede generar una medida de la importancia de los atributos en la predicción, como sí se da para los casos de los clasificadores frecuentistas mencionados anteriormente.

Bayesian Ridge Regression

Entendido como una versión bayesiana de la regresión lineal frecuentista, el algoritmo Bayesian Ridge Regression incorpora la noción de actualización de una estructura informacional en función de nueva evidencia al mecanismo de formas lineales para la predicción de etiquetas de clase. Considérese la estructura general de Ridge Regression desde el paradigma frecuentista, donde las predicciones desde una regresión lineal se basan en la minimización OLS insertando un término de regularización, de la siguiente forma:

$$L(\beta) = \sum_{i=1}^{n} (y_i - X_i\beta)^2 + \lambda||\beta||^2 \quad (7).$$

Donde $L(\beta)$ es la función de pérdida regularizada, X_i son los componentes asociados a los atributos del dataset, y_i es la variable dependiente de la forma lineal, β es el vector de coeficientes de la regresión, y λ es el parámetro de regularización.

El enfoque del algoritmo Bayesian Ridge reemplaza la inferencia frecuentista sobre los coeficientes a estimar, fijando una distribución previa para los mismos. Asumiendo una distribución normal, la siguiente expresión refleja la distribución previa para los coeficientes de regresión lineal:

$$p(\beta|\lambda) = \mathcal{N}(0, \lambda^{-1}I) \quad (8).$$

De esta forma, la verosimilitud para la forma lineal está dada por una distribución normal de la forma $p(y \mid X, \beta, \alpha) = \mathcal{N}(X\beta, \alpha^{-1}I)$, y la distribución posterior de los coeficientes de regresión está dada por la siguiente expresión:

$$p(\beta|X, y, \alpha, \lambda) = \mathcal{N}(\mu, \Sigma) \quad (9).$$

Con α como el inverso de la varianza de la distribución normal, $\mu = (X^T X + \lambda I)^{-1} X^T y$ la media posterior de β, y $\Sigma = (X^T X + \lambda I)^{-1}$ la correspondiente matriz de correlación. Así, en lugar de apelar a la inferencia estadística frecuentista sobre una estimación fija de β, Bayesian Ridge determina sus predicciones desde una distribución posterior para los coeficientes [7].

De cara a problemas de clasificación, las predicciones pueden realizarse fijando umbrales para los valores estimados de y, convirtiendo la salida en una estructura binaria correspondiente a las etiquetas de clase, o extrayendo los valores de probabilidad de la forma lineal empleando la función sigmoide, convirtiendo el algoritmo en un análogo bayesiano de la regresión logística.

Las principales ventajas de este algoritmo son la estimación de incertidumbre que provee la estructura distribucional que otorga a los coeficientes de regresión, su prevención contra el sobreajuste siendo superior a la regresión Ridge frecuentista, y su funcionamiento conveniente para datasets pequeños. Entre sus desventajas resaltan un mayor costo computacional que Naive Bayes, su sensibilidad a los supuestos establecidos para la distribución previa de los coeficientes, y, al igual que Naive Bayes, el hecho de que no puede calcular medidas de importancia de los atributos en las predicciones realizadas.

Si bien la estadística bayesiana no se presenta como la perspectiva de mayor uso, su popularidad viene creciendo entre los practicantes por su flexibilidad metodológica, su capacidad de lidiar con sistemas complejos y cambiantes, y su adaptabilidad a modelos matemáticos y aplicativos computacionales heterogéneos. Contribuciones de la perspectiva bayesiana para el contexto de este proyecto pueden observarse en [8], [9] y [10].

Finalmente, los modelos seleccionados para evaluar el contraste entre ambas perspectivas parten de su adaptación al tipo de estructura de datos a disposición, la cual, como se expondrá más adelante, resalta por la presencia de atributos mayoritariamente categóricos y con una cantidad elevada de instancias. Así, en conjunto con el marco analítico de cadenas de valor, la evaluación de estas perspectivas de modelación determina la orientación para la resolución del problema de investigación planteado.

1.3.3 ANTECEDENTES

En el espectro práctico, el desarrollo de modelos, protocolos y aplicativos para el análisis predictivo y prescriptivo de las operaciones de organizaciones y empresas, puede evidenciarse en distintas contribuciones en el marco de la investigación de operaciones y de las ciencias económicas y de la gestión. Si bien no existe una fijación explícita en la selección de una metodología estadística vinculada a estructuras de cadenas de valor, como lo pretende este proyecto, en las contribuciones expuestas a continuación puede revisarse el uso y contraste de distintos modelos y tecnologías en donde la aplicación de técnicas y procesos de Ciencia de Datos reduce el riesgo estadístico e incrementa la productividad en las operaciones de las organizaciones.

[11] presenta un meta análisis global con respecto al empleo de métodos asociados a Big Data en relación al monitoreo y optimización de las capacidades de las firmas en competencia, considerando 42 estudios en la materia, y hallando como implicaciones principales la relación directamente proporcional entre el grado de tecnificación organizacional del sector analizado y la capacidad de mejora en función de estructuras analíticas de Big Data, y, la fuerte influencia del país en que operan las firmas analizadas de cara a la efectividad de estos métodos y modelos.

La postura de los autores del trabajo mencionado anteriormente puede verse reforzada desde otro punto de vista con la contribución de [12], quienes realizan un análisis cuantitativo considerando 125 organizaciones de 26 países, en busca de una determinación de la composición del valor añadido de las empresas a partir de métodos de Bussiness Analytics (BA). La evidencia hallada respalda que el alto desempeño de las estructuras organizacionales analizadas en el marco de procesos BA parte principalmente de factores sociales, tales como la calidad de los recursos humanos, la capacidad de gestión por parte de la alta gerencia, y una cultura organizacional sana. Nótese que los autores afirman que estos factores sociales predominan en la generación de valor por encima de los aspectos meramente técnicos de las organizaciones.

En cuanto a contribuciones con metodologías específicas y dependientes de sus casos de estudio, se tiene que se emplean diversos modelos de cara a dinámicas de optimización organizacional y sectorial, haciendo uso de herramientas familiares para el practicante de la Ciencia de Datos. Tal es el caso de [13], quien emplea un mecanismo de inferencia causal (causal inference) para contrastar su hipótesis sobre los beneficios de las aglomeraciones económicas de cara a la productividad del sector empresarial en India.

Por otro lado, diversas aplicaciones de Machine Learning pueden apreciarse como mecanismos de optimización de operaciones en distintos sectores, en contribuciones como [14], [15] y [16].

Algunas aplicaciones próximas al paradigma bayesiano, tales como modelos mixtos y esquematizaciones composicionales, pueden observarse en contribuciones asociadas al análisis del desempeño organizacional en función de los incentivos a agentes clave, como es el caso de [17], o [18] y [19] con sus aplicaciones de algoritmos de árboles de decisión para la medición del desempeño de firmas a partir de indicadores financieros.

Finalmente, el uso de modelos de analítica predictiva y prescriptiva en el sentido de la medición y optimización del desempeño organizacional, se puede evaluar en contribuciones que abordan el concepto de Gestión de Cadenas de Suministro (SCM por sus siglas en inglés), el cual es una extensión del concepto de cadenas de valor al estudio de las cadenas de suministro en macro mercados de bienes tangibles y manufactura. El trabajo de [20] presenta una revisión de casos de distintas compañías en donde se contrasta el uso de técnicas predictivas avanzadas de cara a la variación de indicadores multidimensionales de desempeño en la gestión de cadenas de suministro, resaltando las prácticas exitosas de uso de modelos predictivos, así como los retos que se afrontan para la correcta implementación de estas tecnologías.

Por esta misma línea, y a manera de extrapolación, [21], [22] y [23] presentan revisiones bibliométricas y meta análisis reflejando el incremento de investigaciones donde se emplea la modelación predictiva y prescriptiva a través de métodos en Ciencia de Datos, de cara al manejo de shocks en cadenas de suministro en la industria alimentaria, de tal manera que los métodos y protocolos estructurados representan tanto un instrumento en la regulación de las operaciones de entes privados, como una herramienta de contingencia en materia de política industrial y seguridad alimentaria.

Ahora, evocando investigaciones relacionadas al ejercicio específico de aplicación de este proyecto, contribuciones como [24] y [25] presentan ejercicios de clasificación eficiente de candidatos en mercados laborales a través de ejercicios de Deep Choice y predicción de series en datos panel, respectivamente.

Trayendo a contexto la similitud en el grado de alcance macro del caso de Comfacauca con los casos de las contribuciones mencionadas anteriormente, resaltan los impactos potenciales del empleo de métodos de analítica avanzada sobre las actividades de cadenas de valor de alto alcance poblacional, de tal manera que los efectos positivos de la estructuración de protocolos de Ciencia de Datos en sus dinámicas organizacionales se halla directamente relacionada con la mejoría en indicadores socioeconómicos de impacto regional.

2. PROTOCOLO DE INVESTIGACIÓN, DESCRIPCIÓN DE LA CADENA DE VALOR DE LA RUTA DE ATENCIÓN Y CONSECUCIÓN DE ESTRUCTURAS DE DATOS RELEVANTES

En este capítulo se aborda una descripción puntual de la metodología general sobre la cual se diseña e implementa el protocolo flexible que representa el eje de esta investigación, prosiguiendo con la contextualización correspondiente a la cadena de valor de la Agencia de Empleo, y las estructuras de datos que darán forma a los resultados, conclusiones y recomendaciones del proyecto.

2.1 PROTOCOLO PARA EL ANÁLISIS DEL CASO DE ESTUDIO.

El objetivo general de esta investigación radica en la necesidad de la implementación de un protocolo puntual, tanto para una correcta contextualización con respecto al problema de negocio a resolver, como para la realización de ejercicios exploratorios y de modelación que otorguen resultados fiables a la luz de criterios cuantitativos rigurosos. Además, este protocolo también debe comprender una etapa de divulgación de los resultados y recomendaciones obtenidas de manera compatible con el lenguaje estándar de los receptores del proyecto al interior de Comfacauca.

Adicional a lo anterior, el objetivo general del proyecto se ve limitado por el alcance del mismo, el cual restringe la investigación al análisis de las estructuras de datos disponibles en un sentido externo, y no comprende una fase de implementación/evaluación de las recomendaciones obtenidas debido a la limitación normativa del acuerdo de colaboración entre quien realiza el proyecto y la organización (ver Acuerdo de Colaboración en los anexos).

En este sentido, y con el propósito de garantizar la flexibilidad y posterior extrapolación por parte de Comfacauca de los desarrollos obtenidos por el proyecto, se opta por diseñar e implementar un protocolo que integra a los lineamientos estándar de los protocolos de investigación académica [26] las orientaciones para la aplicación de proyectos de Ciencia de Datos de la metodología CRISP-DM [27], teniendo en cuenta la naturaleza de cadena de valor [2] de los servicios que presta Comfacauca, particularmente, de la Agencia de Empleo.

Así, el protocolo propuesto por este proyecto aplicado consiste en la ejecución de un protocolo de investigación estándar [26], con sus respectivas fases: propuesta del proyecto, delimitación del problema de investigación, formulación de objetivos, justificación, marco de referencia, desarrollo metodológico, consecución de resultados, y discusión de conclusiones y recomendaciones.

Ahora, para esta estructura estándar, el protocolo propuesto puntualiza su orientación investigativa al insertar la noción de identificación de la(s) cadena(s) de valor relevante(s) para el análisis [2] y las acciones correspondientes a la metodología CRISP-DM [27] en la fase de desarrollo metodológico: comprensión del problema de negocio a enfrentar, la comprensión y preparación de la estructura de datos disponible, su modelación desde las estructuras teóricas propuestas, y la evaluación de los resultados obtenidos.

Las consignaciones sucesivas en los capítulos y secciones de este documento describen los resultados de la implementación de este protocolo. La descripción del protocolo diseñado desde sus fases, acciones e interlocutores, se presenta como Anexo a la entrega de este documento.

2.2 LA CADENA DE VALOR DE LA RUTA DE ATENCIÓN DEL SERVICIO DE EMPLEO DE COMFACAUCA.

Como se estableció en la metodología consignada en el anteproyecto, el núcleo de este proyecto aplicado consiste en esquematizar las actividades y servicios provistos por la organización Comfacauca en el sentido de cadenas de valor, de tal forma que los procesos fijados por ésta misma puedan entenderse como un conjunto de actividades estructuradas secuencialmente de cara a la entrega de servicios específicos, y así contrastar propuestas de modelación que arrojen resultados congruentes de cara a la optimización de sus problemas de negocio.

En este sentido, y dada la amplia gama de servicios que provee Comfacauca, se decidió priorizar en este ejercicio las actividades de la Agencia de Empleo, el cual actúa en el marco de la Ley 1636 de 2013 y sus complementos para ejercer en el Cauca el Mecanismo de Protección al Cesante (MPC), esencialmente desde la provisión de asistencia en la búsqueda de empleo para los afiliados cesantes, así como el acompañamiento en materia de subsidios y capacitación. Todo lo anterior se ejecuta a través de una Ruta, la cual puede comprenderse como la cadena de valor principal del Servicio de Empleo.

A continuación, se recopilan los problemas de negocio a evaluar en el marco de las actividades del Servicio de Empleo:

- Necesidad de formulación de criterios cuantitativos específicos para el incremento en la eficiencia y eficacia de procesos de remisión y selección en vacantes ofrecidas por las empresas hacia los cesantes que se registran en la Ruta de Atención del Servicio de Empleo.
- Identificación de habilidades y áreas de formación a priorizar en las capacitaciones a llevar a cabo por parte del Servicio de Empleo.

- Caracterización de condiciones laborales generales en el panorama del mercado laboral del Cauca desde la base de datos del Servicio de Empleo.
- Identificación de parámetros como insumos cuantitativos generales para la toma de decisiones por parte del Servicio.

De acuerdo a la priorización de problemas de negocio presentada anteriormente, se obtuvo una esquematización del proceso secuencial a través del cual el Servicio de Empleo administra el Mecanismo de Protección al Cesante en el sentido de una cadena de valor. A manera de fases, esta estructura de cadena de valor puede verse de la siguiente forma:

1. El candidato se registra de manera autónoma o asistida en la plataforma de Bolsa de Empleo de Comfacauca, facilitando el conjunto de atributos estandarizados con los que participará en el mercado laboral en busca de un empleo.

2. Dadas las vacantes publicadas por las empresas adscritas al Servicio, se realiza la remisión del perfil de posibles candidatos a obtener la posición demandada.

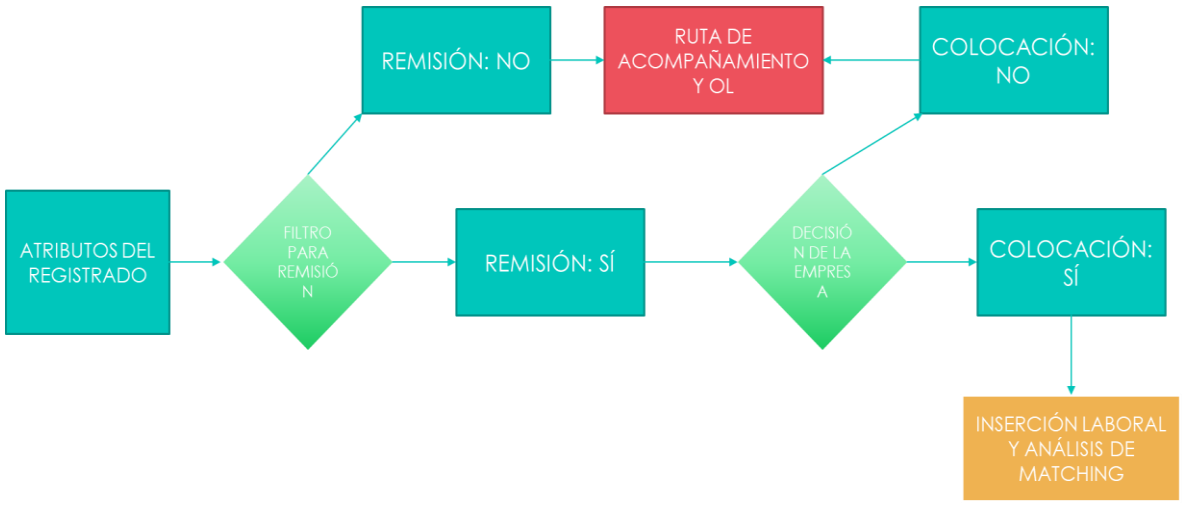
3. Se efectúa el proceso de selección por parte de las empresas, tras lo cual se decide si los candidatos son colocados o no en las vacantes.

4. En caso de ser colocados, el Servicio registra el caso de inserción del candidato en el mercado laboral, acompañado de las condiciones del contrato obtenido.

5. En caso de no ser colocados, el Servicio continúa con el monitoreo de la ruta de atención a través del ofrecimiento de capacitaciones y talleres de orientación en el mercado laboral. Además, se realiza la entrega de subsidios de acuerdo al marco legal existente para este propósito.

La siguiente figura presenta un diagrama de flujo para la cadena de valor de la ruta de atención del Servicio de Empleo.

Figura 1. Cadena de valor de la Ruta de Atención del Servicio de Empleo.



2.3 ESTRUCTURA DE DATOS.

Tras la luz verde para la facilitación de información por parte de la organización, el proceso de recolección de información operacional y de afiliados en el marco de la cadena de valor analizada se acordó en dos fases.

En la primera fase, se realiza la entrega de una muestra inicial de los datos disponibles para el proceso analítico y de modelación, sobre el cual se definen los procedimientos de transformación y pre procesamiento necesarios para el contraste de los modelos, de tal forma que sea posible definir un prototipo de esquema para modelación predictiva y prescriptiva.

Dada la identificación de estos procedimientos, se elevan requerimientos de confección y escalamiento del dataset para la ejecución de los modelos finales en la segunda fase. Con la entrega de los datos a escala real en la segunda fase, se procede a los ejercicios de modelación a escala real.

Así, para la consecución del dataset a escala real, y dada la disponibilidad de información por parte de la organización, se recibió un archivo Excel (xlsx) con 3 sheets relevantes para el ejercicio de limpieza y pre procesamiento, del cual se obtuvieron 87.735 casos de aplicación registrados indexados por sus IDs. Las 3 sheets del archivo base describen los componentes clave de oferta y demanda laboral en la cadena de valor a analizar, de la siguiente forma:

- Sheet 1: Registrados (38 atributos).
- Sheet 2: Colocados (29 atributos).
- Sheet 3: Vacantes (28 atributos).

Dado el archivo de base de datos facilitado, se realizó el pre procesamiento y limpieza del dataset. Los métodos empleados se resumen en los siguientes pasos secuenciales (los cuales se llevaron a cabo empleando R Markdown):

1. Merging de las sheets relacionadas empleando la operación left join, a través del cual se obtuvo un dataset crudo con 87.735 casos de aplicaciones a vacantes.
2. Selección de los atributos de mayor relevancia para los ejercicios necesarios de modelación.
3. Extracción de etiquetas binarias de clase para modelos de aprendizaje supervisado: Colocados y Colocados por sector.
4. Limpieza de dataset: Renombramiento de columnas e imputación de missings.

Así, se obtuvo el dataset listo para incurrir en análisis descriptivo y posteriores procesos de modelación.

2.4 ANÁLISIS EXPLORATORIO

Dado el ejercicio de pre procesamiento y limpieza del dataset crudo, se obtuvo un dataset limpio inicial para modelación, con 87.735 observaciones asociadas a casos de aplicación de los registrados a vacantes. El dataset obtenido se halla descrito por la siguiente tabla.

Tabla 1. Variables seleccionadas para análisis en el dataset.

Atributo	Tipo	Descripción
ID	ID	
Edad	Numérica	Edad
CanalRegistro	Categórica	Medio de registro (asistido o autoregistro)
Genero	Categórica	Género
PercHojaVida	Categórica	Porcentaje de completitud del formulario de hoja de vida
CiudadResidencia	Categórica	Municipio de residencia
NivelEstudio	Categórica	Nivel de estudio
Urb_Rural	Categórica	Vive en zona rural o urbana
AspiracionSalarial.x	Categórica	Aspiración salarial
TituloObtenido	Texto	Título obtenido descrito de manera específica
UltimaExperienciaLaboral.x	Texto	Última experiencia laboral descrita de manera específica
Colocado	Etiqueta binaria de clase	Colocado (Sí o No)
ColocadoBinary	Etiqueta multiclase	Colocado (Colocados por sector económico y no colocados)

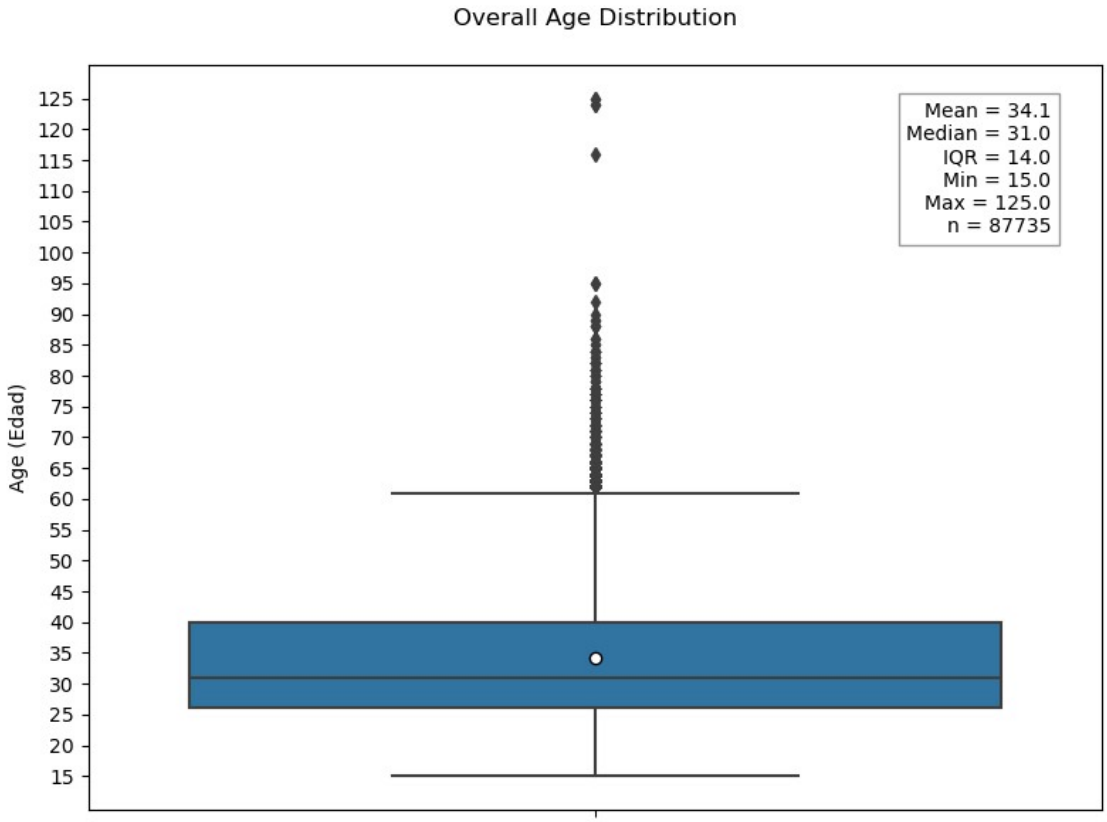
Nótese que los atributos seleccionados se obtuvieron de manera acordada con los encargados de las operaciones del servicio de empleo bajo criterios de relevancia en el problema de negocio y conveniencia estadística. En general, los atributos para el modelo de clasificación son en su mayoría categóricos, y se requirió el ejercicio de un procedimiento de clustering para los dos tributos de texto, en tanto de manera previa se sospecha que son claves para una estructura de correlación que decante en un modelo eficiente de clasificación de candidatos. A continuación, se presenta un análisis descriptivo inicial para cada una de las variables analizadas, lo cual indica la línea base de caracterización de la proporción del mercado laboral que capta el Servicio de Empleo, dado el dataset facilitado

por la organización. En cuanto al análisis de correspondencia requerido para esta fase exploratoria, las salidas puntuales de los test de correlación Chi-cuadrado sobre las variables categóricas del dataset se consignan en el Apéndice.

Edad:

La variable edad es la única variable numérica en el dataset. Para las 87.735 observaciones analizadas, se tiene que de los casos de aplicantes a las vacantes, la edad mediana se ubica sobre los 31 años, existiendo además un rango intercuartílico de entre los 26 y los 40 años. Lo anterior caracteriza el rango medio sobre el cual es posible hallar al grueso de los aplicantes a las vacantes consideradas en el dataset. A manera de hipótesis previa, y, considerando la baja tasa de colocación que se expresará más adelante, se puede observar una alta incidencia del desempleo en edades jóvenes para el departamento del Cauca, sin embargo, también es posible afirmar que, al menos para esta muestra, se entiende que el departamento dispone en gran medida de fuerza laboral joven. El siguiente boxplot resume el comportamiento de la distribución empírica de la variable Edad en el dataset.

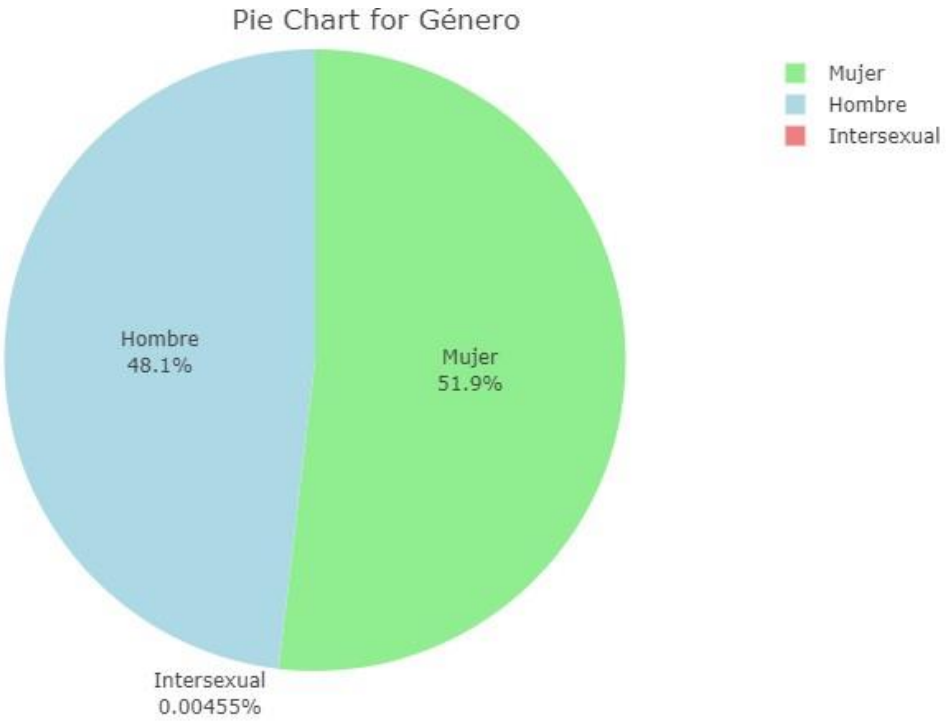
Figura 2. Boxplot para la variable Edad.



Género:

Como lo denota el gráfico de torta desplegado a continuación, el comportamiento de la frecuencia del género en el dataset denota una ligera predominancia de las mujeres en la fuerza laboral para la muestra analizada. Así, puede denotarse un patrón en el cual se evidencia una mayor participación femenina en el mercado laboral. Sin embargo, debe mencionarse que el test de correlación Chi-cuadrado denota una correlación estadísticamente significativa de propensión de los hombres a ser colocados en mayor medida.

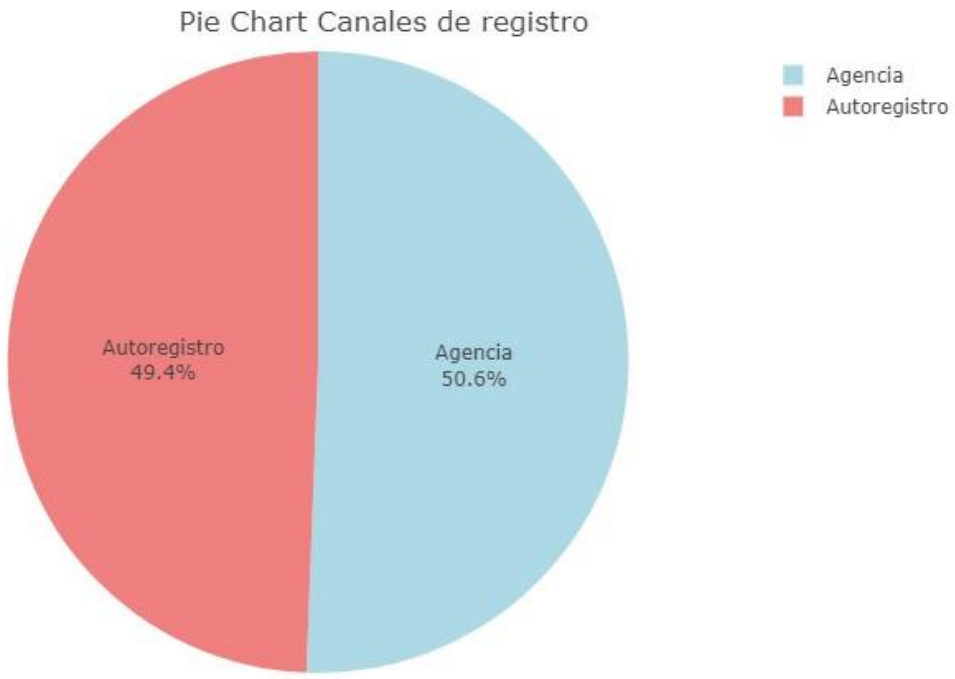
Figura 3. Pie chart para la variable Género.



Medio de Registro:

El comportamiento de la frecuencia del medio empleado para el registro en la bolsa de empleo también puede denotar una aproximación al nivel de manejo de herramientas tecnológicas por parte de los aspirantes al mercado laboral. Existiendo así un patrón de ligera dominancia del registro por medio de la agencia entre los aplicantes. Además, la prueba de correlación Chi-cuadrado denota una alta asociación de los registrados en la agencia con la obtención de empleo.

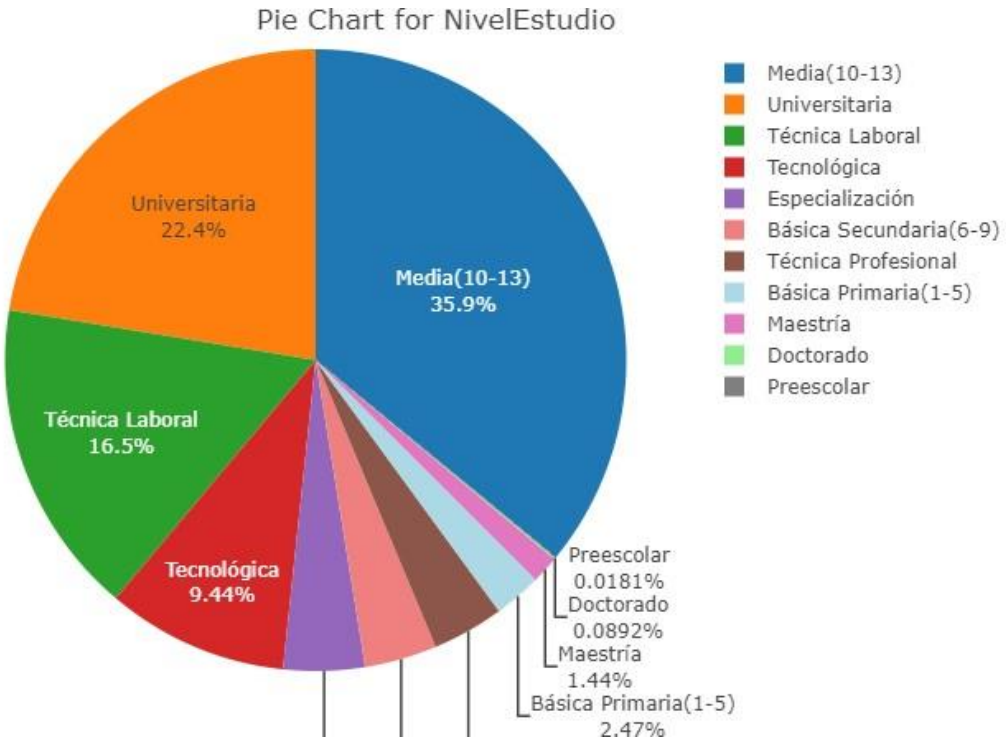
Figura 4. Pie chart para la variable Canal de Registro.



<u>Nivel de Formación Educativa</u>:

En cuanto al nivel de formación educativa de los aplicantes, de acuerdo a las categorías estándar del sistema colombiano, el gráfico de torta desplegado debajo denota una predominancia de la educación media, con una proporción del alrededor del 35.5%, seguido de la educación universitaria (14.5%), la educación técnica laboral (11.4%), y la formación tecnológica (6,51%). De tal forma que las categorías subsiguientes se muestran con proporciones bastante inferiores. Como hipótesis previa es importante resaltar el panorama donde la formación profesional no es una regla dominante entre los aspirantes, lo cual puede influir en gran medida, no sólo sobre una alta tasa de fallos en la aplicación a vacantes requiriendo niveles de formación por encima de la educación media, sino en una idea previa sobre una escasez significativa de oferta laboral calificada.

Figura 5. Pie chart para la variable Nivel de Estudio.

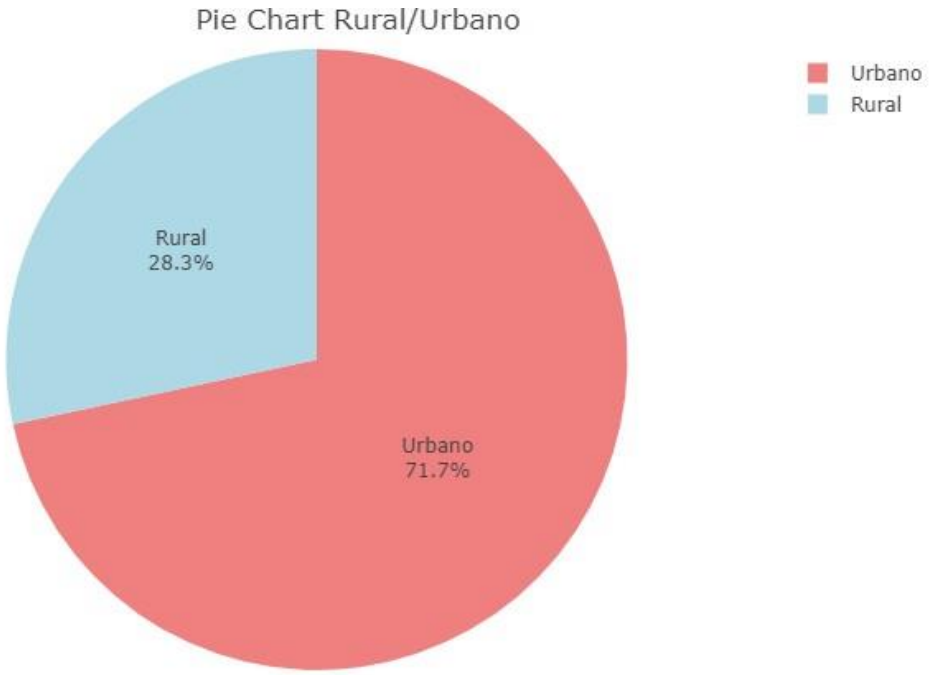


En cuanto al nivel de formación educativa de los aplicantes, de acuerdo a las categorías estándar del sistema colombiano, el gráfico de torta desplegado debajo denota una predominancia de la educación media, con una proporción del alrededor del 35.9%, seguido de la educación universitaria (22,4%), la educación técnica laboral (16.5%), y la formación tecnológica (9,44%). De tal forma que las categorías subsiguientes se muestran con proporciones bastante inferiores. Como hipótesis previa es importante resaltar el panorama donde la formación profesional no es una regla dominante entre los aspirantes, lo cual puede influir en gran medida, no sólo sobre una alta tasa de fallos en la aplicación a vacantes requiriendo niveles de formación por encima de la educación media, sino en una idea previa sobre una escasez significativa de oferta laboral calificada. El test de correlación Chi-Cuadrado denota significancia estadística en la asociación mayoritaria entre la obtención de empleo y el nivel educativo Medio y Técnico.

Ruralidad/Urbanidad:

El gráfico de torta debajo denota un comportamiento esperado, en tanto existe una marcada predominancia de pertenencia a zonas urbanas por parte de los aplicantes. Esto es congruente con la disponibilidad de empleos y fuerza laboral de mayor capacitación en las zonas urbanas del departamento. Esto es congruente con la orientación en los resultados del test de correlación Chi-cuadrado.

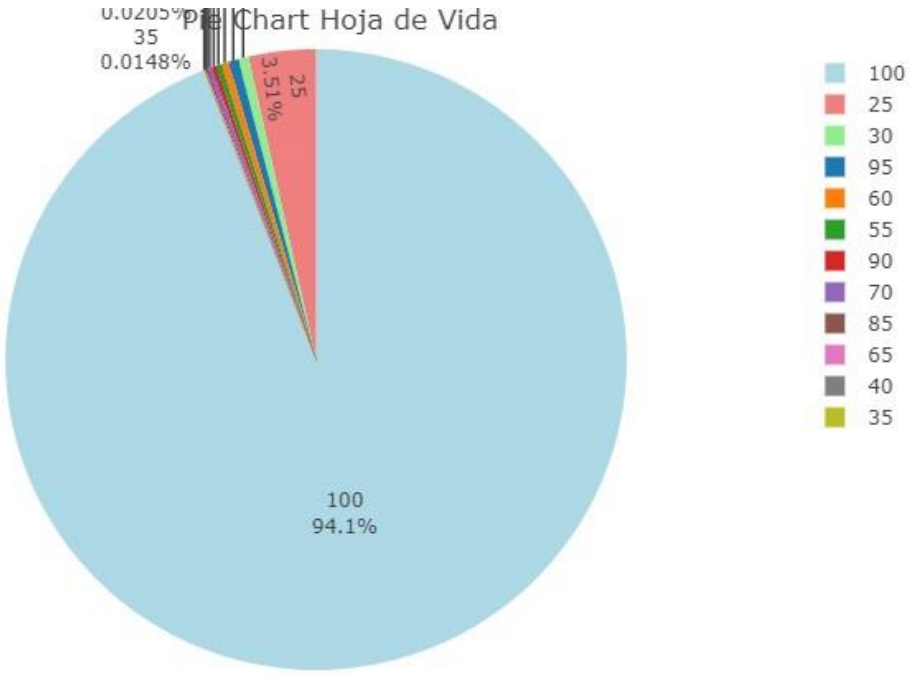
Figura 6. Pie chart para la variable Rural/Urbano.



Hoja de Vida (Porcentaje de completitud en plataforma):

La variable categórica asociada a la completitud de la hoja de vida en la plataforma de Bolsa de Empleo denota una especie de aproximación al grado de manejo de las herramientas tecnológicas que ofrece el Servicio por parte de los aspirantes. Además, esta variable también puede servir como una aproximación a una revisión del bajo nivel de capacitación de la fuerza laboral en el departamento. En congruencia con el gráfico desplegado a continuación, el test de correlación Chi-Cuadrado denota una asociación predominante de la obtención de empleo a las hojas de vida llenas al 100%.

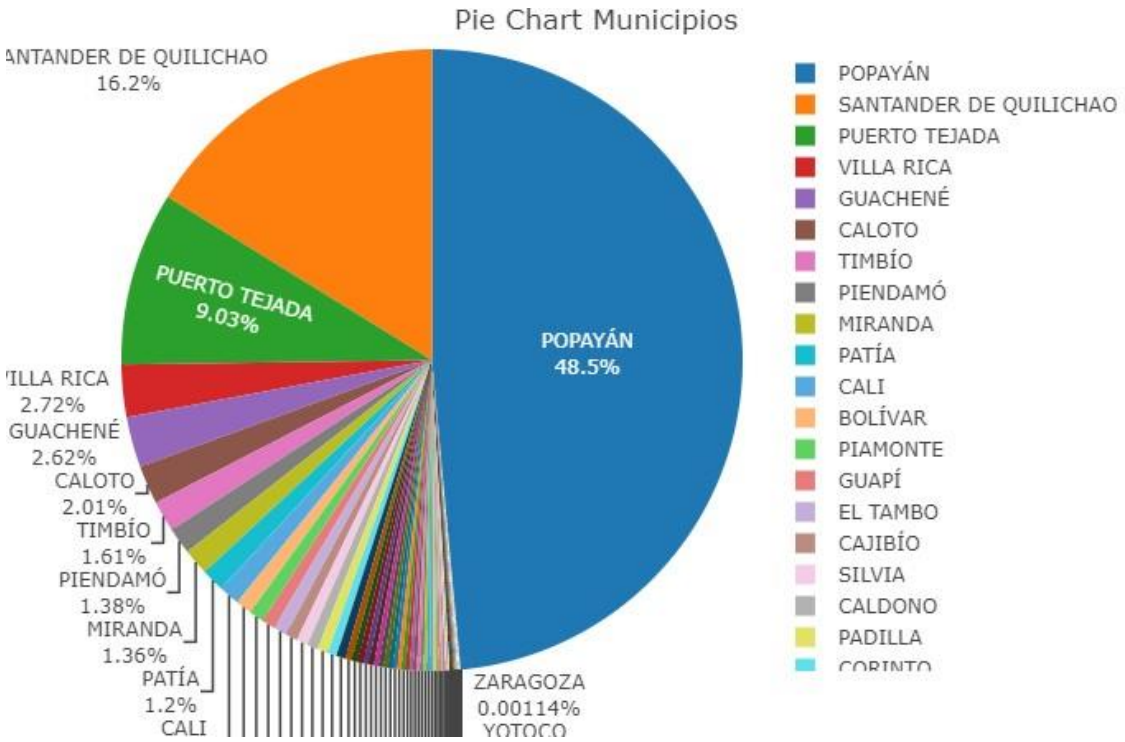
Figura 7. Pie chart para la variable Porcentaje de Hoja de Vida.



Municipios de Residencia:

La distribución de la frecuencia de los aplicantes con respecto a los municipios de residencia denota una característica común en la dinámica económica del departamento, pues, la predominancia de los municipios de Popayán (48.5%), Santander de Quilichao (16.2%), y Puerto Tejada (9.03%), es congruente con la aglomeración de los enclaves industriales y productivos alrededor de la capital del departamento y las zonas francas ubicadas al norte y en frontera con el alto volumen de actividad económica del Valle del Cauca. En este sentido, no sorprendería encontrar que, al formalizar un modelo de clasificación, la residencia en estos municipios sea un factor de alta influencia en la tasa de colocación del mercado laboral. Esta hipótesis previa es soportada por los resultados del test de correlación Chi-cuadrado con respecto a la etiqueta binaria de colocación.

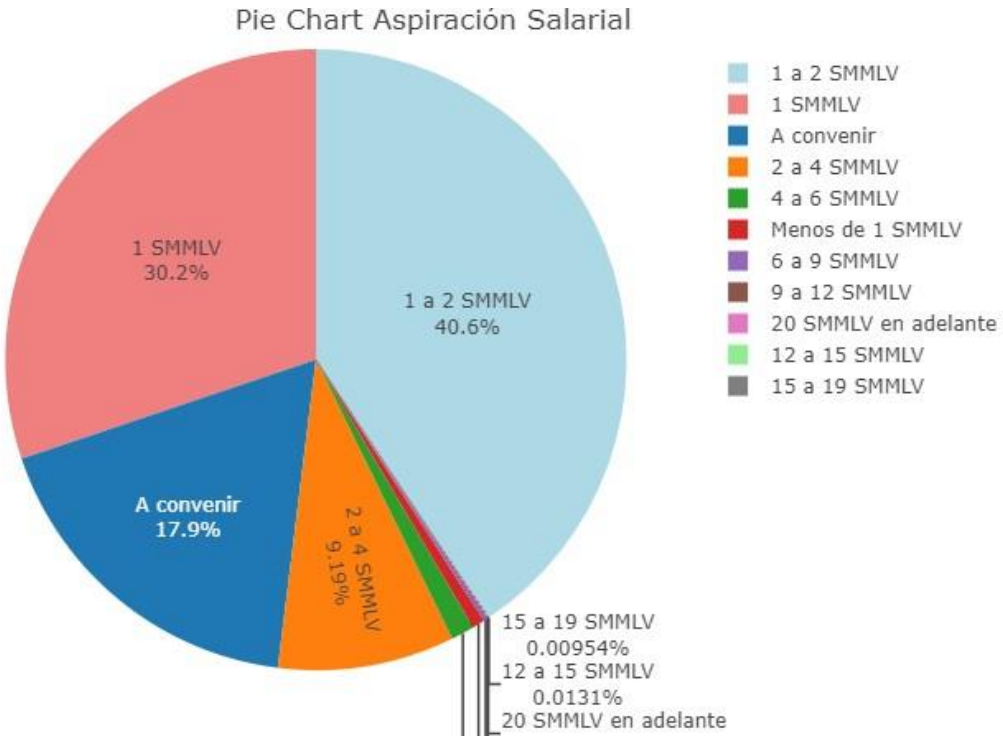
Figura 8. Pie chart para la variable Municipio de Residencia.



<u>Aspiración Salarial</u>:

Como lo muestra el gráfico de torta desplegado debajo, se tiene que la mayoría de los casos de aplicación proyecta a aplicantes con aspiraciones salariales de 1 a 2 SMMLV (40.2%), 1 SMMLV (30.2%), y bajo la categoría libre A convenir (17.9%). De tal forma que aspiraciones salariales de mayor volumen se presentan en proporciones muy poco significativas. Este resultado es congruente con la hipótesis previa provista por el test de correlación Chi-cuadrado, el cual denota una alta asociación de obtención de empleo con aspiraciones de 1 a 2 SMMLV. Nótese cómo la descripción de esta categoría se alinea con la descripción del nivel de estudio de los candidatos.

Figura 9. Pie chart para la variable Aspiración Salarial.

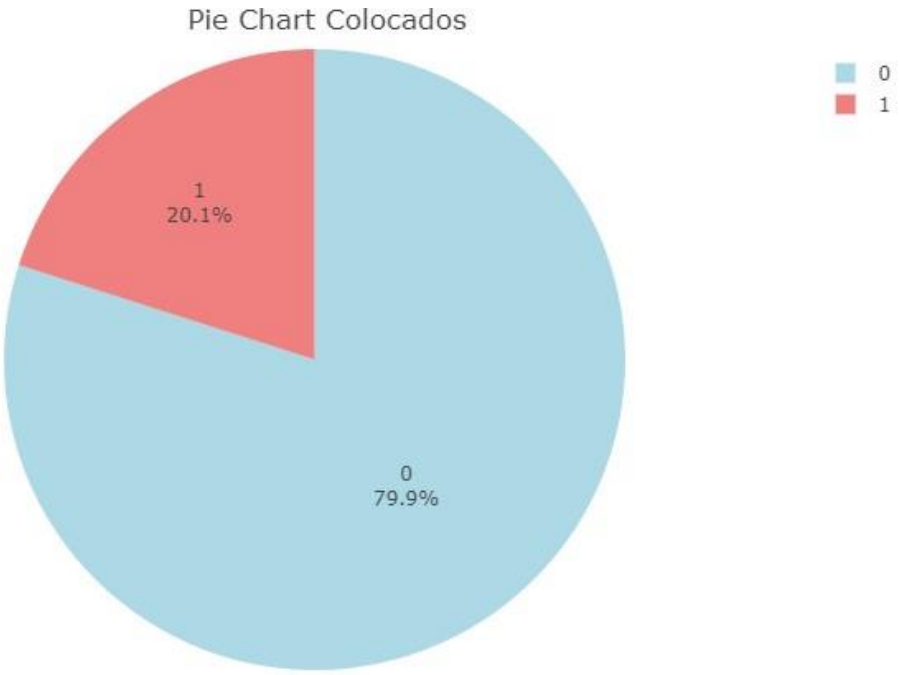


Variables de texto: Título Obtenido y Última Experiencia Laboral.

Las variables de texto en el dataset representan un obstáculo en la etapa exploratoria y de pre procesamiento, en tanto se sospecha que contienen información clave respecto a los efectos diferenciales que determinan la obtención de empleo, pero a la vez implican una enorme dispersión en las posibles respuestas debido a que se trata de entradas libres de texto en la plataforma. Así, fue necesaria la generación de categorías puntuales en función de textos a través de un modelo de clustering. Esto se consigna en el siguiente capítulo.

Finalmente, el gráfico de torta desplegado debajo denota la línea base fundamental del dataset, en tanto describe la frecuencia de obtención de empleo en la etiqueta binaria ColocadoBinary, de cara a la construcción de modelos de clasificación bajo aprendizaje supervisado.

Figura 10. Pie chart para la etiqueta de clase (1: Colocado, 2: No colocado).



Nótese cómo la clase positiva presenta una frecuencia bastante baja con respecto a la clase negativa. Lo anterior es una aproximación a la baja ratio de obtención de empleo debido a la estructura económica captada por la Agencia con respecto al departamento del Cauca. Además, esta situación anuncia de manera previa el sesgo inmerso en la modelación de aprendizaje sobre la clase negativa. Lo anterior evoca el conjunto de recursos y la estructura de resultados a presentar tras la selección de los modelos más eficientes en el siguiente capítulo.

3. MODELACIÓN

En este capítulo se despliegan los pormenores del conjunto de ejercicios de modelación propuestos para el problema de clasificación planteado, de tal forma que es posible establecer un contraste entre las propiedades de los métodos relacionados teóricamente en el capítulo 1.

Además, de manera previa a la exposición de los resultados de los ejercicios de modelación bajo los métodos planteados, es necesario especificar un proceso de modelación previo a la construcción de los clasificadores que determinan esta investigación. Como se mencionó en la sección correspondiente al análisis exploratorio, los atributos Título Obtenido y Última Experiencia Laboral se presentan como entradas de texto libre por parte de los candidatos registrados en la Ruta de Atención. Así, en tanto la especificación del título y experiencia laboral previa del candidato se consideran a priori como un atributo de alta importancia para el problema de clasificación a resolver, es indispensable realizar un ejercicio de modelación que permita el tratamiento normalizado de estas entradas de texto como insumo proveniente del dataset.

3.1 Normalización y agrupamiento de entradas de texto por medio de K-means clustering.

La información contenida en las variables de texto de Título Obtenido y Última Experiencia Laboral se considera como indispensable de cara a la clasificación de mejores candidatos y la determinación de factores clave para la consecución de empleos en el mercado laboral del Departamento.

Sin embargo, la multiplicidad y heterogeneidad de las posibles entradas de texto implica una dificultad para la correcta lectura de estos atributos como insumos numéricos de un clasificador. En este sentido, se optó por aplicar una metodología de clustering sobre las variables de texto, la cual genera columnas de etiquetado de cada observación en un cluster, otorgando una señalización operable para el título específico de los candidatos y su última experiencia laboral, en el sentido de convertir las columnas con entradas de texto a columnas categóricas con múltiples instancias.

El ejercicio de clustering se realizó en Python, ejecutando un pipeline de pre procesamiento de texto y clustering por medio de K-means. Bajo los siguientes pasos secuenciales:

- Carga de datos y selección de columnas de texto.
- Pre procesamiento para entradas de texto en español: remoción de puntuaciones y stopwords, modificación de palabras con acentos, ajuste de espacios en blanco.

- Ejecución de entrenamiento de modelo de clustering por medio de K-means, con ajuste manual de hiperparámetros (cantidad de clusters).
- Salida de resultados y asignación de etiquetas de cluster en las nuevas columnas: EduCluster para TituloObtenido, y ExperienciaCluster para UltimaExperienciaLaboral.
- Evaluación sucesiva de la calidad de clustering por medio de métricas Silhouette index, Davies-Bouldin index y Caliski-Harabasz index; hasta alcanzar un nivel óptimo conjunto.

En general, se optó por la realización de este ejercicio a través del ajuste manual sucesivo del método K-means por su amplia ventaja frente a otros modelos en términos de interpretabilidad, cantidad de hiperparámetros (sólo uno, cantidad de clusters) y eficiencia computacional.

El método K-means consiste en la determinación de clusters a través del ajuste iterativo de las observaciones consideradas a distancias mínimas con respecto a los centroides muestreados por este algoritmo de aprendizaje no supervisado. Este proceso se realiza de manera simple y de fácil interpretación, en tanto su medida de ajuste parte de la evaluación de distancias euclidianas entre las observaciones del dataset [6]. Para el caso de entradas de texto, se requiere la normalización de las mismas a través de columnas binarias que indiquen la ocurrencia de cada palabra considerada.

Es importante resaltar que este método se escogió sin un contraste frente a otras propuestas de modelación por aprendizaje no supervisado, en tanto la tarea a satisfacer consistió en el agrupamiento de palabras clave, no en la identificación de patrones puntuales de aglomeración para el dataset en general (éste suele ser el núcleo de un problema de clustering que amerite un contraste entre distintos tipos de modelos).

Nótese que la simplicidad de este problema de agrupamiento de palabras clave parte de la posibilidad de asumir que existen múltiples títulos y empleos previos que pueden agrupar a los candidatos en el dataset, de esta forma, se tiene que se puede llegar a una convergencia con respecto a las métricas de evaluación de clustering al incrementar sucesivamente la cantidad de clusters como hiperparámetro de K-means, hasta llegar a un límite de eficiencia.

La siguiente tabla resume el proceso de ajuste sucesivo ejecutado, al presentar la cantidad de clusters elegidos y sus correspondientes métricas de evaluación, hasta obtener un nivel óptimo.

Tabla 2. Output de resultados para ejercicio de clustering sobre variables de texto (K-means).

Columna EduCluster (Agrupamiento de TituloObtenido)			
Valor de K	**Silhouette Score**	**Davies-Bouldin index**	**Calinski-Harabasz index**
429	0,7753	1,005	5551,4996
1978	0,8681	0,7583	5547,7639
2480	0,8773	0,6757	6120,1999
Columna ExperienciaCluster (Agrupamiento de UltimaExperienciaLaboral)			
Valor de K	**Silhouette Score**	**Davies-Bouldin index**	**Calinski-Harabasz index**
429	0,9125	0,4747	48566,2581
1978	0,9846	0,0711	88217527,06
2440	0,9917	0,0004	1,28509E+20

Nótese cómo, al incrementar la cantidad del clusters como hiperparámetro de K-means, la eficiencia en términos de las métricas de evaluación mejora hasta acercarse a un límite. El coeficiente de Silhouette determina la eficiencia del grado de agrupamiento dada la cantidad de clusters seleccionada en un escala entre 0 y 1, donde la cercanía a 1 denota la máxima eficiencia posible [7], como se presenta en los resultados para ambas columnas. El índice de Davies-Bouldin determina la eficiencia del agrupamiento al evaluar la compacidad y distancia entre los clusters formados, en una escala para los reales positivos [7], donde la cercanía a cero denota la máxima eficiencia, como se aprecia de forma descendente en los resultados obtenidos. Finalmente, el índice de Calinski-Harabasz realiza una evaluación bajo criterios similares a los de Davies-Bouldin, pero bajo una escala en los reales positivos donde la eficiencia del clustering incrementa en tanto el índice se acerca más a infinito [7], esto también se aprecia de manera sucesiva en los resultados obtenidos.

Así, se tiene que el proceso realizado determinó dos columnas categóricas suplementarias, EduCluster y ExperienciaCluster, con 2481 y 2441 instancias (el valor de cluster 0 se asigna a las observaciones con ruido), respectivamente, las cuales proveen estructuras operables de muestreo y correlación propicias para fijar la información sobre las variables de título obtenido y experiencia laboral previa del candidato como insumo para el entrenamiento del clasificador requerido.

Otro refuerzo al argumento de robustez del método empleado es que muy pocas observaciones del dataset son determinadas como ruido. De 87.735 casos de aplicaciones consideradas, se tiene que sólo 7 para EduCluster, y una para ExperienciaCluster, se etiquetan como ruido tras aplicar K-means en su forma óptima.

Es importante resaltar que el ejercicio de agrupamiento de entradas de texto realizado no sólo se posiciona como insumo para el entrenamiento del modelo central de esta investigación, sino que también implica una herramienta para facilitar el análisis de grupos ocupacionales de manera más rápida debido a la normalización de texto y la identificación de frecuencias de cargos y títulos característicos en el mercado laboral del Departamento. Representando una idea inicial para las recomendaciones para la organización en cuanto al manejo de sus estructuras de datos de cara a propósitos analíticos.

El ejercicio de clustering realizado completa la fase exploratoria del proyecto, y da paso al ejercicio principal de modelación.

3.2 Aprendizaje supervisado: Modelos de clasificación para la caracterización del mercado laboral.

Una vez seleccionados los atributos pertinentes para el contraste de clasificación, así como la generación de clusters consistentes para las variables críticas de texto relacionadas a títulos educativos y experiencia laboral previa de los candidatos, se procede a la construcción del clasificador sobre el cual se basa el análisis de esta investigación. Este ejercicio se realizó por medio de Python, a partir de la construcción de pipelines estructuradas en los siguientes pasos para cada modelo considerado:

- Carga de datos, selección de atributos por tipo de datos, definición de etiqueta binaria de clase.
- Splitting 70-30 para subsets de entrenamiento y prueba con base en la etiqueta binaria de clase. Notando que ejercicios de splitting 80-20 y 90-10 fueron realizados, pero se obtuvieron resultados similares o con peor desempeño que en el caso 70-30.
- Pre procesamiento de datos numéricos y categóricos.
- Balance de observaciones en función de etiqueta de clase para el subset de entrenamiento (según el modelo, se empleó SMOTE o balanced class weights).
- Entrenamiento de modelo.
- Despliegue de resultados, métricas de evaluación y relación de orden de importancia de los atributos (para los modelos que pueden calcular esta propiedad).
- Salida de archivo con candidatos organizados en orden descendente por probabilidad estimada de colocación.

Considerando el contraste de paradigmas que propone este proyecto, se toman como candidatos los modelos explorados en el marco teórico, de la siguiente forma:

- Paradigma frecuentista: Logistic Regression, Random Forest, Gradient Boosting.

- Paradigma bayesiano: Naive Bayes, Bayesian Ridge Regression.

Nótese que la selección de estos modelos parte de su conveniencia en términos computacionales e interpretabilidad con respecto al problema de negocio en cuestión. Además, debe mencionarse que se consideró aplicar modelos multiclase debido a la existencia de diversos sectores productivos para el mercado laboral, sin embargo, estos modelos otorgaron resultados muy pobres debido al sesgo inherente de la clase negativa en el dataset (una baja probabilidad idiosincrática de conseguir empleo), y a la definición laxa de categorías institucionales para las instancias de cada sector.

Regresión Logística

El clasificador construido por medio del algoritmo de regresión logística se entrenó optimizando el espacio de hiperparámetros a través de RandomizedSearchCV, mecanismo de afinamiento de hiperparámetros que halla los mejores valores por medio de validación cruzada de muestras aleatorias en 50 iteraciones, tomando como referencia el score provisto por el área bajo la curva (AUC) del gráfico estándar ROC (Receiver Operating Characteristic), el cual determina el nivel de ajuste de las predicciones del modelo entrenado desde su matriz de confusión. En general, el espacio de búsqueda de hiperparámetros se fijó de manera estandarizada, al considerar un rango puntual para el nivel de regularización para evitar el overfitting, y, adicionalmente, se plantearon dos estrategias adicionales para el sobre muestreo SMOTE, con fin de lidiar con el problema del alto sesgo presente en el dataset para la instancia negativa de la etiqueta de clase. Este esquema de afinamiento se presenta en la siguiente tabla.

Tabla 3. Espacio de afinamiento de hiperparámetros para el entrenamiento del modelo por medio de regresión logística.

Hiperparámetro	**Search space**
Fuerza de regularización (base 10, 20 valores de muestra)	(-4,4)
Tipo de penalización	L2
Estrategias de muestreo SMOTE	50-50, 50% de la clase mayoritaria, 60% de la clase mayoritaria

Como lo muestra la Tabla 4, los valores óptimos decantaron por una estrategia SMOTE de la clase minoritaria como 60% de la clase mayoritaria, y un valor de regularización sobre la potencia 10^2. De esta forma, los resultados obtenidos para este modelo fueron bastante pobres, pues se alcanzó un nivel muy bajo en las predicciones, tanto a nivel promedio de precisión (precisión media ponderada de 0,33 en el umbral estándar de 50%), como en

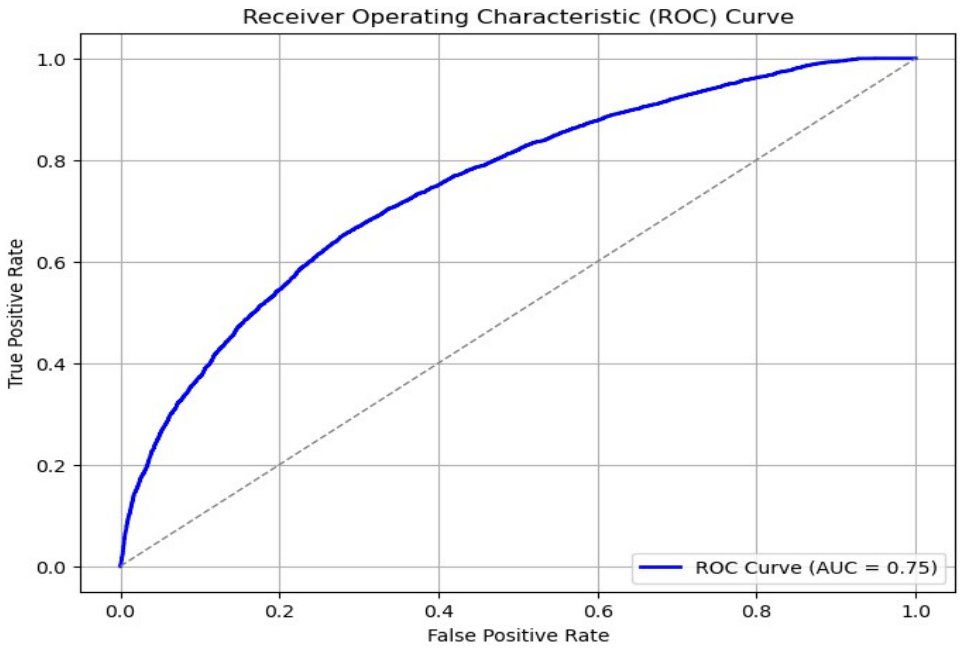
cuanto a las métricas de evaluación puntuales para cada instancia de la etiqueta de clase (colocado vs no colocado), incluso considerando el sesgo inherente sobre la instancia negativa de la etiqueta de clase.

Tabla 4. Reporte de clasificación (umbral de 50%), modelo por medio de regresión logística.

```
Fitting 5 folds for each of 50 candidates, totalling 250 fits
Best Parameters:
 {'smote__sampling_strategy': 0.6, 'classifier__penalty': 'l2', 'class
ifier__C': 545.5594781168514}
Classification Report:
              precision    recall  f1-score   support

  Not Placed       0.96      0.17      0.30     21038
      Placed       0.23      0.97      0.37      5283

    accuracy                           0.33     26321
   macro avg       0.59      0.57      0.33     26321
weighted avg       0.81      0.33      0.31     26321
```

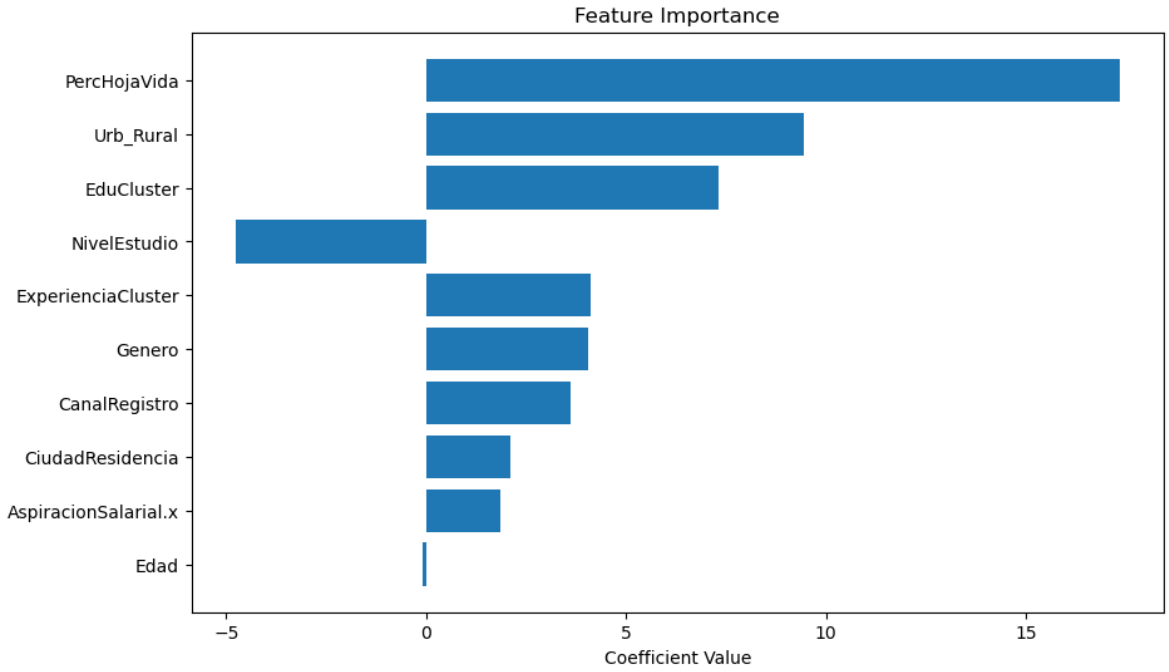
Además, a nivel general, el desempeño de este modelo también se mostró por debajo del de las otras propuestas, obteniéndose un score AUC de 0,75, como se puede apreciar en la Figura 11, la cual también denota una tendencia poco ajustada hacia la esquina superior izquierda.

Figura 11. Curva ROC para el modelo entrenado por medio de Regresión Logística.



A continuación, se presenta el cálculo de orden de importancia de los atributos en la Figura 12, bajo el cual, si bien se sabe que el cálculo de este coeficiente para regresión logística se basa en la magnitud de los coeficientes de la regresión lineal estimada, denotando la orientación de la relación entre la etiqueta de clase y cada atributo (por ejemplo, puede notarse que existe una relación inversamente proporcional entre el ascenso en el nivel de estudio y la probabilidad estimada de colocación), no es posible obtener un análisis claro de las variables de mayor relevancia al caracterizar el mercado laboral considerado en el dataset.

Figura 12. Orden de importancia de atributos para el modelo entrenado por medio de Regresión Logística.



La apreciación general respecto a los fallos de este modelo puede resumirse en que el dataset provisto implica una estructura de correlación no lineal entre los atributos y la etiqueta de clase, lo cual impide un ajuste correcto desde la optimización de coeficientes del método por regresión logística.

Naive Bayes

Si bien el clasificador construido por medio del algoritmo Naive Bayes no posee hiperparámtros fundamentales a optimizar, éste se entrenó optimizando el espacio de

hiperparámetros a través de RandomizedSearchCV (50 iteraciones), explorando un parámetro de regularización sobre la varianza de los atributos, con el fin de prevenir el sobre ajuste y posibles problemas operativos de división por cero. Este esquema de afinamiento se presenta en la siguiente tabla.

Tabla 5. Espacio de afinamiento de hiperparámetros para el entrenamiento del modelo por medio de Naive Bayes.

Hiperparámetro	**Search space**
Variance smoothing parameter (base 10, 100 valores)	(0,-9)

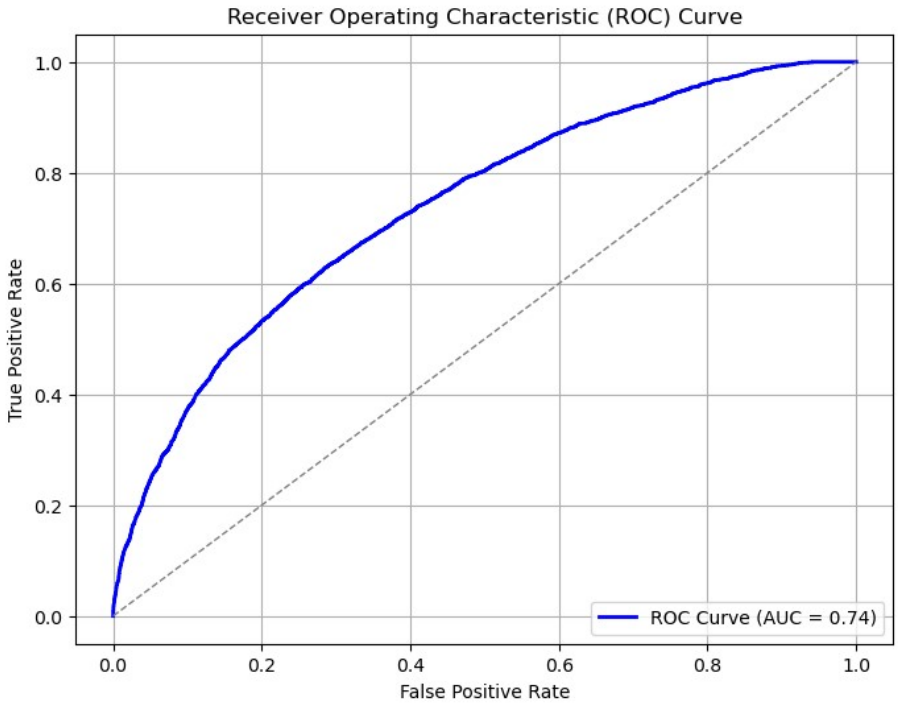
Con un parámetro óptimo regularización del orden 10^{-2}, los resultados del modelo entrenado presentaron una mejora significativa con respecto a la regresión logística, lo anterior en términos de las métricas de evaluación promedio (una precisión media ponderada de 0,81 bajo el umbral estándar de 50%) y de forma puntual con respecto a cada clase, como se observa en la Tabla 6.

Tabla 6. Reporte de clasificación (umbral de 50%), modelo por medio de Naive Bayes.

```
Fitting 5 folds for each of 50 candidates, totalling 250 fits
Best Parameters:
 {'classifier__var_smoothing': 0.01}
Classification Report:
               precision    recall  f1-score   support

  Not Placed       0.83      0.95      0.89     21038
      Placed       0.56      0.24      0.33      5283

    accuracy                           0.81     26321
   macro avg       0.69      0.59      0.61     26321
weighted avg       0.78      0.81      0.78     26321
```

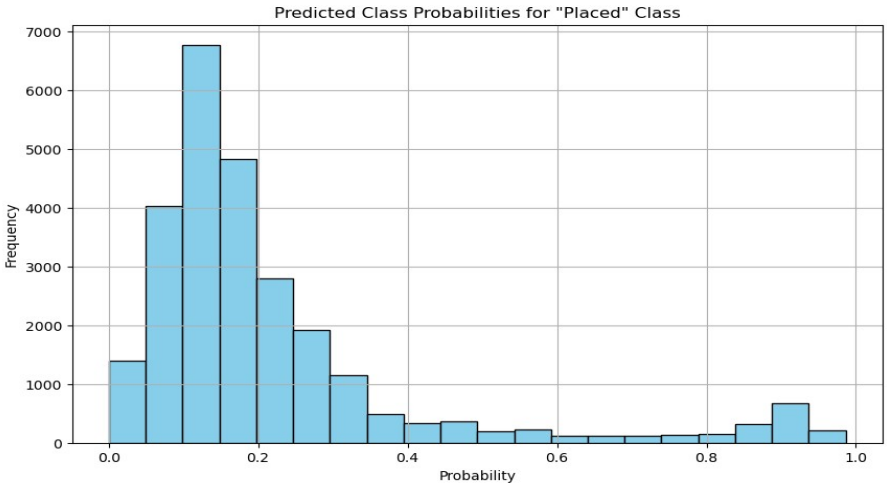
Sin embargo, a nivel general, se tiene que el nivel de ajuste de este modelo no dista mucho de los resultados obtenidos para el caso de regresión logística. Como lo muestra la Figura 13, se obtiene una curva ROC bastante similar a la del modelo anterior, con un valor AUC de 0,74.

Figura 13. Curva ROC para el modelo entrenado por medio de Naive Bayes.



Como se mencionó en el marco teórico, los modelos de clasificación de corte bayesiano considerados para este ejercicio no permiten el cálculo de una medida de importancia de atributos. Sin embargo, es posible proveer un histograma de las probabilidades posteriores estimadas para la clase positiva (colocados), el cual da cuenta de la distribución de la frecuencia con que se estima la probabilidad de que los candidatos obtengan empleos según el comportamiento estadístico de los atributos. Este gráfico se muestra a continuación.

Figura 14. Histograma de frecuencias absolutas para las probabilidades posteriores estimadas de la clase positiva, Naive Bayes.



El histograma presentado denota la verosimilitud de las predicciones realizadas por el clasificador, en tanto éstas exponen una estructura asimétrica asociada al sesgo sobre la instancia negativa de la etiqueta de clase en el conjunto de prueba del dataset (una baja propensión idiosincrática a ser colocado en un empleo), de tal forma que los candidatos con predicciones de probabilidad por encima del 50% son una proporción minoritaria.

Bayesian Ridge Regression

El clasificador por medio del modelo Bayesian Ridge Regression se entrenó optimizando el espacio de hiperparámetros a través de RandomizedSearchCV (50 iteraciones), explorando valores en escala logarítmica para las matrices de precisión (inversos de varianza) tanto de los coeficientes de regresión del modelo (parámetro alpha 1) como del ruido correspondiente a los valores de salida de la regresión lineal (parámetro alpha 2), estos parámetros se hallan relacionados con la expresión α en la ecuación (9). Además, se exploraron valores en escala logarítmica para los parámetros de regularización asociados a estas mismas matrices de precisión, siendo lambda 1 y lambda 2 parámetros asociados con el componente λ en la ecuación (9). Se insertan estos parámetros con el fin de prevenir problemas de sobre ajuste. Este esquema de afinamiento se presenta en la siguiente tabla.

Tabla 7. Espacio de afinamiento de hiperparámetros para el entrenamiento del modelo por medio de Bayesian Ridge Regression.

Hiperparámetro	**Search space**
Distribución previa de alpha 1 (base 10, 13 valores de muestra)	(-6,6)
Distribución previa de alpha 2 (base 10, 13 valores de muestra)	(-6,6)
Fuerza de regularización lambda 1 (base 10, 13 valores de muestra)	(-6,6)
Fuerza de regularización lambda 2 (base 10, 13 valores de muestra)	(-6,6)

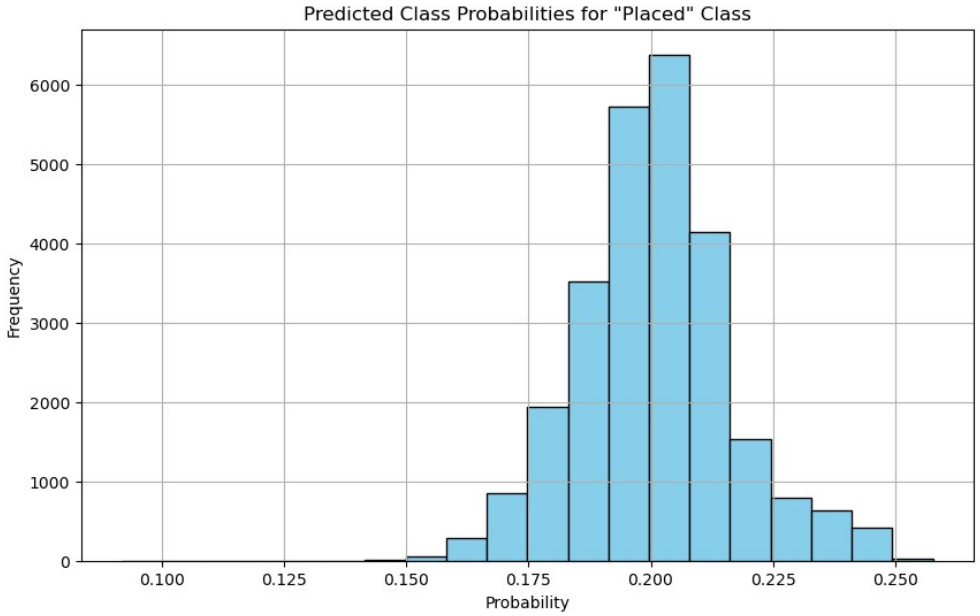
Así, el resultado del afinamiento a través de RandomizedSearchCV otorgó los valores 10^3, 10^3, 10^6, y 10^1, respectivamente, para los hiperparámetros considerados. Los resultados de este ejercicio de clasificación denotan el peor desempeño de entre los modelos considerados, en tanto las métricas de evaluación para el umbral estándar del 50% reflejan predicciones bastante deficientes y excesivamente sesgadas hacia la instancia negativa de la etiqueta de clase, como se muestra en la siguiente tabla.

Tabla 8. Reporte de clasificación (umbral de 50%), modelo por medio de Bayesian Ridge Regression.

```
Best Parameters:
 {'classifier__lambda_2': 10.0, 'classifier__lambda_1': 1000000.0, 'cl
assifier__alpha_2': 1000.0, 'classifier__alpha_1': 10000.0}
Classification Report:
              precision    recall  f1-score   support

  Not Placed       0.80      1.00      0.89     21038
      Placed       0.00      0.00      0.00      5283

    accuracy                           0.80     26321
   macro avg       0.40      0.50      0.44     26321
weighted avg       0.64      0.80      0.71     26321
```

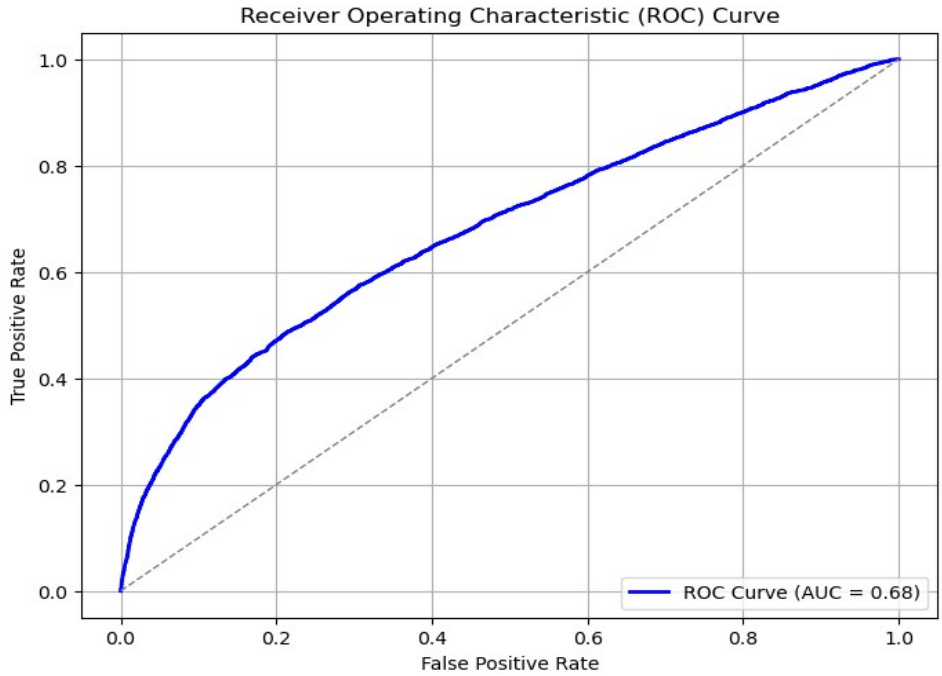
El pronunciado sesgo de las predicciones de este modelo hacia la instancia negativa de la etiqueta de clase puede verse con mayor claridad en la Figura 15, la cual presenta el histograma de probabilidades posteriores predichas. Nótese cómo la frecuencia para predicciones por encima del 50% de probabilidad es nula.

Figura 15. Histograma de frecuencias absolutas para las probabilidades posteriores estimadas de la clase positiva, Bayesian Ridge Regression.



A nivel general, la curva ROC refuerza la idea de descartar este modelo, pues tanto el ajuste visto gráficamente, como un valor AUC de 0,68, reflejan un pésimo desempeño para las predicciones de este clasificador, como se muestra en la Figura 16.

Figura 16. Curva ROC para el modelo entrenado por medio de Bayesian Ridge Regression.



Nótese que el desempeño deficiente de los modelos bayesianos explorados puede deberse al hecho de que los atributos del dataset considerado se hallan altamente correlacionados entre sí, esta desventaja de los modelos bayesianos se mencionó en el marco teórico.

Random Forest

Los métodos de ensamble basados en árboles, como se explicó en el marco teórico, poseen espacios de hiperparámetros de mayor complejidad, lo cual permite que estos modelos puedan captar estructuras de correlación más allá de desarrollos basados en ecuaciones lineales, como es el caso de las propuestas vistas anteriormente. A continuación, la Tabla 9 relaciona el espacio de hiperparámetros diseñado para el clasificador empleado por medio de Random Forest.

Tabla 9. Espacio de afinamiento de hiperparámetros para el entrenamiento del modelo por medio de Random Forest.

Hiperparámetro	**Search space**
Cantidad de estimadores (árboles)	(5, 10, 20, 40, 80, 100, 200, 300, 400, 500, 600)
Profundidad máxima	(ninguna, 1 a 10)
Mínimo de muestras para división de ramas internas	(2, 5, 10)
Mínimo de muestras por nodos hoja	(1, 2, 4)
Uso de bootstrapping	(Sí, No)

En general, el diseño del espacio de hiperparámetros para este modelo se realizó siguiendo especificaciones basadas en evidencia previa, de tal forma que el espacio de hiperparámetros considere el comportamiento de sensibilidad esperada con respecto a parámetros clave, y el contexto del dataset empleado.

El valor de cantidad de estimadores se postula como el hiperparámetro de mayor importancia en Random Forest, existiendo un consenso sobre la afirmación de que la precisión de las predicciones del modelo incrementa con la cantidad de árboles fijados [28]. Así, se determinó como espacio de búsqueda la secuencia consignada en la Tabla 9, con el fin de reflejar granularidad para cantidades bajas de estimadores, y evaluar la sensibilidad en la precisión del modelo en cantidades con incrementos más altos.

La profundidad máxima es considerada como el segundo hiperparámetro en orden de importancia para los métodos basados en árboles, en tanto es el mecanismo principal para regular la complejidad de los modelos entrenados. En ese sentido, se optó por apegarse al contexto del dataset empleado, apelando a la regla estandarizada de considerar la profundidad natural como una posibilidad, y fijar la profundidad máxima en un valor igual a la cantidad de atributos predictores en el dataset [29]. De igual forma, los hiperparámetros de mínimo de muestras para división de ramas internas y nodos hoja son

factores de regulación de complejidad de los árboles entrenados, por lo cual se fijaron escalas crecientes para establecer diversidad en los niveles de complejidad posible para la estimación óptima del modelo final [29].

Adicionalmente, se consideró si el modelo empleado utiliza el método de estimación con bootstrapping, en el cual las muestras consideradas pueden ser extraídas múltiples veces. Nótese que, en el trabajo de [30], se concluye que la eficiencia de Random Forest al emplear bootstrapping sobre datasets con variables categóricas de abundantes instancias, como es el caso de esta investigación, decrece significativamente.

Lo anterior se alinea con los hiperparámetros óptimos tras la ejecución del algoritmo por medio de RandomizedSearchCV, como se muestra en la Tabla 10, junto a los demás hiperparámetros óptimos y el reporte de clasificación para el umbral estándar de 50%.

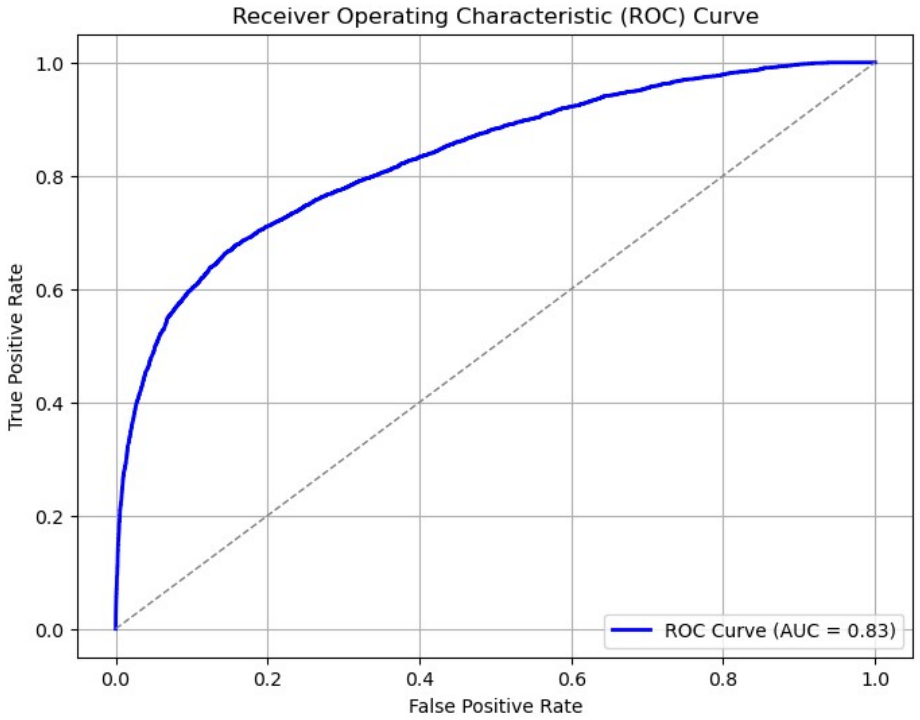
Tabla 10. Reporte de clasificación (umbral de 50%), modelo por medio de Random Forest.

```
Fitting 5 folds for each of 50 candidates, totalling 250 fits
Best Parameters:
 {'classifier__n_estimators': 500, 'classifier__min_samples_split': 2,
'classifier__min_samples_leaf': 4, 'classifier__max_depth': None, 'cla
ssifier__bootstrap': False}
Classification Report:
              precision    recall  f1-score   support

  Not Placed       0.91      0.85      0.88     21038
      Placed       0.53      0.67      0.59      5283

    accuracy                           0.81     26321
   macro avg       0.72      0.76      0.74     26321
weighted avg       0.83      0.81      0.82     26321
```

Se aprecia que los hiperparámetros óptimos obtenidos se alinean con la bibliografía de referencia, en tanto se determina una cantidad elevada de árboles (500), se fija evitar el uso de bootstrapping, y se determinan valores mínimos de muestras para división de ramas y hojas diferentes a los extremos de sus correspondientes espacios de búsqueda. Nótese que el óptimo de la profundidad máxima es la profundidad natural del árbol, lo cual puede estar relacionado con la elevada cantidad de instancias de los atributos cluster obtenidos para loso títulos educativos y experiencia laboral previa de los candidatos, haciendo indispensable el reflejar una alta complejidad en los árboles estimados para obtener una precisión óptima.

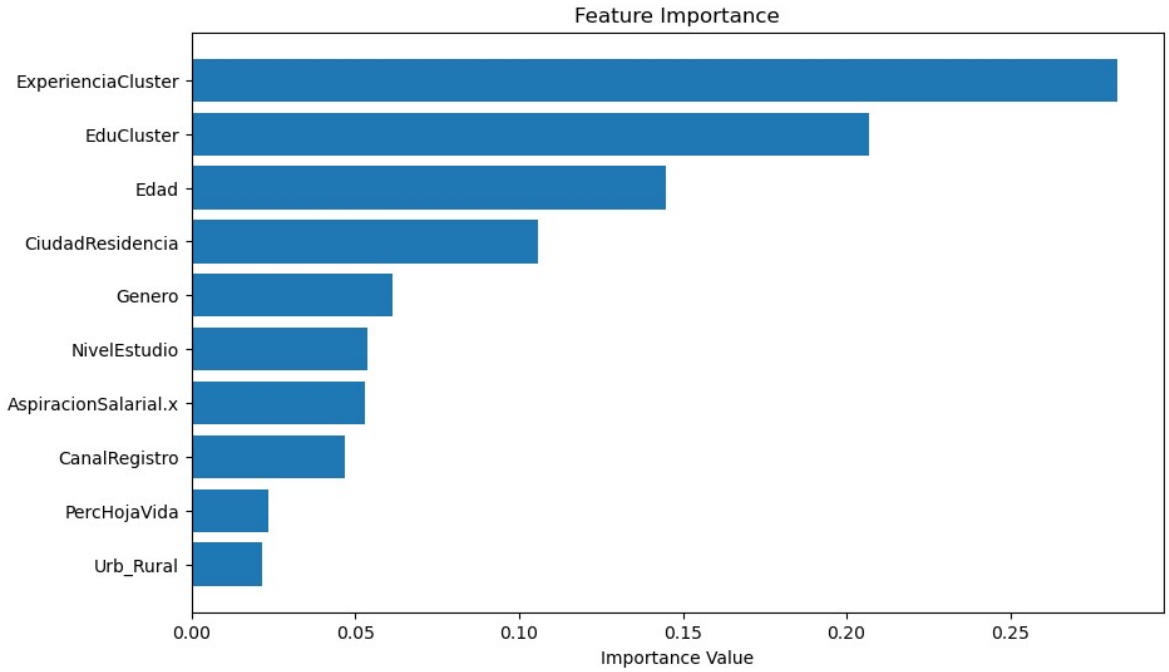
A nivel general, el modelo estimado por medio de Random Forest otorga el mayor nivel de ajuste en las predicciones para todos los umbrales posibles, con un valor AUC de 0,83. Esto se muestra en la Figura 17.

Figura 17. Curva ROC para el modelo entrenado por medio de Random Forest.



Además, se tiene que la eficacia de los resultados del modelo en términos de métricas de evaluación se refuerza con su cálculo de importancia de parámetros, el cual refleja de manera verosímil cómo los atributos cluster asociados a palabras clave de título educativo y experiencia laboral previa son los factores de mayor importancia para la clasificación y caracterización del mercado laboral caucano de acuerdo al dataset obtenido. Esto se muestra en la Figura 18.

Figura 18. Orden de importancia de atributos para el modelo entrenado por medio de Random Forest.



Nótese que, si bien se sabe que el cálculo del valor de importancia de atributos para Random Forest suele sesgarse sobre variables categóricas con bastantes instancias, en tanto éstas determinan mayores niveles de reducción en entropía/Gini en esquemas de árboles, no es el caso de este ejercicio, pues el atributo de Edad es una variable numérica y se postula como tercer atributo en el orden de importancia. Lo anterior se refuerza con el hecho de que, como se explicó previamente, la importancia de ExperienciaCluster y EduCluster para las predicciones de los ejercicios de modelación es crítica, en tanto estos atributos presentan el mayor efecto de diferencia entre los candidatos pertenecientes al dataset.

Es importante resaltar que el cálculo de orden de importancia de atributos para los métodos por árboles difiere de aquel expuesto para el caso de regresión logística, pues en el caso de los métodos basados en árboles, el valor de importancia de los atributos parte de la reducción media de entropía/Gini como un valor absoluto, mientras que, en regresión logística, el valor de importancia de atributos se asocia al valor del coeficiente estimado para estos en la forma lineal óptima del logit entrenado, razón por la cual sólo se encuentran valores positivos en la Figura 18, y existen valores negativos en la Figura 12.

El orden de importancia obtenido para los métodos basados en árboles constituye una parte clave del ejercicio de caracterización del mercado laboral que motiva esta investigación, esta discusión se ampliará en el capítulo de resultados.

Gradient Boosting

El diseño del espacio de afinamiento de hiperparámetros para Gradient Boosting se realizó de manera similar al caso de Random Forest, en tanto éste también es un método por medio de árboles. Así, se apeló a ejercicios de modelación previos en referencias bibliográficas, y se tuvo en cuenta el contexto del dataset empleado. El espacio de hiperparámetros resultante se consigna en la Tabla 11.

Tabla 11. Espacio de afinamiento de hiperparámetros para el entrenamiento del modelo por medio de Gradient Boosting.

Hiperparámetro	**Search space**
Cantidad de etapas de boosting (árboles)	(5, 10, 20, 40, 80, 100, 200, 300, 400, 500, 600)
Tasa de aprendizaje (shrinkage)	(0.01, 0.05, 0.1, 0.2)
Profundidad máxima	(ninguna, 1 a 10)
Mínimo de muestras para división de ramas internas	(2, 5, 10)
Mínimo de muestras por nodos hoja	(1, 2, 4)

Como puede apreciarse, se retoman los rangos para el espacio de hiperparámetros considerados para Random Forest, con excepción del uso de bootstrapping (el cual no puede emplearse en la metodología de aprendizaje de Gradient Boosting), en tanto la regulación de la complejidad de los árboles a estimar se fija bajo los mismos criterios anteriores. Adicionalmente, se fija como parámetro clave el rango relacionado para la tasa de aprendizaje, pues, como se explica en [31], se debe considerar el rango estándar para la tasa de aprendizaje bajo cantidades altas de árboles (0.05 a 0.1), así como valores por encima y por debajo de este rango, con el objetivo de evaluar la sensibilidad del modelo ante estas variaciones.

Bajo el umbral estándar de 50%, el modelo entrenado por medio de Gradient Boosting otorga resultados ligeramente mejores que para el caso de Random Forest, como lo muestran las métricas de evaluación en la Tabla 12.

Tabla 12. Reporte de clasificación (umbral de 50%), modelo por medio de Gradient Boosting.

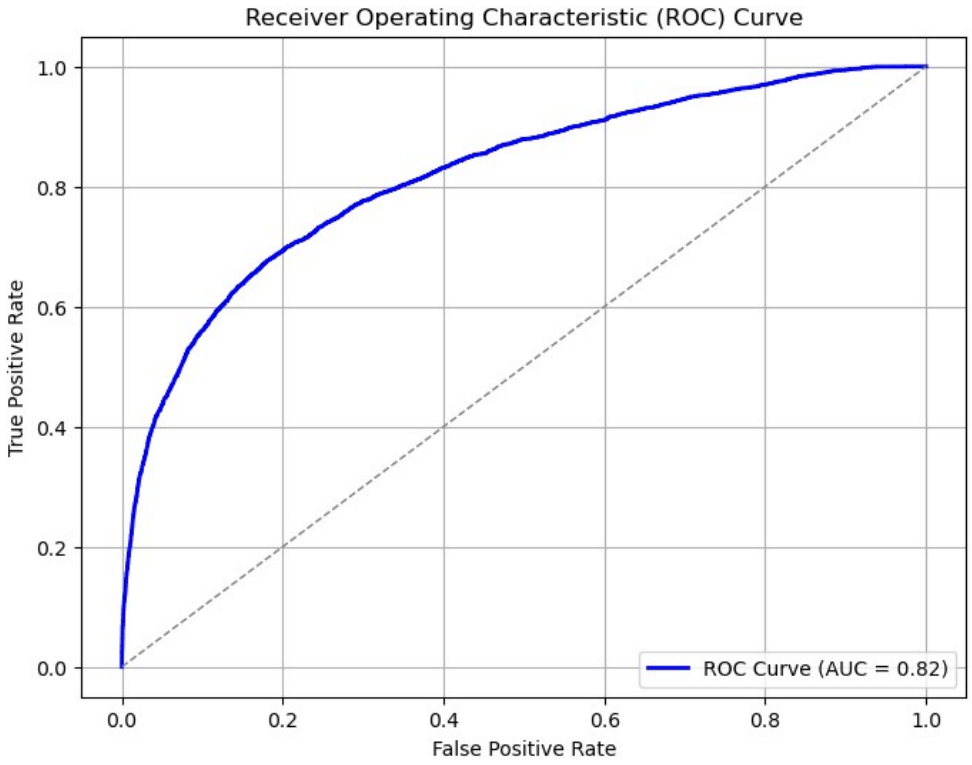
```
Fitting 5 folds for each of 50 candidates, totalling 250 fits
Best Parameters:
 {'classifier__n_estimators': 600, 'classifier__min_samples_split': 2,
'classifier__min_samples_leaf': 1, 'classifier__max_depth': 8, 'classi
fier__learning_rate': 0.05}
Classification Report:
              precision    recall  f1-score   support
```

Not Placed	0.87	0.95	0.91	21038
Placed	0.69	0.43	0.53	5283
accuracy			0.85	26321
macro avg	0.78	0.69	0.72	26321
weighted avg	0.83	0.85	0.83	26321

Nótese cómo algunos de los hiperparámetros óptimos distan de la configuración análoga para el caso de Random Forest, pues se fija un número más alto de árboles, una menor cantidad mínima de muestras para la determinación de hojas, y una profundidad máxima mucho menor (8 niveles de división). Con una tasa de aprendizaje dentro de las referencias consideradas, se tiene que el modelo entrenado por medio de Gradient Boosting obtiene resultados muy similares al caso de Random Forest, pero entrenando una estructura de árboles de menor complejidad.

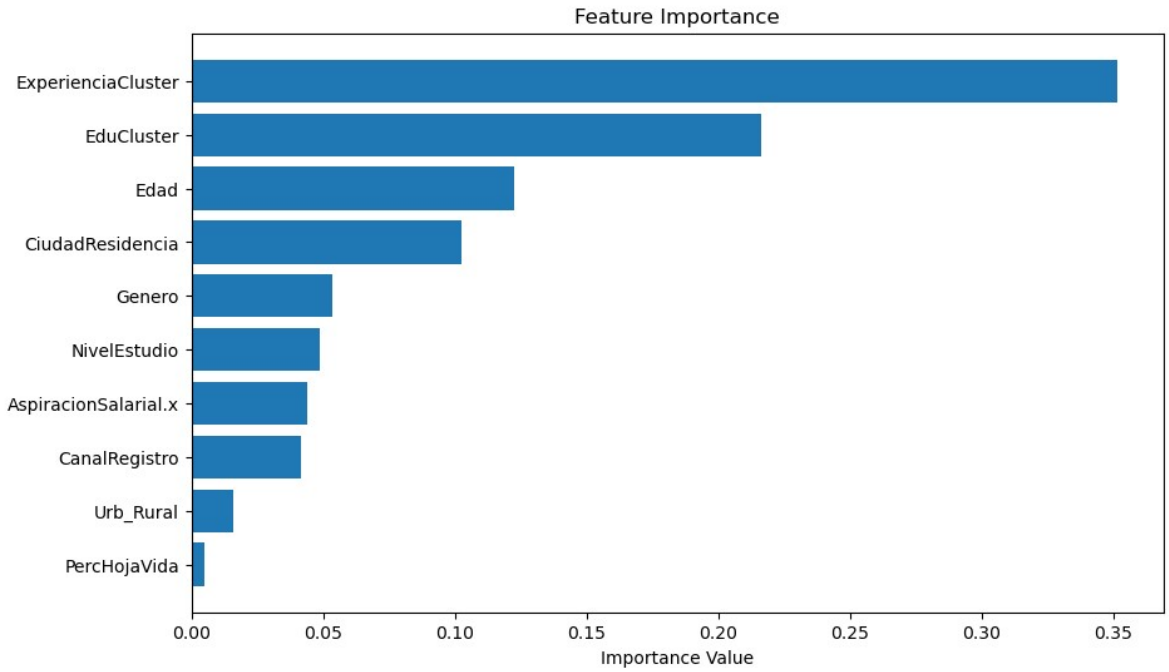
A nivel general, el desempeño de Gradient Boosting se halla un punto por debajo del de Random Forest, obteniéndose un valor AUC de 0,82, con una curva ROC bastante similar a la del caso mencionado. Lo anterior se presenta a continuación.

Figura 19. Curva ROC para el modelo entrenado por medio de Gradient Boosting.



La similitud en los resultados de ambos modelos también puede apreciarse en el orden de importancia de atributos calculado, como lo muestra la Figura 20.

Figura 20. Orden de importancia de atributos para el modelo entrenado por medio de Gradient Boosting.



Finalmente, debe mencionarse que los resultados sobresalientes provistos por los métodos basados en árboles refuerzan la hipótesis previa de su idoneidad para el dataset empleado, en tanto la estructura de evaluaciones condicionales sucesivas de Random Forest y Gradient Boosting es altamente compatible con un conjunto de atributos en el cual predominan las variables categóricas con bastantes instancias.

4. EVALUACIÓN DE EFICIENCIA Y SELECCIÓN DE MEJORES MODELOS

El conjunto de ejercicios de modelación desplegados en el capítulo anterior se caracterizó por la predominancia del paradigma frecuentista. Particularmente, los resultados obtenidos favorecen de forma indiscutible a los métodos basados en árboles, los cuales se hallaron muy por encima de aquellos de los demás modelos candidatos, tanto a nivel de sus valores en las métricas de evaluación, como en un cálculo de orden de importancia de atributos correspondiente con las hipótesis previas formuladas en el análisis exploratorio, y con las perspectivas informales de los miembros de la organización.

En términos del contraste de paradigmas propuesto, se puede afirmar con confianza que la superioridad del paradigma frecuentista para este problema de clasificación yace en su capacidad predictiva, así como en la posibilidad de calcular y monitorear la relevancia de cada atributo del dataset empleado.

Ahora, si bien las persistentes deficiencias de las propuestas bayesianas facilitan la selección del paradigma frecuentista de cara al contexto del problema de clasificación resuelto, resulta necesario ampliar la justificación de por qué los métodos basados en árboles se presentan como modelos idóneos, relacionando las soluciones obtenidas con el problema de negocio que suscita esta investigación.

4.1 La eficiencia técnica de los métodos basados en árboles.

La justificación de la idoneidad de los modelos basados en árboles se vio claramente expuesta en el capítulo anterior, en tanto sus resultados en términos de métricas de evaluación son superiores con gran diferencia a los de los modelos del paradigma frecuentista y de regresión logística.

Este grado de predominancia puede observarse en las métricas de evaluación estándar bajo el umbral de 50% y a nivel general desde la curva ROC y el valor AUC, teniéndose que para ambos niveles de evaluación los modelos Random Forest y Gradient Boosting son bastante similares. La Tabla 13 resume el nivel de ajuste de las predicciones de los modelos contrastados a partir de sus correspondientes valores AUC.

Tabla 13. Comparación de desempeño de los modelos contrastados (valores AUC).

Modelo de clasificación	Desempeño (Valor AUC)
Regresión Logística	0,75
Naive Bayes	0,74
Bayesian Ridge Regression	0,68
Random Forest	**0,83**
Gradient Boosting	0,82

Si bien la selección de los modelos por medio de árboles es indiscutible para este ejercicio, debe recalcarse que para el contexto del dataset empleado y los resultados a presentar en el próximo capítulo, se decidió optar por Random Forest, en tanto sus predicciones poseen un mayor nivel de ajuste por una mínima diferencia con respecto a Gradient Boosting.

Nótese que, dada la estructura de los datos de candidatos al mercado laboral en manos de la Agencia, el ejercicio de clasificación se halla estrechamente relacionado con la capacidad de agrupar la información en texto de títulos obtenidos y experiencias laborales previas en clusters bien definidos, por lo cual se presume que la predominancia de Random Forest se da en correspondencia con la complejidad que emerge de estos dos atributos con múltiples instancias categóricas (EduCluster y ExperienciaCluster), generando predicciones ligeramente superiores a las de Gradient Boosting.

Ahora, de cara a la implementación práctica de un clasificador por parte de la organización dentro de la cadena de valor de la Ruta de Atención, visto como una sofisticación de la modelación realizada en este trabajo, surgen un conjunto de ventajas que resaltan la eficiencia técnica de los métodos por árboles dentro de este contexto práctico.

Una de las fortalezas de los métodos por árboles con respecto a las demás propuestas, con la excepción de regresión logística, es su interpretabilidad. Se tiene que la idea central de los clasificadores basados en arboles de decisión puede ser menos hostil para el público no especializado a comparación, por ejemplo, del Teorema de Bayes. Este hecho puede facilitar la difusión superficial información sobre el funcionamiento e implementación del clasificador para los supervisores competentes en la cadena de valor, e incluso para el personal propenso a emplearlo.

La eficiencia en términos computacionales asociados a tiempos de ejecución y recursos disponibles favorece plenamente a Random Forest, en tanto su estimación paralela de árboles de decisión permite una mayor fluidez con respecto a Gradient Boosting, en el cual la corrección progresiva de los errores de predicción de clasificadores débiles implica mayores tiempos de ejecución y consumo de recursos.

Finalmente, la precisión de los métodos por árboles, ampliamente expuesta en el capítulo anterior, va de la mano con la posibilidad de monitorear la caracterización obtenida del mercado laboral captado, tanto en términos de relevancia de atributos, como de candidatos y profesiones con mayores (o menores) probabilidades estimadas de colocación.

4.2 La eficiencia económica y organizacional de los métodos por árboles en el contexto de la Ruta de Atención de la Agencia de Empleo.

Considerando la estructuración hipotética de un clasificador como herramienta de remisión y preselección al interior de la cadena de valor de la Ruta de Atención de la Agencia de Empleo de Comfacauca, se pueden vislumbrar un conjunto de ventajas críticas para incrementar la eficiencia económica y organizacional de este proceso, las cuales emergen de las propiedades de los métodos por árboles como modelos predilectos en esta investigación.

Como se mencionó anteriormente, Random Forest destaca tanto por su grado de precisión predictiva, como por su eficiencia en relación a Gradient Boosting en términos de costos computacionales de tiempo y recursos. En términos económicos, lo anterior implica una profunda reducción en los tiempos de ejecución del subproceso de remisión de candidatos, en tanto se obtiene un criterio automático y transparente para el empalme entre las necesidades de la demanda y la disposición de la oferta desde la probabilidad estimada de colocación en el mercado laboral. Esto implica una reducción significativa en el costo financiero de largo plazo de la ejecución del proceso completo, además de sugerir una estructuración alternativa del personal, que puede alinearse con funciones adicionales para el cumplimiento de la Misión de la Agencia. Lo anterior puede verse relacionado con los incrementos de eficiencia provistos por la inserción de herramientas analíticas al interior de cadenas de valor y dinámicas organizacionales, como se expone en [12].

Nótese que el grado de transparencia que otorgan las predicciones por medio del clasificador sugerido como una sofisticación de este trabajo, implican una profunda mejora en la eficiencia organizacional del proceso analizado, asegurando la pre selección y remisión de candidatos bajo criterios estadísticos rigurosos, lo cual también hace del clasificador propuesto una herramienta para el control de calidad de la Ruta de Atención.

La simplicidad y eficiencia computacional relativas de Random Forest también implican la facilidad en su potencial implementación como una pipeline alimentada por una estructura ETL o dispositivos similares, la cual pueda proveer a los colaboradores de la agencia una herramienta actualizada y fluida para el proceso de remisión, así como para la evaluación posterior del desempeño de la Ruta de Atención, el monitoreo del comportamiento del

mercado laboral captado por la organización, y la posterior toma de decisiones basadas en el mismo.

Frente a lo anterior, es crítico resaltar el papel de la capacidad de los métodos por árboles para evaluar con precisión la relevancia de los atributos en la predicción de colocaciones, así como de ordenar a los candidatos en función de probabilidades estimadas de colocación y filtros específicos asociados a atributos claves del contexto del problema de negocio.

5. RESULTADOS Y DISCUSIÓN

Dado el problema de investigación formulado y la solución propuesta a través de los objetivos específicos de este proyecto aplicado, resta realizar la presentación de los resultados obtenidos tras el ejercicio de modelación desde el contraste entre los paradigmas frecuentista y bayesiano. En este sentido, este capítulo despliega los resultados obtenidos desde los criterios planteados, en dos fases. En la primera fase, se expone la caracterización del mercado laboral obtenida a partir de la implementación del mejor modelo (Random Forest) sobre el dataset empleado. En la segunda fase, se realiza una descripción de los impactos potenciales de la implementación estructurada de un modelo de clasificación para mejorar la eficiencia multidimensional, tanto de la cadena de valor correspondiente a la Ruta de Atención de la Agencia de Empleo, como para la toma de decisiones de la Comfacauca en general.

5.1 Caracterización del mercado laboral captado por la Agencia de Empleo.

El conjunto de resultados provistos por el modelo Random Forest, anteriormente justificado como la mejor opción de modelación obtenida para este proyecto, permite caracterizar el mercado laboral reflejado en el dataset empleado desde la identificación de los factores de mayor influencia en la probabilidad de obtención de empleo en el departamento del Cauca, permitiendo una descripción directa de las propiedades de la relación entre la oferta y la demanda laboral desde la información disponible. Así, esta sección consiste en la discusión de los hallazgos obtenidos considerando el papel de los atributos de mayor relevancia desde el clasificador construido, como lo expone la Figura 20. Nótese que, dada la suficiencia de las relaciones descritas en la fase de análisis exploratorio, se opta por realizar el análisis propuesto para los atributos del dataset desde los cuales se pueden exponer relaciones críticas, respaldados por sus niveles de importancia para el clasificador, estas variables son el municipio de residencia, la edad, el género, la aspiración salarial, el título educativo obtenido, y el cargo de última experiencia laboral de los candidatos.

Municipios de residencia por probabilidad estimada de colocación

Los resultados obtenidos para la relación entre la probabilidad estimada de colocación en el mercado laboral caucano y el municipio de residencia de los candidatos se apegan a las hipótesis previas formuladas en la fase de análisis exploratorio, en tanto logra denotarse la profunda estructura asimétrica de la distribución del empleo en el departamento, como lo muestran los resultados descritos en la Tabla 14.

Tabla 14. Análisis de probabilidad estimada de colocación por municipios de residencia de los candidatos del mercado laboral, ordenados por coeficiente de fiabilidad (probabilidad media vs cantidad de observaciones).

CiudadResidencia	Avg_Probability	Std_Deviation	N_Observations	Q1	Q3	CI_95_low	CI_95_high	IQR	Reliability_Score
PUERTO TEJADA	0,466507217	0,336594255	7927	0,137238	0,804414	0,459097385	0,473917048	0,667176	4,188315756
SANTANDER DE QUILICHAO	0,436290137	0,304099204	14219	0,153463	0,728269	0,43129167	0,441288604	0,574806	4,171952175
PRADERA	0,851262405	0,266561441	95	0,9186	0,993789	0,797659097	0,904865714	0,075189	3,876544196
FLORIDA	0,640617793	0,349217725	308	0,292004	0,9915	0,601616681	0,679618906	0,699496	3,670803878
GUACHENÉ	0,44882045	0,330344806	2300	0,139128	0,782089	0,435319646	0,462321254	0,642962	3,474168479
VILLA RICA	0,424893572	0,326728486	2390	0,126629	0,74136	0,411794391	0,437992752	0,614731	3,305267764
PALMIRA	0,650741082	0,362607119	149	0,285251	0,994739	0,592517455	0,70896471	0,709488	3,256273436
PIAMONTE	0,491133029	0,316704875	661	0,200235	0,8023	0,466988981	0,515277077	0,602065	3,189296992
CALOTO	0,410507869	0,316119773	1768	0,126071	0,726463	0,39577231	0,425243428	0,600392	3,069615383
POPAYÁN	0,279157397	0,262536246	42562	0,068377	0,422343	0,276663179	0,281651615	0,353966	2,975459726
CANDELARIA	0,550276176	0,358707121	115	0,192287	0,935854	0,484714938	0,615837415	0,743567	2,611023108
MIRANDA	0,31327524	0,284008076	1190	0,077131	0,510908	0,297138593	0,329411886	0,433777	2,218523955
CALI	0,311853619	0,256342869	955	0,106145	0,458289	0,295595301	0,328111937	0,352144	2,139849514
PADILLA	0,327577276	0,293858105	486	0,083115	0,512514	0,301451115	0,353703438	0,429399	2,026461373
JAMUNDÍ	0,278757757	0,236532073	236	0,098645	0,374457	0,248579781	0,308935733	0,275811	1,523085499
PATÍA	0,212122949	0,226813649	1050	0,055391	0,272825	0,198403692	0,225842207	0,217434	1,475642937
SANTA ROSA	0,261220615	0,262651861	262	0,072367	0,354203	0,22941631	0,29302492	0,281836	1,454566375
CORINTO	0,194303006	0,218598677	423	0,043482	0,267057	0,173470888	0,215135124	0,223575	1,175022593
BOGOTÁ, D.C.	0,230466665	0,259076924	146	0,05804	0,297891	0,188441602	0,272491729	0,239851	1,148555199
TIMBÍO	0,157906033	0,199469988	1414	0,033363	0,186818	0,147509012	0,168303054	0,153455	1,145478446
SOTARA	0,237145654	0,30358849	102	0,039899	0,277962	0,178228564	0,296062744	0,238063	1,096792204
CALDONO	0,176691709	0,213383224	495	0,04203	0,215662	0,157893618	0,195489799	0,173632	1,096293912
INZÁ	0,210805493	0,242357344	174	0,046961	0,241734	0,17479429	0,246816696	0,194772	1,087557194
TOTORÓ	0,181992782	0,229787253	226	0,036158	0,230845	0,15203375	0,211951815	0,194687	0,986498246
PIENDAMÓ	0,131440261	0,177781624	1212	0,028144	0,157145	0,121431239	0,141449282	0,129	0,933229422
BUENOS AIRES	0,170316756	0,207819746	227	0,035041	0,213414	0,143281533	0,197351979	0,178373	0,923959888
GUAPÍ	0,135989859	0,211555091	626	0,019577	0,130894	0,119417193	0,152562526	0,111318	0,875686351
ROSAS	0,168087169	0,225438523	160	0,027417	0,192098	0,133155108	0,20301923	0,164681	0,853071598
LA VEGA	0,15762355	0,218996703	163	0,037341	0,158182	0,124003385	0,191243716	0,12084	0,802894992
CAJIBÍO	0,122642631	0,172148738	539	0,028676	0,130255	0,108109288	0,137175974	0,101579	0,771387267
PURACÉ	0,138287022	0,178550852	208	0,030321	0,159669	0,114021685	0,162552359	0,129348	0,738112246
SILVIA	0,112489214	0,175860134	525	0,020508	0,123126	0,09744589	0,127532539	0,102618	0,704564749
EL TAMBO	0,100000125	0,156200972	605	0,014785	0,106193	0,087553199	0,112447051	0,091408	0,640523645
ARGELIA	0,120727371	0,194638373	196	0,021488	0,111657	0,093477999	0,147976743	0,090169	0,637212907
LA SIERRA	0,119400957	0,192728889	156	0,019893	0,122799	0,089156863	0,149645052	0,102905	0,602957642
SAN SEBASTIÁN	0,122681855	0,237663212	136	0,013779	0,096212	0,082738128	0,162625581	0,082433	0,602693613
TORIBÍO	0,112035826	0,147729424	208	0,011014	0,151606	0,091959168	0,132112483	0,140591	0,597995485
BOLÍVAR	0,084030601	0,159856492	737	0,013966	0,07228	0,072489357	0,095571845	0,058314	0,554819428
BALBOA	0,104330436	0,165649215	201	0,019402	0,098019	0,081429806	0,127231066	0,078617	0,553296112
MORALES	0,09715851	0,148562844	237	0,013755	0,102047	0,07824413	0,11607289	0,088292	0,531268575
JAMBALÓ	0,111038616	0,175462926	106	0,016969	0,106549	0,077635383	0,14444185	0,08958	0,517821824
SUÁREZ	0,090229388	0,108401051	188	0,014897	0,129124	0,074733708	0,105725068	0,114227	0,472480952
MERCADERES	0,072192859	0,132252126	410	0,006889	0,076608	0,059391184	0,084994535	0,069719	0,434323588
ALMAGUER	0,082565174	0,146928805	155	0,014452	0,082335	0,059434027	0,105696321	0,067883	0,416411272
TIMBIQUÍ	0,074032359	0,176132016	267	0,007263	0,05706	0,052905315	0,095159404	0,049797	0,413637201
SUCRE	0,099545173	0,139677205	63	0,023969	0,135432	0,065053733	0,134036614	0,111463	0,412429064
PÁEZ	0,073528915	0,121769465	148	0,013944	0,086294	0,053910514	0,093147316	0,07235	0,367439596
FLORENCIA	0,063066242	0,091166337	63	0,009673	0,073085	0,04055392	0,085578565	0,063412	0,261291938

Esta tabla determina la probabilidad media de colocación de los candidatos (el promedio de las probabilidades estimadas desde el modelo de clasificación para los candidatos residentes en cada municipio), y presenta los valores asociados a la descripción de su comportamiento distribucional, ordenando a cada municipio según un score de fiabilidad de la forma $ln(n_i) * ProbMedia_i$ para cada i municipio con al menos 50 observaciones en el dataset, el cual penaliza el valor de probabilidad media calculada al considerar la cantidad de observaciones con las que este promedio se calcula. Para propósitos de orientación del

lector, a los valores de probabilidad media se añade una escala descendente de azul a amarillo, y a los valores de coeficiente de fiabilidad se añade una escala descendente de rojo a amarillo.

El resultado general de la tabla permite confirmar la hipótesis previa de baja generación de empleo en la mayoría de municipios del departamento del Cauca, de tal forma que una minoría de los municipios posee probabilidades medias de colocación de cerca del 40%. Esta minoría, donde puede considerarse que existen probabilidades medias consistentes con un mayor volumen de generación de empleo, se decanta específicamente por los municipios del norte del departamento del Cauca, y algunos municipios del sur del Valle del Cauca, los cual es congruente con la presencia de la conocida dinámica industrial en la región a raíz de la zona franca decretada por la Ley Páez. Nótese la excepción del municipio de Piamonte, donde se percibe una alta tasa de colocación debido a la generación de vacantes para el enclave petrolero existente en esa región.

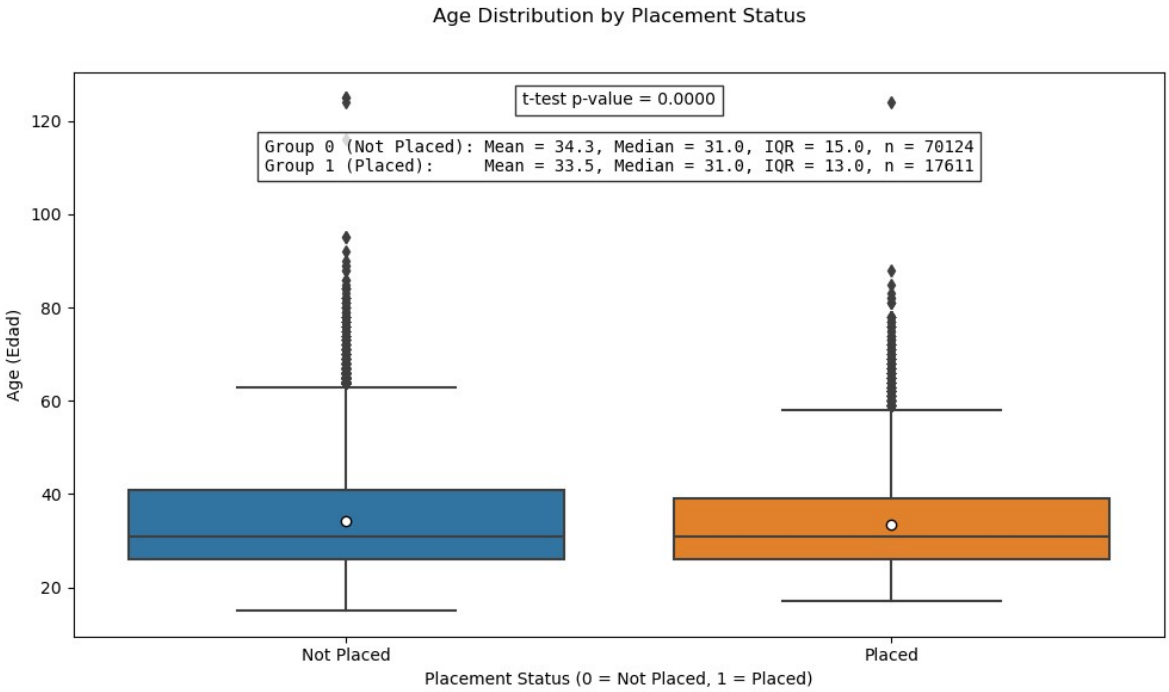
Además, la tabla calculada permite inferir la profunda situación de desempleo en Popayán como capital del departamento. Con 42.562 candidatos, Popayán es, por mucho, el municipio con la mayor cantidad de observaciones en el dataset, sin embargo, su probabilidad media de colocación se ubica sobre 27,9%. Incluso considerando su desviación estándar sobre 26,25%, se tiene que el comportamiento de la colocación en la capital se acentúa sobre una muy baja probabilidad en comparación con los municipios del Norte. Este resultado es congruente con la alta tasa idiosincrática de desempleo en la ciudad de Popayán, a raíz de una estructura sectorial inestable basada en el comercio y servicios, y afectada por la informalidad, como se estudia en [32].

En tanto la situación en la tabla empeora profundamente para los demás municipios del departamento (probabilidades estimadas de entre el 6% y el 21%), los resultados de las estimaciones del clasificador construido reflejan de manera específica cómo el grado de industrialización y generación de empleo formal se concentran sobre el norte del departamento, resaltando una muy baja dinámica económica en la región, y, considerando las muy altas probabilidades estimadas para algunos municipios del Valle del Cauca, poca competencia en la formación de la fuerza laboral apta.

Comparación distribucional de Edad para la muestra de candidatos

El atributo de edad de los candidatos en el dataset se postuló como la tercera variable en el orden de importancia del modelo óptimo de esta investigación, de tal forma que es posible inferir un comportamiento distribucional diferenciado para los candidatos con procesos exitosos de colocación versus los candidatos que permanecen por fuera de la población empleada. Este análisis se presenta en la Figura 21.

Figura 21. Comparación de distribuciones empíricas de Edad para las instancias de la etiqueta de clase del clasificador (Colocados vs No colocados).



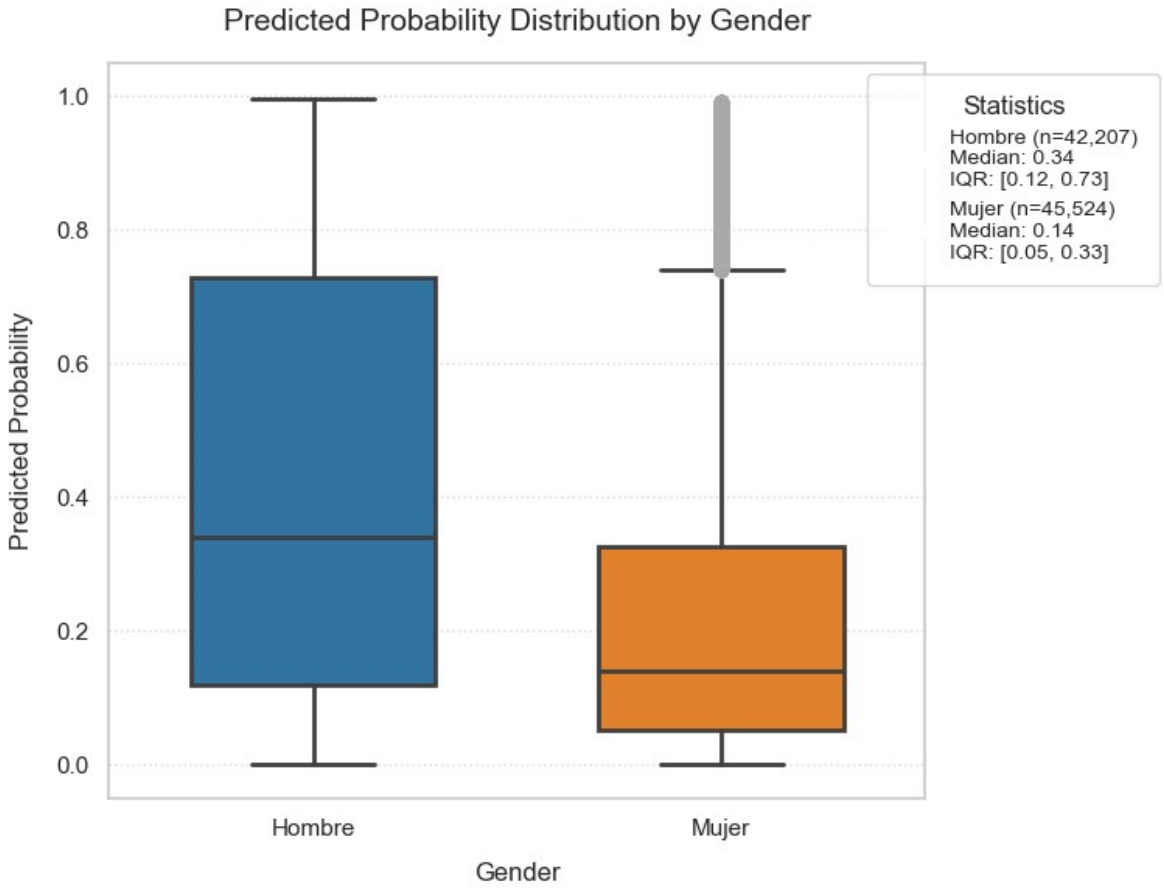
Con un nivel de significancia sobre el 0% en la prueba de hipótesis t-student, se confirma que las distribuciones empíricas de colocados y no colocados presentan un comportamiento estadístico diferenciado, de tal forma que la población con procesos exitosos de colocación es, en promedio, más joven; teniendo una edad media de 33,5 años versus los 34,3 años de los no colocados. Además, la distribución empírica para los colocados presenta menos variabilidad, en tanto su rango intercuartílico acapara 13 años (entre los 26 y los 39 años), lo cual dista al caso de los no colocados, con un rango intercuartílico de 15 años (entre los 26 y los 41 años).

Si bien este resultado permite afirmar que la generación de empleo en el departamento se concentra, en promedio, sobre población un poco más joven, existiendo significancia estadística para la diferencia entre ambos tipos de población; la similitud relativa entre las edades de colocados y no colocados permite reforzar la afirmación con respecto a la proliferación del desempleo en el departamento, en tanto puede observarse que la población en el rango de edad apto para el trabajo se entremezcla entre población empleada y desempleada.

Comparación distribucional de la probabilidad estimada de colocación por género

La hipótesis previa planteada en el análisis exploratorio para el comportamiento de la colocación con respecto al género denotó que, a pesar que existen más mujeres en la muestra, existe una mayor propensión a la colocación de hombres. Esta hipótesis se refuerza con creces en los resultados del clasificador, pues el comportamiento distribucional determinado presenta una mucho mayor probabilidad de colocación para los hombres, así lo muestra la Figura 22.

Figura 22. Comparación de distribuciones empíricas para las probabilidades estimadas de colocación (Hombres vs Mujeres).



Para los hombres, se observa una probabilidad mediana de colocación del 34%, existiendo un alto nivel de variabilidad fundamentado por un rango intercuartílico de entre el 12% y el 73%, de tal forma que la obtención de empleo es una posibilidad dentro del comportamiento estadístico regular. Sin embargo, para las mujeres, se tiene una probabilidad mediana de colocación del 14%, con un rango intercuartílico de entre el 5% y el 33%. En este sentido, se tiene que para las mujeres la obtención de empleo no se halla

dentro del comportamiento estadístico regular. Lo anterior se refuerza al observar en la Figura 22 cómo los valores de probabilidad estimada por encima del 75% son outliers.

Así, el modelo construido y sus predicciones reflejan la profunda asimetría de género existente en el mercado laboral del departamento, como también se expone en [32].

Aspiraciones salariales por probabilidad estimada de colocación

La hipótesis previa con respecto a la correlación entre la aspiración salarial de los candidatos y la colocación afirmó una tendencia a la concentración de ésta sobre aspiraciones salariales de entre 1 y 2 SMMLV. El modelo estimado respalda esta hipótesis, sin embargo, arroja conclusiones adicionales con respecto a la diversidad en aspiraciones salariales y su relación con la obtención de empleo. Así lo muestra la Tabla 15, la cual despliega los resultados correspondientes bajo la misma metodología de la Tabla 14.

Tabla 15. Análisis de probabilidad estimada de colocación por aspiraciones salariales de los candidatos del mercado laboral, ordenados por coeficiente de fiabilidad (probabilidad media vs cantidad de observaciones).

AspiracionSalarial	Avg_Probability	Std_Deviation	N_Observations	Q1	Q3	CI_95_low	CI_95_high	IQR	Reliability_Score
1 SMMLV	0,38446495	0,305507304	25286	0,113859	0,653912	0,38069932	0,388230579	0,540053	3,897708028
A convenir	0,365750609	0,312232759	14978	0,096176	0,638115	0,360750176	0,370751042	0,541939	3,516449886
1 a 2 SMMLV	0,310627253	0,278276325	34018	0,082812	0,490913	0,307670072	0,313584434	0,408101	3,241285134
2 a 4 SMMLV	0,237143179	0,253450232	7694	0,049022	0,32283	0,231479839	0,24280652	0,273809	2,122003666
4 a 6 SMMLV	0,233671024	0,265556426	1016	0,042842	0,308029	0,217341782	0,250000267	0,265187	1,617851393
Menos de 1 SMMLV	0,253782621	0,272735392	547	0,053881	0,364766	0,230926427	0,276638814	0,310885	1,599959538
6 a 9 SMMLV	0,149798811	0,154133671	173	0,038976	0,207152	0,126830422	0,1727672	0,168176	0,771956954
No registra	0,002956201	0,013830686	3941	0	0,000311	0,002524387	0,003388015	0,000311	0,02447495

Puede observarse que las definiciones de aspiraciones salariales con mayor fiabilidad, tanto en términos de tamaño de muestra como de probabilidad media estimada, corresponden a 1 SMMLV y la categoría “a convenir”, la cual denota flexibilidad del candidato con respecto a los salarios ofrecidos en las vacantes.

Así, se observa un comportamiento descrito por una marcada relación inversamente proporcional entre la aspiración salarial y la probabilidad de conseguir empleo, lo cual es esperado, en tanto una aspiración salarial más alta implica que el candidato posea niveles de formación y experiencia especializados y escasos, incrementando el salario esperado, pero disminuyendo la cantidad de puestos posibles.

Adicionalmente, nótese cómo los resultados obtenidos desde el enfoque de aspiración salarial también refuerzan la situación de alto desempleo en el departamento, pues las probabilidades medias estimadas de colocación alcanzan tan sólo un máximo del 38,4%, con un límite superior de rango intercuartílico del 65% para 1 SMMLV. Así, se tiene que la

probabilidad idiosincrática de conseguir empleo es muy baja incluso para las peores expectativas de salario.

Títulos educativos por probabilidad estimada de colocación

En este punto, se entra a discutir los resultados correspondientes a las dos variables con mayor importancia en la estimación de probabilidades de colocación. El comportamiento de los títulos educativos dentro del clasificador, obtenidos a través del agrupamiento de palabras clave en las entradas de texto, suscita la discusión de hechos interesantes que permiten añadir mayor especificidad a la situación general del mercado laboral descrita por las variables anteriores.

En primer lugar, se tiene que las palabras asociadas a los títulos con mayores probabilidades de colocación corresponden a la educación media, lo cual se alinea a los resultados de las aspiraciones salariales bajas para concluir que el grueso de la demanda de empleo en el departamento consiste en empleos requiriendo fuerza laboral con bajos niveles de formación educativa.

Los resultados obtenidos también denotan un comportamiento sectorial inclinado hacia la provisión de servicios varios, pues se hallan probabilidades de colocación altas para palabras asociadas a las disciplinas de contabilidad, enfermería, desarrollo de software, ciencias de la gestión, y, especialmente, formación de electricistas.

Nótese cómo los resultados provistos también exponen una mayor inclinación a la demanda de personal con títulos técnicos y tecnólogos por encima de candidatos con títulos obtenidos en educación superior, los cuales, en la mayoría de los casos, poseen probabilidades de colocación muy bajas. Las disciplinas con formaciones en educación universitaria más perjudicadas en términos de bajas probabilidades estimadas de colocación son derecho, psicología, administración de empresas, pedagogía, ingenierías, profesionales en el sector de construcción, y ciencias sociales. Lo anterior refuerza la conclusión de que la composición sectorial de la economía departamental es muy baja en términos de inserción de ciencia, tecnología y conocimientos especializados, lo cual es congruente con los bajos niveles de expectativas salariales hallados en el análisis anterior.

Finalmente, vale recalcar el hecho de que los candidatos que no registraron su título educativo en la plataforma de la Agencia poseen una probabilidad más alta de colocación que la de la mayoría de los títulos analizados en la tabla. Lo anterior no sólo se alinea con la expectativa de baja formación educativa de la fuerza laboral en la economía departamental, sino que resalta una mucho mayor influencia del cargo de experiencia laboral previa en la posibilidad de conseguir empleo, como se mostrará en la siguiente discusión. Los resultados

se consignan en la Tabla 16, la cual retoma la metodología de las tablas 14 y 15, considerando a los clusters que contienen más de 100 observaciones.

Tabla 16. Análisis de probabilidad estimada de colocación por títulos obtenidos (palabras agrupadas) de los candidatos del mercado laboral, ordenados por coeficiente de fiabilidad (probabilidad media vs cantidad de observaciones).

TituloObtenido	Avg_Probability	Std_Deviation	N_Observations	Q1	Q3	IQR	Reliability_Score
bachiller academico	0,439399844	0,28671577	19289	0,18439	0,694512	0,510122	4,3356858
media	0,552274042	0,303394079	807	0,277512	0,847798	0,570286	3,696548919
bachiller	0,415952612	0,311258064	3710	0,136501	0,703561	0,56706	3,418625982
enfermera	0,540995736	0,305908416	226	0,265926	0,85587	0,589944	2,932486322
bachiller tecnico	0,40835173	0,284466771	957	0,165582	0,636719	0,471138	2,802845991
auxiliar de enfermeria	0,394146117	0,267390467	1069	0,16022	0,605182	0,444962	2,748963778
tecnico en cocina	0,570839403	0,280344559	105	0,319464	0,839943	0,520479	2,656663948
tecnologo en analisis y desarrollo de sistemas de informacion	0,503732668	0,286137704	187	0,259402	0,80488	0,545479	2,635080301
bachiller tecnico ambiental	0,551810044	0,267276072	111	0,297531	0,781582	0,484051	2,598766069
tecnologo en contabilidad y finanzas	0,521513493	0,252789481	140	0,31521	0,707455	0,392245	2,577133202
tecnico electricista	0,533144304	0,245900584	124	0,315373	0,761083	0,44571	2,56990566
no registra	0,27319237	0,34495418	10620	0,00016	0,559137	0,558977	2,532628309
tecnologo en gestion del talento humano	0,521128257	0,258044441	116	0,316921	0,725572	0,408651	2,477230171
tecnologia en gestion empresarial	0,509180401	0,270981746	103	0,28138	0,729245	0,447865	2,359913166
contador publico	0,342047663	0,256419628	809	0,131738	0,5326	0,400861	2,290282372
bachiller comercial	0,416556839	0,315920278	209	0,159157	0,704114	0,544957	2,22538587
bachiller tecnico industrial	0,433432037	0,269603806	146	0,201012	0,661173	0,460161	2,16005477
contadora publica	0,359089442	0,258974004	353	0,144996	0,540509	0,395513	2,10658674
tecnologo en gestion empresarial	0,379287812	0,26809724	244	0,170415	0,546536	0,376121	2,085008909
ingeniero industrial	0,338827724	0,269218184	448	0,120358	0,508418	0,38806	2,068473194
administrador de empresas	0,307546986	0,236078154	625	0,119665	0,430666	0,311	1,979911115
bachiller tecnico agropecuario	0,364093203	0,296180396	189	0,121758	0,574363	0,452605	1,908484462
bachiller agropecuario	0,362664919	0,27762792	185	0,134907	0,587533	0,452625	1,893239924
bachillerato	0,322854597	0,297550684	291	0,09071	0,489352	0,398642	1,831658495
administracion de empresas	0,302503795	0,227785812	422	0,117836	0,420884	0,303048	1,828637048
basica primaria	0,243357591	0,248371768	1528	0,049336	0,352611	0,303274	1,784228494
tecnico en sistemas	0,237476757	0,240369051	1092	0,066326	0,31111	0,244784	1,661331858
tecnologo en entrenamiento deportivo	0,348271395	0,270712501	114	0,146504	0,53716	0,390656	1,649482442
administradora de empresas	0,257981099	0,212395121	455	0,090152	0,372748	0,282596	1,578921056
fisioterapeuta	0,266996727	0,250718724	334	0,075021	0,395465	0,320444	1,551555628
contaduria publica	0,254072276	0,220087519	387	0,080907	0,350029	0,269122	1,513870526
academico	0,283204091	0,282717143	208	0,05812	0,411552	0,353432	1,511612619
ingeniero de sistemas	0,232683537	0,247273556	369	0,054404	0,327155	0,272752	1,375345072
auxiliar en enfermeria	0,226980687	0,207694298	419	0,060225	0,336761	0,276536	1,370480089
trabajador social	0,26678092	0,241259103	156	0,095602	0,332426	0,236824	1,347205229
ingeniero civil	0,242245053	0,28720224	226	0,026428	0,362453	0,336025	1,313097787
bachiller academica	0,268111414	0,243194874	122	0,118003	0,310646	0,192643	1,288012876
basica secundaria	0,189977333	0,202449347	794	0,04066	0,256514	0,215854	1,268494511
tecnico auxiliar de enfermeria	0,23888666	0,245250375	197	0,055608	0,338845	0,283237	1,262086895
psicologo	0,231305727	0,216907622	192	0,091509	0,260753	0,169243	1,216088788
bachiller tecnico comercial	0,248137295	0,249131524	134	0,080325	0,288793	0,208468	1,215336721
ingenieria de sistemas	0,23738408	0,249782527	166	0,04202	0,34466	0,302639	1,213504519
bachiller basico	0,192518568	0,203752432	434	0,052296	0,25455	0,202254	1,169173838
tecnico en asistencia administrativa	0,207998105	0,225454825	269	0,064036	0,254344	0,190308	1,163689367
ingenieria industrial	0,240927504	0,227476146	120	0,079134	0,31993	0,240796	1,153438438
tecnico en contabilizacion de operaciones comerciales y financieras	0,212094432	0,239577796	206	0,050985	0,252599	0,201615	1,130012868
psicologa	0,168936191	0,199347818	755	0,036956	0,2139	0,176944	1,119492454
derecho	0,195993763	0,179582058	202	0,083481	0,240512	0,157031	1,04038736
ingeniera industrial	0,20061606	0,191678412	157	0,063982	0,259329	0,195347	1,01436411
abogado	0,159825563	0,213130525	481	0,023934	0,2084	0,184466	0,987061463
tecnico en auxiliar de enfermeria	0,186425573	0,20516276	187	0,042112	0,266843	0,224731	0,975212419
primaria	0,207249196	0,234139922	110	0,044241	0,258073	0,213831	0,974170775
bachiller en ciencias	0,185585762	0,204377428	129	0,054997	0,233985	0,178988	0,90191199
trabajadora social	0,145614775	0,174561411	444	0,038405	0,1818	0,143396	0,88764212
tecnico auxiliar en enfermeria	0,180991762	0,178464238	131	0,047272	0,233955	0,186683	0,882370554
normalista superior	0,148135992	0,195030696	313	0,044145	0,164997	0,120852	0,851219508
auxiliar administrativo en salud	0,17872339	0,199988417	110	0,045257	0,215056	0,169799	0,840085784
arquitecto	0,15717889	0,200458405	209	0,019805	0,232625	0,21282	0,839702169
especialista en pedagogia	0,17571418	0,206879815	113	0,04953	0,224713	0,175184	0,830669076
tecnico de sistemas	0,16796098	0,165891226	104	0,064589	0,194777	0,130188	0,780076449
trabajo social	0,157110315	0,176947884	113	0,032333	0,206604	0,174271	0,742721391
tecnico laboral en auxiliar en enfermeria	0,14929471	0,177962035	109	0,045309	0,166802	0,121493	0,700393421
licenciada en pedagogia infantil	0,135790119	0,187498781	168	0,03823	0,162009	0,123779	0,695783677
abogada	0,112901166	0,116256801	424	0,038824	0,149582	0,110758	0,683021962
psicologia	0,127337671	0,141935649	117	0,041624	0,160417	0,118793	0,606404139
tecnico en atencion integral a la primera infancia	0,102025919	0,161375994	258	0,021402	0,09835	0,076947	0,566545803
tecnico en atencion a la primera infancia	0,114645732	0,165518533	102	0,02233	0,132382	0,110053	0,530233395

Cargos de última experiencia laboral por probabilidad estimada de colocación

Como se explicó en el análisis de modelación por medio de Random Forest, esta variable se postula como la más importante en la predicción del comportamiento de colocación de los candidatos inmersos en el dataset empleado. De tal forma que los resultados provistos para este atributo son bastante próximos a una descripción de la estructura del mercado laboral del departamento.

Inicialmente, se tiene que los cargos con probabilidades estimadas de colocación se hallan asociados a las dinámicas de producción industrial en el norte del departamento, y al enclave de producción petrolera en el municipio de Piamonte, de tal forma que existe una predominancia en la colocación de cargos de corte operativo. Esto se alinea con las conclusiones previas asociadas a bajas expectativas salariales y el requerimiento de niveles de formación sobre la educación media, técnica y tecnológica.

La presencia de cargos operativos en el sector de construcción también se halla dentro de los rangos con probabilidades estimadas altas, los cual es congruente con la transversalidad de este sector en todos los municipios del departamento.

Seguido de esto, se hallan probabilidades altas de colocación para cargos previos asociados al sector agrícola, y, particularmente, al sector de servicios generales y comercio. En estos últimos resaltan los cargos de electricistas, soldadores, vigilantes, asesores comerciales y de ventas, supernumerarios, e inspectores de calidad.

El sector salud posee una estructura interesante, en tanto se hallan probabilidades de colocación altas para experiencias previas como médicos generales y ayudantes de ambulancia. Sin embargo, se registran probabilidades de colocación bajas para auxiliares de enfermería, fisioterapeutas y fonoaudiólogos.

En general, los cargos asociados a formación en educación superior presentan probabilidades de colocación bajas, con las excepciones de docentes de educación superior. Dado lo anterior, se halla mayor perjuicio para los cargos asociados al derecho, psicología, administración de empresas, pedagogía, ingenierías, profesionales en el sector de construcción, y ciencias sociales.

Esta última conclusión es completamente congruente con el análisis de títulos educativos, por lo cual se tiene que la conclusión general del mercado laboral del departamento como escaso en inserción de ciencia, tecnología y conocimiento se repite en este rubro, añadiendo una mayor especificidad en términos de los cargos más demandados. Los resultados analizados se consignan en la Tabla 17.

Tabla 17. Análisis de probabilidad estimada de colocación por cargos de última experiencia laboral (palabras agrupadas), ordenado por coeficiente de fiabilidad.

UltimaExperienciaLaboral	Avg_Probability	Std_Deviation	N_Observations	Q1	Q3	IQR	Reliability_Score
auxiliar de produccion	0,744835726	0,247696818	2551	0,586177	0,9441	0,357924	5,842670727
ayudante de ambulancia	0,778968485	0,231789965	1604	0,669733	0,962352	0,292619	5,748986671
obrero ayudante produccion en pozos de petroleo y gas	0,687164025	0,236585687	641	0,500956	0,887838	0,386881	4,441161338
operario agricola de mantenimiento del cultivo	0,645954027	0,289858688	492	0,384496	0,936725	0,552229	4,003932286
operario de empaque	0,528473349	0,28113259	1635	0,27107	0,805535	0,534464	3,910384687
operador de maquina fabricacion material plastico	0,697527305	0,269999199	239	0,459645	0,928349	0,468704	3,81998286
ayudante de obra	0,658565953	0,246789566	265	0,459934	0,860157	0,400222	3,674620089
supernumerario	0,773947665	0,256720426	108	0,508216	0,983763	0,475547	3,62372453
auxiliar de centro de distribucion	0,584796813	0,28257033	375	0,328215	0,85022	0,522005	3,466047452
auxiliar bodega	0,637511314	0,240063795	197	0,436941	0,842077	0,405136	3,368102151
ayudante de construccion	0,54548698	0,260627946	471	0,312649	0,782535	0,469887	3,357394956
soldador	0,623358805	0,252182925	185	0,414242	0,861077	0,446835	3,25415477
operador de planta auxiliar de refrigeracion	0,552391068	0,265851001	356	0,315708	0,798084	0,482376	3,24525926
vigilante	0,449533499	0,238347358	1111	0,247757	0,642943	0,395185	3,152585523
auxiliar de descargue	0,640709095	0,246752278	126	0,426569	0,852898	0,426329	3,098649804
asesor comercial	0,426660543	0,258029781	1413	0,208785	0,632353	0,423568	3,094769614
coordinador de empacadores agricola	0,523012867	0,288807771	348	0,268944	0,804961	0,536017	3,060777195
parcelero agricola	0,594507564	0,285386224	146	0,314637	0,873455	0,558818	2,962791835
electricista	0,556975168	0,232511982	196	0,357319	0,792085	0,434766	2,939778801
operador de montacargas	0,608119204	0,251129059	123	0,390922	0,866656	0,475734	2,926381718
oficial de construccion	0,531708715	0,23690284	175	0,331698	0,736351	0,404653	2,746161712
docente de universidad	0,562630108	0,314217347	121	0,269874	0,875003	0,605129	2,698256151
no registra	0,265232897	0,29231327	23235	0,022652	0,4742	0,451548	2,6664964
docente de educacion superior	0,498465911	0,272386203	197	0,261103	0,764517	0,503413	2,63349696
auxiliar de enfermeria	0,368685841	0,260812252	1259	0,151079	0,574855	0,423776	2,631706459
medico general	0,531169509	0,254822507	133	0,297979	0,802485	0,504506	2,597604347
inspector de calidad	0,543862153	0,250431093	114	0,321929	0,760391	0,438462	2,575839087
conductor de vehiculo liviano	0,42967027	0,247944022	326	0,223035	0,614161	0,391126	2,486457763
auxiliar de bodega	0,43418625	0,250152543	279	0,222706	0,618361	0,395655	2,444994726
asesor comercial de ventas	0,373595085	0,240643072	692	0,175509	0,533385	0,357876	2,44315717
operario de produccion	0,433189281	0,256783279	248	0,214953	0,645937	0,430985	2,388358233
auxiliar contable	0,340533783	0,251160749	1063	0,130799	0,532376	0,401576	2,373128983
guardia de seguridad	0,359189203	0,215686561	575	0,186266	0,492318	0,306052	2,282421109
supervisor servicio cajeros automaticos	0,398689896	0,23141866	299	0,203556	0,569284	0,365728	2,272709258
tecnico electricista	0,465873945	0,262824976	127	0,232452	0,703734	0,471282	2,256780547
auxiliar logistico	0,443709199	0,273289524	158	0,220078	0,636854	0,416776	2,246319989
auxiliar aseo	0,382019693	0,259729186	261	0,175706	0,626186	0,45048	2,125756378
auxiliar de almacen y bodega	0,358221581	0,245367685	352	0,151389	0,52024	0,368852	2,100479233
conductor	0,308169146	0,232681422	833	0,113072	0,485596	0,372525	2,072447877
recepcionista	0,432286907	0,271705972	112	0,217062	0,626265	0,409203	2,039745281
instructor de deportes	0,433582545	0,289074701	110	0,208193	0,743212	0,535019	2,038046239
auxiliar administrativo	0,257381311	0,250771929	2584	0,0805	0,337258	0,256758	2,022269119
tecnologo en salud ocupacional	0,420123337	0,213763786	120	0,255434	0,594217	0,338782	2,011337005
enfermera profesional	0,427649183	0,257665414	106	0,209412	0,61082	0,401408	1,994315921
auxiliar de cocina	0,304511346	0,261176904	619	0,097602	0,480766	0,383165	1,957430986
operario de servicios generales	0,400287188	0,255986501	123	0,172547	0,582789	0,410242	1,926255744
auxiliar de farmacia	0,323125801	0,261102935	379	0,115731	0,452543	0,336812	1,918571142
auxiliar de servicios generales	0,275495122	0,215034735	1008	0,081629	0,42246	0,34083	1,905248073
ingeniero de sistemas	0,353872686	0,25853117	213	0,123823	0,513426	0,389602	1,897214862
auxiliar de archivo y registro	0,375194528	0,274198024	138	0,171341	0,477026	0,305685	1,848678619
auxiliar de ventas en caja	0,398364307	0,251522172	102	0,198866	0,59313	0,394264	1,84242409
supervisor	0,383492098	0,275672639	122	0,172848	0,584213	0,411365	1,842304108
fisioterapeuta	0,322407169	0,251358846	277	0,132416	0,48234	0,349923	1,813223565
agente de ventas	0,350730144	0,229310573	167	0,183053	0,437935	0,254882	1,795034707
auxiliar de almacen	0,331464012	0,234777934	187	0,146787	0,475228	0,328441	1,733924249
asesor	0,314597085	0,226102118	226	0,133629	0,423122	0,289494	1,705284511
mercaderista	0,345134046	0,244305066	137	0,164299	0,483783	0,319483	1,698052923
operario de maquinas	0,349239853	0,257547665	121	0,153162	0,498632	0,34547	1,674881183
psicologo organizacional	0,360758916	0,270843299	103	0,156656	0,457027	0,30037	1,672019805
auxiliar operativo	0,320490668	0,232951865	168	0,139993	0,410928	0,270934	1,642182639
vendedor de suministros industriales	0,280568666	0,247999513	344	0,085282	0,404318	0,319036	1,63870104
auxiliar de caja	0,340913737	0,243908557	121	0,15049	0,487511	0,337021	1,634950878
cajero	0,263017477	0,207025306	461	0,096331	0,375555	0,279224	1,613190876
asesor servicio al cliente	0,343772361	0,224664394	106	0,191321	0,428398	0,237077	1,603161466
mesero	0,264350573	0,223287098	325	0,091323	0,396776	0,305453	1,528957503
vendedor	0,240733176	0,193168943	508	0,094993	0,319975	0,224982	1,499883587

UltimaExperienciaLaboral	Avg_Probability	Std_Deviation	N_Observations	Q1	Q3	IQR	Reliability_Score
cajero de supermercado	0,305071429	0,219938845	127	0,129385	0,405838	0,276453	1,477823077
almacenista	0,301819955	0,243407683	112	0,105721	0,397701	0,29198	1,424137115
asistente administrativo	0,22067229	0,210347439	635	0,06738	0,284778	0,217398	1,424136204
auxiliar de atencion al cliente	0,196166196	0,1970772	1284	0,058158	0,252447	0,194289	1,404105743
docente de educacion basica secundaria	0,219038952	0,233283728	529	0,061664	0,289491	0,227827	1,373590736
administrador de empresas	0,22335554	0,21174774	367	0,060917	0,322123	0,261206	1,318995283
auxiliar servicios generales	0,190395137	0,205484576	929	0,041841	0,2743	0,232459	1,301181071
secretaria	0,210054419	0,232547638	467	0,050183	0,257655	0,207472	1,291063621
ingeniero industrial	0,264392805	0,214459645	120	0,108977	0,325997	0,21702	1,26577837
auxiliar de archivo	0,221011434	0,215739751	283	0,071731	0,267453	0,195721	1,247708315
ingeniero civil	0,2399956	0,294866036	157	0,020869	0,423364	0,402494	1,213476747
auxiliar de apoyo administrativo	0,210622692	0,222686126	235	0,062262	0,242648	0,180387	1,149912596
conductor de autobus	0,226443824	0,193098345	153	0,065953	0,319992	0,254039	1,139111601
docente de basica primaria	0,188017584	0,231272738	423	0,03839	0,22585	0,187459	1,137012305
encuestador	0,243021609	0,227319493	103	0,096034	0,293284	0,197251	1,126339298
administrador punto de venta	0,220882777	0,199567932	145	0,081012	0,293922	0,21291	1,099274771
docente de educacion basica primaria	0,176576509	0,236153743	478	0,036066	0,180908	0,144842	1,089408326
auxiliar de caja y ventas	0,215881575	0,23063942	145	0,064286	0,259785	0,195499	1,074385119
docente de educacion media	0,211069225	0,230708894	145	0,059564	0,261704	0,20214	1,050435335
auxiliar de aseo y limpieza	0,18486576	0,200879258	283	0,043647	0,242075	0,198428	1,043649832
asesor comercial ventas no tecnicas	0,209494392	0,206668096	136	0,067935	0,256866	0,188931	1,029173649
coordinador administrativo	0,184455451	0,211745381	264	0,045004	0,230247	0,185243	1,028514208
comunicador social	0,198050724	0,209590898	174	0,047115	0,278569	0,231453	1,021754639
auxiliar enfermeria	0,159712766	0,168904873	559	0,042831	0,213521	0,17069	1,010366833
administrador de ventas y servicios	0,200838776	0,214099477	147	0,064772	0,250414	0,185642	1,002272375
auxiliar de servicio al cliente	0,192907394	0,220205622	180	0,064713	0,220611	0,155899	1,001759775
entrenador deportivo	0,203775623	0,22322185	115	0,044802	0,249833	0,205031	0,966901501
escolta	0,194034708	0,165655231	122	0,083696	0,227422	0,143726	0,932146823
psicologo	0,146281817	0,169986765	504	0,039452	0,184034	0,144581	0,910249765
auxiliar de servicios	0,175687085	0,231100558	164	0,024009	0,208996	0,184987	0,895980668
auxiliar de docente	0,148803478	0,199735414	304	0,039325	0,152351	0,113026	0,850713607
mensajero	0,17140875	0,157716897	113	0,068008	0,192789	0,12478	0,810315639
docente de educacion media academica	0,167935441	0,17779141	102	0,06228	0,17785	0,11557	0,77669685
soporte tecnico en sistemas	0,150848732	0,148434599	146	0,062315	0,18042	0,118105	0,75177074
tecnico de salud ocupacional	0,157986409	0,16355879	113	0,062784	0,188283	0,125499	0,746863023
administrador de negocios	0,157859203	0,169334029	113	0,053374	0,186161	0,132787	0,746261672
trabajador social	0,139654928	0,14658035	202	0,041107	0,18566	0,144553	0,741325744
abogado	0,121568276	0,146758669	333	0,031985	0,14175	0,109765	0,706085868
profesional de trabajo social	0,132098434	0,183486448	201	0,028357	0,151332	0,122976	0,700558274
arquitecto	0,13048595	0,177125304	159	0,021206	0,130422	0,109217	0,661420781
psicologo comunitario	0,134174322	0,162900914	125	0,030828	0,172453	0,141624	0,647835722
docente de jardin y preescolar	0,121368967	0,167998572	202	0,029126	0,12889	0,099763	0,644258965
docente de preescolar	0,118399379	0,167801927	226	0,03202	0,133591	0,101571	0,641787978
psicologo educativo	0,124942945	0,181536664	138	0,027286	0,122006	0,09472	0,615625588
vendedor de almacen	0,115404966	0,157244309	116	0,03267	0,129488	0,096817	0,548587916
auxiliar de facturacion	0,106150102	0,11220464	164	0,039916	0,128843	0,088927	0,541351339
ingeniero ambiental	0,105095784	0,135379004	105	0,028247	0,123407	0,095161	0,489111611
psicologo clinico	0,102682067	0,122461699	103	0,026867	0,125228	0,09836	0,475903551
ingeniero	0,094222664	0,114468471	116	0,027045	0,11694	0,089896	0,44789593
educador de primera infancia	0,073429033	0,12441171	402	0,017984	0,079267	0,061284	0,44031368
colaborador de profesores	0,085368377	0,097974432	170	0,025559	0,106219	0,08066	0,438434778
fonoaudiologo	0,082591158	0,123122135	103	0,020576	0,083368	0,062792	0,382787634

5.2 Los impactos potenciales de un protocolo de Ciencia de Datos para la Agencia de Empleo y las herramientas de planeación de Comfacauca.

Los resultados presentados anteriormente consistieron en una caracterización del mercado laboral reflejado en la información disponible desde la óptica de los principales atributos asociados a los candidatos considerados, de tal forma que fue posible dar forma a las

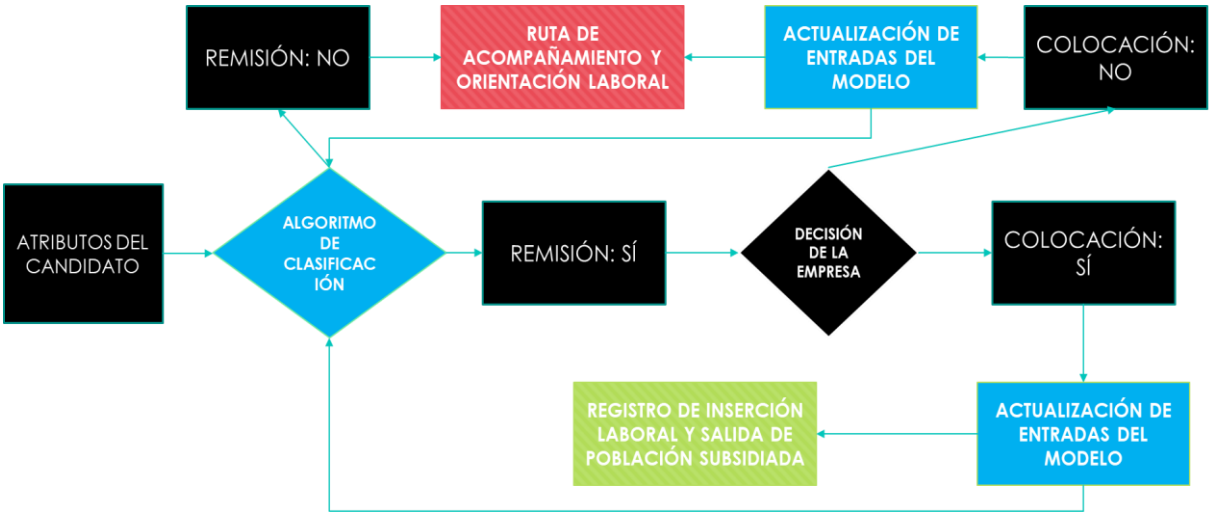
propiedades de la fuerza laboral del departamento y proveer perspectivas con respecto a la situación de la dinámica económica caucana.

Si bien el ejercicio realizado se considera como una caracterización estática desde un modelo de clasificación, en el sentido de que los resultados obtenidos sólo tienen validez para el periodo comprendido entre enero de 2018 y octubre de 2024, la lección principal de este ejercicio investigativo fue cómo la organización está en la plena capacidad de emplear métodos de Ciencia de Datos para monitorear desde el rigor estadístico las actividades y resultados de la Agencia de Empleo, y, en el sentido de una posible extrapolación, de Comfacauca en general.

Así, este proyecto aplicado ha logrado proveer un vistazo tangible a la gama de herramientas que pueden estar a disposición de Comfacauca al aplicar un protocolo de Ciencia de Datos, tanto a nivel de la Agencia de Empleo como para sus dinámicas transversales de planeación.

Las ventajas potenciales a resaltar para las operaciones de la cadena de valor correspondiente a la Ruta de Atención de la Agencia de Empleo radican en la posibilidad de incrementar la eficiencia técnica, financiera y organizacional de manera conjunta, en tanto un protocolo de Ciencia de Datos desde un algoritmo de clasificación significa una aceleración de la fluidez de la cadena de valor, un incremento del alcance de las operaciones desde el monitoreo del mercado laboral, y un incremento progresivo de la eficiencia financiera de los recursos empleados en el mediano y largo plazo. La siguiente figura presenta un diagrama de flujo para una propuesta de integración sofisticada de los métodos empleados en este trabajo con el fin de mejorar la eficiencia multidimensional del proceso de la Ruta de Atención, el cual es una modificación del diagrama de flujo que describe el estado actual de la estructura de funcionamiento de esta cadena de valor (Figura 1).

Figura 23. Diagrama de flujo de la cadena de valor de la Ruta de Atención bajo la integración del protocolo de Ciencia de Datos para la clasificación y caracterización del mercado laboral.



El diagrama de flujo expuesto implica la dinamización del proceso de filtro para la remisión de candidatos a las empresas e instituciones solicitantes, a la vez mejora este subproceso al insertar un criterio transparente fundamentado en el rigor estadístico y el grado de precisión de las predicciones del clasificador construido. Además, a diferencia de la naturaleza estática del modelo construido en esta investigación, esta propuesta involucra una dinámica de actualización basada en los resultados sucesivos de procesos de remisión y colocación según las vacantes, manteniendo la precisión del protocolo y otorgando una caracterización en tiempo real del estado del mercado laboral captado por la Agencia. Nótese que lo anterior implica una herramienta de auto evaluación del enfoque de los recursos de la Agencia, y de su misma estructura operacional.

En términos de impactos positivos potenciales sobre la eficiencia técnica, financiera y organizacional de la Agencia, se resaltan los siguientes puntos clave:

- Disminución potencial de los tiempos de ejecución de remisiones y del proceso general de la Ruta de Atención para cada candidato.
- Auto evaluación de los eslabones de la cadena de valor y la eficiencia de los colaboradores para cada sub proceso.
- Disminución progresiva del costo unitario de acompañamiento por candidato en el mediano y largo plazo.
- Capacidad de monitoreo interno del proceso de la Ruta de Atención y posibilidades de reestructuración del mismo de acuerdo a las necesidades de la organización.
- Tendencia a la estandarización de procesos, tecnologías y estructuras de datos.

- Capacidad de monitoreo externo de la dinámica económica del departamento del Cauca, y evaluación de la eficiencia financiera de los recursos empleados por la organización en mecanismos de subsidios e intervención.
- Conformación de una fuente fiable y constantemente actualizada de información sobre el mercado laboral. Apta para procesos de investigación y colaboración con instituciones a cargo de la política económica y sectorial del Departamento.

Ahora, si bien se ha logrado justificar claramente la necesidad de la implementación de un protocolo de Ciencia de Datos desde sus aspectos positivos, resulta necesario abordar el conjunto de esfuerzos adicionales que implica un proyecto de este tipo para la organización:

- Formulación e implementación del protocolo de Ciencia de Datos a cargo de un equipo competente, el cual pueda compaginar con el estado actual del proceso y realizar su inserción de forma apropiada.
- Implementación transversal de una estructura estandarizada y una política concreta de gobernanza de datos.
- Formulación, implementación y evaluación de modelos de mayor complejidad, capaces de asimilar las características específicas de la cadena de valor de la Ruta de Atención y el carácter dinámico del mercado laboral sobre el que opera.
- Diseño de mecanismos de monitoreo, verificación y evaluación de los impactos reales del proyecto de implementación del protocolo.

Dada la disponibilidad normativa de información de monitoreo por parte de Comfacauca, también es posible vislumbrar las fortalezas de la implementación de protocolos análogos como una extrapolación de este ejercicio, reflejando los beneficios del uso estandarizado de metodologías de Ciencia de Datos como un instrumento riguroso de planeación.

De acuerdo al alcance de la Comfacauca en el Departamento, los resultados presentados, y su sofisticación ejercicios futuros, se postulan como insumos de planeación desde los siguientes aspectos clave:

- Identificación de incentivos para la formulación de políticas y programas de intervención para la movilidad de fuerza laboral disponible fuera de las zonas tradicionales de aglomeración identificadas en este ejercicio (Popayán y norte del Cauca).
- Identificación de oportunidades de inversión desde la localización de recursos y fuerza laboral disponible en zonas segmentadas de la dinámica económica departamental.
- Revisión de títulos y cargos de mayor y menor demanda por probabilidades estimadas de colocación, de cara al (re) planteamiento de programas de capacitación e intervención desde la organización.

- Identificación de la eficiencia marginal (económica y financiera) de los programas de intervención, subsidios y capacitación de Comfacauca bajo un enfoque de rigor estadístico. Planteando la posibilidad de su redistribución bajo un criterio de minimización del riesgo en su ejecución, tanto en materia financiera como socioeconómica.
- Identificación de fortalezas y debilidades de la estructura económica departamental, suscitando la reevaluación de la ejecución de programas y recursos desde una óptica investigativa y formal. Incluso colaborando con personal e instituciones especializadas en este campo alrededor de la región.

Finalmente, en términos de extrapolaciones generales, es importante resaltar que este ejercicio investigativo es una muestra de las amplias capacidades que la implementación de metodologías de modelación en Ciencia de Datos ofrece a las organizaciones, de tal manera que la réplica sofisticada y estandarizada de este tipo de ejercicios constituye una herramienta potencial para la eficiencia multidimensional de las operaciones de Comfacauca de forma transversal. Como se propuso en el inicio de este proyecto aplicado, el poder predictivo y prescriptivo de los protocolos de Ciencia de Datos resalta su carácter como una herramienta indispensable para la organización en el mediano y largo plazo.

A la entrega de este documento de proyecto aplicado se anexa un informe puntual de los resultados hallados, dirigido a los directivos de la organización a cargo de su supervisión.

6. CONCLUSIONES Y TRABAJOS FUTUROS

La formulación y ejecución de este proyecto aplicado ha permitido establecer una caracterización del mercado laboral del departamento del Cauca sin precedentes, en tanto ha sido posible obtener una descripción de su comportamiento estadístico desde la óptica de las metodologías de modelación en Ciencia de Datos, basado en un dataset de alcance poblacional que ha sido un insumo del que no ha podido disponer ninguna investigación previa con un alcance y temática similar.

Esta caracterización se ha obtenido a través del diseño e implementación de un protocolo de investigación en Ciencia de Datos sobre la Agencia de Empleo como núcleo de problemas de negocio. El protocolo realizado retomó y adaptó las orientaciones genéricas de la metodología CRISP-DM [27] bajo la lógica de la cadena de valor de la Ruta de Atención de la Agencia de Empleo, planteando, ejecutando y documentando lineamientos para la comprensión del problema de negocio a enfrentar, la comprensión y preparación de la estructura de datos disponible, su modelación desde las estructuras teóricas propuestas, y la evaluación de los resultados obtenidos. Si bien el alcance de este proyecto no comprendió una fase de despliegue de las soluciones obtenidas, el ejercicio realizado y su correspondiente informe de resultados fijan los cimientos para la inserción sofisticada del protocolo dentro de las actividades de la Agencia.

En general, este ejercicio de modelación predictiva y prescriptiva ha permitido la identificación específica de los factores de mayor influencia en la determinación de la relación entre la oferta y la demanda del mercado laboral, otorgando predicciones y resultados bajo un nivel consistente de rigor estadístico.[1]

Además de resaltar las fortalezas y debilidades de la distribución sectorial de la fuerza laboral en la región, esta investigación ha permitido explicar de forma detallada el funcionamiento e impactos positivos potenciales de la inserción de los métodos de Ciencia de Datos en las operaciones de la cadena de valor de la Ruta de Atención de la Agencia de Empleo, sugiriendo sus beneficios potenciales a nivel interno de su estructura organizacional, a manera de una potencial extrapolación a otras cadenas de valor de Comfacauca, y en términos de los beneficios socioeconómicos que puede implicar el desarrollar herramientas predictivas fiables como herramientas para la planeación de este componente fundamental del Sistema de Seguridad Social del Departamento.

[1] Como anexo a la entrega de este documento a la organización, se relaciona un archivo con el dataset limpio, adicionando la categorización suplementaria de títulos educativos y experiencia laboral previa, además del ordenamiento de los candidatos por su probabilidad de colocación de acuerdo al modelo optimizado, siendo esto un insumo clave para futura toma de decisiones de la Agencia.

Este trabajo no sólo representa una muestra clara de la necesidad de la implementación de protocolos de modelación predictiva y prescriptiva como un insumo de planeación en pro de la eficiencia multidimensional de Comfacauca, sino que es una invitación al uso generalizado de herramientas estadísticas avanzadas para el mejoramiento de las prácticas organizacionales y operativas en las instituciones a cargo de ejecutar las políticas públicas y económicas clave del Estado, como lo es en este caso el sistema de retribuciones de recursos parafiscales.

Como se expuso claramente en la caracterización obtenida para el mercado laboral, el departamento del Cauca enfrenta grandes retos en materia de políticas económicas, sectoriales y de empleo, si es que se propone superar el conjunto de adversidades que se observan a lo largo de la región en términos del volumen de generación de empleo, sus condiciones y distribución geográfica. Por lo cual, las herramientas de monitoreo y evaluación rigurosa que proponen los protocolos de Ciencia de Datos como el que se desarrolla en este trabajo se muestran como indispensables para la alineación de iniciativas de política económica, sus límites normativos, y la certidumbre sobre su efectividad.

No es posible promover el avance de la población en materia socioeconómica si no se dispone de herramientas consistentes para conocer el punto de partida de las iniciativas que buscan este propósito, ni el cómo llegar a los objetivos propuestos. Es ahí donde la práctica de la Ciencia de Datos se postula como una herramienta indispensable para la mitigación del riesgo y la incertidumbre, y la optimización de los recursos disponibles.

Respecto a brechas investigativas importantes que deja este proyecto, éstas pueden considerarse desde tres perspectivas. Inicialmente, se considera la perspectiva técnica y cuantitativa, desde la cual futuros trabajos alrededor del proyecto pueden enfatizar en el desarrollo de estimadores y metodologías de evaluación de impactos sobre la línea de contribuciones como [4] y [5]. Además, en tanto se mencionó que la implementación organizacional del protocolo de clasificación en tiempo real sugiere modelos de mayor complejidad y carácter dinámico, emergen propuestas de modelación predictiva dinámica como lo son las redes neuronales estocásticas y metodologías de clasificación bajo aprendizaje profundo, similares a [14] y [24]. Es más, caben propuestas de modificación del problema fundamental propuesto, de un problema de clasificación como el que se abordó en este trabajo, a un problema de regresión para la estimación del tiempo medio para la colocación. Naturalmente, la formulación e implementación de este tipo de sofisticaciones implica esfuerzos adicionales que decanten en estructuras de datos aptas para estos tipos de modelación. Este es el paso a seguir que se sugiere para la Agencia de Empleo y para Comfacauca a nivel de planeación transversal.

La segunda perspectiva a considerar puede retomar las conclusiones del proyecto desde el análisis teórico y de modelación de redes y cadenas de valor, utilizando las implicaciones

empíricas de este trabajo para ahondar en la construcción de frameworks sobre líneas similares a las de [8], [9] y [17].

Finalmente, la tercera perspectiva para evaluar las implicaciones de este proyecto aplicado en el sentido posibles trabajos futuros se halla próxima al análisis empírico de variables socioeconómicas y su influencia en la dinámica del mercado laboral de la región. Lo anterior bajo el hecho de que el proyecto aplicado realizado se dirige a un análisis de alcance poblacional en la región, con un grado de integración de la información (provisto por la organización) del cual no se tienen precedentes.

En este sentido, y como conclusión final, se espera que los resultados de esta investigación también sean una invitación al trabajo colaborativo de Comfacauca, y demás instituciones ejecutantes de políticas económicas en el Departamento, de la mano con la academia, generando enlaces para la innovación social desde el estudio cualitativo y cuantitativo del mercado laboral y la dinámica económica a través del levantamiento de barreras sobre información clave para la formulación e implementación de alternativas en pro del bienestar de la región.

APÉNDICE

Salida de análisis de correspondencia con etiqueta de clase: Género.

```
                0     1
  Hombre      30518 11745
  Intersexual     4     0
  Mujer       39683  5883

        Pearson's Chi-squared test

data:  tabla1
X-squared = 3027, df = 2, p-value < 2.2e-16
```

Salida de análisis de correspondencia con etiqueta de clase: Nivel educativo.

```
                            0     1
  Básica Primaria(1-5)    1659   250
  Básica Secundaria(6-9)  2643   266
  Doctorado                 53    16
  Especialización         2816   460
  Maestría                 920   192
  Media(10-13)           20441  7364
  Preescolar                14     0
  Técnica Laboral        10361  2367
  Técnica Profesional     2316   574
  Tecnológica             5681  1625
  Universitaria          14793  2556

        Pearson's Chi-squared test

data:  tabla1
X-squared = 1406.7, df = 10, p-value < 2.2e-16
```

Salida de análisis de correspondencia con etiqueta de clase: Municipio de residencia.

```
                              0     1
  ABEJORRAL                   1     0
  ACACÍAS                     1     0
  AGRADO                      0     1
  AGUA DE DIOS                1     0
  AGUAZUL                     2     0
  ALCALÁ                      1     0
  ALMAGUER                  148     7
  ANDALUCÍA                   1     0
  ARGELIA                   179    17
  ARMENIA                    14     1
  BALBOA                    188    14
  BARBACOAS                   1     0
  BARBOSA                     1     0
  BARRANCABERMEJA             1     0
  BARRANQUILLA               12     0
  BELÉN DE LOS ANDAQUÍES      1     0
  BELLO                       7     7
  BOGOTÁ, D.C.              126    20
  BOLÍVAR                   702    35
```

BUCARAMANGA	6	0
BUENAVENTURA	6	4
BUENOS AIRES	208	19
BUGALAGRANDE	1	0
CAICEDONIA	2	0
CAJAMARCA	1	0
CAJIBÍO	500	40
CAJICÁ	1	0
CALARCÁ	1	0
CALDAS	1	0
CALDONO	453	42
CALI	800	155
CALIMA	2	0
CALOTO	1291	477
CANDELARIA	68	47
CARTAGENA DE INDIAS	11	2
CARTAGO	3	0
CHACHAGÜÍ	0	1
CHÍA	2	0
CHINCHINÁ	1	1
COPACABANA	2	1
CÓRDOBA	2	0
CORINTO	385	38
COTA	3	0
CUCUNUBÁ	1	0
CÚCUTA	5	0
CUMBITARA	1	0
DOSQUEBRADAS	7	1
DUITAMA	2	0
EL CARMEN DE VIBORAL	1	0
EL CERRITO	5	2
EL ROSAL	2	0
EL TAMBO	572	33
ENVIGADO	2	0
ESPINAL	1	0
FACATATIVÁ	4	0
FLORENCIA	60	3
FLORIDA	145	163
FONSECA	1	0
FUNZA	2	0
FUSAGASUGÁ	1	0
GARZÓN	1	0
GIRARDOT	1	0
GIRARDOTA	1	0
GIRÓN	1	0
GUACARÍ	3	0
GUACHENÉ	1598	703
GUACHETÁ	1	0
GUADALAJARA DE BUGA	11	3
GUADALUPE	0	1
GUALMATÁN	1	0
GUAPÍ	577	51
GUARNE	1	0
GÜEPSA	1	0
IBAGUÉ	11	2
ILES	1	0
INZÁ	150	24
IPIALES	9	0
ISNOS	4	0
ITAGÜÍ	3	1
JAMBALÓ	101	5
JAMUNDÍ	208	28

LA ARGENTINA	1	0
LA CEJA	1	0
LA CRUZ	3	0
LA CUMBRE	1	0
LA ESTRELLA	1	0
LA PLATA	7	0
LA SIERRA	145	11
LA TEBAIDA	1	0
LA UNIÓN	2	0
LA VEGA	145	18
LA VIRGINIA	1	0
LEIVA	1	0
LÓPEZ DE MICAY	15	3
MADRID	4	0
MANÍ	0	1
MANIZALES	7	0
MARINILLA	0	1
MEDELLÍN	41	4
MERCADERES	392	18
MIRANDA	970	223
MITÚ	1	0
MOCOA	10	1
MONTELÍBANO	1	0
MONTERÍA	1	0
MONTERREY	0	1
MORALES	230	8
MOSQUERA	6	2
NARIÑO	1	0
NEIVA	10	1
ORITO	3	0
OSPINA	1	0
PADILLA	398	88
PÁEZ	144	4
PAICOL	2	0
PALMIRA	69	80
PASTO	25	3
PATÍA	905	145
PELAYA	1	0
PEREIRA	21	5
PIAMONTE	416	249
PIENDAMÓ	1126	87
PITALITO	16	1
POPAYÁN	35547	7081
PRADERA	20	75
PUERRES	1	0
PUERTO ASÍS	4	0
PUERTO CAICEDO	1	0
PUERTO GAITÁN	0	1
PUERTO SALGAR	1	0
PUERTO TEJADA	5358	2572
PUPIALES	1	0
PURACÉ	194	14
QUINCHÍA	1	0
RESTREPO	1	0
RÍO DE ORO	1	0
RIOFRÍO	1	0
RIONEGRO	9	1
ROLDANILLO	1	0
ROSAS	145	16
SABANETA	3	0
SAMANIEGO	1	0
SAN AGUSTÍN	3	1

SAN ANDRÉS DE TUMACO	5	1
SAN ANTERO	1	0
SAN PABLO	1	1
SAN PEDRO	1	0
SAN PEDRO DE CARTAGO	1	0
SAN SEBASTIÁN	127	10
SAN VICENTE DEL CAGUÁN	1	0
SANDONÁ	1	0
SANTA FÉ DE ANTIOQUIA	1	0
SANTA MARTA	2	0
SANTA ROSA	223	39
SANTA ROSA DE CABAL	2	0
SANTACRUZ	1	0
SANTANDER DE QUILICHAO	10211	4016
SARAVENA	1	0
SIBUNDOY	4	0
SILVIA	499	29
SOACHA	8	0
SOTARA	87	16
SUAITA	1	0
SUÁREZ	187	1
SUCRE	60	3
TAME	1	0
TAMINANGO	1	0
TIMBÍO	1294	121
TIMBIQUÍ	256	11
TOCA	1	0
TOCANCIPÁ	1	0
TOLEDO	1	0
TORIBÍO	198	10
TOTORÓ	205	21
TULUÁ	11	2
TUNJA	1	0
TÚQUERRES	4	0
VALLE DEL GUAMUEZ	3	0
VALLEDUPAR	1	0
VIJES	1	0
VILLA DE SAN DIEGO DE UBATÉ	1	0
VILLA RICA	1712	678
VILLAGARZÓN	3	0
VILLAMARÍA	1	0
VILLAVICENCIO	10	0
VITERBO	1	0
YACUANQUER	1	0
YOLOMBÓ	1	0
YOPAL	1	1
YOTOCO	1	0
YUMBO	13	4
ZARAGOZA	1	0
ZARZAL	2	0
ZIPAQUIRÁ	2	0

```
        Pearson's Chi-squared test

data:  tabla1
X-squared = 3957.2, df = 193, p-value < 2.2e-16
```

Salida de análisis de correspondencia con etiqueta de clase: Porcentaje de Hoja de Vida.

```
          0     1
```

```
  100 65044 17613
  25   3078     1
  30    463     0
  35     13     0
  40     18     0
  55    251     0
  60    341     0
  65     65     0
  70    179     8
  85     95     0
  90    195     6
  95    463     0

        Pearson's Chi-squared test

data:  tabla1
X-squared = 1344.4, df = 11, p-value < 2.2e-16
```

Salida de análisis de correspondencia con etiqueta de clase: Canal de Registro.

```
                    0     1
  Agencia       32244 12197
  Autoregistro 37961  5431

        Pearson's Chi-squared test with Yates' continuity correction

data:  tabla1
X-squared = 3049.5, df = 1, p-value < 2.2e-16
```

REFERENCIAS BIBLIOGRÁFICAS

[1] Caja de Compensación Familiar del Cauca - COMFACAUCA, «Informe de gestión social y estados financieros 2023,» COMFACAUCA, Popayán, Cauca, 2023.

[2] M. Porter, Competitive Advantage: Creating and Sustaining Superior Performance, New York: Simon and Schuster, 1985.

[3] J. Barney y H. W, «Organizational economics: understanding de relationship between organizations and economic analysis,» de *The SAGE handbook of organization studies*, 2006, pp. 111-148.

[4] G. Imbens y J. Wooldridge, «Recent developments in the econometrics of program evaluation,» *Journal of Economic Literature, 47(1),* pp. 5-86, 2009.

[5] J. Heckman y E. Vytlacil, «Econometric evaluation of social programs, part I: causal models, structural models and econometric policy evaluation,» de *Handbook of Econometrics, volume 6B*, Amsterdam, North Holland, 2007, pp. 4779-4874.

[6] G. James, D. Witten, T. Hastie, R. Tibshirani y J. Taylor, An introduction to statistical learning, Cham, Switzerland: Springer Texts in Statistics, 2023.

[7] M. Schonlau, Applied Statistical Learning, Waterloo, Canada: Springer Nature, 2023.

[8] Y. Tang y F. Liou, «DOES FIRM PERFORMANCE REVEAL ITS OWN CAUSES? THE ROLE OF BAYESIAN INFERENCE,» *STRATEGIC MANAGEMENT JOURNAL,* vol. 31, nº 1, pp. 39-47, 2010.

[9] D. Saha, T. Young y J. Thacker, «Predicting firm performance and size using machine learning with a Bayesian perspective,» *MACHINE LEARNING WITH APPLICATIONS,* vol. 11, 2023.

[10] R. Subramanian y D. Prabha, «Ensemble Variable Selection for Naive Bayes to Improve Customer Behaviour Analysis,» *COMPUTER SYSTEMS SCIENCE AND ENGINEERING,* vol. 41, nº 1, pp. 339-355, 2022.

[11] K. Ansari y M. Ghasemaghaei, «Big Data Analytics Capability and Firm Performance: Meta-Analysis,» *JOURNAL OF COMPUTER INFORMATION SYSTEMS,* vol. 63, nº 6, pp. 1477-1494, 2023.

[12] T. Oesterreich, E. Anton y F. Teuteberg, «What translates big data into business value? A meta-analysis of the impacts of business analytics on firm performance,» *INFORMATION AND MANAGEMENT,* vol. 59, nº 6, 2022.

[13] S. Gupta, «Model-Selection Inference for Causal Impact of Clusters and Collaboration on MSMEs in India,» *JOURNAL OF QUANTITATIVE ECONOMICS,* 2023.

[14] C. Giap, D. Son y D. Ha, «Firm Performance Prediction for Macroeconomic Diffusion Index using Machine Learning,» *INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS,* vol. 12, nº 6, pp. 87-89, 2021.

[15] V. d'Amato, R. D'Ecclesia y S. Levantesi, «Firms' profitability and ESG score: A machine learning approach,» *APPLIED STOCHASTIC MODELS IN BUSSINESS AND INDUSTRY,* vol. 40, nº 2, pp. 243-261, 2023.

[16] M. Lee, L. GYH y K. Lim, «Machine Learning-Based Causality Analysis of Human Resource Practices on Firm Performance,» *ADMINISTRATIVE SCIENCES,* vol. 14, nº 4, 2024.

[17] J. Campbell y M. Weese, «Compositional Models and Organizational Research: Application of a Mixture Model to Nonexperimental Data in the Context of CEO Pay,» *ORGANIZATIONAL RESEARCH METHODS,* vol. 20, nº 1, pp. 95-120, 2017.

[18] D. Delen, C. Kuzey y A. Uyar, «Measuring firm performance using financial ratios: A decision tree approach,» *EXPERT SYSTEMS WITH APPLICATIONS,* vol. 40, nº 10, pp. 3970-3983, 2013.

[19] R. Manogna y A. Mishra, «Measuring financial performance of Indian manufacturing firms: application of decision tree algorithms,» *MEASURING BUSSINESS EXCELLENCE,* vol. 26, nº 3, pp. 288-307, 2021.

[20] A. Nag, N. Choudhary, D. Sinha, S. AP y S. Mishra, «Predictive analytics - new business intelligence in SCM,» *INTERNATIONAL JOURNAL OF VALUE CHAIN MANAGEMENT,* vol. 14, nº 3, 2023.

[21] P. Tamasiga, E. Ouassou, H. Onyeaka y M. Bakwena, «Forecasting disruptions in global food value chains to tackle food insecurity: The role of AI and big data analytics - A bibliometric and scientometric analysis,» *JOURNAL OF AGRICULTURE AND FOOD RESEARCH,* vol. 14, 2023.

[22] R. Meriton, R. Bhandal, G. Graham y A. Brown, «An examination of the generative mechanisms of value in big data-enabled supply chain management research,» *INTERNATIONAL JOURNAL OF PRODUCTION RESEARCH,* vol. 59, nº 23, pp. 7283-7310, 2021.

[23] R. Dubei, Z. Luo, A. Gunaskeran, S. Akter y B. Hazen, «Big data and predictive analytics in humanitarian supply chains: Enabling visibility and coordination in the presence of swift trust,» *INTERNATIONAL JOURNAL OF LOGISTICS MANAGEMENT,* vol. 29, nº 2, pp. 485-512, 2018.

[24] Y. Ma, Z. Zhang y A. Ihler, «A Deep Choice Model for Hiring Outcome Prediction in Online Labor Markets,» *INTERNATIONAL OF COMPUTERS, COMMUNICATIONS & CONTROL,* vol. 15, nº 2, 2020.

[25] M. Drobotenko y A. Nevecherya, «Forecasting the sectoral structure of population employment,» *EKONOMIKA I MATEMATICESKIE METODY-ECONOMICS AND MATHEMATICAL METHOD,* vol. 59, nº 1, pp. 22-29, 2023.

[26] J. Ortiz-García, «Guía descriptiva para la elaboración de protocolos de investigación,» *Salud en Tabasco,* vol. 12, nº 3, pp. 530-540, 2006.

[27] G. Linoff y M. Berry, «Chapter 3, The Data Mining Process,» de *Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management*, Wiley, 2011, pp. 67-99.

[28] L. Breiman, «Random Forests,» *Machine Learning,* vol. 45, nº 1, pp. 5-32, 2001.

[29] R. Mantovani, J. Rossi, V. Vanschoren, B. Bischl y A. de Carvahlo, «Effectiveness of random search in SVM hyper-parameter tuning,» de *2015 International Joint Conference on Neural Networks*, 2015.

[30] P. Probst, M. Wright y A. Boulesteix, «Hyperparameters and tuning strategies for random forest,» *Wiley Interdiscip.Rev.: Data Min. Knowl. Discov,* vol. 9, nº 3, 2019.

[31] E. Bartz, T. Bartz-Beielstein, M. Zaefferer y O. Mersmann, Hyperparameter Tuning for Machine and Deep Learning with R, Singapore: Springer, 2023.

[32] A. Gómez Sánchez, J. Sarmiento Castillo y C. Fajardo Hoyos, «Análisis de la dinámica del mercado laboral en Popayán - Colombia,» *Económicas CUC,* vol. 37, nº 1, p. 135, 2016.

Anexo 1. LICENCIA DE AUTORIZACIÓN PARA LA PUBLICACIÓN DE OBRAS EN VITELA REPOSITORIO INSTITUCIONAL DE LA UNIVERSIDAD JAVERIANA CALI

Fecha: | 27 | 06 | 2025 |

Señores
Biblioteca General
Pontificia Universidad Javeriana Cali
Cuidad

Por medio del presente documento otorgo (otorgamos) a la Pontificia Universidad Javeriana Cali para que, en perfeccionamiento de la siguiente licencia de uso parcial[1], pueda ejercer sobre mi (nuestra) obra las facultades que se indican a continuación, teniendo en cuenta que, en cualquier caso, la finalidad perseguida será facilitar, difundir y promover el aprendizaje, la enseñanza y la investigación.

En consecuencia, autorizo (autorizamos) a la Pontificia Universidad Javeriana Cali, a los usuarios de la Biblioteca General, así como a los usuarios de las redes, bases de datos y demás sitios web con los que la Universidad tenga perfeccionado un convenio o con los que establezcan redes de colaboración, son:

AUTORIZO (AUTORIZAMOS)	
1.	La conservación de los ejemplares necesarios
2.	La consulta en línea
3.	La reproducción y/o transformación por cualquier formato conocido o por conocer. Igualmente, la edición, o cualquier otra forma de reproducción, incluyendo la posibilidad de trasladarla al sistema o entorno digital.
4.	La difusión y comunicación pública por cualquier medio físico o digital, así como su disposición en Internet a través de índices, buscadores y otros medios conocidos o por conocer.
5.	La publicación en bases de datos y en sitios web, sean éstos onerosos o gratuitos, existiendo con ellos previo acuerdo desarrollado con la Pontificia Universidad Javeriana para efectos de cumplir los fines predichos.
6.	La inclusión en el repositorio institucional de la Pontificia Universidad Javeriana Cali.
7.	La inclusión en cualquier otro formato digital, físico o soporte como multimedia, colecciones, recopilaciones o, en general, servir de base para cualquier otra obra derivada.

La presente licencia parcial se otorga a título gratuito por el máximo tiempo legal colombiano, con el propósito de que en dicho lapso mi (nuestra) obra sea utilizada en las condiciones aquí concertadas y para los fines indicados, respetando siempre la titularidad de los derechos patrimoniales y morales correspondientes, de acuerdo con los usos honrados, de manera proporcional y justificada a la finalidad perseguida, sin ánimo de lucro ni de comercialización.

[1] *De conformidad con lo establecido en el artículo 30 de la Ley 23 de 1982 y el artículo 11 de la Decisión Andina 351 de 1993, "Los derechos morales sobre el trabajo son propiedad de los autores", los cuales son irrenunciables, imprescriptibles, inembargables e inalienables. En consecuencia, la Pontificia Universidad Javeriana está en la obligación de respetarlos y hacerlos respetar, para lo cual tomará las medidas correspondientes para garantizar su observancia.*

Sí autorizo (autorizamos) la publicación: _____x_____

No autorizo (autorizamos) la publicación: _____

De manera complementaria, garantizo (garantizamos) en mi (nuestra) calidad de autor (es) exclusivo (s), que la obra en cuestión, es producto de mi (nuestra) plena autoría, de mi (nuestro) esfuerzo personal intelectual, como consecuencia de mi (nuestra) creación original particular y, por tanto, soy (somos) el (los) único (s) titular (es) de la misma. Además, aseguro (aseguramos) que no contiene citas, ni transcripciones de otras obras protegidas, por fuera de los límites autorizados por la ley, según los usos honrados, y en proporción a los fines previstos; ni tampoco contempla declaraciones difamatorias contra terceros; respetando el derecho a la imagen, intimidad, buen nombre y demás derechos constitucionales. Adicionalmente, manifiesto (manifestamos) que no se incluyeron expresiones contrarias al orden público ni a las buenas costumbres. En consecuencia, la responsabilidad directa en la elaboración, presentación, investigación y, en general, contenidos de la obra es de mí (nuestro) competencia exclusiva, eximiendo de toda responsabilidad a la Pontifica Universidad Javeriana Cali por tales aspectos.

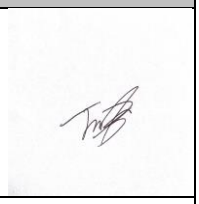
Si el documento ha sido apoyado o financiado por alguna organización, con excepción de la Pontificia Universidad Javeriana Cali, el (los) autor (es) garantiza (n) que se ha cumplido con las obligaciones requeridas por el respectivo acuerdo.

En constancia a lo anterior,

Título de la obra:	
Métodos Bayesianos y Frecuentistas Flexibles para la Caracterización y Análisis Predictivo y Prescriptivo de Cadenas de Valor	
Presentada, sustentada y aprobada en el año: **2025**	
Programa Académico: **Maestría en Ciencia de Datos**	
Facultad: Facultad de Ingeniería y Ciencias	
Premio o distinción: -	
Director (es) y/o asesor (es) de la obra:	
Isabel Cristina García Arboleda	
Evaluador (es):	
Diego Linares	
Jennifer Chilito	
Palabras clave:	**Keywords:**
1. Cadenas de Valor	**1. Value Chains**
2. Aprendizaje Supervisado	**2. Supervised Learning**
3. Mercado Laboral	**3. Labor Market**
4. Estadística Bayesiana	**4. Bayesian Statistics**
5. Estadística Frecuentista	**5. Frequentist Statistics**

Resumen:

La implementación de metodologías de Ciencia de Datos para la caracterización y análisis predictivo y/o prescriptivo de las actividades organizacionales toma fuerza como un instrumento predilecto para la minimización de la incertidumbre y el incremento de la productividad en cada vez más sectores de las economías y la sociedad en general. De tal forma que la ausencia de este tipo de herramientas analíticas puede implicar una limitación importante para el desarrollo de un rango de acción eficiente y planeación efectiva de las organizaciones. Tal es el caso de la Caja de Compensación Familiar del Cauca – Comfacauca, la cual, como figura de administración de recursos parafiscales en el marco del Sistema de Seguridad Social colombiano para el departamento del Cauca, posee un aparato operativo y de caracterización socioeconómica de alcance poblacional que le ha permitido consolidarse como una institución ejemplar a nivel nacional en su medio normativo. Sin embargo, la ausencia de instrumentos de caracterización estadística y optimización para la toma de decisiones al interior de la organización, representa una debilidad que puede comprometer el cumplimiento de sus objetivos estratégicos y metas de cobertura en el mediano y largo plazo. Para resolver este problema, este proyecto aplicado propone tomar como referencia la conceptualización organizacional de cadenas de valor y el contraste entre las metodologías de modelación bayesiana y frecuentista, con el fin de desarrollar un protocolo para la caracterización y análisis de operaciones de la organización en el marco de su Agencia de Empleo y Mecanismo de Protección al Cesante (MPC), el cual pueda servir como insumo para la toma de decisiones óptimas fundamentadas en criterios cuantitativos rigurosos, de cara al incremento de la eficiencia multidimensional y, por ende, del bienestar socioeconómico de la población caucana. En general, esta propuesta sugiere aplicaciones de Ciencia de Datos desde el planteamiento de un problema de clasificación en el contexto de los mecanismos de matching entre oferta y demanda del mercado laboral captado por la organización desde su Agencia. Además, por su esencia metodológica, también propone aplicaciones en las áreas de economía organizacional, construcción de protocolos integrados para la evaluación y monitoreo de vacantes laborales, y el estudio teórico y empírico de cadenas de valor para servicios y productos intangibles.

Nombres y Apellidos completos Autor(es)	No. Documento Identidad	Correo	Firma
Jesús Andrés Burbano Gómez	**1061782182**	**jesus.burbano@javerianacali.edu.co**	

Santiago de Cali, 8 de junio de 2025.

Doctor
Diego Luis Linares O.
Director Maestría en Ciencia de Datos
Facultad de Ingeniería y Ciencias
Pontificia Universidad Javeriana de Cali

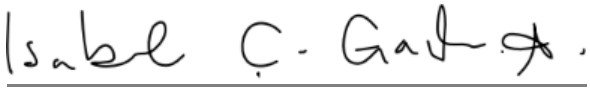
Asunto: Presentación para evaluación del proyecto aplicado

Cordial Saludo,

Con el fin de cumplir con los requisitos exigidos por la Universidad para optar por el título de Magíster en Ciencia de Datos, nos permitimos presentar a su consideración el proyecto denominado "MÉTODOS BAYESIANOS Y FRECUENTISTAS FLEXIBLES PARA LA CARACTERIZACIÓN Y ANÁLISIS PREDICTIVO Y PRESCRIPTIVO DE CADENAS DE VALOR", el cual fue realizado por el estudiante Jesús Andrés Burbano Gómez con código 8992305, perteneciente a la Maestría en Ciencia de Datos, bajo la dirección de la profesora Isabel Cristina García Arboleda.

El suscrito director del Proyecto Aplicado autoriza para que se proceda a hacer la evaluación de este proyecto, toda vez que ha revisado cuidadosamente el documento y avala que ya se encuentra listo para ser presentado y sustentado oficialmente.

Atentamente,



Jesús Andrés Burbano Gómez — C.C. 1061782182 de Popayán, Cauca

Isabel Cristina García Arboleda — C.C. 431097 de Bello, Antioquia

Documentación anexa:
Resumen del Proyecto Aplicado en formato digital (máximo 1 página).
Una copia digital (PDF) del documento del proyecto aplicado

ACUERDO DE COLABORACIÓN

Entre los suscritos a saber, **CAJA DE COMPENSACION FAMILIAR DEL CAUCA "COMFACAUCA",** con NIT. No. 891.500.182-0, representada legalmente por **GLORIA MARISOL VELASCO CHAGUENDO,** identificada con la cédula de ciudadanía No.34´540.290 de Popayán, Jefe Departamento Desarrollo Institucional, Funcionaria Delegada por el **Dr. JUAN CRISTÓBAL VELASCO CAJIAO**, identificado con la cédula de ciudadanía No. 10.519.479 de Popayán, mediante Resolución No. 39 de fecha 16 de septiembre de 2024, quien para efectos del presente acuerdo se llamará **COMFACAUCA,** por una parte **JESÚS ANDRES BURBANO GOMEZ,** identificado con la cédula de ciudadanía No. 1.061.782.182 de Popayán (Cauca), Economista, Aspirante a Magíster en Ciencia de datos, quien se llamará **EL PROFESIONAL**, hemos celebrado el siguiente acuerdo.

JUSTIFICACION DEL ACUERDO

- Que, la Agencia de Empleo de Comfacauca está llevando a cabo la unificación de sus documentos digitales con el fin de optimizar los procesos de análisis de información y aumentar la eficiencia en los diferentes componentes, el presente acuerdo contribuye de manera significativa al logro de este objetivo establecido por la Agencia.

- Que, el economista **Jesús Andrés Burbano Gómez**, identificado con la cédula de ciudadanía No. 1.061.752.152, expedida en Popayán, presentó una propuesta académica para el desarrollo del proyecto titulado "Métodos bayesianos y frecuentitas flexibles para la caracterización, análisis predictivo y prescriptivo de cadenas de valor.", cuyo objeto aparece descrito en la correspondiente cláusula.

- Que el economista Jesús Andrés Burbano Gómez, posee la formación académica necesaria para adelantar las actividades teórico prácticas requeridas para la ejecución del proyecto.

OBJETO: EL PROFESIONAL se obliga a **REALIZAR EL PROYECTO ACADÉMICO APLICADO Y TITULADO "MÉTODOS BAYESIANOS Y FRECUENTISTAS FLEXIBLES PARA LA CARACTERIZACIÓN Y ANÁLISIS PREDICTIVO Y PRESCRIPTIVO DE CADENAS DE VALOR.",** actividades que se llevarán a cabo de manera directa, sin sujeción a la continua subordinación o dependencia respecto a **COMFACAUCA,** manteniendo la autonomía, técnica, administrativa e iniciativa en la ejecución del proyecto, de conformidad con la propuesta académica presentada y aprobada el 22 de mayo de 2024, documento que se anexa y forma parte integral del presente acuerdo.

DESCRIPCIÓN DEL ACUERDO DE COLABORACIÓN:

Para lograr el objetivo se prevé la realización de las actividades que a continuación se enumeran:

- Diseñar y aplicar un protocolo para la caracterización y optimización estadística del desempeño de cadenas de valor específicas asociadas a la estructura organizacional y actividades de la Caja de Compensación Familiar del Cauca.
- Caracterizar las cadenas de valor relacionadas con las actividades de la Agencia de Empleo de Comfacauca, consolidando la información en redes compuestas por sub-actividades interconectadas que conducen a la provisión de bienes o servicios finales.
- Modelar el comportamiento estadístico de las cadenas de valor mediante el contraste de eficiencia entre métodos bayesianos (modelos mixtos) y frecuentitas (ML, inferencia causal).
- Utilizar la modelación desarrollada para fines predictivos y prescriptivos, facilitando la toma de decisiones óptimas para la organización.
- Entregar informe de resultados que incluya una descripción clara y comprensible de los impactos potenciales en materia socioeconómica y de eficiencia (técnica, financiera y organizacional).

PARAGRAFO: La ejecución de dichas actividades serán de conformidad con el cronograma que las partes acuerden.

TÉRMINO: La duración del presente acuerdo es del **VEINTICINCO (25) DE OCTUBRE DE DOS MIL VEINTICUATRO (2024) AL TREINTA (30) DE JUNIO DE DOS MIL VEINTICINCO (2025).**

DERECHOS Y OBLIGACIONES DE LAS PARTES INTERVINIENTES.

COMFACAUCA entregará al **PROFESIONAL** el material que sea necesario para el cumplimiento de las obligaciones acordadas en este documento. Por tratarse de información **CONFIDENCIAL** el **PROFESIONAL** se compromete a,

- Utilizar dicha información de forma reservada.
- No divulgar ni comunicar la información técnica facilitada por **COMFACAUCA,** la cual se entrega anonimizada.
- Impedir la copia o revelación de esa información a terceros, salvo que gocen de aprobación escrita de **COMFACAUCA,** y únicamente en los términos de tal aprobación.
- No utilizar la información o fragmentos de ésta para fines distintos de la ejecución de este acuerdo
- Dar estricto cumplimiento a lo establecido en la Ley 1581 de 2012, sobre protección de datos personales.

DERECHOS PREVIOS SOBRE LA INFORMACIÓN

Toda información puesta a disposición del **PROFESIONAL** es de propiedad exclusiva de **COMFACAUCA** y no es precisa la concesión de licencia para dicho intercambio. **EL PROFESIONAL** no utilizará información confidencial previa para su propio uso, salvo que se autorice lo contrario por parte de **COMFACAUCA.**

La información que se proporciona no da derecho o licencia para quien la recibe sobre las marcas, derechos de autor o patentes que pertenezcan a quien la proporciona.

La divulgación de información no implica transferencia o cesión de derechos, a menos que se redacte expresamente alguna disposición al respecto.

PARÁGRAFO PRIMERO: COMFACAUCA La parte receptora se compromete a manejar la información marcada como **CONFIDENCIAL** exclusivamente con los empleados que se requieran para dar cumplimiento al objeto del presente acuerdo

PARÁGRAFO SEGUNDO: **EL PROFESIONAL,** acepta calificar su relación de alta confidencialidad en razón al acceso de información por **COMFACAUCA** a tecnología, secretos, documentos y conceptos comerciales, financieros, empresariales, administrativos y de cualquier otro tipo de propiedad intelectual de **COMFACAUCA.** El incumplimiento a ésta cláusula será justa causa para dar por terminado este acuerdo, sin perjuicio de las acciones civiles o penales a que hubiere lugar.

PROTECCIÓN DE DATOS

En el desarrollo del presente acuerdo, **EL PROFESIONAL** podría tener acceso a datos de carácter personal, protegidos por la Ley 1581 de 2012, por lo que, se compromete a que el uso y tratamiento de los datos afectados serán acorde a las actuaciones que resulten necesarias para el desarrollo del objeto acordado, según las instrucciones facilitadas en cada momento, esto en cumplimiento del decreto 1377 de 2013 el cual reglamenta parcialmente la ley 1581 de 2012.

CLÁUSULA PENAL PECUNIARIA: COMFACAUCA podrá adelantar acciones de carácter legal por incumplimiento en el manejo de la información y el presente acuerdo presta mérito ejecutivo. Para el caso de incumplimiento por parte del Sr Burbano Gómez, se fija como cláusula penal pecuniaria la suma equivalente a **DIEZ SALARIOS MINIMOS MENSUALES LEGALES VIGENTES (10SMMLV),** independientemente de la indemnización plena de los perjuicios a que hubiere lugar de conformidad con el artículo 1.613 del Código Civil.

GARANTÍAS. – EL PROFESIONAL se obliga a suscribir a favor de **COMFACAUCA** una letra de cambio en blanco con su respectiva carta de instrucción, como garantía de cumplimiento del acuerdo, por el término del mismo y tres (3) meses más.

PARÁGRAFO PRIMERO: La letra debe suscribirse y entregarse a **COMFACAUCA** previamente.

PARÁGRAFO SEGUNDO: EL PROFESIONAL se obliga a constituir y presentar debidamente legalizada la garantía exigida en un término de ocho (08) días hábiles, contados a partir de la fecha de suscripción del presente acuerdo; de no cumplirse con esta exigencia el acuerdo no podrá iniciar su ejecución.

PARÁGRAFO TERCERO: La garantía no podrá ser cancelada sin la autorización de **COMFACAUCA.**

SUPERVISIÓN: El presente acuerdo será supervisado por la CP **JULIEDT ORDOÑEZ GARZÓN,** Coordinadora de Gestión, y/o quien haga sus veces, y quien verificará el estricto cumplimiento de los requisitos de perfeccionamiento o ejecución, de acuerdo con la costumbre mercantil y las cláusulas que sean obligatorias en atención a su naturaleza. La supervisora del acuerdo será responsable de que se dé estricto cumplimiento a la presente previsión, en los términos del principio de responsabilidad, de acuerdo con el art. 20 del Manual de Contratación de **COMFACAUCA**, actualizado en Febrero de 2020, quien deberá rendir informes mensuales respecto al cumplimiento de las obligaciones asumidas por **EL PROFESIONAL**, debiendo reportar a la Jefatura del Departamento Desarrollo Institucional, cualquier novedad que pueda afectar la ejecución del presente acuerdo, igualmente deberá informar a la Oficina Jurídica de **COMFACAUCA,** la terminación del acuerdo o vencimiento del mismo.

NATURALEZA JURIDICA: El presente acuerdo de carácter atípico en la legislación colombiana, se regulará por los principios del derecho civil, dada la regularización de **COMFACAUCA,** Corporación de derecho privado, con domicilio principal en la ciudad de Popayán, departamento del Cauca, sin ánimo de lucro, con personería jurídica que le fue conferida por medio de la Resolución No. 0133 de 2 de noviembre de 1966, otorgada por la Gobernación del Departamento del Cauca, vigilada por la Superintendencia del Subsidio Familiar, de conformidad con lo dispuesto en el literal 6 del artículo 12 del Decreto 2150 de 1992. Este acuerdo es de naturaleza civil, por lo tanto, no genera relación laboral alguna.

CESIÓN: EL PROFESIONAL no podrá ceder, ni en todo ni en parte las obligaciones que adquiere en virtud de este acuerdo a persona alguna, natural o jurídica, nacional o extranjera el respectivo acuerdo sin el consentimiento previo y escrito de **COMFACAUCA** pudiendo éste negar la autorización de la cesión.

INHABILIDADES, INCOMPATIBILIDADES Y CONFLICTOS DE INTERÉS. - Al suscribir el presente acuerdo, las partes manifiestan no encontrarse incursas en ninguna de las causales de conflicto de intereses previstas en el artículo 23 de la Ley 789 de 2.002, ni en las causales de inhabilidad e incompatibilidad contempladas en la ley, en especial en el Decreto 2463 de 1981 o en los Estatutos de la Caja.

CAUSALES DE TERMINACION: antes de su vencimiento:

a. Por mutuo acuerdo.
b. Por incumplimiento comprobado de cualquiera de las obligaciones.
c. Por decisión Administrativa.
d. Por muerte del **PROFESIONAL**

OBLIGACIÓN: EL PROFESIONAL. Se compromete a estar afiliado y al día en el pago de los aportes al Sistema de Seguridad Social Integral (S.S.S.I) en salud, pensiones y riesgos laborales y parafiscales (Cajas de Compensación Familiar, Sena e ICBF de conformidad con el Decreto 723 de 2013 y el Artículo 2 de la Ley 1562 de 2012 y el artículo 50 de la Ley 789 de 2002.

SOLUCIÓN DE CONTROVERSIAS CONTRACTUALES. En caso de presentarse controversias o diferencias en la ejecución del acuerdo de Prestación de Servicios Profesionales, se recurrirá en primera instancia a los mecanismos alternativos de solución de conflictos de acuerdo a los

procedimientos legales establecidos para tal efecto. Cuando existan diferencias entre **LAS PARTES** de carácter técnico, administrativo, o financiero, éstas serán dirimidas, en principio, por el Director Administrativo de **COMFACAUCA,** Para ello, tanto **LAS PARTES**, enviarán dentro de los cinco **(5)** días calendarios siguientes a que se suscite la controversia, sendos documentos en los cuales expresan el contenido de la discrepancia y las razones de su posición y se definirá el punto en un plazo máximo de treinta (**30**) días calendario, salvo cuando el contenido de la controversia exija un término superior, el cual, en todo caso, será comunicado por **COMFACAUCA**, al **SUPERVISOR** y al **PROFESIONAL,** lo anterior sin perjuicio de que sí un dictamen de **COMFACAUCA** no es aceptable para **EL PROFESIONAL,** se pueda acudir a **AMIGABLE COMPONEDOR TÉCNICO** para una decisión definitiva y obligatoria para las partes, de acuerdo con el procedimiento descrito en los parágrafos siguientes:

PARÁGRAFO PRIMERO: AMIGABLE COMPOSICIÓN para Aspectos Técnicos, sin perjuicio de lo previsto en el clausulado del presente acuerdo, cualquier diferencia relacionada con la ejecución del mismo, asociada a aspectos técnicos, será resuelta a través del mecanismo de la amigable composición, de acuerdo con los procedimientos establecidos en la Ley 1563 de 2012, que en los artículos 59, 60 y 61 reguló nuevamente la Amigable Composición, o las normas que los reemplacen, modifiquen o adicionen, por el Amigable Componedor Técnico.

PARÁGRAFO SEGUNDO: **CONTINUIDAD EN LA EJECUCIÓN DEL ACUERDO. -** La intervención del Amigable Componedor Técnico, no suspenderá la ejecución del acuerdo, salvo aquellos aspectos cuya ejecución dependa necesariamente de la solución de la controversia. Las partes acuerdan que el juez del acuerdo será la Jurisdicción ordinaria.

LEGALIZACIÓN. - El presente acuerdo para su legalización y perfeccionamiento, requiere la firma de las partes. El presente acuerdo es la única relación actual entre las partes, y como tal deja sin vigencia o efectos y/o se declara que se está a paz y salvo por todo concepto con **COMFACAUCA** por quien suscribe el acuerdo.

PREVENCIÓN DE LAVADO DE ACTIVOS Y FINANCIACIÓN DEL TERRORISMO. - Las **PARTES** se obligan expresamente a entregarse entre sí, la información veraz y verificable que se exija para el cumplimiento de la normatividad relacionada con prevención de lavado de activos y la financiación del terrorismo – SARLAFT y, a actualizar sus datos por lo menos dos veces al año, suministrando la totalidad de los soportes que cada una requiera. En el evento en que no se cumpla con la obligación consagrada en la presente cláusula, la PARTE cumplida tendrá la facultad de dar por terminada la relación jurídica surgida. Para efectos de la terminación referida bastará con el envío de una comunicación en ese sentido a la dirección registrada como de correspondencia.

PARÁGRAFO PRIMERO. - DECLARACIONES. Los abajo firmantes, obrando en nombre propio y en representación legal de cada una de las PARTES y de manera voluntaria, realizamos individualmente las siguientes declaraciones: 1. En cumplimiento de la normatividad vigente sobre lavado de activos y financiación del terrorismo – SARLAFT, me obligo a adoptar las medidas de control apropiadas y suficientes, orientadas a evitar que mis negociaciones y/o actuaciones de cualquier tipo puedan ser utilizadas como instrumentos para el ocultamiento, manejo, inversión o aprovechamiento en cualquier forma de dinero u otros bienes provenientes de actividades delictivas o, para dar apariencia de legalidad a dichas actividades o a las transacciones y fondos vinculados con las mismas. 2. Me obligo a realizar toda la debida diligencia, para la obtención correspondiente de la información sobre la fuente de los dineros que están siendo negociados o recibidos, absteniéndome de recibir el dinero cuando provenga de alguna de las actividades arriba mencionadas. 3. Me obligo a informar a la otra PARTE, en el caso de tener sospecha de que los fondos objeto de las negociaciones con mis clientes provengan de una actividad delictiva de las que se indican a continuación, y a no advertir a los clientes sobre el reporte realizado. 4. Conozco que el dinero ilegal no se circunscribe al negocio del narcotráfico, ya que puede tratarse también de: hurto, estafa, abuso de confianza, transferencia ilegal de cheques, cultivo y conservación de plantaciones ilícitas, tráfico, fabricación o porte de estupefacientes, secuestro, extorsión, enriquecimiento ilícito, delitos contra la administración pública, la corrupción, el contrabando, el hurto de vehículos, el tráfico de niños, la trata de blancas, y el tráfico ilegal de armas.

PARÁGRAFO SEGUNDO. - EL PROFESIONAL certifica que las personas con las que mantienen relaciones comerciales no se encuentran incursas en actividades relacionadas con el lavado de activos y la financiación del terrorismo – SARLAFT.

AUTORIZACION PARA EL TRATAMIENTO DE DATOS PERSONALES.- PRINCIPIO DE CONFIDENCIALIDAD.- Los datos registrados en este acuerdo cuenta con el principio de confidencialidad según lo establecido en la ley 1581 de 2012 y el Decreto 1377 de 2013, por lo cual **EL PROFESIONAL** garantiza la reserva de la información y certifica que esta solo será usada para los fines de conocimiento del cliente y proveedores para el proceso de Prevención de Lavado de Activos y Financiación del Terrorismo a lo que está obligada de acuerdo a las normas vigentes.

DOCUMENTOS ANEXOS. - Los siguientes documentos forman parte integral del presente contrato:

- Solicitud No 8354 de fecha 10 de octubre de 2024
- Copia del oficio de Justificación de Contratación Directa
- Oficio de fecha 22 de mayo y de 12 de junio de 2024 del Profesional sobre propuesta del Proyecto Aplicado
- Oficio de fecha 17 de junio de 2024 de Comfacauca sobre aceptación del Desarrollo del Proyecto, dirigido a la Universidad Javeriana de Cali.
- Fotocopia de la cédula de ciudadanía del Profesional.
- Copia del RUT actualizado, con fecha de impresión 05 de septiembre de 2024
- Certificado actualizado de Antecedentes Penales y Judiciales de Policía Nacional de fecha 10 de octubre de 2024
- Certificado actualizado de Antecedentes Fiscales de la Contraloría General de la República de fecha de fecha 10 de octubre de 2024
- Certificado actualizado de Antecedentes Disciplinarios de la Procuraduría General de la Nación del Representa de fecha 10 de octubre de 2024
- Certificado de fecha 22 de octubre de 2024, reporte de búsqueda 304 listas
- Copia de la planilla cancelada de la seguridad social del mes de octubre de 2024
- Soportes del título profesional
- Resolución 39 de 16 de septiembre de 2024 delegación COMFACAUCA (01 folio)

DOMICILIO. - Para todos los efectos legales del presente contrato, se fija como domicilio la ciudad de Popayán.

Para constancia se firma en Popayán a veintidós (22) días del mes de octubre de dos mil veinticuatro (2024).

COMFACAUCA:

GLORIA MARISOL VELASCO CHAGUENDO
Jefe Departamento Desarrollo Institucional
Funcionaria Delegada

EL CONTRATISTA:

Jesús Andrés Burbano Gómez

JESÚS ANDRES BURBANO GOMEZ
C.C. No. 1.061.752.152 de Popayán (C)

Acuerdo Octubre2024
Harold C/Shisamo
Solicitud 8354

MÉTODOS BAYESIANOS Y FRECUENTISTAS FLEXIBLES PARA LA CARACTERIZACIÓN Y ANÁLISIS PREDICTIVO Y PRESCRIPTIVO DE CADENAS DE VALOR

INFORME DE RESULTADOS PARA LA CAJA DE COMPENSACIÓN FAMILIAR DEL CAUCA - COMFACAUCA

Jesús Andrés Burbano Gómez

Proyecto Aplicado para optar al título de

Magister en Ciencia de Datos de la Pontificia Universidad Javeriana Sede Cali

Directora

PhD. Isabel Cristina García Arboleda

POPAYÁN, MAYO 15 DE 2025

RESUMEN

La implementación de metodologías de Ciencia de Datos para la caracterización y análisis predictivo y/o prescriptivo de las actividades organizacionales toma fuerza como un instrumento predilecto para la minimización de la incertidumbre y el incremento de la productividad en cada vez más sectores de las economías y la sociedad en general. De tal forma que la ausencia de este tipo de herramientas analíticas puede implicar una limitación importante para el desarrollo de un rango de acción eficiente y planeación efectiva de las organizaciones. Tal es el caso de la Caja de Compensación Familiar del Cauca – Comfacauca, la cual, como figura de administración de recursos parafiscales en el marco del Sistema de Seguridad Social colombiano para el departamento del Cauca, posee un aparato operativo y de caracterización socioeconómica de alcance poblacional que le ha permitido consolidarse como una institución ejemplar a nivel nacional en su medio normativo. Sin embargo, la ausencia de instrumentos de caracterización estadística y optimización para la toma de decisiones al interior de la organización, representa una debilidad que puede comprometer el cumplimiento de sus objetivos estratégicos y metas de cobertura en el mediano y largo plazo. Para resolver este problema, este proyecto aplicado propone tomar como referencia la conceptualización organizacional de cadenas de valor y el contraste entre las metodologías de modelación bayesiana y frecuentista, con el fin de desarrollar un protocolo para la caracterización y análisis de operaciones de la organización en el marco de su Agencia de Empleo y Mecanismo de Protección al Cesante (MPC), el cual pueda servir como insumo para la toma de decisiones óptimas fundamentadas en criterios cuantitativos rigurosos, de cara al incremento de la eficiencia multidimensional y, por ende, del bienestar socioeconómico de la población caucana. En general, esta propuesta sugiere aplicaciones de Ciencia de Datos desde el planteamiento de un problema de clasificación en el contexto de los mecanismos de matching entre oferta y demanda del mercado laboral captado por la organización desde su Agencia, con el fin de identificar sus ventajas para el incremento de la eficiencia y transparencia de sus procesos, así como las ventajas potenciales para las dinámicas de planeación de Comfacauca en general.

TABLA DE CONTENIDO

LISTA DE ANEXOS

- Acuerdo de colaboración para la autorización de uso de datos por parte de la Caja de Compensación Familiar del Cauca.
- Archivo máster de candidatos ordenados por probabilidades estimadas de colocación (Excel).
- Documento extensivo final del proyecto aplicado.

LISTA DE TABLAS

LISTA DE FIGURAS

INTRODUCCIÓN

La Caja de Compensación Familiar del Cauca – Comfacauca es la organización encargada de la administración de fondos de aportes parafiscales para el departamento del Cauca en el marco del Sistema de Seguridad Social Colombiano, a través de la prestación de servicios sociales y el otorgamiento de subsidios de distinta índole a trabajadores afiliados y sus grupos familiares, además de la asignación de recursos específicos entre la población caucana para la protección social en vivienda, atención integral a la primera infancia, jornada escolar complementaria, fomento al empleo y protección al cesante [1].

A pesar de poseer una estructura organizacional sólida, una infraestructura operacional de alcance departamental para el cumplimiento de su propósito misional, y, consecuentemente, una amplia base de datos de caracterización socioeconómica, Comfacauca carece de herramientas para la caracterización estadística de sus actividades de cara al empleo de dinámicas de análisis predictivo y/o prescriptivo para la toma de decisiones.

Si bien se reconoce la consolidación de Comfacauca desde una gestión ejemplar de su fuero normativo como Caja de Compensación Familiar, la ausencia de herramientas para la caracterización y optimización estadística de sus operaciones significa una limitación importante para el incremento de su eficiencia técnica, económica y financiera en el mediano y largo plazo, lo cual puede incluso comprometer el cumplimiento de metas específicas enmarcadas en sus objetivos estratégicos y Visión para 2025. En este sentido, se justifica la pertinencia del desarrollo de protocolos específicos para la medición estadística y análisis predictivo y/o prescriptivo en el marco de las operaciones de la organización.

Este proyecto aplicado aborda esta labor desde la práctica de la Ciencia de Datos, a través de la conceptualización organizacional de cadenas de valor y el contraste entre las metodologías de modelación bayesiana y frecuentista, con el fin de establecer una rutina analítica sobre las operaciones asociadas a la cadena de valor de la Agencia de Empleo y Mecanismo de Protección al Cesante (MPC) de la organización, y proveer como resultado un informe puntual de caracterización y recomendaciones para el desarrollo de escenarios óptimos, las cuales sirvan como insumo de toma de decisiones en busca de mejoras significativas en la cobertura y eficiencia de la Agencia, y, por ende, en una proporción significativa del mercado laboral del departamento del Cauca.

El capítulo inicial de este documento provee la definición del problema a abordar desde su planteamiento y formulación, desglosando la solución planteada desde el objetivo general y objetivos específicos de la investigación, su alcance, justificación y marco de referencia. El capítulo 2 aborda la contextualización del caso de estudio a la luz del concepto de cadena de valor de la Agencia de Empleo, y la estructura de datos a emplear. El capítulo 3 aborda

de manera resumida la modelación contrastadas, para así proveer una justificación concreta de los modelos de clasificación más eficientes para los propósitos de este trabajo en el capítulo 4. Los resultados y su discusión en términos de una caracterización del mercado laboral captado por la Agencia se presentan en el capítulo 5. Dada la justificación de la importancia crítica de este tipo de protocolos para la Agencia de Empleo y Comfacauca en general, el sexto capítulo propone un despliegue del alcance y limitaciones de un proyecto de implementación de métodos de Ciencia de Datos en la Ruta de Atención. Se finaliza con conclusiones y el relacionamiento de trabajos futuros potenciales en el capítulo 7.

1. CONTEXTUALIZACIÓN DEL PROYECTO

1.1 DEFINICIÓN DEL PROBLEMA

1.1.1 PLANTEAMIENTO DEL PROBLEMA

En el contexto de la Comfacauca como caso de referencia, el problema de investigación consiste en la ausencia de un protocolo flexible para la caracterización estadística de la relación entre las estructuras de datos asociadas a las operaciones de la organización, y las cadenas de valor definidas para la provisión de servicios en el marco de sus funciones como Caja de Compensación en el departamento del Cauca, descritas por los manuales de procesos y procedimientos existentes para el manejo de los Fondos a su cargo.

Comfacauca ha venido desempeñando su labor de manera exitosa, siendo encargada de la administración de fondos asociados a la prestación de servicios sociales y el otorgamiento de subsidios a trabajadores afiliados y sus grupos familiares, además de la asignación de recursos específicos entre la población caucana para la protección social en vivienda, atención integral a la primera infancia, jornada escolar complementaria, fomento al empleo y protección al cesante. Siendo una organización consolidada en su funcionamiento y alcance a nivel departamental durante 57 años, Comfacauca cuenta con una estructura organizacional amplia e infraestructura adecuada para el desarrollo y registro de sus operaciones, lo cual le ha permitido aumentar progresivamente su cobertura en el marco de su misión como Caja de Compensación del Sistema Colombiano de Seguridad Social, existiendo un incremento del 9% en la provisión de subsidios con respecto a 2022 [1].

Sin embargo, es de resaltarse que, a pesar de poseer una estructura operativa sólida y bases de datos de alcance poblacional en el Departamento del Cauca, la organización no cuenta con un protocolo analítico para la medición de sus operaciones y modelación estadística de escenarios para la toma de decisiones en función de éstas. Esto representa una limitación importante al potencial de acción de la organización en el departamento, en tanto no posee una herramienta cuantitativa de análisis para la planeación de operaciones que permita reconocer patrones y tendencias de cara al desarrollo de estrategias bajo un riesgo estadístico fiable.

La causa principal de esta situación puede hallarse en la naturaleza monopólica de la organización, pues es la única figura jurídica autorizada para la provisión de subsidios y servicios provenientes de recursos parafiscales en el departamento. La ausencia de “competencia” en su entorno, así como la predisposición normativa a obtener un alcance poblacional en sus operaciones, implican la ausencia previa de incentivos para el desarrollo

de tecnologías y protocolos analíticos que incrementen orgánicamente la demanda por sus servicios o la eficiencia organizacional.

Sin embargo, el hecho de no poseer competidores en su misión no es sinónimo de poseer una cobertura plena en el Departamento o de alcanzar la máxima eficiencia en su estructura organizacional, aspectos clave a mejorar que se reconocen en los objetivos estratégicos de la organización, y, particularmente, en su visión para 2025, la cual propone alcanzar una cobertura con al menos un servicio en el 80% de los afiliados y el 40% de la población del Cauca [1].

Dado el contexto de dispersión del tejido social en el departamento del Cauca y la consecuente vulnerabilidad de una proporción significativa de su población, es notorio el hecho de que el carácter monopólico de la organización no es suficiente para la satisfacción de la meta planteada para el 2025, haciéndose necesario el empleo de recursos adicionales asociados al uso estructurado de información, con el fin de segmentar la distribución de servicios y subsidios de manera más eficiente y con un mayor alcance.

De esta manera, es posible incluso afirmar que la ausencia de protocolos de análisis predictivo y/o prescriptivo en el marco de las dinámicas operacionales de la organización puede representar un impedimento para la progresión de su eficiencia en el futuro, y para el cumplimiento de sus objetivos en el mediano y largo plazo.

1.1.2 FORMULACIÓN DEL PROBLEMA

La necesidad de un protocolo para la medición y modelación estadística en función de las operaciones de la organización se expresó con claridad en el planteamiento del problema que evoca este proyecto aplicado. Ahora, en cuanto a la estructura del problema y posterior proceder con respecto a éste, la siguiente pregunta de investigación retoma el concepto transversal de cadena de valor [2] para determinar su formulación, restringiéndose al caso de la Ruta de Atención de la Agencia de Empleo de Comfacauca:

¿Cómo establecer un protocolo para la optimización estadística del desempeño de la cadena de valor asociada a la Ruta de Atención de la Agencia de Empleo de la Caja de Compensación Familiar del Cauca?

La pregunta fundamental de investigación formulada determina la orientación principal del proyecto aplicado a realizar. Además, el problema formulado implica una especificación en función de preguntas de sistematización.

Inicialmente, es necesario considerar en un sentido descriptivo a la cadena de valor seleccionada para un análisis de la dimensión de este proyecto aplicado, así como su

comportamiento. Para este propósito, la pregunta de sistematización formulada toma una perspectiva técnica del concepto de cadenas de valor analizada en [2]:

¿Cómo identificar y describir el comportamiento de la cadena de valor de la Ruta de Atención de la Agencia de Empleo, por sub actividades sucesivas interconectadas hacia la provisión de bienes o servicios finales?

La siguiente pregunta de sistematización indaga sobre la selección de un modelo(s) estadístico(s) idóneo(s) para los propósitos de análisis predictivo/prescriptivo de este proyecto aplicado:

¿Cuáles son los mejores modelos a emplear para el protocolo de análisis predictivo/prescriptivo de la cadena de valor evaluada?

Además, es necesario considerar el conjunto de criterios técnicos, económicos y contextuales para la determinación del modelo(s) idóneo(s) de análisis predictivo de la cadena de valor:

¿Qué modelo presenta los mejores resultados considerando los criterios de eficiencia técnica, estadística y económica planteados bajo el contexto de la organización?

Finalmente, los resultados del ejercicio práctico a realizar en el marco del proyecto aplicado implican una comunicación efectiva hacia las partes interesadas en este problema de negocio, lo cual evoca la última pregunta de sistematización de este problema:

¿Cómo comunicar efectivamente los resultados del proyecto y sus implicaciones potenciales a las partes interesadas de la organización?

1.2 OBJETIVOS DEL PROYECTO

Como solución al problema planteado, la implementación de un protocolo analítico que fundamente la toma de decisiones en hechos estadísticos, puede significar una herramienta clave para la proliferación de mejoras en la eficiencia multidimensional de la Agencia de Empleo, y de Comfacauca en general, convirtiéndose en un instrumento indispensable para su funcionamiento en el largo plazo. El desarrollo de tal protocolo puede proveer un marco estandarizado para la medición de desempeño de cadenas de valor, y posterior mejoramiento de la cobertura, eficiencia (técnica, financiera y organizacional) y planeación de las actividades de la organización, a través del análisis de escenarios óptimos provisto por métodos formales en el marco de la Ciencia de Datos.

En términos puntuales, el proyecto aplicado propuesto abordará el problema formulado a través de la caracterización de la cadena de valor de la Agencia de empleo desde la identificación de las estructuras de datos clave en las actividades encadenadas de la Ruta de Atención, y el contraste de métodos bayesianos y frecuentistas de aprendizaje automático, en función de criterios de eficiencia técnica, computacional y restricciones normativas de la organización. Nótese que las respuestas concretas a la pregunta central de investigación y sus preguntas auxiliares de investigación, se plasman respectivamente en el objetivo general del proyecto y sus objetivos específicos.

1.2.1 OBJETIVO GENERAL

Diseñar y aplicar un protocolo para la optimización estadística del desempeño de la cadena de valor de la Ruta de Atención de la Agencia de Empleo de la Caja de Compensación Familiar del Cauca - Comfacauca.

1.2.2 OBJETIVOS ESPECÍFICOS

- Caracterizar la cadena de valor asociada a las actividades de la Ruta de Atención de la Agencia de Empleo de Comfacauca, desde el registro de los candidatos al mercado laboral, hasta la evaluación de sus resultados con respecto a las vacantes emitidas por la demanda laboral del Departamento del Cauca.
- Modelar el comportamiento estadístico de la cadena de valor de la Ruta de Atención por medio del contraste de eficiencia entre métodos bayesianos y frecuentistas de aprendizaje supervisado.

- Evaluar la eficiencia de los modelos seleccionados bajo criterios técnicos, económicos y asociados al contexto de la organización.
- Proveer un informe de resultados con una descripción directa y comprensible de sus impactos potenciales en materia socioeconómica y de eficiencia (técnica, financiera y organizacional).

1.2.3 ALCANCE

El proyecto aplicado a desarrollar resume su alcance de manera sucesiva en función del siguiente conjunto de etapas de la investigación:

- Planteamiento y contextualización de la cadena de valor analizada y sus problemas de negocio.
- Caracterización y análisis exploratorio de variables relevantes en la cadena de valor de la Ruta de Atención de la Agencia de Empleo.
- Selección e implementación de modelos estadísticos.
- Contraste y validación de modelos ejecutados.
- Determinación de resultados finales y desarrollo de informe entregable con conclusiones y recomendaciones estratégicas para la organización.

En este sentido, el alcance del proyecto aplicado propuesto va desde el planteamiento del problema de negocio alrededor de la cadena de valor de la Ruta de Atención de la Agencia de Empleo de Comfacauca, hasta la entrega de recomendaciones en función del análisis cuantitativo realizado desde las estructuras de modelación estadística empleadas.

Es necesario reconocer dos restricciones clave del caso analizado:

- El alcance de los patrones a analizar se limita al departamento del Cauca y sus subdivisiones, y a la información relevante disponible en las bases de datos de la Agencia.
- El grado de acceso a información sobre los afiliados será regido por las políticas de protección de datos de la organización, y se hará de forma completamente anonimizada.

Es pertinente enfatizar que la implementación práctica por parte de la organización de las recomendaciones generadas en función de esta investigación no se halla dentro del alcance de este proyecto aplicado. Así, una evaluación de impactos post-proyecto se presenta como una sofisticación potencial al trabajo propuesto, tanto para la organización y su dinámica institucional en términos prácticos y de eficiencia, como para el abordaje de implicaciones interesantes desde la academia.

1.2.4 JUSTIFICACIÓN

La justificación esencial para el desarrollo de este proyecto aplicado se basa en la propiedad básica que los métodos de Ciencia de Datos otorgan a las organizaciones en la industria, y, en general, en la sociedad. Ésta propiedad es la capacidad de formular instrumentos específicos para la cuantificación y monitoreo de las actividades y operaciones de manera estructurada, permitiendo el incremento de la eficiencia multidimensional desde la reducción de la incertidumbre y del riesgo en la ejecución de recursos valiosos.

La idea anterior se encuentra perfectamente alineada con el planteamiento y formulación del problema desarrollados anteriormente, pues se reconoce plenamente que la ausencia de un protocolo analítico próximo a la Ciencia de Datos como disciplina en Comfacauca implica la pérdida de los beneficios que puede traer un instrumento para la reducción de la incertidumbre de cara a la mejora en el desempeño de sus actividades misionales, particularmente considerando las metas cuantitativas enmarcadas en la visión de la organización.

La idoneidad del proyecto no sólo se evidencia desde su utilidad potencial como herramienta de la organización, también se justifica desde la disponibilidad de la información necesaria para llevarlo a cabo correctamente, pues el alcance departamental de Comfacauca implica que tanto sus operaciones como sus bases de datos de afiliados se hallan inmersas en la actividad socioeconómica, productiva y laboral de una porción mayoritaria de la población del Cauca. Lo anterior hace del proyecto aplicado una propuesta viable, y, además, sujeta a sofisticaciones y continuaciones que pueden arrojar conclusiones y otros instrumentos que maximicen el beneficio de tal base de datos de corte socioeconómico en el departamento.

Finalmente, la implementación de un protocolo en el sentido del proyecto aplicado propuesto no sólo puede significar una mejora global en la eficiencia de las operaciones de la Agencia de Empleo, sino que, dada la misión y naturaleza jurídica de la organización, tal mejoramiento generalizado en la eficiencia de Comfacauca se hallaría directamente traducido en impactos socioeconómicos positivos para la región, implicando beneficios potenciales en rubros clave de la economía departamental, específicamente, en el mercado laboral.

1.3 MARCO DE REFERENCIA

Con sede principal ubicada en Popayán, Cauca, Comfacauca es la organización asignada como Caja de Compensación Familiar para el departamento del Cauca, regida por la Ley 21 de 1982, la cual dicta las disposiciones para la administración de fondos de subsidio familiar y demás rubros correspondientes a los recursos provenientes de aportes parafiscales en el marco del Sistema de Seguridad Social Colombiano.

Con poco más de 57 años de operaciones, la organización viene desempeñando su rol de administradora de fondos de aportes parafiscales a través de la prestación de servicios sociales y el otorgamiento de subsidios a trabajadores afiliados y sus grupos familiares, además de la asignación de recursos específicos entre la población caucana para la protección social en vivienda, atención integral a la primera infancia, jornada escolar complementaria, fomento al empleo y protección al cesante; reportando para el año 2023 un total de ingresos por aportes parafiscales de $147.646 millones por parte de 8.936 empleadores afiliados, los cuales fueron administrados para la provisión de servicios y subsidios entre 313.665 afiliados entre trabajadores, cónyuges y sus personas a cargo.

A continuación, se presenta un resumen de las operaciones de los rubros principales de la organización sobre los fondos a su cargo para el año 2023 según [1], con el fin de ilustrar la pluralidad de servicios, y, por ende, de cadenas de valor estructuradas por Comfacauca:

- En general, la organización realizó entrega de 968 mil cuotas monetarias, equivalentes a $20.483 millones para subsidiar la prestación de servicios en educación, recreación, capacitación y crédito, así como $10.948 millones de pesos destinados a la entrega de subsidios en especie.
- En el marco del Fondo de Vivienda de Interés social – FOVIS la organización asignó 247 subsidios de vivienda, equivalentes a un valor de $8.328 millones al 92% de los trabajadores con salarios por debajo de los 2 SMMLV.
- En el marco de su Mecanismo de Protección al Cesante – MPC, se registraron desembolsos de subsidios de desempleo por $14.069 millones a 3.738 beneficiarios, además, se registró la capacitación de 33.372 trabajadores activos de manera completamente subsidiada. Adicionalmente, la Agencia de Empleo de la organización registró 1.117 empresas que inscribieron 6.032 vacantes, ocupadas por 4.824 oferentes.
- En cuanto a servicios complementarios en educación y cultura, la organización financia o administra educación de educación media y superior en Popayán y el norte del Cauca.
- En materia de sus programas de Atención Integral a la Primera Infancia y Jornada Escolar Complementaria, se beneficiaron 16.846 jóvenes de 34 municipios del

departamento del Cauca, siendo esto un referente regional y nacional en materia de atención institucional a población vulnerable.

- Finalmente, en materia de recreación, deportes y turismo, la organización registró 1.283.968 usos en sus 6 centros recreativos y deportivos, ubicados en Popayán, Puerto Tejada y Santander de Quilichao.

Como puede apreciarse, el grueso de operaciones y cadenas de valor para la provisión de bienes y servicios específicos por parte de la organización se ve descrito por una ramificación de múltiples ramas y componentes, decantando en un complejo de servicios de alcance poblacional en el departamento, el cual se halla estrechamente vinculado con información de base del mercado laboral y sus características socioeconómicas.

Dado el problema planteado en secciones anteriores, el escenario que propone Comfacauca se muestra como propicio para el uso de métodos de Ciencia de Datos con orientaciones predictivas y/o prescriptivas, este proyecto provee un ejercicio piloto para la implementación de estos métodos en la cadena de valor correspondiente a la Ruta de Atención de la Agencia de Empleo de la organización.

Como subdivisión encargada de la administración del Mecanismo de Protección al Cesante, la Agencia de Empleo de Comfacauca basa sus funciones en el despliegue de un conjunto de subprocesos secuenciales establecidos con el fin de subsidiar a la población desempleada del departamento desde un conjunto de actividades que, principalmente, decantan en el facilitar la consecución de nuevos empleos en función de vacantes disponibles en el mercado laboral captado por la organización, o en la gestión y rastreo de las etapas de desempleo de los candidatos registrados a través de capacitaciones y/o subsidios monetarios. Este conjunto de subprocesos se estructura en una Ruta de Atención, la cual determina la cadena de valor puntual a analizar en este proyecto aplicado.

Dado lo anterior, y como se expondrá más adelante en la descripción de la cadena de valor que fundamenta esta investigación, es crítico resaltar que el eje de la Ruta de Atención de la Agencia de Empleo gira en torno a la determinación de los candidatos al mercado laboral en el departamento como colocados, es decir, como receptores de nuevos empleos vacantes, o como no colocados, es decir, como población que aún permanece desempleada.

De esta manera, se tiene que los resultados esperados de esta investigación y sus implicaciones se basan específicamente en la aplicación de métodos de Ciencia de Datos para obtener un marco estadístico sobre el cuál predecir de manera confiable la posibilidad de los candidatos de obtener un empleo dada la estructura de la demanda laboral del departamento a partir de los atributos inmersos en las bases de datos de la Agencia, y, dada la construcción de este marco estadístico, caracterizar cuantitativamente el panorama del mercado laboral Caucano para generar recomendaciones, tanto para el proceso establecido por la organización para la Ruta de Atención de la Agencia en el sentido de una cadena de

valor, como para el análisis normativo del mercado laboral del departamento y sus implicaciones socioeconómicas. Así, es pertinente resaltar que el núcleo de esta investigación, dado su marco contextual, consiste en un problema de clasificación de los candidatos al mercado laboral en función de la información disponible con respecto a los mismos en las bases de datos de la organización.

2. DESCRIPCIÓN DE LA CADENA DE VALOR DE LA RUTA DE ATENCIÓN Y CONSECUCIÓN DE ESTRUCTURAS DE DATOS RELEVANTES

2.1 LA CADENA DE VALOR DE LA RUTA DE ATENCIÓN DEL SERVICIO DE EMPLEO DE COMFACAUCA.

El núcleo de este proyecto aplicado consiste en esquematizar las actividades y servicios provistos por la organización Comfacauca en el sentido de cadenas de valor, de tal forma que los procesos fijados por ésta misma puedan entenderse como un conjunto de actividades estructuradas secuencialmente de cara a la entrega de servicios específicos, y así contrastar propuestas de modelación que arrojen resultados congruentes de cara a la optimización de sus problemas de negocio.

En este sentido, y dada la amplia gama de servicios que provee Comfacauca, se decidió priorizar en este ejercicio las actividades de la Agencia de Empleo, el cual actúa en el marco de la Ley 1636 de 2013 y sus complementos para ejercer en el Cauca el Mecanismo de Protección al Cesante (MPC), esencialmente desde la provisión de asistencia en la búsqueda de empleo para los afiliados cesantes, así como el acompañamiento en materia de subsidios y capacitación. Todo lo anterior se ejecuta a través de una Ruta, la cual puede comprenderse como la cadena de valor principal del Servicio de Empleo.

A continuación, se recopilan los problemas de negocio a evaluar en el marco de las actividades del Servicio de Empleo:

- Necesidad de formulación de criterios cuantitativos específicos para el incremento en la eficiencia y eficacia de procesos de remisión y selección en vacantes ofrecidas por las empresas hacia los cesantes que se registran en la ruta de atención del Servicio de Empleo.
- Identificación de habilidades y áreas de formación a priorizar en las capacitaciones a llevar a cabo por parte del Servicio de Empleo.
- Caracterización de condiciones laborales generales en el panorama del mercado laboral del Cauca desde la base de datos del Servicio de Empleo.
- Identificación de parámetros como insumos cuantitativos generales para la toma de decisiones por parte del Servicio.

De acuerdo a la priorización de problemas de negocio presentada anteriormente, se obtuvo una esquematización del proceso secuencial a través del cual el Servicio de Empleo administra el Mecanismo de Protección al Cesante en el sentido de una cadena de valor. A manera de fases, esta estructura de cadena de valor puede verse de la siguiente forma:

1. El candidato se registra de manera autónoma o asistida en la plataforma de Bolsa de Empleo de Comfacauca, facilitando el conjunto de atributos estandarizados con los que participará en el mercado laboral en busca de un empleo.

2. Dadas las vacantes publicadas por las empresas adscritas al Servicio, se realiza la remisión del perfil de posibles candidatos a obtener la posición demandada.

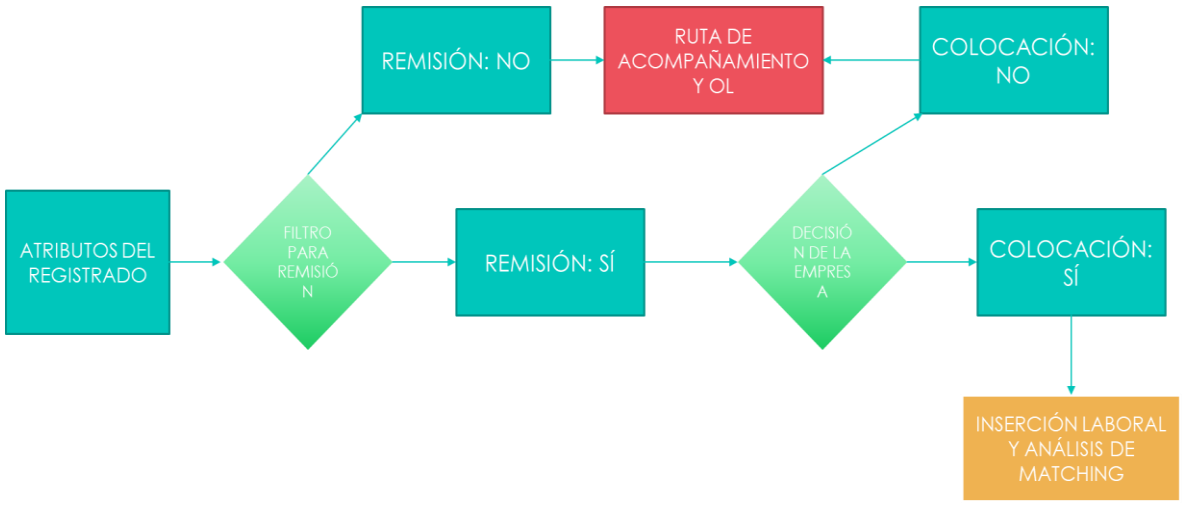
3. Se efectúa el proceso de selección por parte de las empresas, tras lo cual se decide si los candidatos son colocados o no en las vacantes.

4. En caso de ser colocados, el Servicio registra el caso de inserción del candidato en el mercado laboral, acompañado de las condiciones del contrato obtenido.

5. En caso de no ser colocados, el Servicio continúa con el monitoreo de la ruta de atención a través del ofrecimiento de capacitaciones y talleres de orientación en el mercado laboral. Además, se realiza la entrega de subsidios de acuerdo al marco legal existente para este propósito.

La siguiente figura presenta un diagrama de flujo para la cadena de valor de la ruta de atención del Servicio de Empleo.

Figura 1. Cadena de valor de la Ruta de Atención del Servicio de Empleo.



2.2 ESTRUCTURA DE DATOS.

Tras la luz verde para la facilitación de información por parte de la organización, el proceso de recolección de información operacional y de afiliados en el marco de la cadena de valor analizada se acordó en dos fases.

En la primera fase, se realiza la entrega de una muestra inicial de los datos disponibles para el proceso analítico y de modelación, sobre el cual se definen los procedimientos de transformación y pre procesamiento necesarios para el contraste de los modelos, de tal forma que sea posible definir un prototipo de esquema para modelación predictiva y prescriptiva.

Dada la identificación de estos procedimientos, se elevan requerimientos de confección y escalamiento del dataset para la ejecución de los modelos finales en la segunda fase. Con la entrega de los datos a escala real en la segunda fase, se procede a los ejercicios de modelación a escala real.

Así, para la consecución del dataset a escala real, y dada la disponibilidad de información por parte de la organización, se recibió un archivo Excel (xlsx) con 3 sheets relevantes para el ejercicio de limpieza y pre procesamiento, del cual se obtuvieron 87.735 casos de aplicación registrados indexados por sus IDs. Las 3 sheets del archivo base describen los componentes clave de oferta y demanda laboral en la cadena de valor a analizar, de la siguiente forma:

- Sheet 1: Registrados (38 atributos).
- Sheet 2: Colocados (29 atributos).
- Sheet 3: Vacantes (28 atributos).

Dado el archivo de base de datos facilitado, se realizó el pre procesamiento y limpieza del dataset. Los métodos empleados se resumen en los siguientes pasos secuenciales (los cuales se llevaron a cabo empleando R Markdown):

1. Merging de las sheets relacionadas empleando la operación left join, a través del cual se obtuvo un dataset crudo con 87.735 casos de aplicaciones a vacantes.
2. Selección de los atributos de mayor relevancia para los ejercicios necesarios de modelación.
3. Extracción de etiquetas binarias de clase para modelos de aprendizaje supervisado: Colocados y Colocados por sector.
4. Limpieza de dataset: Renombramiento de columnas e imputación de missings.

Así, se obtuvo el dataset listo para incurrir en análisis descriptivo y posteriores procesos de modelación.

2.3 ANÁLISIS EXPLORATORIO

Dado el ejercicio de pre procesamiento y limpieza del dataset crudo, se obtuvo un dataset limpio inicial para modelación, con 87.735 observaciones asociadas a casos de aplicación de los registrados a vacantes. El dataset obtenido se halla descrito por la siguiente tabla.

Tabla 1. Variables seleccionadas para análisis en el dataset.

Atributo	Tipo	Descripción
ID	ID	
Edad	Numérica	Edad
CanalRegistro	Categórica	Medio de registro (asistido o autoregistro)
Genero	Categórica	Género
PercHojaVida	Categórica	Porcentaje de completitud del formulario de hoja de vida
CiudadResidencia	Categórica	Municipio de residencia
NivelEstudio	Categórica	Nivel de estudio
Urb_Rural	Categórica	Vive en zona rural o urbana
AspiracionSalarial.x	Categórica	Aspiración salarial
TituloObtenido	Texto	Título obtenido descrito de manera específica
UltimaExperienciaLaboral.x	Texto	Última experiencia laboral descrita de manera específica
Colocado	Etiqueta binaria de clase	Colocado (Sí o No)
ColocadoBinary	Etiqueta multiclase	Colocado (Colocados por sector económico y no colocados)

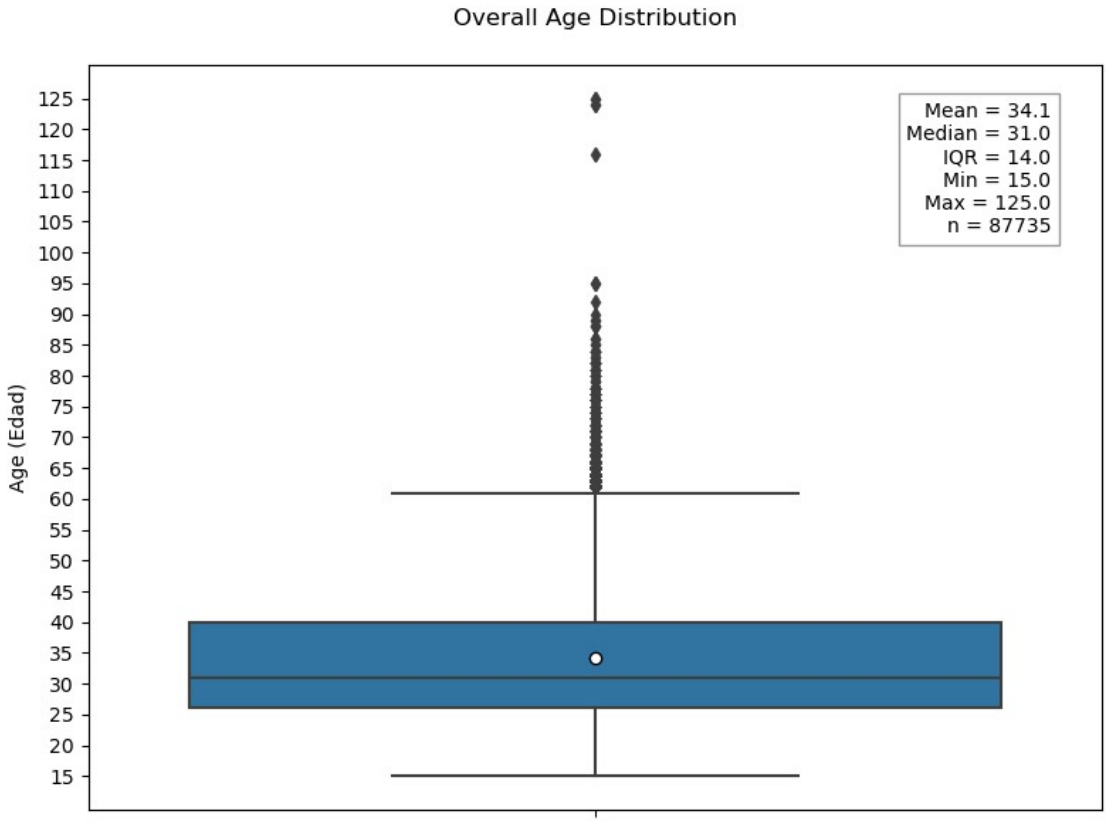
Nótese que los atributos seleccionados se obtuvieron de manera acordada con los encargados de las operaciones del servicio de empleo bajo criterios de relevancia en el problema de negocio y conveniencia estadística. En general, los atributos para el modelo de clasificación son en su mayoría categóricos, y se requirió el ejercicio de un procedimiento de clustering para los dos tributos de texto, en tanto de manera previa se sospecha que son claves para una estructura de correlación que decante en un modelo eficiente de clasificación de candidatos. A continuación, se presenta un análisis descriptivo inicial para cada una de las variables analizadas, lo cual indica la línea base de caracterización de la proporción del mercado laboral que capta el Servicio de Empleo, dado el dataset facilitado

por la organización. En cuanto al análisis de correspondencia requerido para esta fase exploratoria, las salidas puntuales de los test de correlación Chi-cuadrado sobre las variables categóricas del dataset se consignan en el Apéndice.

Edad:

La variable edad es la única variable numérica en el dataset. Para las 87.735 observaciones analizadas, se tiene que de los casos de aplicantes a las vacantes, la edad mediana se ubica sobre los 31 años, existiendo además un rango intercuartílico de entre los 26 y los 40 años. Lo anterior caracteriza el rango medio sobre el cual es posible hallar al grueso de los aplicantes a las vacantes consideradas en el dataset. A manera de hipótesis previa, y, considerando la baja tasa de colocación que se expresará más adelante, se puede observar una alta incidencia del desempleo en edades jóvenes para el departamento del Cauca, sin embargo, también es posible afirmar que, al menos para esta muestra, se entiende que el departamento dispone en gran medida de fuerza laboral joven. El siguiente boxplot resume el comportamiento de la distribución empírica de la variable Edad en el dataset.

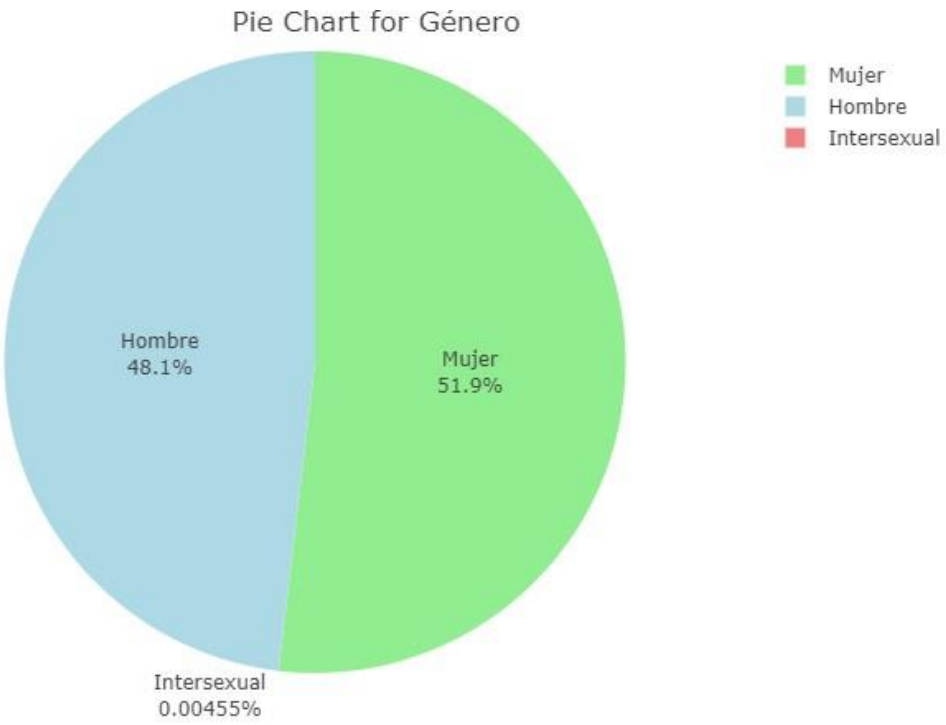
Figura 2. Boxplot para la variable Edad.



Género:

Como lo denota el gráfico de torta desplegado a continuación, el comportamiento de la frecuencia del género en el dataset denota una ligera predominancia de las mujeres en la fuerza laboral para la muestra analizada. Así, puede denotarse un patrón en el cual se evidencia una mayor participación femenina en el mercado laboral. Sin embargo, debe mencionarse que el test de correlación Chi-cuadrado denota una correlación estadísticamente significativa de propensión de los hombres a ser colocados en mayor medida.

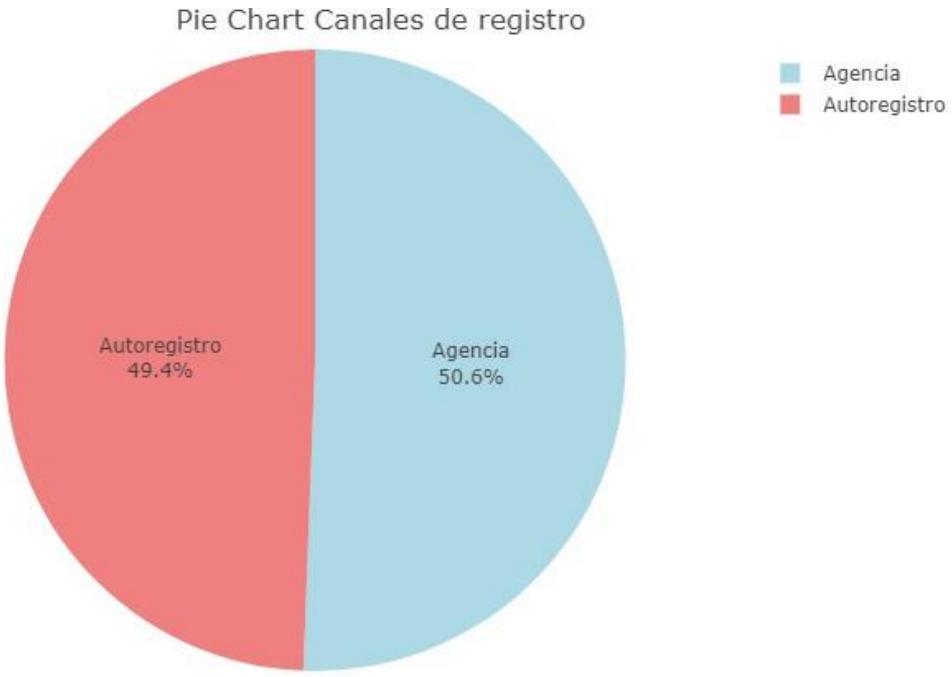
Figura 3. Pie chart para la variable Género.



<u>Medio de Registro</u>:

El comportamiento de la frecuencia del medio empleado para el registro en la bolsa de empleo también puede denotar una aproximación al nivel de manejo de herramientas tecnológicas por parte de los aspirantes al mercado laboral. Existiendo así un patrón de ligera dominancia del registro por medio de la agencia entre los aplicantes. Además, la prueba de correlación Chi-cuadrado denota una alta asociación de los registrados en la agencia con la obtención de empleo.

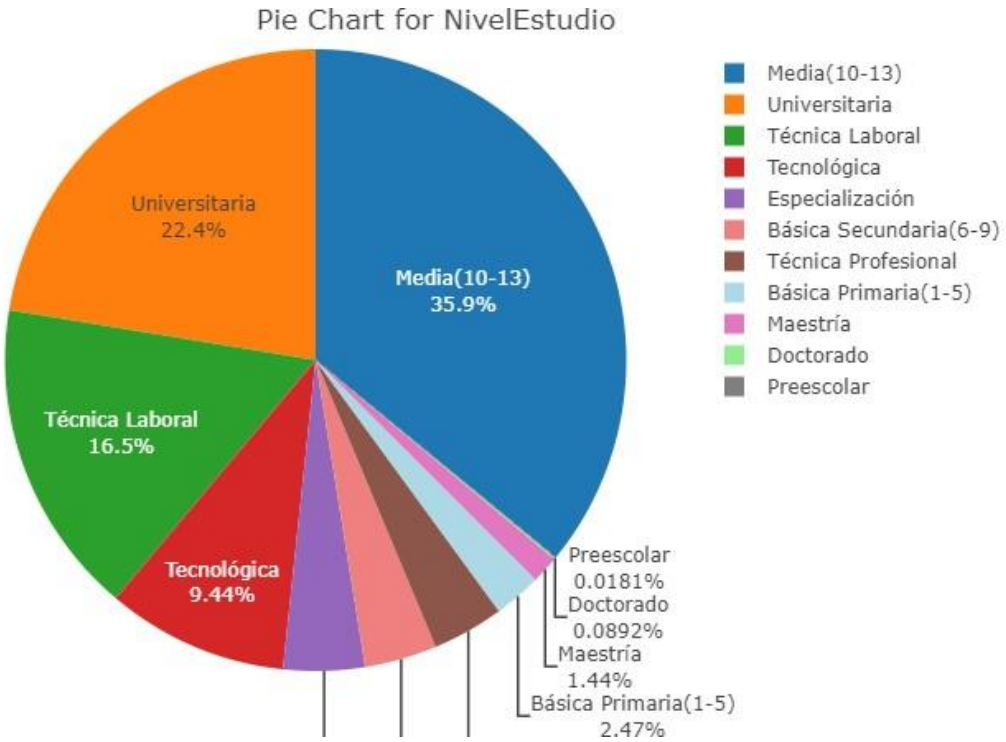
Figura 4. Pie chart para la variable Canal de Registro.



<u>Nivel de Formación Educativa</u>:

En cuanto al nivel de formación educativa de los aplicantes, de acuerdo a las categorías estándar del sistema colombiano, el gráfico de torta desplegado debajo denota una predominancia de la educación media, con una proporción del alrededor del 35.5%, seguido de la educación universitaria (14.5%), la educación técnica laboral (11.4%), y la formación tecnológica (6,51%). De tal forma que las categorías subsiguientes se muestran con proporciones bastante inferiores. Como hipótesis previa es importante resaltar el panorama donde la formación profesional no es una regla dominante entre los aspirantes, lo cual puede influir en gran medida, no sólo sobre una alta tasa de fallos en la aplicación a vacantes requiriendo niveles de formación por encima de la educación media, sino en una idea previa sobre una escasez significativa de oferta laboral calificada.

Figura 5. Pie chart para la variable Nivel de Estudio.

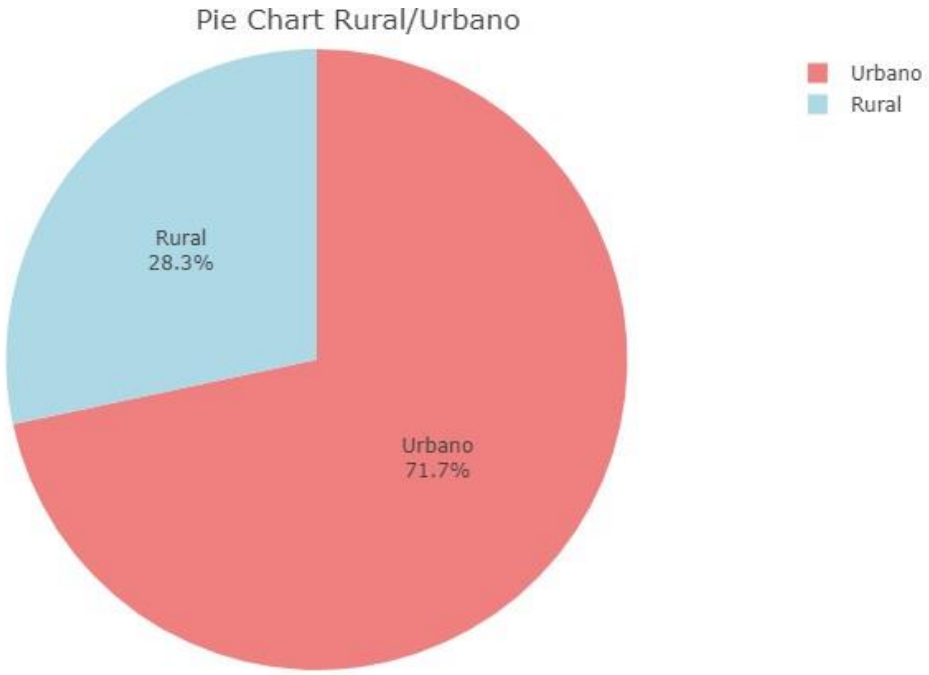


En cuanto al nivel de formación educativa de los aplicantes, de acuerdo a las categorías estándar del sistema colombiano, el gráfico de torta desplegado debajo denota una predominancia de la educación media, con una proporción del alrededor del 35.9%, seguido de la educación universitaria (22,4%), la educación técnica laboral (16.5%), y la formación tecnológica (9,44%). De tal forma que las categorías subsiguientes se muestran con proporciones bastante inferiores. Como hipótesis previa es importante resaltar el panorama donde la formación profesional no es una regla dominante entre los aspirantes, lo cual puede influir en gran medida, no sólo sobre una alta tasa de fallos en la aplicación a vacantes requiriendo niveles de formación por encima de la educación media, sino en una idea previa sobre una escasez significativa de oferta laboral calificada. El test de correlación Chi-Cuadrado denota significancia estadística en la asociación mayoritaria entre la obtención de empleo y el nivel educativo Medio y Técnico.

Ruralidad/Urbanidad:

El gráfico de torta debajo denota un comportamiento esperado, en tanto existe una marcada predominancia de pertenencia a zonas urbanas por parte de los aplicantes. Esto es congruente con la disponibilidad de empleos y fuerza laboral de mayor capacitación en las zonas urbanas del departamento. Esto es congruente con la orientación en los resultados del test de correlación Chi-cuadrado.

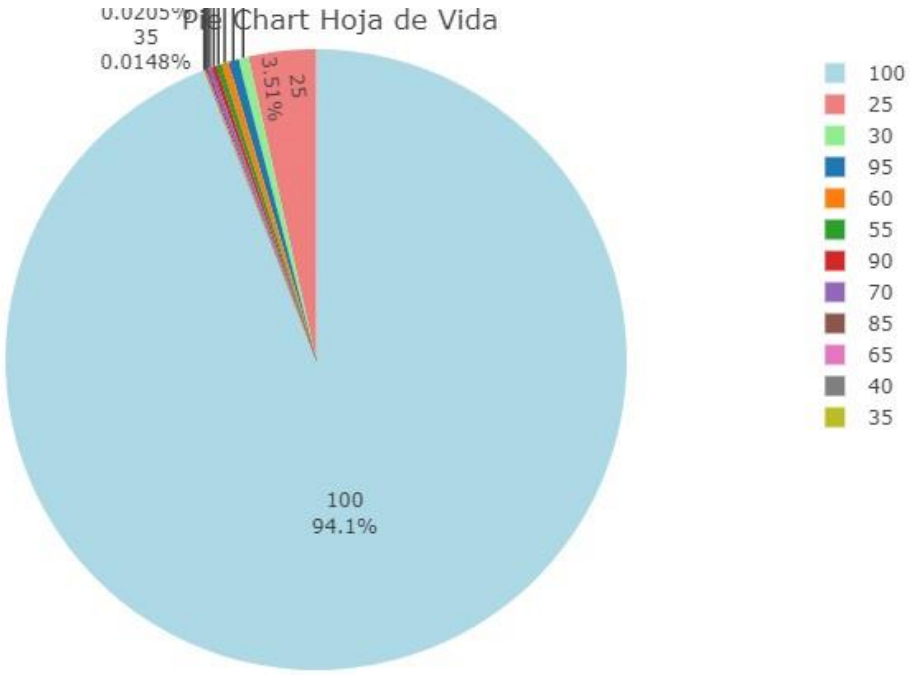
Figura 6. Pie chart para la variable Rural/Urbano.



Hoja de Vida (Porcentaje de completitud en plataforma):

La variable categórica asociada a la completitud de la hoja de vida en la plataforma de Bolsa de Empleo denota una especie de aproximación al grado de manejo de las herramientas tecnológicas que ofrece el Servicio por parte de los aspirantes. Además, esta variable también puede servir como una aproximación a una revisión del bajo nivel de capacitación de la fuerza laboral en el departamento. En congruencia con el gráfico desplegado a continuación, el test de correlación Chi-Cuadrado denota una asociación predominante de la obtención de empleo a las hojas de vida llenas al 100%.

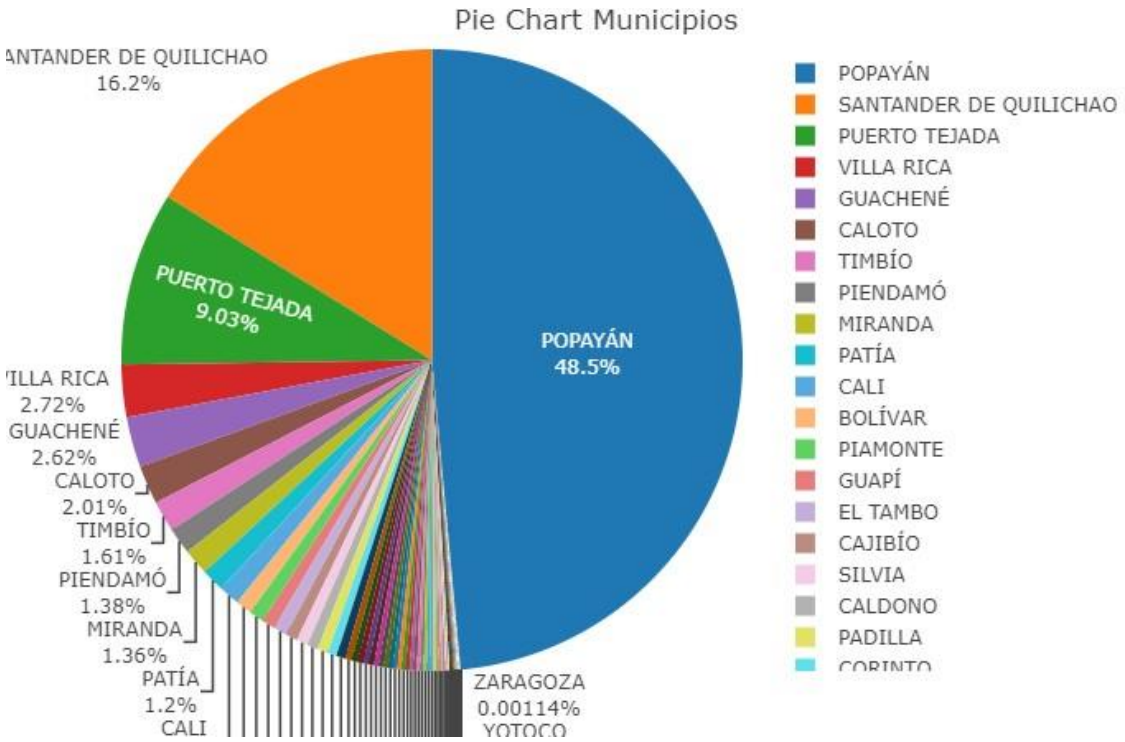
Figura 7. Pie chart para la variable Porcentaje de Hoja de Vida.



Municipios de Residencia:

La distribución de la frecuencia de los aplicantes con respecto a los municipios de residencia denota una característica común en la dinámica económica del departamento, pues, la predominancia de los municipios de Popayán (48.5%), Santander de Quilichao (16.2%), y Puerto Tejada (9.03%), es congruente con la aglomeración de los enclaves industriales y productivos alrededor de la capital del departamento y las zonas francas ubicadas al norte y en frontera con el alto volumen de actividad económica del Valle del Cauca. En este sentido, no sorprendería encontrar que, al formalizar un modelo de clasificación, la residencia en estos municipios sea un factor de alta influencia en la tasa de colocación del mercado laboral. Esta hipótesis previa es soportada por los resultados del test de correlación Chi-cuadrado con respecto a la etiqueta binaria de colocación.

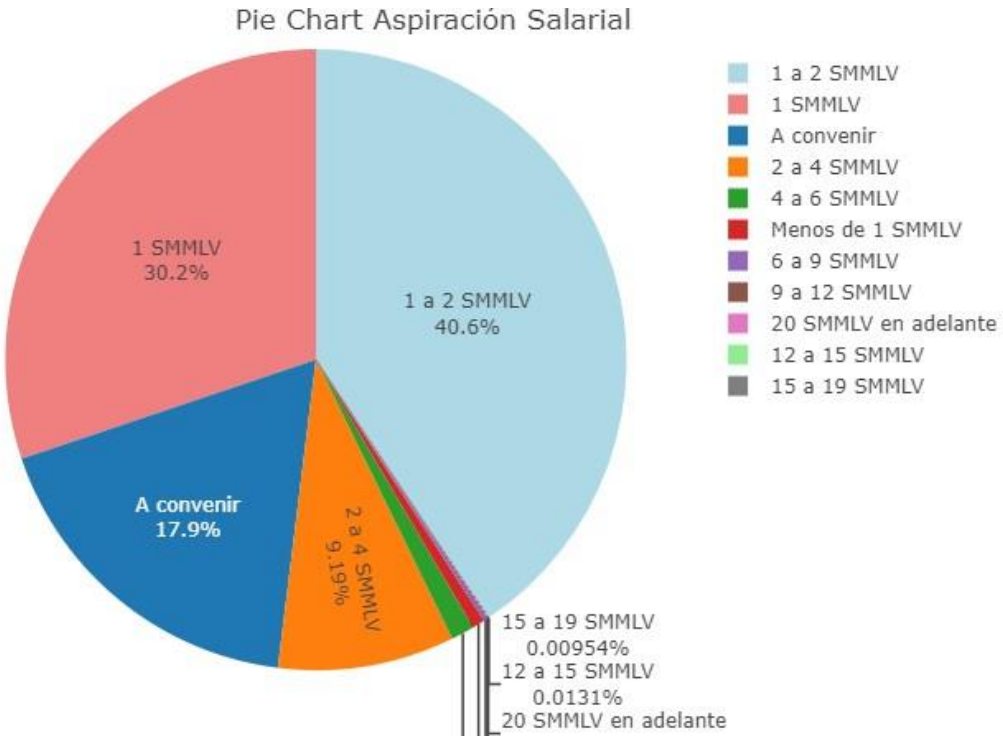
Figura 8. Pie chart para la variable Municipio de Residencia.



<u>Aspiración Salarial</u>:

Como lo muestra el gráfico de torta desplegado debajo, se tiene que la mayoría de los casos de aplicación proyecta a aplicantes con aspiraciones salariales de 1 a 2 SMMLV (40.2%), 1 SMMLV (30.2%), y bajo la categoría libre A convenir (17.9%). De tal forma que aspiraciones salariales de mayor volumen se presentan en proporciones muy poco siginificativas. Este resultado es congruente con la hipótesis previa provista por el test de correlación Chi-cuadrado, el cual denota una alta asociación de obtención de empleo con aspiraciones de 1 a 2 SMMLV. Nótese cómo la descripción de esta categoría se alinea con la descripción del nivel de estudio de los candidatos.

Figura 9. Pie chart para la variable Aspiración Salarial.

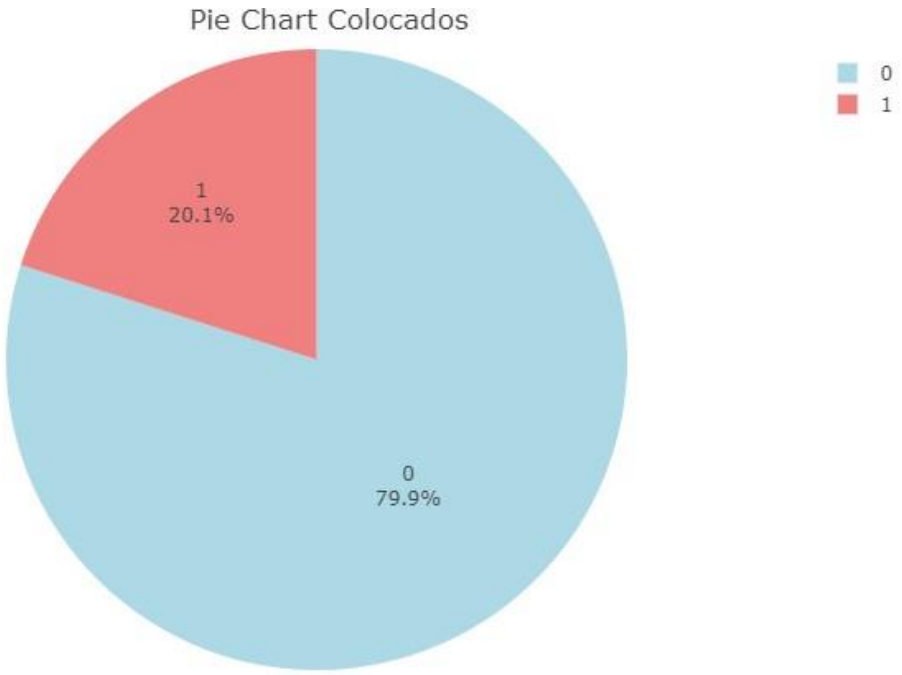


Variables de texto: Título Obtenido y Última Experiencia Laboral.

Las variables de texto en el dataset representan un obstáculo en la etapa exploratoria y de pre procesamiento, en tanto se sospecha que contienen información clave respecto a los efectos diferenciales que determinan la obtención de empleo, pero a la vez implican una enorme dispersión en las posibles respuestas debido a que se trata de entradas libres de texto en la plataforma. Así, fue necesaria la generación de categorías puntuales en función de textos a través de un modelo de clustering. Esto se consigna en el siguiente capítulo.

Finalmente, el gráfico de torta desplegado debajo denota la línea base fundamental del dataset, en tanto describe la frecuencia de obtención de empleo en la etiqueta binaria ColocadoBinary, de cara a la construcción de modelos de clasificación bajo aprendizaje supervisado.

Figura 10. Pie chart para la etiqueta de clase (1: Colocado, 2: No colocado).



Nótese cómo la clase positiva presenta una frecuencia bastante baja con respecto a la clase negativa. Lo anterior es una aproximación a la baja ratio de obtención de empleo debido a la estructura económica captada por la Agencia con respecto al departamento del Cauca. Además, esta situación anuncia de manera previa el sesgo inmerso en la modelación de aprendizaje sobre la clase negativa. Lo anterior evoca el conjunto de recursos y la estructura de resultados a presentar tras la selección de los modelos más eficientes en el siguiente capítulo.

3. METODOLOGÍAS DE MODELACIÓN

En este capítulo se despliega un resumen de los ejercicios de modelación propuestos para el problema de clasificación planteado. Además, se especifica un proceso de modelación previo a la construcción de los clasificadores que determinaron la orientación principal esta investigación. Como se mencionó en la sección correspondiente al análisis exploratorio, los atributos Título Obtenido y Última Experiencia Laboral se presentan como entradas de texto libre por parte de los candidatos registrados en la Ruta de Atención. Así, en tanto la especificación del título y experiencia laboral previa del candidato se consideran a priori como un atributo de alta importancia para el problema de clasificación a resolver, es indispensable realizar un ejercicio de modelación que permita el tratamiento normalizado de estas entradas de texto como insumo proveniente del dataset.

3.1 Normalización y agrupamiento de entradas de texto por medio de K-means clustering.

La información contenida en las variables de texto de Título Obtenido y Última Experiencia Laboral se considera como indispensable de cara a la clasificación de mejores candidatos y la determinación de factores clave para la consecución de empleos en el mercado laboral del Departamento.

Sin embargo, la multiplicidad y heterogeneidad de las posibles entradas de texto implica una dificultad para la correcta lectura de estos atributos como insumos numéricos de un clasificador. En este sentido, se optó por aplicar una metodología de clustering sobre las variables de texto, la cual genera columnas de etiquetado de cada observación en un cluster, otorgando una señalización operable para el título específico de los candidatos y su última experiencia laboral, en el sentido de convertir las columnas con entradas de texto a columnas categóricas con múltiples instancias.

El ejercicio de clustering se realizó en Python, ejecutando un pipeline de pre procesamiento de texto y clustering por medio de K-means. Bajo los siguientes pasos secuenciales:

- Carga de datos y selección de columnas de texto.
- Pre procesamiento para entradas de texto en español: remoción de puntuaciones y stopwords, modificación de palabras con acentos, ajuste de espacios en blanco.
- Ejecución de entrenamiento de modelo de clustering por medio de K-means, con ajuste manual de hiperparámetros (cantidad de clusters).

- Salida de resultados y asignación de etiquetas de cluster en las nuevas columnas: EduCluster para TituloObtenido, y ExperienciaCluster para UltimaExperienciaLaboral.
- Evaluación sucesiva de la calidad de clustering por medio de métricas Silhouette index, Davies-Bouldin index y Caliski-Harabasz index; hasta alcanzar un nivel óptimo conjunto.

En general, se optó por la realización de este ejercicio a través del ajuste manual sucesivo del método K-means por su amplia ventaja frente a otros modelos en términos de interpretabilidad, cantidad de hiperparámetros (sólo uno, cantidad de clusters) y eficiencia computacional.

El método K-means consiste en la determinación de clusters a través del ajuste iterativo de las observaciones consideradas a distancias mínimas con respecto a los centroides muestreados por este algoritmo de aprendizaje no supervisado. Este proceso se realiza de manera simple y de fácil interpretación, en tanto su medida de ajuste parte de la evaluación de distancias euclidianas entre las observaciones del dataset [3]. Para el caso de entradas de texto, se requiere la normalización de las mismas a través de columnas binarias que indiquen la ocurrencia de cada palabra considerada.

Es importante resaltar que este método se escogió sin un contraste frente a otras propuestas de modelación por aprendizaje no supervisado, en tanto la tarea a satisfacer consistió en el agrupamiento de palabras clave, no en la identificación de patrones puntuales de aglomeración para el dataset en general (éste suele ser el núcleo de un problema de clustering que amerite un contraste entre distintos tipos de modelos).

Nótese que la simplicidad de este problema de agrupamiento de palabras clave parte de la posibilidad de asumir que existen múltiples títulos y empleos previos que pueden agrupar a los candidatos en el dataset, de esta forma, se tiene que se puede llegar a una convergencia con respecto a las métricas de evaluación de clustering al incrementar sucesivamente la cantidad de clusters como hiperparámetro de K-means, hasta llegar a un límite de eficiencia.

La siguiente tabla resume el proceso de ajuste sucesivo ejecutado, al presentar la cantidad de clusters elegidos y sus correspondientes métricas de evaluación, hasta obtener un nivel óptimo.

Tabla 2. Output de resultados para ejercicio de clustering sobre variables de texto (K-means).

Columna EduCluster (Agrupamiento de TituloObtenido)			
Valor de K	**Silhouette Score**	**Davies-Bouldin index**	**Calinski-Harabasz index**
429	0,7753	1,005	5551,4996
1978	0,8681	0,7583	5547,7639
2480	0,8773	0,6757	6120,1999
Columna ExperienciaCluster (Agrupamiento de UltimaExperienciaLaboral)			
Valor de K	**Silhouette Score**	**Davies-Bouldin index**	**Calinski-Harabasz index**
429	0,9125	0,4747	48566,2581
1978	0,9846	0,0711	88217527,06
2440	0,9917	0,0004	1,28509E+20

Nótese cómo, al incrementar la cantidad del clusters como hiperparámetro de K-means, la eficiencia en términos de las métricas de evaluación mejora hasta acercarse a un límite. El coeficiente de Silhouette determina la eficiencia del grado de agrupamiento dada la cantidad de clusters seleccionada en un escala entre 0 y 1, donde la cercanía a 1 denota la máxima eficiencia posible [4], como se presenta en los resultados para ambas columnas. El índice de Davies-Bouldin determina la eficiencia del agrupamiento al evaluar la compacidad y distancia entre los clusters formados, en una escala para los reales positivos [4], donde la cercanía a cero denota la máxima eficiencia, como se aprecia de forma descendente en los resultados obtenidos. Finalmente, el índice de Calinski-Harabasz realiza una evaluación bajo criterios similares a los de Davies-Bouldin, pero bajo una escala en los reales positivos donde la eficiencia del clustering incrementa en tanto el índice se acerca más a infinito [4], esto también se aprecia de manera sucesiva en los resultados obtenidos.

Así, se tiene que el proceso realizado determinó dos columnas categóricas suplementarias, EduCluster y ExperienciaCluster, con 2481 y 2441 instancias (el valor de cluster 0 se asigna a las observaciones con ruido), respectivamente, las cuales proveen estructuras operables de muestreo y correlación propicias para fijar la información sobre las variables de título obtenido y experiencia laboral previa del candidato como insumo para el entrenamiento del clasificador requerido.

Otro refuerzo al argumento de robustez del método empleado es que muy pocas observaciones del dataset son determinadas como ruido. De 87.735 casos de aplicaciones consideradas, se tiene que sólo 7 para EduCluster, y una para ExperienciaCluster, se etiquetan como ruido tras aplicar K-means en su forma óptima.

Es importante resaltar que el ejercicio de agrupamiento de entradas de texto realizado no sólo se posiciona como insumo para el entrenamiento del modelo central de esta investigación, sino que también implica una herramienta para facilitar el análisis de grupos ocupacionales de manera más rápida debido a la normalización de texto y la identificación de frecuencias de cargos y títulos característicos en el mercado laboral del Departamento. Representando una idea inicial para las recomendaciones para la organización en cuanto al manejo de sus estructuras de datos de cara a propósitos analíticos.

El ejercicio de clustering realizado completa la fase exploratoria del proyecto, y da paso al ejercicio principal de modelación.

3.2 Aprendizaje supervisado: Modelos de clasificación para la caracterización del mercado laboral.

Una vez seleccionados los atributos pertinentes para el contraste de clasificación, así como la generación de clusters consistentes para las variables críticas de texto relacionadas a títulos educativos y experiencia laboral previa de los candidatos, se procede a la construcción del clasificador sobre el cual se basa el análisis de esta investigación. Este ejercicio se realizó por medio de Python, a partir de la construcción de pipelines estructuradas en los siguientes pasos para cada modelo considerado:

- Carga de datos, selección de atributos por tipo de datos, definición de etiqueta binaria de clase.
- Splitting 70-30 para subsets de entrenamiento y prueba con base en la etiqueta binaria de clase. Notando que ejercicios de splitting 80-20 y 90-10 fueron realizados, pero se obtuvieron resultados similares o con peor desempeño que en el caso 70-30.
- Pre procesamiento de datos numéricos y categóricos.
- Balance de observaciones en función de etiqueta de clase para el subset de entrenamiento (según el modelo, se empleó SMOTE o balanced class weights).
- Entrenamiento de modelo.
- Despliegue de resultados, métricas de evaluación y relación de orden de importancia de los atributos (para los modelos que pueden calcular esta propiedad).
- Salida de archivo con candidatos organizados en orden descendente por probabilidad estimada de colocación.

Considerando el contraste de paradigmas que propone este proyecto, se toman como candidatos los modelos explorados en el marco teórico, de la siguiente forma:

- Paradigma frecuentista: Logistic Regression, Random Forest, Gradient Boosting.

- Paradigma bayesiano: Naive Bayes, Bayesian Ridge Regression.

Nótese que la selección de estos modelos parte de su conveniencia en términos computacionales e interpretabilidad con respecto al problema de negocio en cuestión. Además, debe mencionarse que se consideró aplicar modelos multiclase debido a la existencia de diversos sectores productivos para el mercado laboral, sin embargo, estos modelos otorgaron resultados muy pobres debido al sesgo inherente de la clase negativa en el dataset (una baja probabilidad idiosincrática de conseguir empleo), y a la definición laxa de categorías institucionales para las instancias de cada sector.

Para una exploración de los pormenores de los ejercicios de modelación realizados, referirse al documento extensivo de este proyecto aplicado, el cual se anexa a la entrega de este informe.

4. EVALUACIÓN DE EFICIENCIA Y SELECCIÓN DE MEJORES MODELOS

El conjunto de ejercicios de modelación realizados se caracterizó por la predominancia del paradigma frecuentista. Particularmente, los resultados obtenidos favorecen de forma indiscutible a los métodos basados en árboles, los cuales se hallaron muy por encima de aquellos de los demás modelos candidatos, tanto a nivel de sus valores en las métricas de evaluación, como en un cálculo de orden de importancia de atributos correspondiente con las hipótesis previas formuladas en el análisis exploratorio, y con las perspectivas informales de los miembros de la organización.

En términos del contraste de paradigmas propuesto, se puede afirmar con confianza que la superioridad del paradigma frecuentista para este problema de clasificación yace en su capacidad predictiva, así como en la posibilidad de calcular y monitorear la relevancia de cada atributo del dataset empleado.

Ahora, si bien las persistentes deficiencias de las propuestas bayesianas facilitan la selección del paradigma frecuentista de cara al contexto del problema de clasificación resuelto, resulta necesario ampliar la justificación de por qué los métodos basados en árboles se presentan como modelos idóneos, relacionando las soluciones obtenidas con el problema de negocio que suscita esta investigación.

4.1 La eficiencia técnica de los métodos basados en árboles.

La justificación de la idoneidad de los modelos basados en árboles es bastante directa, en tanto sus resultados en términos de métricas de evaluación son superiores con gran diferencia a los de los modelos del paradigma frecuentista y de regresión logística.

Este grado de predominancia puede observarse en las métricas de evaluación estándar bajo el umbral de 50% y a nivel general desde la curva ROC y el valor AUC, teniéndose que para ambos niveles de evaluación los modelos Random Forest y Gradient Boosting son bastante similares. La Tabla 3 resume el nivel de ajuste de las predicciones de los modelos contrastados a partir de sus correspondientes valores AUC.

Tabla 3. Comparación de desempeño de los modelos contrastados (valores AUC).

Modelo de clasificación	Desempeño (Valor AUC)
Regresión Logística	0,75
Naive Bayes	0,74
Bayesian Ridge Regression	0,68
Random Forest	**0,83**
Gradient Boosting	0,82

Si bien la selección de los modelos por medio de árboles es indiscutible para este ejercicio, debe recalcarse que para el contexto del dataset empleado y los resultados a presentar en el próximo capítulo, se decidió optar por Random Forest, en tanto sus predicciones poseen un mayor nivel de ajuste por una mínima diferencia con respecto a Gradient Boosting.

Nótese que, dada la estructura de los datos de candidatos al mercado laboral en manos de la Agencia, el ejercicio de clasificación se halla estrechamente relacionado con la capacidad de agrupar la información en texto de títulos obtenidos y experiencias laborales previas en clusters bien definidos, por lo cual se presume que la predominancia de Random Forest se da en correspondencia con la complejidad que emerge de estos dos atributos con múltiples instancias categóricas (EduCluster y ExperienciaCluster), generando predicciones ligeramente superiores a las de Gradient Boosting.

Ahora, de cara a la implementación práctica de un clasificador por parte de la organización dentro de la cadena de valor de la Ruta de Atención, visto como una sofisticación de la modelación realizada en este trabajo, surgen un conjunto de ventajas que resaltan la eficiencia técnica de los métodos por árboles dentro de este contexto práctico.

Una de las fortalezas de los métodos por árboles con respecto a las demás propuestas es su interpretabilidad. Se tiene que la idea central de los clasificadores basados en arboles de decisión puede ser menos hostil para el público no especializado a comparación, por ejemplo, del Teorema de Bayes. Este hecho puede facilitar la difusión superficial información sobre el funcionamiento e implementación del clasificador para los supervisores competentes en la cadena de valor, e incluso para el personal propenso a emplearlo.

La eficiencia en términos computacionales asociados a tiempos de ejecución y recursos disponibles favorece plenamente a Random Forest, en tanto su estimación paralela de árboles de decisión permite una mayor fluidez con respecto a Gradient Boosting, en el cual la corrección progresiva de los errores de predicción de clasificadores débiles implica mayores tiempos de ejecución y consumo de recursos.

Finalmente, la precisión de los métodos por árboles va de la mano con la posibilidad de monitorear la caracterización obtenida del mercado laboral captado, tanto en términos de relevancia de atributos, como de candidatos y profesiones con mayores (o menores) probabilidades estimadas de colocación.

4.2 La eficiencia económica y organizacional de los métodos por árboles en el contexto de la Ruta de Atención de la Agencia de Empleo.

Considerando la estructuración hipotética de un clasificador como herramienta de remisión y preselección al interior de la cadena de valor de la Ruta de Atención de la Agencia de Empleo de Comfacauca, se pueden vislumbrar un conjunto de ventajas críticas para incrementar la eficiencia económica y organizacional de este proceso, las cuales emergen de las propiedades de los métodos por árboles como modelos predilectos en esta investigación.

Como se mencionó anteriormente, Random Forest destaca tanto por su grado de precisión predictiva, como por su eficiencia en relación a Gradient Boosting en términos de costos computacionales de tiempo y recursos. En términos económicos, lo anterior implica una profunda reducción en los tiempos de ejecución del subproceso de remisión de candidatos, en tanto se obtiene un criterio automático y transparente para el empalme entre las necesidades de la demanda y la disposición de la oferta desde la probabilidad estimada de colocación en el mercado laboral. Esto implica una reducción significativa en el costo financiero de largo plazo de la ejecución del proceso completo, además de sugerir una estructuración alternativa del personal, que puede alinearse con funciones adicionales para el cumplimiento de la Misión de la Agencia. Lo anterior puede verse relacionado con los incrementos de eficiencia provistos por la inserción de herramientas analíticas al interior de cadenas de valor y dinámicas organizacionales, como se expone en [5].

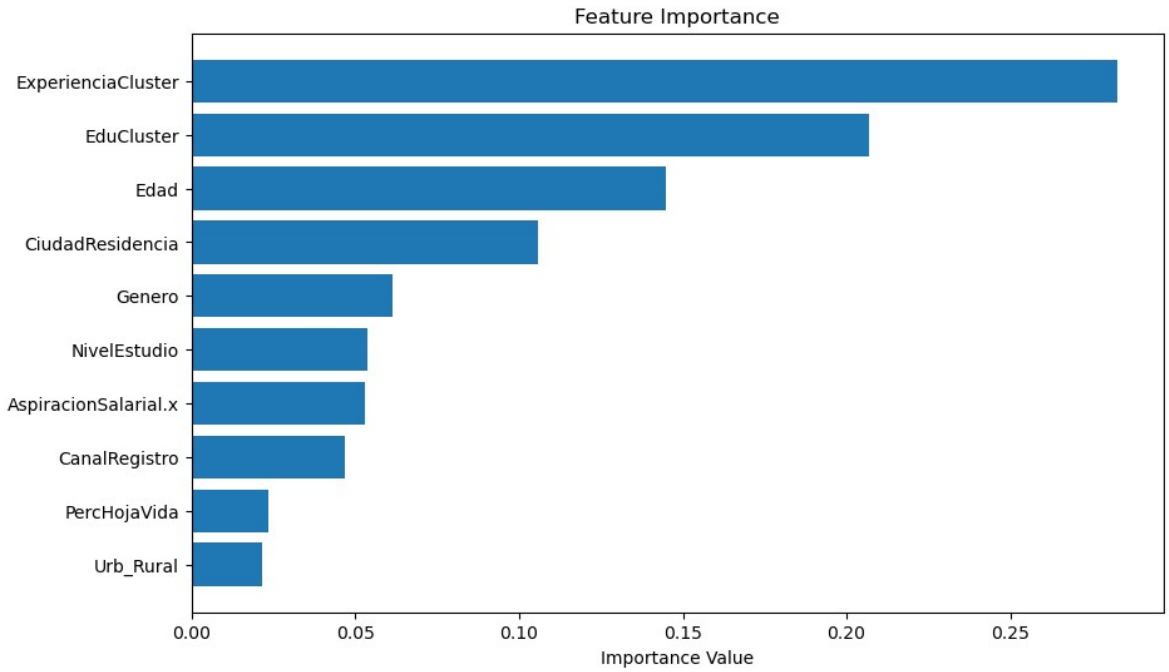
Nótese que el grado de transparencia que otorgan las predicciones por medio del clasificador sugerido como una sofisticación de este trabajo, implican una profunda mejora en la eficiencia organizacional del proceso analizado, asegurando la pre selección y remisión de candidatos bajo criterios estadísticos rigurosos, lo cual también hace del clasificador propuesto una herramienta para el control de calidad de la Ruta de Atención.

La simplicidad y eficiencia computacional relativas de Random Forest también implican la facilidad en su potencial implementación como una pipeline alimentada por una estructura ETL o dispositivos similares, la cual pueda proveer a los colaboradores de la agencia una herramienta actualizada y fluida para el proceso de remisión, así como para la evaluación posterior del desempeño de la Ruta de Atención, el monitoreo del comportamiento del

mercado laboral captado por la organización, y la posterior toma de decisiones basadas en el mismo.

Frente a lo anterior, es crítico resaltar el papel de la capacidad de los métodos por árboles para evaluar con precisión la relevancia de los atributos en la predicción de colocaciones, así como de ordenar a los candidatos en función de probabilidades estimadas de colocación y filtros específicos asociados a atributos claves del contexto del problema de negocio. La Figura 11 presenta el orden de importancia de los atributos considerados para el modelo óptimo construido bajo Random Forest.

Figura 11. Orden de importancia de atributos para el modelo entrenado por medio de Random Forest.



5. RESULTADOS Y DISCUSIÓN

Dado el problema de investigación formulado y la solución propuesta a través de los objetivos específicos de este proyecto aplicado, resta realizar la presentación de los resultados obtenidos tras el ejercicio de modelación desde el contraste entre los paradigmas frecuentista y bayesiano. En este sentido, este capítulo despliega los resultados obtenidos desde los criterios planteados, en dos fases. En la primera fase, se expone la caracterización del mercado laboral obtenida a partir de la implementación del mejor modelo (Random Forest) sobre el dataset empleado. En la segunda fase, se realiza una descripción de los impactos potenciales de la implementación estructurada de un modelo de clasificación para mejorar la eficiencia multidimensional, tanto de la cadena de valor correspondiente a la Ruta de Atención de la Agencia de Empleo, como para la toma de decisiones de la Comfacauca en general.

5.1 Caracterización del mercado laboral captado por la Agencia de Empleo.

El conjunto de resultados provistos por el modelo Random Forest, anteriormente justificado como la mejor opción de modelación obtenida para este proyecto, permite caracterizar el mercado laboral reflejado en el dataset empleado desde la identificación de los factores de mayor influencia en la probabilidad de obtención de empleo en el departamento del Cauca, permitiendo una descripción directa de las propiedades de la relación entre la oferta y la demanda laboral desde la información disponible. Así, esta sección consiste en la discusión de los hallazgos obtenidos considerando el papel de los atributos de mayor relevancia desde el clasificador construido, como lo expone la Figura 11. Nótese que, dada la suficiencia de las relaciones descritas en la fase de análisis exploratorio, se opta por realizar el análisis propuesto para los atributos del dataset desde los cuales se pueden exponer relaciones críticas, respaldados por sus niveles de importancia para el clasificador. Estas variables son el municipio de residencia, la edad, el género, la aspiración salarial, el título educativo obtenido, y el cargo de última experiencia laboral de los candidatos.

Municipios de residencia por probabilidad estimada de colocación

Los resultados obtenidos para la relación entre la probabilidad estimada de colocación en el mercado laboral caucano y el municipio de residencia de los candidatos se apegan a las hipótesis previas formuladas en la fase de análisis exploratorio, en tanto logra denotarse la profunda estructura asimétrica de la distribución del empleo en el departamento, como lo muestran los resultados descritos en la Tabla 4.

Tabla 4. Análisis de probabilidad estimada de colocación por municipios de residencia de los candidatos del mercado laboral, ordenados por coeficiente de fiabilidad (probabilidad media vs cantidad de observaciones).

CiudadResidencia	Avg_Probability	Std_Deviation	N_Observations	Q1	Q3	CI_95_low	CI_95_high	IQR	Reliability_Score
PUERTO TEJADA	0,466507217	0,336594255	7927	0,137238	0,804414	0,459097385	0,473917048	0,667176	4,188315756
SANTANDER DE QUILICHAO	0,436290137	0,304099204	14219	0,153463	0,728269	0,43129167	0,441288604	0,574806	4,171952175
PRADERA	0,851262405	0,266561441	95	0,9186	0,993789	0,797659097	0,904865714	0,075189	3,876544196
FLORIDA	0,640617793	0,349217725	308	0,292004	0,9915	0,601616681	0,679618906	0,699496	3,670803878
GUACHENÉ	0,44882045	0,330344806	2300	0,139128	0,782089	0,435319646	0,462321254	0,642962	3,474168479
VILLA RICA	0,424893572	0,326728486	2390	0,126629	0,74136	0,411794391	0,437992752	0,614731	3,305267764
PALMIRA	0,650741082	0,362607119	149	0,285251	0,994739	0,592517455	0,70896471	0,709488	3,256273436
PIAMONTE	0,491133029	0,316704875	661	0,200235	0,8023	0,466988981	0,515277077	0,602065	3,189296992
CALOTO	0,410507869	0,316119773	1768	0,126071	0,726463	0,39577231	0,425243428	0,600392	3,069615383
POPAYÁN	0,279157397	0,262536246	42562	0,068377	0,422343	0,276663179	0,281651615	0,353966	2,975459726
CANDELARIA	0,550276176	0,358707121	115	0,192287	0,935854	0,484714938	0,615837415	0,743567	2,611023108
MIRANDA	0,31327524	0,284008076	1190	0,077131	0,510908	0,297138593	0,329411886	0,433777	2,218523955
CALI	0,311853619	0,256342869	955	0,106145	0,458289	0,295595301	0,328111937	0,352144	2,139849514
PADILLA	0,327577276	0,293858105	486	0,083115	0,512514	0,301451115	0,353703438	0,429399	2,026461373
JAMUNDÍ	0,278757757	0,236532073	236	0,098645	0,374457	0,248579781	0,308935733	0,275811	1,523085499
PATÍA	0,212122949	0,226813649	1050	0,055391	0,272825	0,198403692	0,225842207	0,217434	1,475642937
SANTA ROSA	0,261220615	0,262651861	262	0,072367	0,354203	0,22941631	0,29302492	0,281836	1,454566375
CORINTO	0,194303006	0,218598677	423	0,043482	0,267057	0,173470888	0,215135124	0,223575	1,175022593
BOGOTÁ, D.C.	0,230466665	0,259076924	146	0,05804	0,297891	0,188441602	0,272491729	0,239851	1,148555199
TIMBÍO	0,157906033	0,199469988	1414	0,033363	0,186818	0,147509012	0,168303054	0,153455	1,145478446
SOTARA	0,237145654	0,30358849	102	0,039899	0,277962	0,178228564	0,296062744	0,238063	1,096792204
CALDONO	0,176691709	0,213383224	495	0,04203	0,215662	0,157893618	0,195489799	0,173632	1,096293912
INZÁ	0,210805493	0,242357344	174	0,046961	0,241734	0,17479429	0,246816696	0,194772	1,087557194
TOTORÓ	0,181992782	0,229787253	226	0,036158	0,230845	0,15203375	0,211951815	0,194687	0,986498246
PIENDAMÓ	0,131440261	0,177781624	1212	0,028144	0,157145	0,121431239	0,141449282	0,129	0,933229422
BUENOS AIRES	0,170316756	0,207819746	227	0,035041	0,213414	0,143281533	0,197351979	0,178373	0,923959888
GUAPÍ	0,135989859	0,211555091	626	0,019577	0,130894	0,119417193	0,152562526	0,111318	0,875686351
ROSAS	0,168087169	0,225438523	160	0,027417	0,192098	0,133155108	0,20301923	0,164681	0,853071598
LA VEGA	0,15762355	0,218996703	163	0,037341	0,158182	0,124003385	0,191243716	0,12084	0,802894992
CAJIBÍO	0,122642631	0,172148738	539	0,028676	0,130255	0,108109288	0,137175974	0,101579	0,771387267
PURACÉ	0,138287022	0,178550852	208	0,030321	0,159669	0,114021685	0,162552359	0,129348	0,738112246
SILVIA	0,112489214	0,175860134	525	0,020508	0,123126	0,09744589	0,127532539	0,102618	0,704564749
EL TAMBO	0,100000125	0,156200972	605	0,014785	0,106193	0,087553199	0,112447051	0,091408	0,640523645
ARGELIA	0,120727371	0,194638373	196	0,021488	0,111657	0,093477999	0,147976743	0,090169	0,637212907
LA SIERRA	0,119400957	0,192728889	156	0,019893	0,122799	0,089156863	0,149645052	0,102905	0,602957642
SAN SEBASTIÁN	0,122681855	0,237663212	136	0,013779	0,096212	0,082738128	0,162625581	0,082433	0,602693613
TORIBÍO	0,112035826	0,147729424	208	0,011014	0,151606	0,091959168	0,132112483	0,140591	0,597995485
BOLÍVAR	0,084030601	0,159856492	737	0,013966	0,07228	0,072489357	0,095571845	0,058314	0,554819428
BALBOA	0,104330436	0,165649215	201	0,019402	0,098019	0,081429806	0,127231066	0,078617	0,553296112
MORALES	0,09715851	0,148562844	237	0,013755	0,102047	0,07824413	0,11607289	0,088292	0,531268575
JAMBALÓ	0,111038616	0,175462926	106	0,016969	0,106549	0,077635383	0,14444185	0,08958	0,517821824
SUÁREZ	0,090229388	0,108401051	188	0,014897	0,129124	0,074733708	0,105725068	0,114227	0,472480952
MERCADERES	0,072192859	0,132252126	410	0,006889	0,076608	0,059391184	0,084994535	0,069719	0,434323588
ALMAGUER	0,082565174	0,146928805	155	0,014452	0,082335	0,059434027	0,105696321	0,067883	0,416411272
TIMBIQUÍ	0,074032359	0,176132016	267	0,007263	0,05706	0,052905315	0,095159404	0,049797	0,413637201
SUCRE	0,099545173	0,139677205	63	0,023969	0,135432	0,065053733	0,134036614	0,111463	0,412429064
PÁEZ	0,073528915	0,121769465	148	0,013944	0,086294	0,053910514	0,093147316	0,07235	0,367439596
FLORENCIA	0,063066242	0,091166337	63	0,009673	0,073085	0,04055392	0,085578565	0,063412	0,261291938

Esta tabla determina la probabilidad media de colocación de los candidatos (el promedio de las probabilidades estimadas desde el modelo de clasificación para los candidatos residentes en cada municipio), y presenta los valores asociados a la descripción de su comportamiento distribucional, ordenando a cada municipio según un score de fiabilidad de la forma $ln(n_i) * ProbMedia_i$ para cada i municipio con al menos 50 observaciones en el dataset, el cual penaliza el valor de probabilidad media calculada al considerar la cantidad de observaciones con las que este promedio se calcula. Para propósitos de orientación del

lector, a los valores de probabilidad media se añade una escala descendente de azul a amarillo, y a los valores de coeficiente de fiabilidad se añade una escala descendente de rojo a amarillo.

El resultado general de la tabla permite confirmar la hipótesis previa de baja generación de empleo en la mayoría de municipios del departamento del Cauca, de tal forma que una minoría de los municipios posee probabilidades medias de colocación de cerca del 40%. Esta minoría, donde puede considerarse que existen probabilidades medias consistentes con un mayor volumen de generación de empleo, se decanta específicamente por los municipios del norte del departamento del Cauca, y algunos municipios del sur del Valle del Cauca, los cual es congruente con la presencia de la conocida dinámica industrial en la región a raíz de la zona franca decretada por la Ley Páez. Nótese la excepción del municipio de Piamonte, donde se percibe una alta tasa de colocación debido a la generación de vacantes para el enclave petrolero existente en esa región.

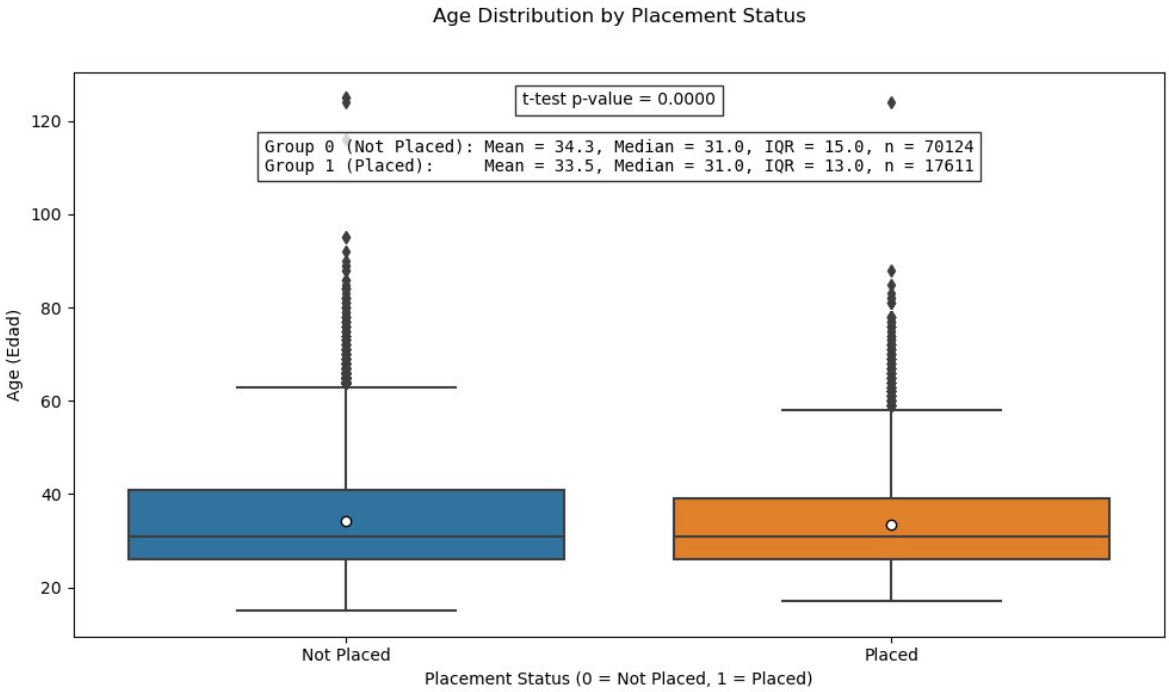
Además, la tabla calculada permite inferir la profunda situación de desempleo en Popayán como capital del departamento. Con 42.562 candidatos, Popayán es, por mucho, el municipio con la mayor cantidad de observaciones en el dataset, sin embargo, su probabilidad media de colocación se ubica sobre 27,9%. Incluso considerando su desviación estándar sobre 26,25%, se tiene que el comportamiento de la colocación en la capital se acentúa sobre una muy baja probabilidad en comparación con los municipios del Norte. Este resultado es congruente con la alta tasa idiosincrática de desempleo en la ciudad de Popayán, a raíz de una estructura sectorial inestable basada en el comercio y servicios, y afectada por la informalidad, como se estudia en [6].

En tanto la situación en la tabla empeora profundamente para los demás municipios del departamento (probabilidades estimadas de entre el 6% y el 21%), los resultados de las estimaciones del clasificador construido reflejan de manera específica cómo el grado de industrialización y generación de empleo formal se concentran sobre el norte del departamento, resaltando una muy baja dinámica económica en la región, y, considerando las muy altas probabilidades estimadas para algunos municipios del Valle del Cauca, poca competencia en la formación de la fuerza laboral apta.

Comparación distribucional de Edad para la muestra de candidatos

El atributo de edad de los candidatos en el dataset se postuló como la tercera variable en el orden de importancia del modelo óptimo de esta investigación, de tal forma que es posible inferir un comportamiento distribucional diferenciado para los candidatos con procesos exitosos de colocación versus los candidatos que permanecen por fuera de la población empleada. Este análisis se presenta en la Figura 12.

Figura 12. Comparación de distribuciones empíricas para las instancias de la etiqueta de clase del clasificador (Colocados vs No colocados).



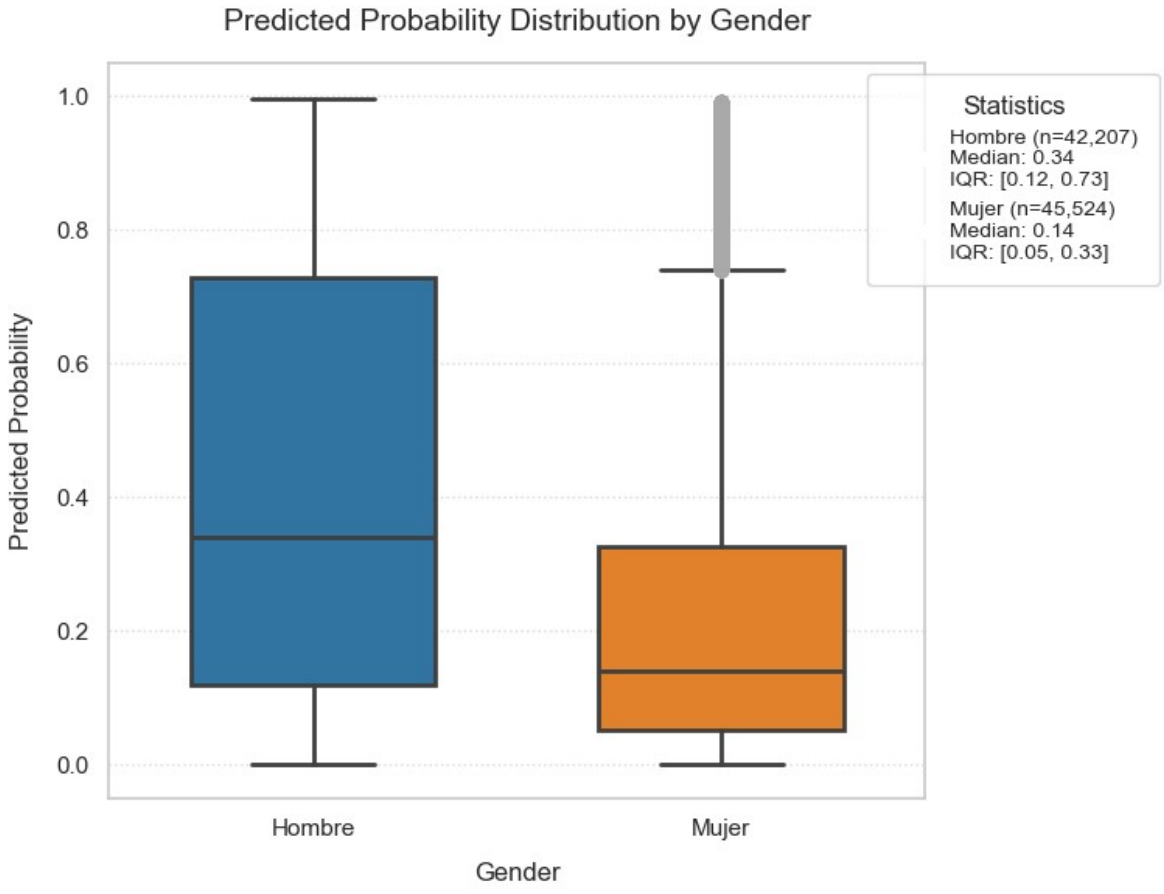
Con un nivel de significancia sobre el 0% en la prueba de hipótesis t-student, se confirma que las distribuciones empíricas de colocados y no colocados presentan un comportamiento estadístico diferenciado, de tal forma que la población con procesos exitosos de colocación es, en promedio, más joven; teniendo una edad media de 33,5 años versus los 34,3 años de los no colocados. Además, la distribución empírica para los colocados presenta menos variabilidad, en tanto su rango intercuartílico acapara 13 años (entre los 26 y los 39 años), lo cual dista al caso de los no colocados, con un rango intercuartílico de 15 años (entre los 26 y los 41 años).

Si bien este resultado permite afirmar que la generación de empleo en el departamento se concentra, en promedio, sobre población un poco más joven, existiendo significancia estadística para la diferencia entre ambos tipos de población; la similitud relativa entre las edades de colocados y no colocados permite reforzar la afirmación con respecto a la proliferación del desempleo en el departamento, en tanto puede observarse que la población en el rango de edad apto para el trabajo se entremezcla entre población empleada y desempleada.

Comparación distribucional de la probabilidad estimada de colocación por género

La hipótesis previa planteada en el análisis exploratorio para el comportamiento de la colocación con respecto al género denotó que, a pesar que existen más mujeres en la muestra, existe una mayor propensión a la colocación de hombres. Esta hipótesis se refuerza con creces en los resultados del clasificador, pues el comportamiento distribucional determinado presenta una mucho mayor probabilidad de colocación para los hombres, así lo muestra la Figura 13.

Figura 13. Comparación de distribuciones empíricas para las probabilidades estimadas de colocación (Hombres vs Mujeres).



Para los hombres, se observa una probabilidad mediana de colocación del 34%, existiendo un alto nivel de variabilidad fundamentado por un rango intercuartílico de entre el 12% y el 73%, de tal forma que la obtención de empleo es una posibilidad dentro del comportamiento estadístico regular. Sin embargo, para las mujeres, se tiene una probabilidad mediana de colocación del 14%, con un rango intercuartílico de entre el 5% y el 33%. En este sentido, se tiene que para las mujeres la obtención de empleo no se halla

dentro del comportamiento estadístico regular. Lo anterior se refuerza al observar en la Figura 13 cómo los valores de probabilidad estimada por encima del 75% son outliers.

Así, el modelo construido y sus predicciones reflejan la profunda asimetría de género existente en el mercado laboral del departamento, como también se expone en [6].

Aspiraciones salariales por probabilidad estimada de colocación

La hipótesis previa con respecto a la correlación entre la aspiración salarial de los candidatos y la colocación afirmó una tendencia a la concentración de ésta sobre aspiraciones salariales de entre 1 y 2 SMMLV. El modelo estimado respalda esta hipótesis, sin embargo, arroja conclusiones adicionales con respecto a la diversidad en aspiraciones salariales y su relación con la obtención de empleo. Así lo muestra la Tabla 5, la cual despliega los resultados correspondientes bajo la misma metodología de la Tabla 4.

Tabla 5. Análisis de probabilidad estimada de colocación por aspiraciones salariales de los candidatos del mercado laboral, ordenados por coeficiente de fiabilidad (probabilidad media vs cantidad de observaciones).

AspiracionSalarial	Avg_Probability	Std_Deviation	N_Observations	Q1	Q3	CI_95_low	CI_95_high	IQR	Reliability_Score
1 SMMLV	0,38446495	0,305507304	25286	0,113859	0,653912	0,38069932	0,388230579	0,540053	3,897708028
A convenir	0,365750609	0,312232759	14978	0,096176	0,638115	0,360750176	0,370751042	0,541939	3,516449886
1 a 2 SMMLV	0,310627253	0,278276325	34018	0,082812	0,490913	0,307670072	0,313584434	0,408101	3,241285134
2 a 4 SMMLV	0,237143179	0,253450232	7694	0,049022	0,32283	0,231479839	0,24280652	0,273809	2,122003666
4 a 6 SMMLV	0,233671024	0,265556426	1016	0,042842	0,308029	0,217341782	0,250000267	0,265187	1,617851393
Menos de 1 SMMLV	0,253782621	0,272735392	547	0,053881	0,364766	0,230926427	0,276638814	0,310885	1,599959538
6 a 9 SMMLV	0,149798811	0,154133671	173	0,038976	0,207152	0,126830422	0,1727672	0,168176	0,771956954
No registra	0,002956201	0,013830686	3941	0	0,000311	0,002524387	0,003388015	0,000311	0,02447495

Puede observarse que las definiciones de aspiraciones salariales con mayor fiabilidad, tanto en términos de tamaño de muestra como de probabilidad media estimada, corresponden a 1 SMMLV y la categoría “a convenir”, la cual denota flexibilidad del candidato con respecto a los salarios ofrecidos en las vacantes.

Así, se observa un comportamiento descrito por una marcada relación inversamente proporcional entre la aspiración salarial y la probabilidad de conseguir empleo, lo cual es esperado, en tanto una aspiración salarial más alta implica que el candidato posea niveles de formación y experiencia especializados y escasos, incrementando el salario esperado, pero disminuyendo la cantidad de puestos posibles.

Adicionalmente, nótese cómo los resultados obtenidos desde el enfoque de aspiración salarial también refuerzan la situación de alto desempleo en el departamento, pues las probabilidades medias estimadas de colocación alcanzan tan sólo un máximo del 38,4%, con un límite superior de rango intercuartílico del 65% para 1 SMMLV, así se tiene que la

probabilidad idiosincrática de conseguir empleo es muy baja incluso para las peores expectativas de salario.

Títulos educativos por probabilidad estimada de colocación

En este punto, se entra a discutir los resultados correspondientes a las dos variables con mayor importancia en la estimación de probabilidades de colocación. El comportamiento de los títulos educativos dentro del clasificador, obtenidos a través del agrupamiento de palabras clave en las entradas de texto, suscita la discusión de hechos interesantes que permiten añadir mayor especificidad a la situación general del mercado laboral descrita por las variables anteriores.

En primer lugar, se tiene que las palabras asociadas a los títulos con mayores probabilidades de colocación corresponden a la educación media, lo cual se alinea a los resultados de las aspiraciones salariales bajas para concluir que el grueso de la demanda de empleo en el departamento consiste en empleos requiriendo fuerza laboral con bajos niveles de formación educativa.

Los resultados obtenidos también denotan un comportamiento sectorial inclinado hacia la provisión de servicios varios, pues se hallan probabilidades de colocación altas para palabras asociadas a las disciplinas de contabilidad, enfermería, desarrollo de software, ciencias de la gestión, y, especialmente, formación de electricistas.

Nótese cómo los resultados provistos también exponen una mayor inclinación a la demanda de personal con títulos técnicos y tecnólogos por encima de candidatos con títulos obtenidos en educación superior, los cuales, en la mayoría de los casos, poseen probabilidades de colocación muy bajas. Las disciplinas con formaciones en educación universitaria más perjudicadas en términos de bajas probabilidades estimadas de colocación son derecho, psicología, administración de empresas, pedagogía, ingenierías, profesionales en el sector de construcción, y ciencias sociales. Lo anterior refuerza la conclusión de que la composición sectorial de la economía departamental es muy baja en términos de inserción de ciencia, tecnología y conocimientos especializados, lo cual es congruente con los bajos niveles de expectativas salariales hallados en el análisis anterior.

Finalmente, vale recalcar el hecho de que los candidatos que no registraron su título educativo en la plataforma de la Agencia poseen una probabilidad más alta de colocación que la de la mayoría de los títulos analizados en la tabla. Lo anterior no sólo se alinea con la expectativa de baja formación educativa de la fuerza laboral en la economía departamental, sino que resalta una mucho mayor influencia del cargo de experiencia laboral previa en la posibilidad de conseguir empleo, como se mostrará en la siguiente discusión. Los resultados

se consignan en la Tabla 6, la cual retoma la metodología de las tablas 4 y 5, considerando a los clusters que contienen más de 100 observaciones.

Tabla 6. Análisis de probabilidad estimada de colocación por títulos obtenidos (palabras agrupadas) de los candidatos del mercado laboral, ordenados por coeficiente de fiabilidad (probabilidad media vs cantidad de observaciones).

TituloObtenido	Avg_Probability	Std_Deviation	N_Observations	Q1	Q3	IQR	Reliability_Score
bachiller academico	0,439399844	0,28671577	19289	0,18439	0,694512	0,510122	4,3356858
media	0,552274042	0,303394079	807	0,277512	0,847798	0,570286	3,696548919
bachiller	0,415952612	0,311258064	3710	0,136501	0,703561	0,56706	3,418625982
enfermera	0,540995736	0,305908416	226	0,265926	0,85587	0,589944	2,932486322
bachiller tecnico	0,40835173	0,284466771	957	0,165582	0,636719	0,471138	2,802845991
auxiliar de enfermeria	0,394146117	0,267390467	1069	0,16022	0,605182	0,444962	2,748963778
tecnico en cocina	0,570839403	0,280344559	105	0,319464	0,839943	0,520479	2,656663948
tecnologo en analisis y desarrollo de sistemas de informacion	0,503732668	0,286137704	187	0,259402	0,80488	0,545479	2,635080301
bachiller tecnico ambiental	0,551810044	0,267276072	111	0,297531	0,781582	0,484051	2,598766069
tecnologo en contabilidad y finanzas	0,521513493	0,252789481	140	0,31521	0,707455	0,392245	2,577133202
tecnico electricista	0,533144304	0,245900584	124	0,315373	0,761083	0,44571	2,56990566
no registra	0,27319237	0,34495418	10620	0,00016	0,559137	0,558977	2,532628309
tecnologo en gestion del talento humano	0,521128257	0,258044441	116	0,316921	0,725572	0,408651	2,477230171
tecnologia en gestion empresarial	0,509180401	0,270981746	103	0,28138	0,729245	0,447865	2,359913166
contador publico	0,342047663	0,256419628	809	0,131738	0,5326	0,400861	2,290282372
bachiller comercial	0,416556839	0,315920278	209	0,159157	0,704114	0,544957	2,22538587
bachiller tecnico industrial	0,433432037	0,269603806	146	0,201012	0,661173	0,460161	2,16005477
contadora publica	0,359089442	0,258974004	353	0,144996	0,540509	0,395513	2,10658674
tecnologo en gestion empresarial	0,379287812	0,26809724	244	0,170415	0,546536	0,376121	2,085008909
ingeniero industrial	0,338827724	0,269218184	448	0,120358	0,508418	0,38806	2,068473194
administrador de empresas	0,307546986	0,236078154	625	0,119665	0,430666	0,311	1,979911115
bachiller tecnico agropecuario	0,364093203	0,296180396	189	0,121758	0,574363	0,452605	1,908484462
bachiller agropecuario	0,362664919	0,27762792	185	0,134907	0,587533	0,452625	1,893239924
bachillerato	0,322854597	0,297550684	291	0,09071	0,489352	0,398642	1,831658495
administracion de empresas	0,302503795	0,227785812	422	0,117836	0,420884	0,303048	1,828637048
basica primaria	0,243357591	0,248371768	1528	0,049336	0,352611	0,303274	1,784228494
tecnico en sistemas	0,237476757	0,240369051	1092	0,066326	0,31111	0,244784	1,661331858
tecnologo en entrenamiento deportivo	0,348271395	0,270712501	114	0,146504	0,53716	0,390656	1,649482442
administradora de empresas	0,257981099	0,212395121	455	0,090152	0,372748	0,282596	1,578921056
fisioterapeuta	0,266996727	0,250718724	334	0,075021	0,395465	0,320444	1,551555628
contaduria publica	0,254072276	0,220087519	387	0,080907	0,350029	0,269122	1,513870526
academico	0,283204091	0,282717143	208	0,05812	0,411552	0,353432	1,511612619
ingeniero de sistemas	0,232683537	0,247273556	369	0,054404	0,327155	0,272752	1,375345072
auxiliar en enfermeria	0,226980687	0,207694298	419	0,060225	0,336761	0,276536	1,370480089
trabajador social	0,26678092	0,241259103	156	0,095602	0,332426	0,236824	1,347205229
ingeniero civil	0,242245053	0,28720224	226	0,026428	0,362453	0,336025	1,313097787
bachiller academica	0,268111414	0,243194874	122	0,118003	0,310646	0,192643	1,288012876
basica secundaria	0,189977333	0,202449347	794	0,04066	0,256514	0,215854	1,268494511
tecnico auxiliar de enfermeria	0,23888666	0,245250375	197	0,055608	0,338845	0,283237	1,262086895
psicologo	0,231305727	0,216907622	192	0,091509	0,260753	0,169243	1,216088788
bachiller tecnico comercial	0,248137295	0,249131524	134	0,080325	0,288793	0,208468	1,215336721
ingenieria de sistemas	0,23738408	0,249782527	166	0,04202	0,34466	0,302639	1,213504519
bachiller basico	0,192518568	0,203752432	434	0,052296	0,25455	0,202254	1,169173838
tecnico en asistencia administrativa	0,207998105	0,225454825	269	0,064036	0,254344	0,190308	1,163689367
ingenieria industrial	0,240927504	0,227476146	120	0,079134	0,31993	0,240796	1,153438438
tecnico en contabilizacion de operaciones comerciales y financieras	0,212094432	0,239577796	206	0,050985	0,252599	0,201615	1,130012868
psicologa	0,168936191	0,199347818	755	0,036956	0,2139	0,176944	1,119492454
derecho	0,195993763	0,179582058	202	0,083481	0,240512	0,157031	1,04038736
ingeniera industrial	0,20061606	0,191678412	157	0,063982	0,259329	0,195347	1,01436411
abogado	0,159825563	0,213130525	481	0,023934	0,2084	0,184466	0,987061463
tecnico en auxiliar de enfermeria	0,186425573	0,20516276	187	0,042112	0,266843	0,224731	0,975212419
primaria	0,207249196	0,234139922	110	0,044241	0,258073	0,213831	0,974170775
bachiller en ciencias	0,185585762	0,204377428	129	0,054997	0,233985	0,178988	0,90191199
trabajadora social	0,145614775	0,174561411	444	0,038405	0,1818	0,143396	0,88764212
tecnico auxiliar en enfermeria	0,180991762	0,178464238	131	0,047272	0,233955	0,186683	0,882370554
normalista superior	0,148135992	0,195030696	313	0,044145	0,164997	0,120852	0,851219508
auxiliar administrativo en salud	0,17872339	0,199988417	110	0,045257	0,215056	0,169799	0,840085784
arquitecto	0,15717889	0,200458405	209	0,019805	0,232625	0,21282	0,839702169
especialista en pedagogia	0,17571418	0,206879815	113	0,04953	0,224713	0,175184	0,830669076
tecnico de sistemas	0,16796098	0,165891226	104	0,064589	0,194777	0,130188	0,780076449
trabajo social	0,157110315	0,176947884	113	0,032333	0,206604	0,174271	0,742721391
tecnico laboral en auxiliar en enfermeria	0,14929471	0,177962035	109	0,045309	0,166802	0,121493	0,700393421
licenciada en pedagogia infantil	0,135790119	0,187498781	168	0,03823	0,162009	0,123779	0,695783677
abogada	0,112901166	0,116256801	424	0,038824	0,149582	0,110758	0,683021962
psicologia	0,127337671	0,141935649	117	0,041624	0,160417	0,118793	0,606404139
tecnico en atencion integral a la primera infancia	0,102025919	0,161375994	258	0,021402	0,09835	0,076947	0,566545803
tecnico en atencion a la primera infancia	0,114645732	0,165518533	102	0,02233	0,132382	0,110053	0,530233395

Cargos de última experiencia laboral por probabilidad estimada de colocación

Como se explicó en el análisis de modelación por medio de Random Forest, esta variable se postula como la más importante en la predicción del comportamiento de colocación de los candidatos inmersos en el dataset empleado. De tal forma que los resultados provistos para este atributo son bastante próximos a una descripción de la estructura del mercado laboral del departamento.

Inicialmente, se tiene que los cargos con probabilidades estimadas de colocación se hallan asociados a las dinámicas de producción industrial en el norte del departamento, y al enclave de producción petrolera en el municipio de Piamonte, de tal forma que existe una predominancia en la colocación de cargos de corte operativo. Esto se alinea con las conclusiones previas asociadas a bajas expectativas salariales y el requerimiento de niveles de formación sobre la educación media, técnica y tecnológica.

La presencia de cargos operativos en el sector de construcción también se halla dentro de los rangos con probabilidades estimadas altas, los cual es congruente con la transversalidad de este sector en todos los municipios del departamento.

Seguido de esto, se hallan probabilidades altas de colocación para cargos previos asociados al sector agrícola, y, particularmente, al sector de servicios generales y comercio. En estos últimos resaltan los cargos de electricistas, soldadores, vigilantes, asesores comerciales y de ventas, supernumerarios, e inspectores de calidad.

El sector salud posee una estructura interesante, en tanto se hallan probabilidades de colocación altas para experiencias previas como médicos generales y ayudantes de ambulancia. Sin embargo, se registran probabilidades de colocación bajas para auxiliares de enfermería, fisioterapeutas y fonoaudiólogos.

En general, los cargos asociados a formación en educación superior presentan probabilidades de colocación bajas, con las excepciones de docentes de educación superior. Dado lo anterior, se halla mayor perjuicio para los cargos asociados al derecho, psicología, administración de empresas, pedagogía, ingenierías, profesionales en el sector de construcción, y ciencias sociales.

Esta última conclusión es completamente congruente con el análisis de títulos educativos, por lo cual se tiene que la conclusión general del mercado laboral del departamento como escaso en inserción de ciencia, tecnología y conocimiento se repite en este rubro, añadiendo una mayor especificidad en términos de los cargos más demandados. Los resultados analizados se consignan en la Tabla 7.

Tabla 7. Análisis de probabilidad estimada de colocación por cargos de última experiencia laboral (palabras agrupadas), ordenado por coeficiente de fiabilidad probabilidad.

UltimaExperienciaLaboral	Avg_Probability	Std_Deviation	N_Observations	Q1	Q3	IQR	Reliability_Score
auxiliar de produccion	0,744835726	0,247696818	2551	0,586177	0,9441	0,357924	5,842670727
ayudante de ambulancia	0,778968485	0,231789965	1604	0,669733	0,962352	0,292619	5,748986671
obrero ayudante produccion en pozos de petroleo y gas	0,687164025	0,236585687	641	0,500956	0,887838	0,386881	4,441161338
operario agricola de mantenimiento del cultivo	0,645954027	0,289858688	492	0,384496	0,936725	0,552229	4,003932286
operario de empaque	0,528473349	0,28113259	1635	0,27107	0,805535	0,534464	3,910384687
operador de maquina fabricacion material plastico	0,697527305	0,269999199	239	0,459645	0,928349	0,468704	3,81998286
ayudante de obra	0,658565953	0,246789566	265	0,459934	0,860157	0,400222	3,674620089
supernumerario	0,773947665	0,256720426	108	0,508216	0,983763	0,475547	3,62372453
auxiliar de centro de distribucion	0,584796813	0,28257033	375	0,328215	0,85022	0,522005	3,466047452
auxiliar bodega	0,637511314	0,240063795	197	0,436941	0,842077	0,405136	3,368102151
ayudante de construccion	0,54548698	0,260627946	471	0,312649	0,782535	0,469887	3,357394956
soldador	0,623358805	0,252182925	185	0,414242	0,861077	0,446835	3,25415477
operador de planta auxiliar de refrigeracion	0,552391068	0,265851001	356	0,315708	0,798084	0,482376	3,24525926
vigilante	0,449533499	0,238347358	1111	0,247757	0,642943	0,395185	3,152585523
auxiliar de descargue	0,640709095	0,246752278	126	0,426569	0,852898	0,426329	3,098649804
asesor comercial	0,426660543	0,258029781	1413	0,208785	0,632353	0,423568	3,094769614
coordinador de empacadores agricola	0,523012867	0,288807771	348	0,268944	0,804961	0,536017	3,060777195
parcelero agricola	0,594507564	0,285386224	146	0,314637	0,873455	0,558818	2,962791835
electricista	0,556975168	0,232511982	196	0,357319	0,792085	0,434766	2,939778801
operador de montacargas	0,608119204	0,251129059	123	0,390922	0,866656	0,475734	2,926381718
oficial de construccion	0,531708715	0,23690284	175	0,331698	0,736351	0,404653	2,746161712
docente de universidad	0,562630108	0,314217347	121	0,269874	0,875003	0,605129	2,698256151
no registra	0,265232897	0,29231327	23235	0,022652	0,4742	0,451548	2,6664964
docente de educacion superior	0,498465911	0,272386203	197	0,261103	0,764517	0,503413	2,63349696
auxiliar de enfermeria	0,368685841	0,260812252	1259	0,151079	0,574855	0,423776	2,631706459
medico general	0,531169509	0,254822507	133	0,297979	0,802485	0,504506	2,597604347
inspector de calidad	0,543862153	0,250431093	114	0,321929	0,760391	0,438462	2,575839087
conductor de vehiculo liviano	0,42967027	0,247944022	326	0,223035	0,614161	0,391126	2,486457763
auxiliar de bodega	0,43418625	0,250152543	279	0,222706	0,618361	0,395655	2,444994726
asesor comercial de ventas	0,373595085	0,240643072	692	0,175509	0,533385	0,357876	2,44315717
operario de produccion	0,433189281	0,256783279	248	0,214953	0,645937	0,430985	2,388358233
auxiliar contable	0,340533783	0,251160749	1063	0,130799	0,532376	0,401576	2,373128983
guardia de seguridad	0,359189203	0,215686561	575	0,186266	0,492318	0,306052	2,282421109
supervisor servicio cajeros automaticos	0,398689896	0,23141866	299	0,203556	0,569284	0,365728	2,272709258
tecnico electricista	0,465873945	0,262824976	127	0,232452	0,703734	0,471282	2,256780547
auxiliar logistico	0,443709199	0,273289524	158	0,220078	0,636854	0,416776	2,246319989
auxiliar aseo	0,382019693	0,259729186	261	0,175706	0,626186	0,45048	2,125756378
auxiliar de almacen y bodega	0,358221581	0,245367685	352	0,151389	0,52024	0,368852	2,100479233
conductor	0,308169146	0,232681422	833	0,113072	0,485596	0,372525	2,072447877
recepcionista	0,432286907	0,271705972	112	0,217062	0,626265	0,409203	2,039745281
instructor de deportes	0,433582545	0,289074701	110	0,208193	0,743212	0,535019	2,038046239
auxiliar administrativo	0,257381311	0,250771929	2584	0,0805	0,337258	0,256758	2,022269119
tecnologo en salud ocupacional	0,420123337	0,213763786	120	0,255434	0,594217	0,338782	2,011337005
enfermera profesional	0,427649183	0,257665414	106	0,209412	0,61082	0,401408	1,994315921
auxiliar de cocina	0,304511346	0,261176904	619	0,097602	0,480766	0,383165	1,957430986
operario de servicios generales	0,400287188	0,255986501	123	0,172547	0,582789	0,410242	1,926255744
auxiliar de farmacia	0,323125801	0,261102935	379	0,115731	0,452543	0,336812	1,918571142
auxiliar de servicios generales	0,275495122	0,215034735	1008	0,081629	0,42246	0,34083	1,905248073
ingeniero de sistemas	0,353872686	0,25853117	213	0,123823	0,513426	0,389602	1,897214862
auxiliar de archivo y registro	0,375194528	0,274198024	138	0,171341	0,477026	0,305685	1,848678619
auxiliar de ventas en caja	0,398364307	0,251522172	102	0,198866	0,59313	0,394264	1,84242409
supervisor	0,383492098	0,275672639	122	0,172848	0,584213	0,411365	1,842304108
fisioterapeuta	0,322407169	0,251358846	277	0,132416	0,48234	0,349923	1,813223565
agente de ventas	0,350730144	0,229310573	167	0,183053	0,437935	0,254882	1,795034707
auxiliar de almacen	0,331464012	0,234777934	187	0,146787	0,475228	0,328441	1,733924249
asesor	0,314597085	0,226102118	226	0,133629	0,423122	0,289494	1,705284511
mercaderista	0,345134046	0,244305066	137	0,164299	0,483783	0,319483	1,698052923
operario de maquinas	0,349239853	0,257547665	121	0,153162	0,498632	0,34547	1,674881183
psicologo organizacional	0,360758916	0,270843299	103	0,156656	0,457027	0,30037	1,672019805
auxiliar operativo	0,320490668	0,232951865	168	0,139993	0,410928	0,270934	1,642182639
vendedor de suministros industriales	0,280568666	0,247999513	344	0,085282	0,404318	0,319036	1,63870104
auxiliar de caja	0,340913737	0,243908557	121	0,15049	0,487511	0,337021	1,634950878
cajero	0,263017477	0,207025306	461	0,096331	0,375555	0,279224	1,613190876
asesor servicio al cliente	0,343772361	0,224664394	106	0,191321	0,428398	0,237077	1,603161466
mesero	0,264350573	0,223287098	325	0,091323	0,396776	0,305453	1,528957503
vendedor	0,240733176	0,193168943	508	0,094993	0,319975	0,224982	1,499883587

UltimaExperienciaLaboral	Avg_Probability	Std_Deviation	N_Observations	Q1	Q3	IQR	Reliability_Score
cajero de supermercado	0,305071429	0,219938845	127	0,129385	0,405838	0,276453	1,477823077
almacenista	0,301819955	0,243407683	112	0,105721	0,397701	0,29198	1,424137115
asistente administrativo	0,22067229	0,210347439	635	0,06738	0,284778	0,217398	1,424136204
auxiliar de atencion al cliente	0,196166196	0,1970772	1284	0,058158	0,252447	0,194289	1,404105743
docente de educacion basica secundaria	0,219038952	0,233283728	529	0,061664	0,289491	0,227827	1,373590736
administrador de empresas	0,22335554	0,21174774	367	0,060917	0,322123	0,261206	1,318995283
auxiliar servicios generales	0,190395137	0,205484576	929	0,041841	0,2743	0,232459	1,301181071
secretaria	0,210054419	0,232547638	467	0,050183	0,257655	0,207472	1,291063621
ingeniero industrial	0,264392805	0,214459645	120	0,108977	0,325997	0,21702	1,26577837
auxiliar de archivo	0,221011434	0,215739751	283	0,071731	0,267453	0,195721	1,247708315
ingeniero civil	0,2399956	0,294866036	157	0,020869	0,423364	0,402494	1,213476747
auxiliar de apoyo administrativo	0,210622692	0,222686126	235	0,062262	0,242648	0,180387	1,149912596
conductor de autobus	0,226443824	0,193098345	153	0,065953	0,319992	0,254039	1,139111601
docente de basica primaria	0,188017584	0,231272738	423	0,03839	0,22585	0,187459	1,137012305
encuestador	0,243021609	0,227319493	103	0,096034	0,293284	0,197251	1,126339298
administrador punto de venta	0,220882777	0,199567932	145	0,081012	0,293922	0,21291	1,099274771
docente de educacion basica primaria	0,176576509	0,236153743	478	0,036066	0,180908	0,144842	1,089408326
auxiliar de caja y ventas	0,215881575	0,23063942	145	0,064286	0,259785	0,195499	1,074385119
docente de educacion media	0,211069225	0,230708894	145	0,059564	0,261704	0,20214	1,050435335
auxiliar de aseo y limpieza	0,18486576	0,200879258	283	0,043647	0,242075	0,198428	1,043649832
asesor comercial ventas no tecnicas	0,209494392	0,206668096	136	0,067935	0,256866	0,188931	1,029173649
coordinador administrativo	0,184455451	0,211745381	264	0,045004	0,230247	0,185243	1,028514208
comunicador social	0,198050724	0,209590898	174	0,047115	0,278569	0,231453	1,021754639
auxiliar enfermeria	0,159712766	0,168904873	559	0,042831	0,213521	0,17069	1,010366833
administrador de ventas y servicios	0,200838776	0,214099477	147	0,064772	0,250414	0,185642	1,002272375
auxiliar de servicio al cliente	0,192907394	0,220205622	180	0,064713	0,220611	0,155899	1,001759775
entrenador deportivo	0,203775623	0,22322185	115	0,044802	0,249833	0,205031	0,966901501
escolta	0,194034708	0,165655231	122	0,083696	0,227422	0,143726	0,932146823
psicologo	0,146281817	0,169986765	504	0,039452	0,184034	0,144581	0,910249765
auxiliar de servicios	0,175687085	0,231100558	164	0,024009	0,208996	0,184987	0,895980668
auxiliar de docente	0,148803478	0,199735414	304	0,039325	0,152351	0,113026	0,850713607
mensajero	0,17140875	0,157716897	113	0,068008	0,192789	0,12478	0,810315639
docente de educacion media academica	0,167935441	0,17779141	102	0,06228	0,17785	0,11557	0,77669685
soporte tecnico en sistemas	0,150848732	0,148434599	146	0,062315	0,18042	0,118105	0,75177074
tecnico de salud ocupacional	0,157986409	0,16355879	113	0,062784	0,188283	0,125499	0,746863023
administrador de negocios	0,157859203	0,169334029	113	0,053374	0,186161	0,132787	0,746261672
trabajador social	0,139654928	0,14658035	202	0,041107	0,18566	0,144553	0,741325744
abogado	0,121568276	0,146758669	333	0,031985	0,14175	0,109765	0,706085868
profesional de trabajo social	0,132098434	0,183486448	201	0,028357	0,151332	0,122976	0,700558274
arquitecto	0,13048595	0,177125304	159	0,021206	0,130422	0,109217	0,661420781
psicologo comunitario	0,134174322	0,162900914	125	0,030828	0,172453	0,141624	0,647835722
docente de jardin y preescolar	0,121368967	0,167998572	202	0,029126	0,12889	0,099763	0,644258965
docente de preescolar	0,118399379	0,167801927	226	0,03202	0,133591	0,101571	0,641787978
psicologo educativo	0,124942945	0,181536664	138	0,027286	0,122006	0,09472	0,615625588
vendedor de almacen	0,115404966	0,157244309	116	0,03267	0,129488	0,096817	0,548587916
auxiliar de facturacion	0,106150102	0,11220464	164	0,039916	0,128843	0,088927	0,541351339
ingeniero ambiental	0,105095784	0,135379004	105	0,028247	0,123407	0,095161	0,489111611
psicologo clinico	0,102682067	0,122461699	103	0,026867	0,125228	0,09836	0,475903551
ingeniero	0,094222664	0,114468471	116	0,027045	0,11694	0,089896	0,44789593
educador de primera infancia	0,073429033	0,12441171	402	0,017984	0,079267	0,061284	0,44031368
colaborador de profesores	0,085368377	0,097974432	170	0,025559	0,106219	0,08066	0,438434778
fonoaudiologo	0,082591158	0,123122135	103	0,020576	0,083368	0,062792	0,382787634

5.2 Los impactos potenciales de un protocolo de Ciencia de Datos para la Agencia de Empleo y las herramientas de planeación de Comfacauca.

Los resultados presentados anteriormente consistieron en una caracterización del mercado laboral reflejado en la información disponible desde la óptica de los principales atributos asociados a los candidatos considerados, de tal forma que fue posible dar forma a las

propiedades de la fuerza laboral del departamento y proveer perspectivas con respecto a la situación de la dinámica económica caucana.

Si bien el ejercicio realizado se considera como una caracterización estática desde un modelo de clasificación, en el sentido de que los resultados obtenidos sólo tienen validez para el periodo comprendido entre enero de 2018 y octubre de 2024, la lección principal de este ejercicio investigativo fue cómo la organización está en la plena capacidad de emplear métodos de Ciencia de Datos para monitorear desde el rigor estadístico las actividades y resultados de la Agencia de Empleo, y, en el sentido de una posible extrapolación, de la Comfacauca en general.

Así, este proyecto aplicado ha logrado proveer un vistazo tangible a la gama de herramientas que pueden estar a disposición de Comfacauca al aplicar un protocolo de Ciencia de Datos, tanto a nivel de la Agencia de Empleo como para sus dinámicas transversales de planeación.

Las ventajas potenciales a resaltar para las operaciones de la cadena de valor correspondiente a la Ruta de Atención de la Agencia de Empleo radican en la posibilidad de incrementar la eficiencia técnica, financiera y organizacional de manera conjunta, en tanto un protocolo de Ciencia de Datos desde un algoritmo de clasificación significa una aceleración de la fluidez de la cadena de valor, un incremento del alcance de las operaciones desde el monitoreo del mercado laboral, y un incremento progresivo de la eficiencia financiera de los recursos empleados en el mediano y largo plazo.

La siguiente figura presenta un diagrama de flujo para una propuesta de integración sofisticada de los métodos empleados en este trabajo con el fin de mejorar la eficiencia multidimensional del proceso de la Ruta de Atención, el cual es una modificación del diagrama de flujo que describe el estado actual de la estructura de funcionamiento de esta cadena de valor (Figura 1).

Figura 14. Diagrama de flujo de la cadena de valor de la Ruta de Atención bajo la integración del protocolo de Ciencia de Datos para la clasificación y caracterización del mercado laboral.

El diagrama de flujo expuesto implica la dinamización del proceso de filtro para la remisión de candidatos a las empresas e instituciones solicitantes, a la vez mejora este subproceso al insertar un criterio transparente fundamentado en el rigor estadístico y el grado de precisión de las predicciones del clasificador construido. Además, a diferencia de la naturaleza estática del modelo construido en esta investigación, esta propuesta involucra una dinámica de actualización basada en los resultados sucesivos de procesos de remisión y colocación según las vacantes, manteniendo la precisión del protocolo y otorgando una caracterización en tiempo real del estado del mercado laboral captado por la Agencia. Nótese que lo anterior implica una herramienta de auto evaluación del enfoque de los recursos de la Agencia, y de su misma estructura operacional.

En términos de impactos positivos potenciales sobre la eficiencia técnica, financiera y organizacional de la Agencia, se resaltan los siguientes puntos clave:

- Disminución potencial de los tiempos de ejecución de remisiones y del proceso general de la Ruta de Atención para cada candidato.
- Auto evaluación de los eslabones de la cadena de valor y la eficiencia de los colaboradores para cada sub proceso.
- Disminución progresiva del costo unitario de acompañamiento por candidato en el mediano y largo plazo.
- Capacidad de monitoreo interno del proceso de la Ruta de Atención y posibilidades de reestructuración del mismo de acuerdo a las necesidades de la organización.
- Tendencia a la estandarización de procesos, tecnologías y estructuras de datos.

- Capacidad de monitoreo externo de la dinámica económica del departamento del Cauca, y evaluación de la eficiencia financiera de los recursos empleados por la organización en mecanismos de subsidios e intervención.
- Conformación de una fuente fiable y constantemente actualizada de información sobre el mercado laboral. Apta para procesos de investigación y colaboración con instituciones a cargo de la política económica y sectorial del Departamento.

Ahora, si bien se ha logrado justificar claramente la necesidad de la implementación de un protocolo de Ciencia de Datos desde sus aspectos positivos, resulta necesario abordar el conjunto de esfuerzos adicionales que implica un proyecto de este tipo para la organización:

- Formulación e implementación del protocolo de Ciencia de Datos a cargo de un equipo competente, el cual pueda compaginar con el estado actual del proceso y realizar su inserción de forma apropiada.
- Implementación transversal de una estructura estandarizada y una política concreta de gobernanza de datos.
- Formulación, implementación y evaluación de modelos de mayor complejidad, capaces de asimilar las características específicas de la cadena de valor de la Ruta de Atención y el carácter dinámico del mercado laboral sobre el que opera.
- Diseño de mecanismos de monitoreo, verificación y evaluación de los impactos reales del proyecto de implementación del protocolo.

Dada la disponibilidad normativa de información de monitoreo por parte de Comfacauca, también es posible vislumbrar las fortalezas de la implementación de protocolos análogos como una extrapolación de este ejercicio, reflejando los beneficios del uso estandarizado de metodologías de Ciencia de Datos como un instrumento riguroso de planeación.

De acuerdo al alcance de la Comfacauca en el Departamento, los resultados presentados y su sofisticación bajo un protocolo puntual se postulan como insumos de planeación desde los siguientes aspectos clave:

- Identificación de incentivos para la formulación de políticas y programas de intervención para la movilidad de fuerza laboral disponible fuera de las zonas tradicionales de aglomeración identificadas en este ejercicio (Popayán y norte del Cauca).
- Identificación de oportunidades de inversión desde la localización de recursos y fuerza laboral disponible en zonas segmentadas de la dinámica económica departamental.
- Revisión de títulos y cargos de mayor y menor demanda por probabilidades estimadas de colocación, de cara al (re) planteamiento de programas de capacitación e intervención desde la organización.

- Identificación de la eficiencia marginal (económica y financiera) de los programas de intervención, subsidios y capacitación de la Caja bajo un enfoque de rigor estadístico. Planteando la posibilidad de su redistribución bajo un criterio de minimización del riesgo en su ejecución, tanto en materia financiera como socioeconómica.
- Identificación de fortalezas y debilidades de la estructura económica departamental, suscitando la reevaluación de la ejecución de programas y recursos desde una óptica investigativa y formal. Incluso colaborando con personal e instituciones especializadas en este campo alrededor de la región.

Finalmente, en términos de extrapolaciones generales, es importante resaltar que este ejercicio investigativo es una muestra de las amplias capacidades que la implementación de metodologías de modelación en Ciencia de Datos ofrece a las organizaciones, de tal manera que la réplica sofisticada y estandarizada de este tipo de ejercicios constituye una herramienta potencial para la eficiencia multidimensional de las operaciones de Comfacauca de forma transversal. Como se propuso en el inicio de este proyecto aplicado, el poder predictivo y prescriptivo de los protocolos de Ciencia de Datos resalta su carácter como una herramienta indispensable para la organización en el mediano y largo plazo.

6. ALCANCE Y LIMITACIONES PARA LA INSERCIÓN DE UN PROTOCOLO DE CIENCIA DE DATOS EN LA AGENCIA DE EMPLEO

En este capítulo se realiza una propuesta de delimitación del alcance para la ejecución de un proyecto de inserción de un equipo de Ciencia de Datos al interior de la cadena de valor de la Agencia de Empleo de Comfacauca, de acuerdo a la guía estándar provista por la Universidad Carnegie Mellon para el manejo gerencial y organizacional de proyectos de Ciencia de Datos [7].

6.1 Título del Proyecto:

Implementación de un Protocolo Analítico para la Optimización de la Ruta de Atención en la Agencia de Empleo de Comfacauca.

6.2 Nombre de la Organización:

Agencia de Empleo de la Caja de Compensación Familiar del Cauca – Comfacauca.

6.3 Descripción del Problema

La Agencia de Empleo de Comfacauca no cuenta con un protocolo analítico estandarizado para gestionar los datos relacionados con la atención y acompañamiento a la población desempleada del Cauca. Esto limita la transparencia, eficiencia y capacidad de caracterización dinámica del mercado laboral, afectando negativamente la toma de decisiones institucionales y las políticas públicas de empleo en el departamento.

6.3.1 ¿Cuál es el problema de negocio o de política pública que están enfrentando?

- Población desempleada del departamento del Cauca.
- Empresas del Cauca que requieren talento humano.
- Personal operativo y directivo de la Agencia de Empleo sin herramientas analíticas a su disposición.
- Entidades públicas como el SENA, MinTrabajo y DANE, interesadas en datos laborales confiables, pero no existe una fuente de datos generalizada y unificada de manera confiable.

6.3.2 ¿Quién o qué se ve afectado por este problema?

- La Agencia de Empleo atiende a miles de personas desempleadas al año en el departamento.
- Al no contar con datos normalizados y procesables, la eficiencia del emparejamiento oferta-demanda laboral se ve comprometida.
- Existe un impacto significativo en la toma de decisiones institucionales, ya que se pierde la posibilidad de analizar dinámicamente el mercado laboral y optimizar servicios.
- La trazabilidad y calidad de los datos afecta la coordinación con actores externos, dificultando el diseño de políticas públicas efectivas en la región.

6.3.3 ¿Cuántas personas/organizaciones/lugares/etc. se ven afectadas por el problema y en qué magnitud?

Durante el año 2023, la Agencia de Empleo de Comfacauca registró 12.903 hojas de vida, logró la colocación laboral de 4.824 personas y asignó 3.738 subsidios al desempleo. Estas cifras reflejan el volumen significativo de ciudadanos desempleados del departamento del Cauca que dependen de los servicios de intermediación laboral ofrecidos por la Agencia [1].

La ausencia de un protocolo analítico para procesar y utilizar estratégicamente esta información afecta directamente la calidad de los servicios prestados. Esto limita la capacidad de:

- Identificar tendencias del mercado laboral.
- Segmentar poblaciones con riesgo de desempleo prolongado.
- Optimizar el emparejamiento entre candidatos y vacantes.

Las empresas locales, que actúan como demandantes de talento, también se ven afectadas por la falta de información estructurada que permita mejorar la pertinencia y oportunidad de los perfiles remitidos

Más de 12.000 personas y decenas de empresas son impactadas anualmente por este problema, el cual obstaculiza el desarrollo de una política de empleo basada en datos confiables, actualizados y accionables.

6.3.4 ¿Por qué resolver este problema es una prioridad para su organización en este momento?

Resolver este problema es prioritario porque actualmente la Agencia de Empleo de Comfacauca carece de una estructura analítica formal, lo cual limita su capacidad para tomar decisiones informadas, medir el impacto de sus intervenciones y responder de forma ágil a las necesidades del mercado laboral.

Además, la organización ha manifestado una alta disposición institucional al cambio y ha reconocido la necesidad de fortalecer la gobernanza de datos, la infraestructura tecnológica y las capacidades analíticas, aspectos que son fundamentales para avanzar hacia una gestión moderna y basada en evidencia.

Esta transformación es clave para mejorar los servicios ofrecidos a las personas desempleadas del departamento del Cauca y para contribuir a políticas públicas de empleo más efectivas y focalizadas.

6.3.5 ¿Cómo han intentado abordar este problema y cuál ha sido el resultado de sus esfuerzos?

La Agencia ha manejado sus operaciones utilizando herramientas básicas como Excel, sin herramientas tecnológicas avanzadas ni procesos automatizados.

Se han utilizado visualizadores como Looker Studio y Power BI, pero sin una base estructurada de datos ni un equipo especializado en analítica.

Las decisiones se toman de forma manual y descentralizada, y no existe un protocolo estandarizado para remitir candidatos a vacantes.

El resultado ha sido una baja capacidad de trazabilidad, poca reutilización de los datos recolectados y dependencia del conocimiento tácito del personal operativo, lo cual dificulta tanto la eficiencia como la sostenibilidad del proceso.

6.3.6 ¿Qué otros grupos o partes interesadas, dentro y fuera de su organización, deben estar involucrados en la delimitación e implementación de este proyecto?

Dentro de la organización:

- Dirección de la Agencia de Empleo (liderazgo y toma de decisiones).
- Área de Tecnología (infraestructura y soporte).
- Personal operativo (registro de datos y ejecución de procesos).

- Futuro equipo de analítica (a conformar para análisis y visualización).

Fuera de la organización:

- SENA (información sobre oferta formativa).
- Ministerio del Trabajo (políticas y normativas).
- DANE (indicadores del mercado laboral).
- Empresas privadas que reportan vacantes y procesos de selección.
- Población desempleada, como usuarios finales del sistema.

6.4 Objetivos puntuales.

6.4.1 ¿Cuáles son sus objetivos sociales, de política pública o de negocio, y qué restricciones enfrentan?

Objetivos:

- Caracterizar dinámicamente el mercado laboral en el Cauca a partir de datos recolectados en la Ruta de Atención.
- Optimizar la remisión y colocación de candidatos a vacantes mediante herramientas analíticas.
- Establecer una cultura de toma de decisiones basada en datos dentro de la Agencia.
- Reducir la dependencia del conocimiento individual y asegurar continuidad operativa con procesos estandarizados.
- Proveer insumos analíticos útiles para el diseño de políticas públicas de empleo.

Restricciones:

- Bajo nivel de alistamiento organizacional, con limitaciones en gobernanza, infraestructura tecnológica y talento humano.
- Falta de un equipo de ciencia de datos especializado.
- Recursos limitados en términos de tiempo, presupuesto y personal.
- Ausencia de procesos normalizados de recolección y transformación de datos.

La relación de objetivos y sus restricciones se presenta en la Tabla 8.

Tabla 8. Objetivos vs restricciones para un proyecto de implementación de un protocolo de Ciencia de Datos en la Agencia de Empleo.

Objetivo	Tipo de Objetivo	Restricciones asociadas
Implementar un sistema analítico que optimice la remisión de candidatos a vacantes	Eficiencia	Falta de personal capacitado, dependencia de procesos manuales, carencia de infraestructura tecnológica
Caracterizar dinámicamente el mercado laboral del Cauca para apoyar la formulación de políticas	Eficiencia	Datos no normalizados, baja calidad de información, falta de interoperabilidad con otras entidades
Garantizar el acceso equitativo a los servicios de orientación y colocación para poblaciones vulnerables	Equidad	Limitaciones presupuestales, falta de indicadores segmentados por grupo, escasa automatización del proceso

6.4.2 ¿Qué compensaciones existen entre estos objetivos?

Automatizar procesos mejora la eficiencia, pero puede dejar por fuera a poblaciones vulnerables si no están bien representadas en los datos. A su vez, enfocar esfuerzos en la equidad puede requerir más tiempo y recursos, reduciendo la eficiencia operativa.

A nivel interno en la Agencia, se recomienda dar prioridad a la eficacia, fortaleciendo la capacidad de análisis del mercado laboral. Esto permitirá tomar mejores decisiones y luego avanzar hacia mejoras en eficiencia y equidad, con una base técnica más sólida. Sin embargo, las modificaciones estructurales a la cadena de valor de la Ruta de Atención que emerjan de estas recomendaciones pueden ser adversas para el personal no capacitado, en tanto pueden suscitarse dinámicas de reemplazo de personal por tecnología (e.g. algortimos de clasificación y selección vs psicólogos ocupacionales), y variación de la composición tecnológica en perjuicio del personal que ya opera en la Agencia.

En general, una implementación ideal de la cadena de valor de la Ruta de Atención puede incrementar exponencialmente su alcance, eficiencia, transparencia, y nivel de integración con agentes internos y externos, pero puede reducir significativamente la cantidad de colaboradores y la estructura de su organigrama.

6.5 Acciones

- Registro y acompañamiento a personas desempleadas mediante la Ruta de Atención.
- Remisión de perfiles a vacantes con base en criterios operativos (actualmente sin modelo de priorización).
- Asignación de subsidios al desempleo.
- Uso limitado de herramientas como Power BI para reportes básicos.
- Planeación de un equipo de analítica de datos que estructure procesos de caracterización laboral y genere recomendaciones automatizadas.
- Diseño de una base de datos estructurada y protocolos de calidad de datos para mejorar la trazabilidad y análisis.

La siguiente tabla desglosa las acciones fundamentales propuestas para la ejecución del proyecto.

***Tabla 9. Relación de acciones fundamentales propuestas para la ejecución de un proyecto de Ciencia de Datos en la Agencia de Empleo*.**

	Acción 1	**Acción 2**	**Acción 3**
¿Cuál es la acción?	Crear un equipo de analítica de datos con roles definidos	Diseñar e implementar una base de datos estructurada y normalizada	Automatizar tableros e informes para la toma de decisiones
¿Qué objetivo ayuda a alcanzar esta acción?	Mejorar la caracterización del mercado laboral y la toma de decisiones (eficacia)	Aumentar la eficiencia en la gestión de datos y prestación de servicios	Aumentar la transparencia y reducir la dependencia de procesos manuales
¿Quién ejecuta esta acción?	Dirección de la Agencia y área de Talento Humano	Área de Tecnología y consultores externos	Equipo de analítica de datos con apoyo de Tecnología
¿Sobre quién o qué se realiza la acción?	Estructura interna y organización del personal	Registros de datos de la Ruta de Atención	Procesos operativos e informes de desempeño
¿Con qué frecuencia se toma la decisión de realizar esta acción?	Una sola vez en la fase inicial, con evaluación anual	Una vez durante el diseño del sistema; se actualiza según necesidad	Actualizaciones semanales o mensuales

¿Qué canales se usan o pueden usarse para realizar esta acción?	Reuniones internas de planeación, resoluciones institucionales	Herramientas digitales, plataformas en la nube, procesos ETL	Herramientas de BI (Power BI, Looker Studio), tableros web
¿Existen restricciones de recursos o capacidad para esta acción?	Sí. Se requiere contratar o reasignar personal con perfil analítico, lo cual depende del presupuesto y procesos administrativos internos.	Sí. Es necesario contar con infraestructura tecnológica (servidores, nube, licencias) y tiempo del equipo técnico.	Sí. Requiere conocimientos específicos en herramientas de BI y tiempo de desarrollo continuo.
¿Qué problemas éticos pueden estar asociados a esta acción?	Riesgo de conformar un equipo sin criterios técnicos claros o con sesgos en la selección.	Riesgo de mal manejo de datos sensibles si no se aplican políticas claras de protección de datos.	Visualizar información sin segmentación podría invisibilizar a grupos vulnerables.
¿Puede brindar otra información útil sobre esta acción?	Esta acción es fundamental para garantizar sostenibilidad del proyecto. Requiere aprobación institucional.	La base de datos estructurada permitirá análisis más precisos y decisiones basadas en evidencia.	Se debe capacitar a los usuarios para que usen correctamente los dashboards e informes generados.

6.6 Datos

La Agencia de Empleo de Comfacauca utiliza como fuente principal los datos administrativos recolectados a lo largo de la Ruta de Atención al desempleado, incluyendo información del registro inicial, servicios prestados (orientación, formación, colocación), y resultados de inserción laboral.

Sin embargo, estos datos actualmente no están normalizados ni organizados para análisis estructurado. Para fortalecer sus capacidades, se plantea un proceso de estandarización para propósitos analíticos, así como la integración de otras fuentes secundarias, con el fin de buscar validez externa para una caracterización congruente del mercado laboral. Algunas de las fuentes externas son:

- Datos del SENA (oferta formativa disponible).
- Indicadores del DANE (tasas de desempleo, informalidad).
- Normativas y registros del Ministerio del Trabajo.

6.6.1 ¿Qué fuentes de datos tiene internamente?

Internamente, la Agencia de Empleo de Comfacauca cuenta con las siguientes fuentes de datos administrativas:

- Formularios de registro de los usuarios atendidos (datos personales, nivel educativo, experiencia laboral, etc.).
- Historial de atención dentro de la Ruta de Atención (actividades realizadas, fechas, tipo de servicio recibido).
- Remisiones a vacantes y resultado del proceso (si fue seleccionado o no).
- Asignación de subsidios al desempleo (tipo, duración, beneficiario).
- Registros de visitas y actividades de orientación, formación y acompañamiento.

La siguiente tabla desglosa las propiedades de las fuentes de datos internas consideradas.

Tabla 10. Relación de fuentes de datos internas para un proyecto de Ciencia de Datos en la Agencia de Empleo.

	Fuente de datos 1	**Fuente de datos 2**	**Fuente de datos 3**
¿Cuál es el nombre de la fuente de datos?	Registro de usuarios (Ruta de Atención)	Historial de remisiones a vacantes	Asignación de subsidios al desempleo
¿Qué contiene?	Datos personales, nivel educativo, experiencia laboral, datos de contacto	Información sobre vacantes, fechas de remisión, empresa, resultado del proceso	Nombre del beneficiario, tipo de subsidio, duración, condiciones de otorgamiento
¿Qué nivel de detalle tiene la fuente de datos?	Nivel individual (por cada persona registrada)	Nivel transaccional (por cada remisión a vacante)	Nivel individual (por cada beneficiario de subsidio)
¿Hasta qué fecha histórica se remonta la fuente de datos?	Actualizada	Actualizada	Actualizada

¿Con qué frecuencia se recopilan o actualizan los datos?	De forma continua, cada vez que se atiende a un usuario	Cada vez que se realiza una remisión a vacante	Mensualmente, según el proceso de asignación del subsidio
¿Los datos tienen identificadores confiables y únicos que puedan vincularse con otras fuentes?	Sí, número de documento de identidad del usuario	Sí, se registra el mismo número de documento del candidato	Sí, se usa el número de documento del beneficiario
¿Quién es el responsable interno de los datos?	Coordinación de la Ruta de Atención	Área de Gestión Empresarial / Intermediación Laboral	Área de Subsidios y Compensación
¿Cuáles son los problemas éticos asociados con el uso de esta fuente de datos	Es necesario garantizar que su uso respete los principios de transparencia y no discriminación.	Pueden surgir riesgos si se desagrega demasiado; deben respetarse condiciones de uso y confidencialidad	No requiere consentimiento individual porque es información pública y agregada.
¿Puede proporcionarnos alguna otra información útil sobre esta fuente de datos?	Transformaciones para variables de aproximación por métodos de modelación y encadenamiento.	Relaciones temporales y tiempos de espera promedio para la colocación en el mercado laboral.	Métricas d evaluación financiera para la distribución de subsidios al desempleo.

6.2. ¿Qué datos se pueden obtener de fuentes externas, privadas o públicas?

La siguiente tabla despliega las propiedades de las fuentes de datos externas para potencial empleo en el proyecto.

Tabla 11. Relación de fuentes de datos externas para un proyecto de Ciencia de Datos en la Agencia de Empleo.

	Fuente de datos 1	Fuente de datos 2	Fuente de datos 3
¿Cuál es el nombre de la fuente de datos?	**Observatorio Laboral del Ministerio del Trabajo**	**Sistema Nacional de Formación para el Trabajo (SENA**	**Gran Encuesta Integrada de Hogares (GEIH) – DANE**
¿Qué contiene?	Indicadores de empleo, informalidad, ocupación, vacantes, cobertura poblacional, políticas laborales	Oferta educativa por nivel, modalidad, edad, género y ubicación geográfica	Condiciones laborales, tasas de desempleo y subempleo, variables sociodemográficas
¿Qué nivel de detalle tiene la fuente de datos?	Nivel departamental y municipal	Nivel municipal e institucional	Nivel de hogar y persona (edad, género, zona)
¿Con qué frecuencia se recopilan o actualizan los datos?	Trimestral o semestral, según boletines oficiales	Según cronograma de convocatorias y actualizaciones internas	Trimestral para indicadores; anual para microdatos
¿Los datos tienen identificadores únicos para vincularse con otras fuentes?	No a nivel individual, pero sí por región, municipio o sector	Sí, en algunos casos: códigos de centro, municipio y programa	Sí, anonimizados; permiten vinculación geográfica y demográfica
¿Quién es el responsable interno de los datos?	Dirección de Empleo y Trabajo – MinTrabajo	Dirección del Sistema Nacional de Formación – SENA	Departamento Administrativo Nacional de Estadística (DANE)
¿Cómo se almacenan?	En PDF y Excel descargables desde el portal institucional	En portales institucionales, en formato Excel y bases abiertas	En portales institucionales, en formato Excel y bases abiertas
¿Cuáles son los problemas éticos asociados con el uso	Aunque son públicos, deben usarse con	Algunos reportes incluyen beneficiarios. Se	Aunque están anonimizados, deben respetarse

de esta fuente de datos?	responsabilidad para evitar decisiones basadas en inferencias erróneas o excluyentes.	deben anonimizar y manejar con criterios de uso ético.	los términos de uso y no cruzarse con otras bases sin control.
¿Puedes proporcionar otra información útil sobre esta fuente de datos?	Es útil para analizar tendencias laborales y evaluar la alineación de la Agencia con políticas nacionales.	Permite conectar la oferta educativa con la demanda laboral observada por la Agencia.	Es la fuente más robusta del país en empleo y condiciones de vida. Útil para establecer líneas base e indicadores de impacto

6.6.3 En un mundo ideal, ¿qué datos adicionales relevantes para este problema le gustaría tener?

- Trazabilidad de los procesos de selección de las empresas, incluyendo fases superadas por los candidatos, razones de rechazo y tiempos de respuesta.
- Datos actualizados en tiempo real sobre nuevas vacantes y perfiles demandados, mediante integraciones con portales de empleo públicos y privados.
- Ubicación geográfica exacta (georreferenciada) de los usuarios y de los centros de empleo/formación para análisis de cobertura territorial y accesibilidad.
- Historial laboral completo y no fragmentado, incluyendo experiencia formal e informal (posiblemente conectado con PILA, RUT u otras bases).
- Indicadores predictivos de riesgo de desempleo prolongado, construidos a partir de inteligencia artificial aplicada a trayectorias laborales.

6.7 Análisis.

El proceso a establecer desde la implementación de un proyecto de Ciencia de Datos en la Agencia de Empleo de Comfacauca contendrá 3 niveles de análisis, a saber: análisis exploratorio y de validación, análisis predictivo, y análisis prescriptivo.

6.7.1 ¿Qué análisis se completarán para informar sobre las acciones del proyecto?

La siguiente tabla contiene una descripción específica de la orientación de cada nivel de análisis propuesto.

Tabla 12. Relación de niveles de análisis para un proyecto de Ciencia de Datos en la Agencia de Empleo.

	Análisis 1	**Análisis 2**	**Análisis 3**
¿Cuál es el tipo de análisis?	**Análisis descriptivo y de detección**	**Análisis predictivo**	**Análisis prescriptivo**
¿Cuál es el propósito del análisis?	Validar la consistencia y homogeneidad de los datos disponibles, así como obtener una descripción del comportamiento de los atributos considerados según el problema de negocio a afrontar.	Establecer predicciones con rigor estadístico para el comportamiento del mercado laboral captado y el nivel de eficiencia organizacional alrededor de la cadena de valor de los procesos de la Agencia.	Determinar prescripciones multidimensionales para el mejoramiento de los procesos a nivel organizacional, y, para el mejoramiento a nivel socioeconómico del Departamento desde la misión de Comfacauca y la Agencia.
¿Sobre qué acciones informará este análisis?	Validación de cumplimiento de normas de calidad en estructuras de datos, limpieza y tipos de atributos considerados. Determinación de líneas base para modelación y planteamiento de objetivos.	Predicciones de comportamiento de la oferta y la demanda del mercado laboral departamental, así como de las métricas de evaluación de la cadena de valor de la Agencia.	Acciones sugeridas bajo rigor estadístico para la toma de decisiones con respecto al estado de sucesos del mercado laboral y el desempeño de la Agencia.
¿Cómo se validará este análisis empleando datos	Se emplearán estándares de validación de	Se emplearán métricas de evaluación de	Cada esquema de análisis prescriptivo tomará un conjunto

existentes? ¿Qué metodología y qué métricas de evaluación se emplearán? ¿Cómo se hará comparación con las líneas base?	estructuras de datos, así como algoritmos de evaluación lógica con respecto a los tipos de atributos y sus rangos de valores en el contexto de los modelos a emplear. En cuanto a las métricas de evaluación, se emplearán compliance ratios para propósitos de validación y limpieza, y estándares puntuales sobre los archivos de salida de las estructuras de datos del proceso. La comparación con líneas base se hará de manera fija con respecto a los estándares que requieran las estructuras de modelación a emplear.	aprendizaje (supervisado y no supervisado) sobre los resultados de los modelos predictivos, unidas a la evaluación contextual sobre estándares propuestos para la Agencia. La comparación con líneas base se hará desde la evaluación sucesiva de propuestas de modelación a contrastar. Según el propósito de las predicciones requeridas.	de indicadores como referencia, de tal forma que la evaluación y validación de las prescripciones se hará sobre el comportamiento diferencial de los mismos, adicionando el conjunto de estándares cualitativos que se propongan para la Agencia desde su mismo contexto, a nivel interno y externo.
¿Qué consideraciones éticas pueden estar asociadas a la ejecución de este análisis?	Manejo de la privacidad de usuarios inscritos y su correspondiente información de identificación y atributos socioeconómicos.	Sensibilidad ante la alteración de predicciones con propósitos no éticos. Posible inserción de grupos de interés	Sensibilidad de las prescripciones a realizar con respecto a sus efectos potenciales en la estructura organizacional (personal y

	Sensibilidad de la construcción de líneas base a la alteración con propósitos no éticos.	sobre el desarrollo de predicciones y su divulgación.	recursos), así como en los grupos de interés afectados por las políticas y dinámicas de intervención de la Caja y la Agencia (población, sectores económicos).

6.8 Consideraciones Éticas.

El conjunto de consideraciones éticas a las que puede estar sujeta la implementación de un proyecto de Ciencia de Datos en la agencia giran en torno a la sensibilidad de las acciones propuestas al grado de relevancia institucional y socioeconómica que da el rol de Comfacauca en el Cauca, así como alrededor del tratamiento restringido de información sensible de los usuarios del servicio.

6.8.1 Privacidad, Confidencialidad y Seguridad de la Información.

El conjunto de procesos propuestos para el proyecto se halla inmerso sobre el grado de información sensible sobre la población departamental de la que dispone la Agencia. Frente a lo anterior, la política de tratamiento de datos para el proyecto se adscribe al marco legal que regula a la Agencia y a Comfacauca, siendo éste el marco legal Habeas Data de la Ley 1581 de 2012 y sus correspondientes regulaciones.

6.8.2 Transparencia.

En cuanto al grado de transparencia de las operaciones, se tienen tres niveles de desclasificación necesarios, según los agentes interesados en los efectos de las labores de la Agencia.

- **Usuarios del servicio:** Se proveerá información con respecto a los atributos evaluados de los usuarios y el grado de rigor, precisión y transparencia de los procesos automatizados de selección tras la provisión de su información competente para el mercado laboral. Haciendo un énfasis adicional en el conjunto de restricciones legales a las que está sujeta la Agencia, con el fin de proteger su información sensible.
- **Colaboradores de la Agencia:** Se especificarán las fases del proceso, así como la divulgación de sus propiedades y el conjunto de fases sobre las cuales se garantiza la ausencia de sesgos internos y externos sobre los resultados a obtener para la

caracterización del mercado laboral como una base para la ejecución de operaciones y la toma de decisiones.

- **Supervisores por parte de Comfacauca y Agentes Institucionales Externos:** Se especificarán las propiedades de los procesos del equipo de Ciencia de Datos, las fuentes de información primaria relacionadas y las fechas de corte para las mismas, así como esquemas técnicos para públicos especializados, de acuerdo a las restricciones del marco legal al que se sujeta el proceso.

6.8.3 Discriminación/Equidad.

El proyecto se adscribe a la política de promoción de equidad sobre la que opera Comfacauca y la Agencia, decantando a la consideración de programas para poblaciones vulnerables específicas que suscita la Ley 21 de 1982 como marco normativo para la personería jurídica de las Cajas de Compensación.

6.8.4 Responsabilidad.

El esquema de responsabilidad del tratamiento correcto de información en el contexto del proyecto está dado por la Ley 1581 de 2012 y sus correspondientes decretos de regulación.

6.8.6 Licencia Social.

La naturaleza de la Agencia de Empleo como adscrita a la Caja de Compensación Comfacauca, y su vínculo normativo con el Sistema de Seguridad Social de la Nación, justifica su posesión del grueso de información laboral y socioeconómica del departamento que puede nutrir la iniciativa del proyecto de Ciencia de Datos. Sin embargo, se recomienda que la implementación de este proyecto vaya de la mano con una campaña de divulgación del mismo para las partes interesadas, con el fin de evitar posibles discrepancias y disputas alrededor de su ejercicio.

6.8.7 Consideraciones éticas adicionales.

Con el fin de resguardar a la Agencia de posibles disputas legales en función del manejo de información sensible, se recomienda recubrir la implementación del proyecto con la consignación de acuerdos de consentimiento en todas sus fases, desde el registro de información laboral de candidatos, hasta su difusión en niveles externos superiores de Comfacauca y grupos o instituciones externos interesados.

7. CONCLUSIONES Y TRABAJOS FUTUROS

La formulación y ejecución de este proyecto aplicado ha permitido establecer una caracterización del mercado laboral del departamento del Cauca sin precedentes, en tanto ha sido posible obtener una descripción de su comportamiento estadístico desde la óptica de las metodologías de modelación en Ciencia de Datos, basado en un dataset de alcance poblacional que ha sido un insumo del que no ha podido disponer ninguna investigación previa con un alcance y temática similar.

En general, este ejercicio de modelación predictiva y prescriptiva ha permitido la identificación específica de los factores de mayor influencia en la determinación de la relación entre la oferta y la demanda del mercado laboral, otorgando predicciones y resultados bajo un nivel consistente de rigor estadístico.[1]

Además de resaltar las fortalezas y debilidades de la distribución sectorial de la fuerza laboral en la región, esta investigación ha permitido explicar de forma detallada el funcionamiento e impactos positivos potenciales de la inserción de los métodos de Ciencia de Datos en las operaciones de la cadena de valor de la Ruta de Atención de la Agencia de Empleo, sugiriendo sus beneficios potenciales a nivel interno de su estructura organizacional, a manera de una potencial extrapolación a otras cadenas de valor de Comfacauca, y en términos de los beneficios socioeconómicos que puede implicar el desarrollar herramientas predictivas fiables como herramientas para la planeación de este componente fundamental del Sistema de Seguridad Social del Departamento.

Este trabajo no sólo representa una muestra clara de la necesidad de la implementación de protocolos de modelación predictiva y prescriptiva como un insumo de planeación en pro de la eficiencia multidimensional de Comfacauca, sino que es una invitación al uso generalizado de herramientas estadísticas avanzadas para el mejoramiento de las prácticas organizacionales y operativas en las instituciones a cargo de ejecutar las políticas públicas y económicas clave del Estado, como lo es en este caso el sistema de retribuciones de recursos parafiscales.

Como se expuso claramente en la caracterización obtenida para el mercado laboral, el departamento del Cauca enfrenta grandes retos en materia de políticas económicas, sectoriales y de empleo, si es que se propone superar el conjunto de adversidades que se observan a lo largo de la región en términos del volumen de generación de empleo, sus condiciones y distribución geográfica. Por lo cual, las herramientas de monitoreo y

[1] Como anexo a la entrega de este documento a la organización, se relaciona un archivo con el dataset limpio, adicionando la categorización suplementaria de títulos educativos y experiencia laboral previa, además del ordenamiento de los candidatos por su probabilidad de colocación de acuerdo al modelo optimizado, siendo esto un insumo clave para futura toma de decisiones de la Agencia.

evaluación rigurosa que proponen los protocolos de Ciencia de Datos como el que se desarrolla en este trabajo se muestran como indispensables para la alineación de iniciativas de política económica, sus límites normativos, y la certidumbre sobre su efectividad.

No es posible promover el avance de la población en materia socioeconómica si no se dispone de herramientas consistentes para conocer el punto de partida de las iniciativas que buscan este propósito, ni el cómo llegar a los objetivos propuestos. Es ahí donde la práctica de la Ciencia de Datos se postula como una herramienta indispensable para la mitigación del riesgo y la incertidumbre, y la optimización de los recursos disponibles.

Respecto a brechas investigativas importantes que deja este proyecto, éstas pueden considerarse desde tres perspectivas. Inicialmente, se considera la perspectiva técnica y cuantitativa, desde la cual futuros trabajos alrededor del proyecto pueden enfatizar en el desarrollo de estimadores y metodologías de evaluación de impactos sobre la línea de contribuciones como [8] y [9]. Además, en tanto se mencionó que la implementación organizacional del protocolo de clasificación en tiempo real sugiere modelos de mayor complejidad y carácter dinámico, emergen propuestas de modelación predictiva dinámica como lo son las redes neuronales estocásticas y metodologías de clasificación bajo aprendizaje profundo, similares a [10] y [11]. Es más, caben propuestas de modificación del problema fundamental propuesto, de un problema de clasificación como el que se abordó en este trabajo, a un problema de regresión para la estimación del tiempo medio para la colocación. Naturalmente, la formulación e implementación de este tipo de sofisticaciones implica esfuerzos adicionales que decanten en estructuras de datos aptas para estos tipos de modelación. Este es el paso a seguir que se sugiere para la Agencia de Empleo y para Comfacauca a nivel de planeación transversal.

La segunda perspectiva a considerar puede retomar las conclusiones del proyecto desde el análisis teórico y de modelación de redes y cadenas de valor, utilizando las implicaciones empíricas de este trabajo para ahondar en la construcción de frameworks sobre líneas similares a las de [12], [13] y [14].

Finalmente, la tercera perspectiva para evaluar las implicaciones de este proyecto aplicado en el sentido posibles trabajos futuros se halla próxima al análisis empírico de variables socioeconómicas y su influencia en la dinámica del mercado laboral de la región. Lo anterior bajo el hecho de que el proyecto aplicado realizado se dirige a un análisis de alcance poblacional en la región, con un grado de integración de la información (provisto por la organización) del cual no se tienen precedentes.

En este sentido, y como conclusión final, se espera que los resultados de esta investigación también sean una invitación al trabajo colaborativo de Comfacauca, y demás instituciones ejecutantes de políticas económicas en el Departamento, de la mano con la academia, generando enlaces para la innovación social desde el estudio cualitativo y cuantitativo del

mercado laboral y la dinámica económica a través del levantamiento de barreras sobre información clave para la formulación e implementación de alternativas en pro del bienestar de la región.

APÉNDICE

Salida de análisis de correspondencia con etiqueta de clase: Género.

```
                0     1
  Hombre      30518 11745
  Intersexual     4     0
  Mujer       39683  5883

        Pearson's Chi-squared test

data:  tabla1
X-squared = 3027, df = 2, p-value < 2.2e-16
```

Salida de análisis de correspondencia con etiqueta de clase: Nivel educativo.

```
                            0     1
  Básica Primaria(1-5)    1659   250
  Básica Secundaria(6-9)  2643   266
  Doctorado                 53    16
  Especialización         2816   460
  Maestría                 920   192
  Media(10-13)           20441  7364
  Preescolar                14     0
  Técnica Laboral        10361  2367
  Técnica Profesional     2316   574
  Tecnológica             5681  1625
  Universitaria          14793  2556

        Pearson's Chi-squared test

data:  tabla1
X-squared = 1406.7, df = 10, p-value < 2.2e-16
```

Salida de análisis de correspondencia con etiqueta de clase: Municipio de residencia.

```
                              0     1
  ABEJORRAL                   1     0
  ACACÍAS                     1     0
  AGRADO                      0     1
  AGUA DE DIOS                1     0
  AGUAZUL                     2     0
  ALCALÁ                      1     0
  ALMAGUER                  148     7
  ANDALUCÍA                   1     0
  ARGELIA                   179    17
  ARMENIA                    14     1
  BALBOA                    188    14
  BARBACOAS                   1     0
  BARBOSA                     1     0
  BARRANCABERMEJA             1     0
  BARRANQUILLA               12     0
  BELÉN DE LOS ANDAQUÍES      1     0
  BELLO                       7     7
  BOGOTÁ, D.C.              126    20
  BOLÍVAR                   702    35
```

BUCARAMANGA	6	0
BUENAVENTURA	6	4
BUENOS AIRES	208	19
BUGALAGRANDE	1	0
CAICEDONIA	2	0
CAJAMARCA	1	0
CAJIBÍO	500	40
CAJICÁ	1	0
CALARCÁ	1	0
CALDAS	1	0
CALDONO	453	42
CALI	800	155
CALIMA	2	0
CALOTO	1291	477
CANDELARIA	68	47
CARTAGENA DE INDIAS	11	2
CARTAGO	3	0
CHACHAGÜÍ	0	1
CHÍA	2	0
CHINCHINÁ	1	1
COPACABANA	2	1
CÓRDOBA	2	0
CORINTO	385	38
COTA	3	0
CUCUNUBÁ	1	0
CÚCUTA	5	0
CUMBITARA	1	0
DOSQUEBRADAS	7	1
DUITAMA	2	0
EL CARMEN DE VIBORAL	1	0
EL CERRITO	5	2
EL ROSAL	2	0
EL TAMBO	572	33
ENVIGADO	2	0
ESPINAL	1	0
FACATATIVÁ	4	0
FLORENCIA	60	3
FLORIDA	145	163
FONSECA	1	0
FUNZA	2	0
FUSAGASUGÁ	1	0
GARZÓN	1	0
GIRARDOT	1	0
GIRARDOTA	1	0
GIRÓN	1	0
GUACARÍ	3	0
GUACHENÉ	1598	703
GUACHETÁ	1	0
GUADALAJARA DE BUGA	11	3
GUADALUPE	0	1
GUALMATÁN	1	0
GUAPÍ	577	51
GUARNE	1	0
GÜEPSA	1	0
IBAGUÉ	11	2
ILES	1	0
INZÁ	150	24
IPIALES	9	0
ISNOS	4	0
ITAGÜÍ	3	1
JAMBALÓ	101	5
JAMUNDÍ	208	28

LA ARGENTINA	1	0
LA CEJA	1	0
LA CRUZ	3	0
LA CUMBRE	1	0
LA ESTRELLA	1	0
LA PLATA	7	0
LA SIERRA	145	11
LA TEBAIDA	1	0
LA UNIÓN	2	0
LA VEGA	145	18
LA VIRGINIA	1	0
LEIVA	1	0
LÓPEZ DE MICAY	15	3
MADRID	4	0
MANÍ	0	1
MANIZALES	7	0
MARINILLA	0	1
MEDELLÍN	41	4
MERCADERES	392	18
MIRANDA	970	223
MITÚ	1	0
MOCOA	10	1
MONTELÍBANO	1	0
MONTERÍA	1	0
MONTERREY	0	1
MORALES	230	8
MOSQUERA	6	2
NARIÑO	1	0
NEIVA	10	1
ORITO	3	0
OSPINA	1	0
PADILLA	398	88
PÁEZ	144	4
PAICOL	2	0
PALMIRA	69	80
PASTO	25	3
PATÍA	905	145
PELAYA	1	0
PEREIRA	21	5
PIAMONTE	416	249
PIENDAMÓ	1126	87
PITALITO	16	1
POPAYÁN	35547	7081
PRADERA	20	75
PUERRES	1	0
PUERTO ASÍS	4	0
PUERTO CAICEDO	1	0
PUERTO GAITÁN	0	1
PUERTO SALGAR	1	0
PUERTO TEJADA	5358	2572
PUPIALES	1	0
PURACÉ	194	14
QUINCHÍA	1	0
RESTREPO	1	0
RÍO DE ORO	1	0
RIOFRÍO	1	0
RIONEGRO	9	1
ROLDANILLO	1	0
ROSAS	145	16
SABANETA	3	0
SAMANIEGO	1	0
SAN AGUSTÍN	3	1

SAN ANDRÉS DE TUMACO	5	1
SAN ANTERO	1	0
SAN PABLO	1	1
SAN PEDRO	1	0
SAN PEDRO DE CARTAGO	1	0
SAN SEBASTIÁN	127	10
SAN VICENTE DEL CAGUÁN	1	0
SANDONÁ	1	0
SANTA FÉ DE ANTIOQUIA	1	0
SANTA MARTA	2	0
SANTA ROSA	223	39
SANTA ROSA DE CABAL	2	0
SANTACRUZ	1	0
SANTANDER DE QUILICHAO	10211	4016
SARAVENA	1	0
SIBUNDOY	4	0
SILVIA	499	29
SOACHA	8	0
SOTARA	87	16
SUAITA	1	0
SUÁREZ	187	1
SUCRE	60	3
TAME	1	0
TAMINANGO	1	0
TIMBÍO	1294	121
TIMBIQUÍ	256	11
TOCA	1	0
TOCANCIPÁ	1	0
TOLEDO	1	0
TORIBÍO	198	10
TOTORÓ	205	21
TULUÁ	11	2
TUNJA	1	0
TÚQUERRES	4	0
VALLE DEL GUAMUEZ	3	0
VALLEDUPAR	1	0
VIJES	1	0
VILLA DE SAN DIEGO DE UBATÉ	1	0
VILLA RICA	1712	678
VILLAGARZÓN	3	0
VILLAMARÍA	1	0
VILLAVICENCIO	10	0
VITERBO	1	0
YACUANQUER	1	0
YOLOMBÓ	1	0
YOPAL	1	1
YOTOCO	1	0
YUMBO	13	4
ZARAGOZA	1	0
ZARZAL	2	0
ZIPAQUIRÁ	2	0

```
        Pearson's Chi-squared test

data:  tabla1
X-squared = 3957.2, df = 193, p-value < 2.2e-16
```

Salida de análisis de correspondencia con etiqueta de clase: Porcentaje de Hoja de Vida.

```
          0     1
```

```
  100 65044 17613
  25   3078     1
  30    463     0
  35     13     0
  40     18     0
  55    251     0
  60    341     0
  65     65     0
  70    179     8
  85     95     0
  90    195     6
  95    463     0

        Pearson's Chi-squared test

data:  tabla1
X-squared = 1344.4, df = 11, p-value < 2.2e-16
```

Salida de análisis de correspondencia con etiqueta de clase: Canal de Registro.

```
                    0     1
  Agencia       32244 12197
  Autoregistro 37961  5431

        Pearson's Chi-squared test with Yates' continuity correction

data:  tabla1
X-squared = 3049.5, df = 1, p-value < 2.2e-16
```

REFERENCIAS BIBLIOGRÁFICAS

[1] Caja de Compensación Familiar del Cauca - COMFACAUCA, «Informe de gestión social y estados financieros 2023,» COMFACAUCA, Popayán, Cauca, 2023.

[2] M. Porter, Competitive Advantage: Creating and Sustaining Superior Performance, New York: Simon and Schuster, 1985.

[3] G. James, D. Witten, T. Hastie, R. Tibshirani y J. Taylor, An introduction to statistical learning, Cham, Switzerland: Springer Texts in Statistics, 2023.

[4] M. Schonlau, Applied Statistical Learning, Waterloo, Canada: Springer Nature, 2023.

[5] T. Oesterreich, E. Anton y F. Teuteberg, «What translates big data into business value? A meta-analysis of the impacts of business analytics on firm performance,» *INFORMATION AND MANAGEMENT,* vol. 59, nº 6, 2022.

[6] A. Gómez Sánchez, J. Sarmiento Castillo y C. Fajardo Hoyos, «Análisis de la dinámica del mercado laboral en Popayán - Colombia,» *Económicas CUC,* vol. 37, nº 1, p. 135, 2016.

[7] Carnegie Mellon University, «Project Scoping Worksheet,» 1 Octubre 2021. [En línea]. Available: https://datasciencepublicpolicy.org/wp-content/uploads/2021/09/ProjectScopingWorksheetBlank.pdf. [Último acceso: 13 Mayo 2025].

[8] G. Imbens y J. Wooldridge, «Recent developments in the econometrics of program evaluation,» *Journal of Economic Literature, 47(1),* pp. 5-86, 2009.

[9] J. Heckman y E. Vytlacil, «Econometric evaluation of social programs, part I: causal models, structural models and econometric policy evaluation,» de *Handbook of Econometrics, volume 6B*, Amsterdam, North Holland, 2007, pp. 4779-4874.

[10] C. Giap, D. Son y D. Ha, «Firm Performance Prediction for Macroeconomic Diffusion Index using Machine Learning,» *INTERNATIONAL JOURNAL OF ADVANCED COMPUTER SCIENCE AND APPLICATIONS,* vol. 12, nº 6, pp. 87-89, 2021.

[11] Y. Ma, Z. Zhang y A. Ihler, «A Deep Choice Model for Hiring Outcome Prediction in Online Labor Markets,» *INTERNATIONAL OF COMPUTERS, COMMUNICATIONS & CONTROL,* vol. 15, nº 2, 2020.

[12] Y. Tang y F. Liou, «DOES FIRM PERFORMANCE REVEAL ITS OWN CAUSES? THE ROLE OF BAYESIAN INFERENCE,» *STRATEGIC MANAGEMENT JOURNAL,* vol. 31, nº 1, pp. 39-47, 2010.

[13] D. Saha, T. Young y J. Thacker, «Predicting firm performance and size using machine learning with a Bayesian perspective,» *MACHINE LEARNING WITH APPLICATIONS,* vol. 11, 2023.

[14] J. Campbell y M. Weese, «Compositional Models and Organizational Research: Application of a Mixture Model to Nonexperimental Data in the Context of CEO Pay,» *ORGANIZATIONAL RESEARCH METHODS,* vol. 20, nº 1, pp. 95-120, 2017.