Uso de técnicas de Machine Learning para la predicción de las tasas de desempleo y ocupación en tres ciudades de Colombia: Cali, Medellín y Popayán.

Julieth Stefens Cerón y Emerson Trujillo

Nota de Aceptación

Certificamos que el presente Trabajo de Grado Satisface, en alcances y calidad, todos los requisitos que demanda un Trabajo de Grado de Maestría.

David Arango Londoño

Director

Luis Eduardo Tobón

Valentina Corchuelo

Aprobado en cumplimiento de los requisitos exigidos por la Pontificia Universidad Javeriana Cali, para optar el título de Magister en Ciencia de Datos.

Camilo Rocha

HERNÁN CAMILO ROCHA NIÑO Ph. D.
Decano Facultad de Ingeniería y Ciencias

JUAN CARLOS MARTÍNEZ ARIAS
Director Posgrados de Ingeniería y Ciencias

Santiago de Cali, 06 de julio 2023.

Acta de Correcciones al Documento de Trabajo de Grado

Santiago de Cali, 06 de julio 2023

Autores: Julieth Cerón y Emerson Trujillo

Título del Trabajo de Grado: "Uso de técnicas de Machine Learning para la predicción de las tasas de desempleo y ocupación en tres ciudades de Colombia: Cali, Medellín y Popayán".

Director: David Arango

Como indica el artículo 2.13 de las Directrices para Trabajo de Grado de Maestría, he verificado que el estudiante indicado arriba ha implementado todas las correcciones que los Jurados del Proyecto de Trabajo de Grado definieron que se efectuaran, como consta en el Acta de Evaluación correspondiente.

David Arango Londoño

Firma del Director del Trabajo de Grado

Santiago de Cali, 06 de julio 2023

Ingeniero:
Juan Carlos Martínez Arias
Director Posgrados de Ingeniería
Facultad de Ingeniería y Ciencias
Pontificia Universidad Javeriana - Cali

Con el fin de cumplir con los requisitos exigidos por la Universidad para llevar a cabo el Trabajo de Grado y posteriormente optar por el título de Magíster en Ciencia de Datos, nos permitimos presentar a su consideración el proyecto de Trabajo de Grado denominado Uso de técnicas de Machine Learning para la predicción de las tasas de desempleo y ocupación en tres ciudades de Colombia: Cali, Medellín y Popayán, el cual será realizado por los estudiantes Julieth Stefens Cerón con código 0218220 y Emerson Trujillo con código 0060313, bajo la dirección del profesor David Arango.

El suscrito director del Trabajo de Grado autoriza para que se proceda a hacer la evaluación de este Proyecto ante el Tribunal que para el efecto se designe, toda vez que ha revisado cuidadosamente el documento y avala que ya se encuentra listo para ser presentado oficialmente.

Atentamente,

Julieth Stefens Cerón	**Emerson Trujillo**	**David Arango**
C.C. 1.143.861.514 de Cali	C.C. 16.843.432 de Jamundí	C.C. 1.130.586.950 de Cali

Pontificia Universidad JAVERIANA Cali

Maestría en Ciencia de Datos
Facultad de Ingeniería y Ciencias

FICHA RESUMEN -TRABAJO DE GRADO DE MAESTRÍA

TITULO: "Uso de técnicas de Machine Learning para la predicción de las tasas de desempleo y ocupación en tres ciudades de Colombia: Cali, Medellín y Popayán."

1. ÉNFASIS: N/A
2. TIPO DE PROYECTO: Investigación
3. ÁREA DE TRABAJO: Ingeniería y Economía
4. ESTUDIANTE (S): Julieth Stefens Cerón y Emerson Trujillo
5. CORREO ELECTRÓNICO: stefens07@javerianacali.edu.co, etrujillo@javerianacali.edu.co
6. DIRECCIÓN Y TELÉFONO: Calle 13 oeste 24 D 40/3146912314
7. DIRECTOR: David Arango
8. VINCULACIÓN DEL DIRECTOR (en la universidad): Planta
9. CORREO ELECTRÓNICO DEL DIRECTOR: Docente planta de la Facultad de Ingeniería y Ciencias
10. CODIRECTOR(ES) (Si aplica): Lya Paola Sierra
11. GRUPO O EMPRESA QUE LO AVALA (Si aplica): N/A
12. OTROS GRUPOS O EMPRESAS: N/A
13. PALABRAS CLAVE (al menos 5): Machine Learning, indicador del mercado laboral, covid-19, desempleo, predicción.
14. ODS QUE APLICA EL PROYECTO (Agenda 2030): Trabajo decente y crecimiento económico y, educación de calidad.
15. FECHA DE INICIO (Desarrollo del proyecto): 1/11/2022
16. RESUMEN (máximo 400 palabras).

En los últimos dos años, la economía regional en Colombia ha sufrido choques económicos y sociales sin precedentes debido a la pandemia del Covid-19 y el paro nacional. En consecuencia, las técnicas econométricas tradicionales de pronóstico del mercado laboral pueden resultar inadecuadas o insuficientes para capturar las nuevas condiciones y tendencias macroeconómicas. Este proyecto aplicado combina variables del mercado laboral, búsquedas en Google Trends y el Indicador Mensual de Actividad Económica (IMAE) como variable macroeconómica, para estimar un indicador del mercado laboral en tres ciudades en Colombia: Cali, Medellín y Popayán utilizando técnicas de Machine Learning. Con el uso de Máquinas de Soporte Vectorial para Regresión y Redes Neuronales se pronosticaron las tasas de desempleo y ocupación laboral para anticipar los datos oficiales proporcionados por el Departamento Administrativo Nacional de Estadística (DANE) en 1 mes. Los resultados de este estudio muestran que los errores de pronóstico de los modelos propuestos son bajos, que la previsión mejora con relación al modelo de referencia tradicional ARIMA y que las estimaciones se adaptan rápidamente a los cambios estructurales en el mercado laboral regional.

Uso de técnicas de Machine Learning para la predicción de las tasas de desempleo y ocupación en tres ciudades de Colombia: Cali, Medellín y Popayán.

Julieth Cerón Ordoñez -0218220

Emerson Trujillo Sierra -0060313

Proyecto Aplicado para optar al título de

Magister en Ciencia de Datos

Director

David Arango

Codirectora

Lya Paola Sierra Suárez

FACULTAD DE INGENIERÍA Y CIENCIAS

MAESTRÍA EN CIENCIA DE DATOS

SANTIAGO DE CALI, JULIO 06 DE 2023.

Tabla de contenido

LISTA DE TABLAS

LISTA DE FIGURAS

INTRODUCCIÓN

En los últimos dos años Colombia ha experimentado choques estructurales de gran relevancia. La pandemia del covid-19 en 2020 tuvo repercusiones económicas y sociales sin precedentes. El mercado laboral tanto local como regional, se vio fuertemente afectado [1]. Durante el segundo trimestre del año 2020, las ciudades al interior de Colombia presentaron las tasas de ocupación más bajas y las tasas de desempleo más altas de la historia económica y social regional.

La tasa de ocupación en Popayán se redujo hasta un 35%, en Cali un 40% y en Medellín hasta un 43,6% (en los tres casos una reducción trimestral entre 11 y 20 puntos porcentuales respecto a febrero 2020). De forma similar, la tasa de desempleo reflejó el mayor impacto de la pandemia del covid-19 y las medidas restrictivas para contener este flagelo. En Popayán, el dato de desempleo se incrementó hasta un 31,5%, en Cali 30,1% y en Medellín hasta un 26,7%. Después de este efecto adverso, la tendencia de recuperación en estos indicadores claves del mercado laboral regional empezó a evidenciarse. No obstante, en el segundo trimestre del año 2021, principalmente, el suroccidente del país tuvo que afrontar el impacto generado por el paro nacional, los bloqueos y alteraciones de orden público desencadenados por más de dos meses en esta región del país. En este periodo, el mercado laboral, nuevamente se vio afectado, sobre todo en Cali donde la tasa de desempleo nuevamente incrementó y la tasa de ocupación disminuyó. Luego en 2022, continuó la recuperación gradual de estos principales indicadores del mercado de trabajo regional, pero a un ritmo pausado y gradual.

Bajo ese contexto, los formuladores de política, entes públicos y privados y, los trabajadores están buscando una recuperación rápida y sostenible principalmente en las tasas de desempleo, participación en la fuerza laboral y ocupación, que son las tres medidas principales que indican la salud y evolución del mercado de trabajo. No obstante, los departamentos y ciudades al interior de Colombia carecen de un sistema de información sobre el mercado laboral, que reporte de manera concisa y precisa su estado actual. Las métricas laborales son de carácter anual a nivel de departamentos y aunque a nivel municipal son trimestrales, existe un rezago en su publicación.

Todo ello imposibilita tener un conocimiento en tiempo real de la situación económica regional, lo cual perjudica la eficacia y necesaria prontitud de la toma de decisiones de todos los agentes económicos. Cuando se conocen los datos del mercado de trabajo, dados los rezagos que contiene dicha información, ya son pocas las acciones que se pueden considerar. La toma de decisiones, por tanto, tiende a actuar tarde y corrigiendo los errores pasados.

Teniendo en cuenta que la pandemia y el paro nacional al interior de Colombia crearon condiciones sin precedentes junto con respuestas políticas inesperadas, los enfoques tradicionales de previsión pueden no ser eficaces debido a su alto grado de sensibilidad a la especificación del modelo, o a sus elevados requisitos de datos [2]. Esto, por lo tanto, ofrece una oportunidad para mejorar la predicción macroeconómica aprovechando los datos provenientes de internet y los recientes avances en técnicas de Machine Learning (ML), véase [3] y [4].

Por ello, en este proyecto de grado se combinan las métricas tradicionales y los datos de búsquedas en Google para construir un nuevo indicador del mercado laboral y predecir las tasas de ocupación y desempleo utilizando técnicas de Machine Learning en tres ciudades de Colombia: Cali, Medellín y Popayán. A nivel internacional hay varios artículos que predicen las tasas de desempleo utilizando datos de búsqueda de Google y haciendo uso de técnicas de ML (por ejemplo [5], [6], [7], [8], [9]). En Colombia, por el contrario, no existen trabajos de este tipo. Sólo existen unos pocos trabajos que predicen el empleo usando los datos de Google Trends (por ejemplo, [10] y [11]).

Para la construcción del indicador del mercado laboral, se utilizan técnicas de selección o de reducción (Regresión Lasso y Análisis de Componentes principales (ACP)). Este indicador proporciona información relevante y oportuna para una mejor comprensión del ciclo económico y social de las economías regionales, proporcionando a las empresas y a los hogares información que reducirá la incertidumbre y permitirá reaccionar con rapidez a los cambios en las tendencias económicas. Los hacedores de política tendrán otra métrica a su disposición, construida a partir de un rango más amplio de variables clave que podrán reflejar el desempeño del mercado laboral

en el corto plazo (ver una aplicación para Colombia: [12]).

Para la predicción de corto plazo en la tasa de ocupación y desempleo regional, se aplican Máquinas de Soporte Vectorial para Regresión (SVR, por sus siglas en inglés) y redes neuronales (RN). El pronóstico se realiza para el primer mes del año 2023 en las tres ciudades objeto de estudio. Se observa que los modelos aplicados se adaptaron rápidamente a los cambios en los mercados laborales regionales y las políticas. También, se compara el rendimiento de previsión de estos modelos con una alternativa de referencia (modelo tradicional ARIMA) durante el período de evaluación (2018:04-2022:12). El rendimiento del pronóstico se examinó con el error cuadrático medio de predicción fuera de muestra (RMSE) y el error medio absoluto (MAE) para previsiones con un paso de antelación.

El resto de este documento está organizado de la siguiente forma: la sección 1, define el problema; la sección 2 los objetivos del trabajo de grado; la sección 3 relaciona el marco teórico y la revisión de literatura; la sección 4 y 5 describe la metodología y especificación del modelo; la sección 6 describe los datos y los pasos para la formulación de los modelos propuestos; la sección 7 da los resultados de previsión y análisis; y la sección 8 contiene las conclusiones.

1. DEFINICIÓN DEL PROBLEMA

1.1 PLANTEAMIENTO DEL PROBLEMA

El comportamiento del mercado laboral juega un papel fundamental para conocer la realidad económica, social y política de un territorio. La tasa de desempleo y la tasa de ocupación son dos de las métricas más relevantes del mercado laboral en Colombia y son calculadas por el Departamento Administrativo Nacional de Estadística (DANE) en trimestres móviles y con un rezago en su publicación mayor a 30 días, lo que imposibilita la toma de decisiones oportuna, eficiente y en tiempo real de todos los agentes económicos. Por su impacto e implicaciones en las finanzas públicas, en la estabilidad macroeconómica y el desarrollo social, para los hacedores de la política pública es importante poder prever a tiempo los cambios en la tasa de desempleo y ocupación, a fin de aplicar las políticas pertinentes que se requieren; para el sector privado, por su parte es relevante conocer la evolución del mercado de trabajo para evaluar las diferentes estrategias de generación y retención de empleos y, para los hogares, es un tema crucial que afecta directamente sus condiciones de vida y desarrollo [13].

En este sentido, es importante para las diferentes regiones al interior de Colombia contar con una predicción de estas dos métricas en un tiempo inferior a 30 días y en la cual se sintetice la evolución actualizada y las perspectivas del mercado laboral a partir de ponderar apropiadamente las diferentes variables relacionadas con el empleo y de considerar la evolución del crecimiento económico de las ciudades a partir del Indicador Mensual de Actividad Económica (IMAE). Esto es fundamental para la política pública y el desenvolvimiento del sector privado, sobre todo en la actual coyuntura donde la dinámica del ciclo económico regional en los últimos dos años estuvo permeada por dos choques económicos y sociales sin precedentes y que han sido de gran envergadura e impacto: la crisis generada por el Covid-19 y las medidas restrictivas para contener su propagación y, los bloqueos permanentes generados por el paro nacional del 28 de abril 2021.

1.2 FORMULACIÓN DEL PROBLEMA

En este contexto, este proyecto de grado se destinó a responder las siguientes preguntas**:** ¿Las técnicas de Machine Learning permiten realizar pronósticos de corto plazo eficientes y óptimos para la tasa de desempleo y ocupación laboral a nivel regional?, ¿Las búsquedas en Google Trends relacionadas con el mercado de trabajo pueden contener información valiosa para la predicción de la tasa de desempleo y ocupación laboral a nivel regional? ¿Los modelos de Machine Learning mejoran la previsión de las tasas de desempleo y ocupación regionales dada por modelos tradicionales en series de tiempo como el ARIMA?

2. OBJETIVOS DEL PROYECTO

2.1 OBJETIVO GENERAL

Predecir con técnicas de Machine Learning la tasa de desempleo y la tasa de ocupación laboral en la ciudad de Cali, Medellín y Popayán para contribuir en tiempo real a la toma de decisiones por parte de todos los agentes económicos (empresas, Estado y hogares).

2.2 OBJETIVOS ESPECÍFICOS

- Construir un indicador mensual de seguimiento del mercado laboral para Cali, Medellín y Popayán respectivamente, con métricas propias del mercado laboral y búsquedas relacionadas al mercado laboral en Google Trends.

- Evaluar el rendimiento de los modelos de Machine Learning en series temporales para el pronóstico de corto plazo de la tasa de desempleo y ocupación laboral de Cali, Medellín y Popayán respectivamente.

3. MARCO TEÓRICO Y REVISIÓN DE LITERATURA

3.1. MARCO TEÓRICO

En esta sección, se presentan las teorías, conceptos y/o modelos relevantes que se utilizaron para desarrollar este proyecto de grado.

En Colombia, el DANE realiza la Encuesta Integrada de Hogares (GEIH) en la cual recoge información mensual a nivel nacional, sobre el mercado laboral de la población del país, y de las características sociodemográficas de la población colombiana. A partir de esta información se construye la tasa de desempleo que equivale al número total de personas desocupados que están buscando trabajo sobre la población económicamente activa (PEA) y la tasa de ocupación que corresponde al número total de personas ocupadas sobre la población en edad de trabajar (PET) [14].

Hacer seguimiento a estas métricas es de gran relevancia porque son indicadores claves del estado de la economía de una región. Si la tasa de desempleo es alta, puede ser una señal de que la economía no está creciendo lo suficiente como para crear empleos nuevos, lo que puede tener efectos negativos en otros aspectos de la economía. Además, estos indicadores son utilizados por los gobiernos y las empresas para tomar decisiones informadas sobre políticas y estrategias económicas. En suma, hacer seguimiento a la tasa de ocupación y desempleo es importante para entender el estado de la economía, el bienestar de las personas, la toma de decisiones y la evaluación de políticas públicas y programas relacionados con el empleo.

A nivel regional son pocos o casi nulos los trabajos existentes en pronósticos inmediatos del mercado laboral o en indicadores de empleo implementando técnicas de Machine Learning. Esta herramienta es uno de los campos que más desarrollo ha tenido en los últimos años debido a la gran cantidad de aplicaciones exitosas que pueden encontrarse en diferentes disciplinas. Estas técnicas permiten automatizar tareas, tomar decisiones basadas en datos, personalizar experiencias, detectar patrones y anomalías, y promover avances en investigación y ciencia. El

ML tiene aplicaciones en una amplia gama de industrias y campos, y su adopción continúa creciendo a medida que se reconocen sus beneficios y su potencial para mejorar la eficiencia y el rendimiento, véase [3].

La familia de machine learning comprende un conjunto de algoritmos y estructuras matemáticas amplio, como los perceptrones multicapa, arquitecturas de redes neuronales recurrentes hasta arquitecturas más sofisticadas de Deep Learning como las unidades LSTM (por sus siglas en inglés, Long Short Term Memory) las cuales son un tipo particular de redes neuronales recurrentes [5]; además sistemas de inferencia neuro difusa, máquinas de soporte vectorial (SVM, por sus siglas en inglés), modelos de redes dinámicas bayesianas, bosques aleatorios, árboles de decisión, Lasso, redes elásticas, entre otras ([3]), [4]), [15] y [16])

En otro ámbito, el acceso más fácil a internet durante los últimos años ha generado que cada vez más personas realicen un gran número de actividades virtuales, entre las cuales una de las más comunes resulta la búsqueda de información. Esta fuente de datos es importante actualmente porque proporcionan información abundante, revelan tendencias y comportamientos del usuario, impulsan la investigación y el desarrollo, respaldan la toma de decisiones informadas y fomentan la innovación y la creación de valor [17]. En suma, el acceso a estos datos permite a las personas y las organizaciones aprovechar el poder de la información para obtener ventajas competitivas, generar conocimiento y tomar decisiones más acertadas [18].

En ese contexto, una fuente de datos de internet que es de gran relevancia para la literatura económica es Google Trends, que muestra la popularidad y tendencias de búsqueda de palabras clave específicas en un período de tiempo y para determinados temas.como la llegada de turistas a Colombia [19]), las ventas de automóviles en Estados Unidos ([20]; [21]) y las solicitudes de seguro de desempleo [21]. Además, Google Trends se ha utilizado para medir el racismo a nivel estatal en Estados Unidos [22] y los efectos del seguro de desempleo en la búsqueda de empleo [23]. Otros estudios han proporcionado más ejemplos de pronósticos actuales de la actividad económica utilizando Google Trends [24] y [25].

Por su parte, el uso de modelos de selección de variables se ha convertido en una técnica importante para reducir la complejidad de los modelos predictivos y mejorar su precisión. En este proyecto de grado se utiliza el modelo Backward y la Regresión Lasso, dos técnicas ampliamente utilizadas para seleccionar variables en la construcción de modelos predictivos. El modelo Backward se basa en la eliminación secuencial de variables, comenzando con el modelo completo y eliminando una a una las variables que no son estadísticamente significativas. El modelo final se construye utilizando solo las variables estadísticamente significativas. Esta técnica tiene la ventaja de ser fácil de implementar y entender, y puede ser utilizada en conjunción con cualquier modelo predictivo [26].

La regresión Lasso, por otro lado, es una técnica de selección de variables que utiliza la regularización para reducir el número de variables en un modelo. La regresión Lasso se basa en la penalización de los coeficientes de regresión, lo que conduce a la eliminación de variables que no contribuyen significativamente a la predicción. Esta técnica tiene la ventaja de manejar adecuadamente conjuntos de datos con un gran número de variables, evitando la sobrecarga de datos y el sobreajuste, [27].

Para la creación del indicador de mercado laboral, en este proyecto de grado también se utiliza el análisis de componentes principales (ACP), que es una técnica estadística utilizada para reducir la dimensionalidad de los datos. En el contexto del mercado laboral, esta técnica se puede utilizar para crear indicadores que capturen la variabilidad de múltiples variables relacionadas con el empleo. Los indicadores del mercado laboral son medidas que reflejan el estado del empleo en una economía. Estos indicadores pueden incluir la tasa de desempleo, la tasa de participación en la fuerza laboral, la tasa de subempleo, entre otros. Sin embargo, a menudo hay muchas otras variables que influyen en el mercado laboral, como el género, la edad, la educación, la experiencia laboral, la región geográfica, entre otros.

En este sentido, el ACP se utiliza para reducir la dimensionalidad de estos datos al identificar las

variables que tienen una alta correlación entre sí. El objetivo es simplificar los datos y obtener información relevante a partir de las variables originales. El ACP busca identificar una serie de componentes principales que resumen la mayor parte de la variabilidad en los datos originales. Estos componentes se calculan mediante la combinación lineal de las variables originales y se ordenan por su contribución a la variabilidad total de los datos. Una vez que se han identificado los componentes principales, se pueden utilizar como base para crear indicadores del mercado laboral. Por ejemplo, si un componente principal tiene una alta carga en variables como la tasa de desempleo y la tasa de subempleo, se puede crear un indicador que capture la variabilidad en estas variables. Este indicador puede ser más representativo de la situación del mercado laboral que cualquiera de las variables individuales.

Para la predicción de la tasa de ocupación y desempleo, se hace uso de tres modelos. El primero, es un modelo tradicional ARIMA, que es un enfoque estadístico utilizado para el análisis y la predicción de series de tiempo. Combina componentes autorregresivos (AR), de promedio móvil (MA) y de integración (I) en un solo modelo. El modelo ARIMA es ampliamente utilizado en diversas áreas, como economía, finanzas, meteorología y ciencias sociales. El modelo ARIMA se define mediante tres parámetros principales: p, d y q; "p" representa el orden del componente autoregresivo, que captura las relaciones lineales entre la variable en estudio y sus valores pasados, "q" representa el orden del componente de media móvil, que captura las relaciones lineales entre la variable en estudio y los errores de predicción pasados y "d" representa el orden de diferenciación aplicada a la serie de tiempo. La diferenciación se utiliza para hacer estacionaria a una serie de tiempo, es decir, eliminar las tendencias y las estructuras de dependencia de orden superior. Estos tres parámetros se ajustan y seleccionan durante el proceso de modelado para obtener un modelo ARIMA adecuado para la serie de tiempo en cuestión.

El segundo, son las máquinas de soporte vectorial para regresión (SVR), que son un conjunto de algoritmos de aprendizaje supervisado utilizados tanto para problemas de clasificación como para regresión. En el caso de este proyecto de grado, se hace uso del enfoque de la regresión, para modelar relaciones no lineales entre variables independientes y dependientes. Estos modelos se

basan en la idea de encontrar un hiperplano que se ajuste lo mejor posible a los puntos de datos, minimizando el error de predicción.

Para abordar la sensibilidad a los datos atípicos y el ruido, las SVR introducen el concepto de un margen epsilon-insensible. Este margen establece una banda alrededor de la función de regresión, dentro de la cual no se penalizan los errores de predicción. Los puntos de datos que caen fuera de este margen se consideran errores y se penalizan en la función de pérdida. La función de pérdida utilizada en las SVR se basa en el error absoluto o en el error cuadrático, dependiendo de la elección del problema. La función de pérdida objetivo se combina con un término de regularización que controla la complejidad del modelo. El objetivo es minimizar la suma de los errores de predicción y el término de regularización.

El tercer modelo utilizado en este proyecto de grado son las redes neuronales para regresión, que son un tipo de modelo de aprendizaje automático utilizado para predecir valores numéricos continuos. Estas redes se basan en la arquitectura de redes neuronales, que consiste en capas de neuronas interconectadas. En el contexto de la regresión, una red neuronal típica consta de tres tipos de capas: capa de entrada, capa oculta y capa de salida. La red neuronal se entrena ajustando los pesos y sesgos de las conexiones entre las neuronas para minimizar una función de pérdida que mide la discrepancia entre las predicciones de la red y los valores reales de los datos de entrenamiento.

3.2. REVISIÓN DE LITERATURA

Tras una revisión de la literatura, se encuentra que las metodologías de Machine Learning (ML) más populares utilizadas para pronosticar indicadores del mercado laboral son redes neuronales con diferentes arquitecturas (como MLP, RN recurrente y PSI sigma Network), Máquina de Soporte Vectorial para regresión (SVR), regresión Lasso, bosques aleatorios y splines adaptativos multivariados, entre otras (ver [3], [7], [28], [29], [30], [31], [32]).

En el campo de la predicción, existe un creciente interés en la aplicación de técnicas de machine learning (ML). En [31], se llevó a cabo una revisión de los modelos de pronóstico de series temporales, abarcando tanto modelos paramétricos como modelos de ML no paramétricos. Se sugiere el uso de validación cruzada (VC) para evaluar los modelos de ML, que incluyen redes neuronales (RN), vectores de soporte de regresión (SVR) y Long Short-Term Memory (LSTM), que es un tipo de red neuronal recurrente. En el estudio, se comparó el rendimiento de los modelos de ML con los modelos estadísticos, utilizando medidas de precisión como el error cuadrático medio (MAE) y el coeficiente U de Theil. Se realizaron pruebas tanto en series de tiempo sintéticas como en series de tiempo reales. En concreto, este estudio ofrece una visión general de los modelos de pronóstico utilizados, con énfasis en los modelos de ML, y proporciona información sobre su rendimiento, ventajas y desventajas, y las características de los datos.

En [34], se utilizaron técnicas de machine learning (ML), como boosting, ridge regression, elastic net y least angle regression, para modelar y predecir variables macroeconómicas. También se comparó el rendimiento de estos modelos de ML con un modelo autoregresivo (AR) y un modelo bayesiano para promediar en diferentes horizontes de pronóstico. Se encontró que, en algunos de los horizontes de pronóstico, la implementación de modelos de ML mejoró los resultados en comparación con los otros modelos utilizados.

También, se han realizado varios estudios que implementan modelos de ML para predecir las tasas de ocupación y desempleo. En [13], se pronostica la tasa de desempleo de la zona euro, teniendo en cuenta 36 variables explicativas que se introducen en tres metodologías de ML: árboles de decisión, bosques aleatorios (RF) y máquinas de vectores soporte para regresión (SVR), mientras que en el área de la econometría se utiliza un modelo de regresión logística (logit) de red elástica. Los resultados de este análisis muestran que el modelo SVR superan a los demás modelos alcanzando una precisión de previsión del conjunto de datos completo del 88,5% y del 85,4% fuera de la muestra.

En [33], se llevó a cabo una comparación de la efectividad de diferentes modelos, como redes neuronales (RN) y vectores de soporte de regresión (SVR), con modelos autoregresivos (AR), autoregresivos de media móvil (ARMA) y autoregresivos de transición suave (AR de transición suave), para predecir cambios mensuales en la tasa de desempleo de Estados Unidos. Los resultados demostraron que la combinación de cuantificación vectorial de aprendizaje con RN o SVR ofreció resultados más precisos en comparación con otros modelos de predicción.

En [30], se llevó a cabo un pronóstico de la tasa de desempleo en Estados Unidos utilizando redes neuronales (RN) y regresión Lasso. Los resultados confirmaron que estos modelos eran superiores a las predicciones realizadas por encuestadores profesionales y al enfoque ingenuo de pronóstico. Esto se evaluó utilizando la raíz del error cuadrático medio (RMSE).

Por otro lado, en [29], se realizó una comparación del rendimiento de modelos tradicionales de series temporales como ARIMA, ARFIMA y GARCH, junto con modelos de machine learning (ML) como redes neuronales (RN), máquinas de soporte vectorial (SVM) y MARS (Multivariate Adaptive Regression Splines), para pronosticar la tasa de desempleo en 22 países diferentes con varios pasos de anticipación. El estudio también comparó el rendimiento utilizando el RMSE y el error absoluto medio (MAE). Además, se encontró que las redes neuronales capturaban mejor las no linealidades en comparación con los modelos tradicionales de series temporales.

De manera similar, en [28], se compararon las predicciones fuera de muestra para el empleo utilizando modelos de redes neuronales (RN) y modelos autorregresivos (AR). Se encontró que los modelos de redes neuronales superaron a los modelos autorregresivos al considerar la

diferenciación de los datos. Otro estudio relevante es el de [32], que comparó tres arquitecturas de RN diferentes (MLP, RN recurrente, PSI sigma Network) y pronosticó utilizando una combinación de SVR y filtro Kalman, encontró que SVR predijo la tasa de empleo de Estados Unidos mejor que cualquier otro modelo. En esta misma línea, destaca el estudio de [35], quienes desarrollan un conjunto de herramientas de minería de datos que incluye redes neuronales (NN) y regresiones de vectores de soporte (SVR) para predecir la tendencia del desempleo en China. Los resultados de este análisis muestran que el marco propuesto supera claramente a los enfoques de predicción tradicionales, y que la regresión de vectores de soporte (SVR)con núcleo de función de base radical (RBF) es dominante para la predicción de la tasa de desempleo.

De este modo, este proyecto de grado explora el uso de dos técnicas de machine learning, SVR (Máquinas de Vectores Soporte para la Regresión, SVR) y RN (Redes Neuronales), para pronosticar las tasas de ocupación y desempleo en las ciudades de Cali, Medellín y Popayán. Estos dos modelos son ampliamente aplicados en los estudios revisados. La SVR es una técnica de machine learning utilizada para problemas de regresión en los que la variable respuesta es continua. Es similar a un modelo de regresión lineal, pero utiliza una función no lineal llamada kernel para transformar las entradas. Para obtener una descripción más detallada, se puede consultar [36].

Por otro lado, las RN son modelos estadísticos no lineales utilizados en la previsión financiera y macroeconómica. Son flexibles y capaces de modelar cualquier dinámica no lineal en una serie temporal. Este tipo de modelos se utilizan ampliamente en diferentes aplicaciones, como se ha visto en los estudios mencionados [29], [28], [30], [33] y [38]. Además, las RN son modelos avanzados que pueden integrar eficientemente una variedad de información y permiten estructuras de modelos complejos, como no linealidades e interacciones. Esto los hace superiores a otros modelos de machine learning e incluso a los modelos autorregresivos tradicionales [37]. Esta capacidad de manejar la complejidad es especialmente valiosa al realizar pronósticos en periodos de incertidumbre o sin precedentes, como fue el caso en 2020 con la pandemia del Covid-19.

Es importante mencionar que, en Colombia no existe evidencia de estudios empíricos que combinen las búsquedas en Google Trends y las técnicas de Machine Learning para predecir el

mercado de trabajo local y regional. La mayoría de los artículos encontrados en la predicción de la Tasa de Desempleo (TD) no utilizan variables explicativas y en su lugar utilizan rezagos de la serie temporal de la TD. Otros utilizan los datos de Google Trends para predecir la TD con modelos tradicionales. El trabajo de [10], utiliza la información de la plataforma Google Trends para mejorar las predicciones de corto plazo para la tasa de desempleo en Colombia. Para esto, se seleccionan los términos de búsqueda de Google Trends que más se relacionan con la tasa de desempleo y se estiman modelos de regresión lineal simple, modelos autorregresivos integrados de media móvil (ARIMA) y su versión ampliada por variables exógenas (ARIMAX). Encuentran que el volumen de consultas mejora el ajuste de los modelos y en particular que las búsquedas de los términos "Trabajo", "Ofertas de trabajo" y "Busco trabajo" mejoran los pronósticos del comportamiento del mercado laboral, lo que muestra su potencial como fuente de información complementaria en el análisis del mercado laboral. Bajo esta misma línea de investigación, se encuentra el estudio de [11], quien propone un modelo de regresión lineal simple y la fuente de datos de Google Trends para pronosticar de manera semanal la tasa de desempleo en Colombia.

4. METODOLOGÍA

Para la creación del indicador del mercado laboral regional, primero se hace una selección de variables del catastro total construido (ver Anexo 1), utilizando técnicas de reducción. En concreto, se utilizaron el método de Regresión Lasso (proviene del acrónimo en inglés, "Least Absolute Shrinkage and Selection Operator") y el método de eliminación hacia atrás (en inglés, Backward elimination). Ambos son métodos comúnmente utilizados para seleccionar variables en modelos de regresión y muy efectivos en la identificación de un subconjunto óptimo de variables para la construcción de un modelo de [39], Ver Anexo 2.

La primera, tiene la capacidad de hacer una elección automática de variables al reducir los coeficientes de regresión a cero para variables irrelevantes, lo que simplifica el modelo y mejora su interpretabilidad. Además, ayuda a controlar el sobreajuste al limitar la complejidad. Esto es especialmente útil cuando se trabaja con conjuntos de datos de alta dimensionalidad o con un número grande de variables predictoras como es el caso de este análisis. También, Lasso tiene la capacidad de seleccionar una única variable representativa de un grupo de variables altamente correlacionadas [39]. De forma similar, el segundo enfoque se basa en un proceso iterativo que comienza con un modelo que incluye todas las variables independientes y en cada paso elimina sistemáticamente las variables menos relevantes [26].

Después de elegir el método que selecciona el número de variables óptimo y las de mayor relevancia[1] se utiliza el Análisis de Componentes Principales (ACP) para crear los indicadores del mercado laboral. El ACP toma los ejes de mayor varianza de los datos y crea una nueva característica independiente y ortogonal llamada componentes principales, que resume la información del conjunto de datos más extenso. Esta técnica ha sido popular en Colombia para la construcción de este tipo de indicadores [12]. También, resalta que el método ACP ha sido utilizado por otros autores ([40] y [41]) quienes han seleccionado los primeros componentes principales como indicadores de la condición del mercado laboral, el nivel de actividad o la

[1] Se elige el método Lasso porque selecciona un menor número de variables en dos de las tres ciudades de análisis.

velocidad de mejora. Una vez aplicado el ACP para las variables seleccionadas en cada ciudad, se obtienen los indicadores del mercado laboral como primer y segundo componente principal.

El siguiente paso consiste en integrar esos indicadores a otras variables de datos macroeconómicos con el fin de construir una proyección 1 mes adelante de las tasas de ocupación y desempleo. Los modelos, Máquina de Soporte Vectorial para Regresión y redes neuronales se aplican específicamente a la previsión de series temporales, adaptados a un problema de regresión con una respuesta cuantitativa Y. También se estimó un modelo ARIMA como modelo de referencia para comparar el rendimiento y precisión de los modelos propuestos de machine learning.

5. ESPECIFICACIÓN DEL MODELO

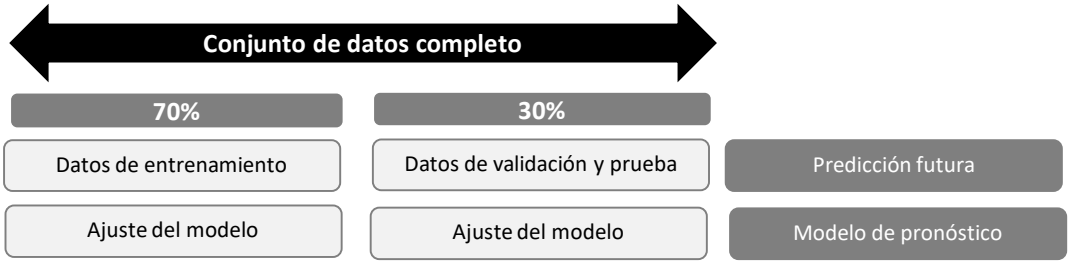
En el proceso de ajuste, la variable objetivo Y se toma como entrada para la función de pérdida. Si $f(X)$ representa el modelo con las variables de entrada X, una función de pérdida de uso común y conveniente es el error cuadrático medio (ECM), dado como:

$$L(Y, f(X)) = \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i^2 = \frac{1}{n}\sum_{i=1}^{n} (Y_i - f(x_i, \beta))^2 \quad (1)$$

Donde ε_i representa el error entre el valor observado y el valor ajustado, β corresponde a los parámetros del modelo, que se estiman tras un proceso de optimización. x_i es un vector con p variables de entrada en el momento i. La variable de respuesta Y es un vector de dimensión $(n \times 1)$y el tamaño del conjunto de datos es $n \times p$. Así, el proceso de estimación de estos modelos consiste en minimizar la función de pérdida y encontrar el vector de parámetros que penaliza los errores de predicción.

En la Figura 1, se muestra la representación del marco en el proceso completo de entrenamiento, evaluación y predicción. El modelo se estima en los datos de entrenamiento, que corresponden al 70% del conjunto total de datos. A continuación, se realiza la evaluación en el 30% restante de los datos, donde se prueba el modelo prediciendo 1 mes adelante, utilizando una ventana expansiva hacia la derecha, y reentrenando el modelo en cada paso, de acuerdo con la validación cruzada para series temporales, ver [42]. La predicción futura es el valor pronosticado para el mes fuera del conjunto de datos completo.

Figura 1. Representación de un marco de entrenamiento-prueba para proyecciones de series temporales



Fuente: Elaboración propia de los autores.

5.1 MODELOS DE PREDICCIÓN

Modelo Autorregresivo integrado de media móvil (ARIMA)

El modelo ARIMA de orden (p, d, q), se utiliza como referencia para comparar el rendimiento de los modelos de RN y SVR para cada serie temporal.

Se parte de un ARIMA en donde la tasa de desempleo y la tasa de ocupación para cada ciudad es modelada en función de sus rezagos y un proceso de ruido blanco $\{\varepsilon_t\}^{T_{t=1}}$, bajo la estructura presentada en (2):

$$\Delta^d y_t = \emptyset_0 + \sum_{i=1}^{P} \emptyset_i \Delta^d y_t + \sum_{i=0}^{P} \theta_i \varepsilon_{t-i} \qquad (2)$$

donde d corresponde al orden de integración de la tasa de desempleo o la tasa de ocupación, $\emptyset_1$, ..., $\emptyset p$ a los ponderadores del polinomio sobre los rezagos de y_t y $\emptyset_1$, ..., $\emptyset_p$ a los coeficientes correspondientes al polinomio sobre las innovaciones del proceso, es decir $\{\varepsilon_t\}^{T_{t=1}}$.

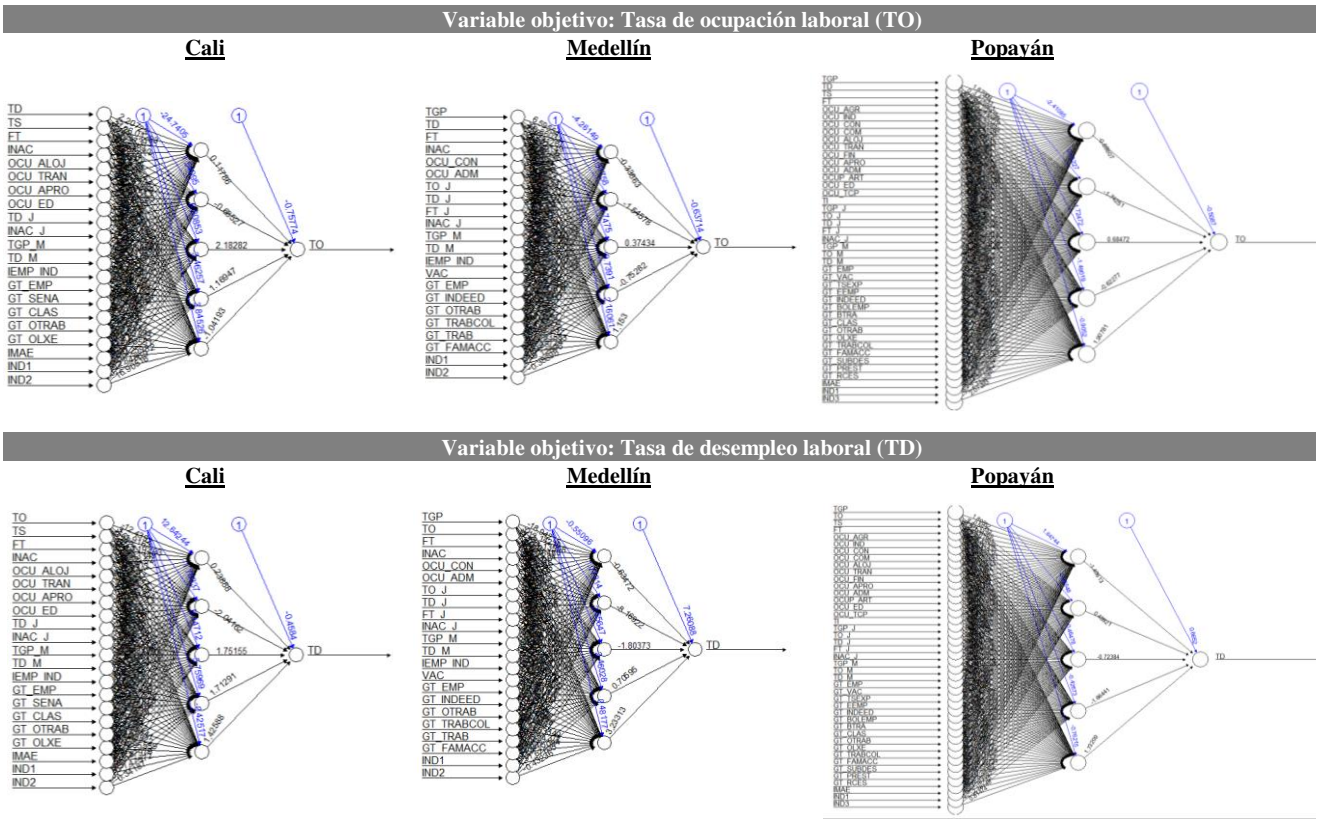
Redes neuronales (RN)

Las RN son modelos estadísticos no lineales utilizados para explotar la dinámica de las series temporales. Su naturaleza adaptativa las convierte en una herramienta ideal para explorar las relaciones entre los datos. En el último año, se han encontrado muchas aplicaciones de las RN en la previsión financiera y macroeconómica ([32] y [37]).

La arquitectura de la RN que se utiliza en este proyecto de grado es un Perceptrón Multicapa (por sus siglas en inglés, Multi-Layer Perceptron, MLP), con al menos tres capas de nodos: 1) capa de entrada, que consta de nodos que representan las variables predictoras o características de los datos, 2) capas ocultas: estas capas se encuentran entre la capa de entrada y la capa de salida. Cada capa oculta está formada por un conjunto de nodos o neuronas interconectadas. Cada nodo o neurona utiliza una función de activación no lineal (tangente hiperbólica). Destaca que, el número de nodos de cada capa oculta indica la complejidad del modelo. En el caso específico de este análisis se utilizó una única capa oculta con 5 unidades o nodos y, 3) capa de salida, que consiste en un solo nodo que representa la variable objetivo o la variable que se desea predecir, en este caso TD o TO. Esta arquitectura, es una forma de red neuronal alimentada hacia adelante, lo que significa que la información fluye en una dirección, desde la capa de entrada hasta la capa de salida, sin ciclos o retroalimentación. En general, la función de activación del nodo de salida es lineal, dado que es un problema adaptado a regresión. Se debe contemplar que, más capas ocultas, o nodos en las capas ocultas, implica más parámetros que estimar.

En la Figura 2 se muestran las representaciones gráficas del MLP en cada ciudad de análisis. En la capa de entrada, se tuvieron en cuenta 21 variables para Cali, 22 variables para Medellín y 43 para Popayán. Para las tres ciudades se tiene una capa oculta con cinco nodos y un nodo en la capa de salida (variable objetivo: TO o TD según corresponda).

Figura 2. Representación de una red neuronal (MLP) con una capa de entrada, una única capa de salida y una capa oculta con cinco nodos para Cali, Medellín y Popayán.



Fuente: Elaboración propia de los autores.

Máquinas de Soporte Vectorial para Regresión (SVR)

Las máquinas de soporte vectorial para regresión (SVR), es un algoritmo de aprendizaje supervisado utilizado para modelos de regresión no lineales. Esta técnica busca encontrar una función de regresión que se ajuste a los datos de entrenamiento y, al mismo tiempo, minimice la violación de un margen de tolerancia alrededor de los puntos objetivo.

El objetivo de la SVR es encontrar una función de regresión que esté dentro de una banda de tolerancia alrededor de los puntos objetivo. Los puntos objetivo que caen dentro de la banda de tolerancia se consideran correctamente ajustados, mientras que los que están fuera de la banda se

consideran errores. La idea principal es encontrar la función que minimice la violación del margen, es decir, los puntos que están fuera de la banda de tolerancia.

El proceso de entrenamiento de SVR implica la optimización de una función de costo que tiene en cuenta tanto la cantidad de violaciones del margen como la magnitud de las violaciones. Para esto, se utiliza un enfoque basado en la técnica de programación cuadrática y se resuelve un problema de optimización convexa.

Supongamos que los datos de entrenamiento son $\{(x_1, y_1), \dots, (x_n, y_n)\}$, donde cada x_i representa un vector de características, y el objetivo $y_i$es un número real, con $i = 1{:}\,n$.

El objetivo de una regresión ε-SV es encontrar una función $f(x)$ con una desviación del valor real $y_i$menor que ε, para todos los datos de entrenamiento. Una función SVR simple se define como:

$$f(x) = w^T \phi(x) + b \tag{3}$$

Donde w y b son los vectores de parámetros de regresión para la función y $\phi(x)$ es la función no lineal que mapea el vector de datos de entrada en un espacio de características donde los datos de entrenamiento son lineales.

La función de pérdida insensible a ε, L_ε, encuentra los puntos de predicción que se encuentran dentro del tubo creado por dos variables de holgura ξ_i y ξ_i^*:

$$L_\varepsilon(x_i) = \begin{cases} 0, & Si\ |y_i - f(x_i)| \le \varepsilon \\ |y_i - f(x_i)| - \varepsilon, & \text{en otro caso} \end{cases} \tag{4}$$

Tal y como se indica en [43] , este problema de optimización puede formularse de la siguiente forma:

$$\text{Minimizar } \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i + \xi_i^*) \text{ sujeto a } \begin{cases}\xi_i, \xi_i^* \geq 0 \\ C > 0\end{cases} \text{ y } \begin{cases} y_i - w^T\phi(x_i) - b \leq \varepsilon + \xi_i \\ w^T\phi(x_i) + b - y_i \leq \varepsilon + \xi_i^*\end{cases} \quad (5)$$

El parámetro C determina el equilibrio entre la planitud de f y la medida en que se toleran desviaciones mayores que ε .Los multiplicadores de Lagrange proporcionan la solución a este problema. Para más detalles, véase [32].

5.2 AJUSTE DEL MODELO

Se utiliza el método de validación cruzada (VC) para el proceso de ajuste de los modelos. Es una técnica utilizada en machine learning y estadística para evaluar el rendimiento de un modelo y estimar cómo se generalizará a datos no vistos. Proporciona una evaluación más confiable y robusta del modelo al utilizar múltiples divisiones de los datos en conjuntos de entrenamiento y prueba.

Al utilizar diferentes divisiones de los datos en cada iteración de validación cruzada, se puede evaluar la capacidad del modelo para generalizar y adaptarse a diferentes subconjuntos de datos. Esto ayuda a identificar si el modelo está sobreajustando (overfitting) o subajustando (underfitting) los datos.

Este método se aplica para ajustar los dos modelos de machine learning propuestos, redes neuronales y máquinas de soporte vectorial para regresión, SVR y un modelo de series temporales convencional (ARIMA) dado un conjunto de datos, así como para encontrar predicciones en las variables objetivo, TO y TD.

El modelo SVR se implementó en R utilizando la biblioteca “Caret”. En la configuración del control de entrenamiento para el ajuste del modelo SVR, se hicieron 200 iteraciones en VC para brindar una estimación más precisa del rendimiento del modelo y reducir la variabilidad asociada con una sola partición de datos en conjuntos de entrenamiento y validación. El modelo se ajusta mediante la función de entrenamiento (train), donde se especifica la fórmula que relaciona la

variable objetivo (TO cuando se va a predecir tasa de ocupación y TD cuando se va a predecir tasa de desempleo) con todas las variables predictoras disponibles en el conjunto de datos.

La red neuronal, MLP, se implementa en R utilizando la biblioteca “neuralnet”. Los hiperparámetros se fijaron en 5 unidades ocultas o neuronas en una única capa oculta, que es responsable de aprender patrones complejos en los datos y ayudar en la predicción de la variable objetivo. Se estableció un algoritmo de retropropagación, una tasa de aprendizaje entre 0,020 y 0,025 y una función de activación no lineal (tangente hiperbólica, tanh), que ayuda a transformar los datos de entrada y capturar relaciones no lineales entre las variables. Además, se definió el número máximo de pasos de entrenamiento en 150.000. Esto es útil para controlar cuánto tiempo tiene el modelo para ajustar sus pesos y buscar un mejor ajuste a los datos de entrenamiento.

El número de unidades ocultas se seleccionó tras una búsqueda de prueba y error y en función de los datos disponibles y los parámetros totales a estimar, teniendo en cuenta que, a mayor número de capas o unidades ocultas, se tendrían más parámetros a estimar. Por otro lado, la tasa de aprendizaje se seleccionó tras probar una lista de valores potenciales, entrenar el modelo y comparar los valores que producen los mejores resultados según el RMSE por fuera de muestra.

En este caso, la configuración utilizada permite ajustar el modelo de red neuronal para lograr la precisión deseada. Esto se logra mediante la combinación del número de capas ocultas (hidden) y el número máximo de iteraciones (stepmax), asegurando que el modelo converja sin sobre ajustarse. Al ajustar el número de capas ocultas y el número máximo de iteraciones, se buscó un equilibrio entre la capacidad del modelo para capturar patrones complejos y evitar el sobreajuste. De esta manera, se puede obtener un modelo que converge y se ajusta de manera óptima a los datos de entrenamiento, sin comprometer la capacidad de generalización a nuevos datos

Por último, para el modelo ARIMA, se utilizó la biblioteca “Forecast”. Con la función auto.arima se automatizó el proceso de selección del modelo ARIMA óptimo para las series de tiempo de interés, en este caso, Tasa de desempleo (TD) y Tasa de ocupación (TO) en cada ciudad. Esta técnica, utiliza criterios como el AIC (Criterio de Información de Akaike) y el BIC (Criterio de Información Bayesiano) para seleccionar el modelo ARIMA con el mejor ajuste para los datos. En este caso, no se incluyeron covariables, dado que es un análisis univariante tradicional.

Para Cali se utilizó un modelo de orden (2,0,2) en la tasa de ocupación y un modelo de orden (4,0,2) para la tasa de desempleo. Para Medellín, un modelo (1,0,3) para la tasa de ocupación y un modelo (2,0,3) para la tasa de desempleo. En Popayán se usó un modelo de orden (3,0,1) y (0,0,3) para la tasa de ocupación y desempleo respectivamente.

Los modelos en general se entrenan mediante una previsión directa. Para el horizonte 1 se desarrolló la estructura siguiente:

Predicción para $y_{t+j}, \hat{y}_{t+j} = \mathbf{modelo_j}(Y_t, X_t)$

Donde $\mathbf{modelo_j}$ es el modelo entrenado para predecir el valor de la variable objetivo en el tiempo $t + j$ y cada Y_t es un vector con las observaciones retardadas de la variable Y hasta el tiempo t y X_t es el vector de todas las características consideradas en el tiempo t.

En este análisis, el modelo para una previsión 1 mes adelante sería:

Predicción para $y_{t+1}, \hat{y}_{t+1} = \mathbf{modelo_1}(Y_t, X_t)$

En este caso, también se utiliza el método de validación cruzada, que proporciona una predicción fuera de muestra para cada punto de datos de la muestra de prueba.

5.3 MÉTRICAS DE EVALUACIÓN

La evaluación de cada modelo de pronóstico para un mes adelante se realiza con el Error Cuadrático Medio, RMSE (por sus siglas en inglés, Root Mean Squared) dentro de la muestra, mientras que fuera de la muestra se realiza utilizando el RMSE y el Error Medio Absoluto, MAE (por sus siglas en inglés, Mean Absolute Error). Estas medidas nos permiten evaluar el rendimiento y precisión de las previsiones de los modelos.

El MAE se calcula como la media de las diferencias absolutas entre las predicciones y los valores reales. Es una medida de la magnitud promedio de los errores de predicción. Un valor de MAE más bajo indica una mejor precisión del modelo, donde un MAE de 0 indica una predicción perfecta.

Por su parte, el RMSE se calcula como la raíz cuadrada de la media de los errores al cuadrado. El RMSE también es una medida de la magnitud de los errores de predicción, pero penaliza más los errores grandes debido a la operación de elevar al cuadrado. Al igual que el MAE, un valor de RMSE más bajo indica una mejor precisión del modelo, y un RMSE de 0 indica una predicción perfecta.

En el contexto de la evaluación del modelo, el MAE y el RMSE se utilizan para medir la precisión dentro y fuera de muestra. Dentro de muestra, se refiere a la evaluación de la precisión del modelo en los datos utilizados para ajustar o entrenar el modelo. El MAE y el RMSE dentro de muestra proporcionan una medida de cuán bien el modelo se ajusta a los datos de entrenamiento.

Mientras que fuera de muestra, se refiere a la evaluación de la precisión del modelo en datos que no se utilizaron durante el entrenamiento. Estos datos representan nuevos datos o datos futuros que el modelo no ha visto antes. El MAE y el RMSE fuera de muestra proporcionan una medida de cuán bien el modelo generaliza a datos no vistos previamente.

Las medidas se calculan como se muestra a continuación:

$$RMSE_{en\ muestra,i} = \sqrt{\frac{1}{n}\sum_{t=1}^{n} e_{i,t+j|t}^2} \tag{6}$$

$$RMSE_{fuera\ de\ muestra} = \sqrt{\frac{1}{n}\sum_{t=1}^{n} e_{t+j|t}^2} \tag{7}$$

$$MAE_{fuera\ de\ muestra} = \sqrt{\frac{1}{n}\sum_{t=1}^{n} |\, e_{t+j|t}|} \tag{8}$$

Donde n es el número de pronósticos, $e_{i,t+j|t}$ es el error de pronóstico en el horizonte temporal $t+j$ dada la información hasta el momento t. En la muestra de entrenamiento, i representa la submuestra en la metodología de VC.

Para comparar los modelos, se calculó el RMSE de los modelos de machine learning y del modelo de referencia (ARIMA) en el horizonte de previsión, a 1 mes, así como la relación o ratio $\left(\frac{RMSE_{modelo}}{RMSE_{referencia}}\right)$ entre ellos. Los valores de ratio inferiores a uno significan que los modelos machinen learning propuestos superan al modelo de referencia, ver [44].

6. DATOS Y FORMULACIÓN DEL MODELO

En esta sección se describen los datos, las fuentes consultadas y los detalles de la formulación del modelo de predicción.

6.1 VARIABLES

Para este análisis se construye inicialmente un catastro de datos con variables relacionadas de forma directa e indirecta con el mercado laboral en cada una de las tres ciudades de estudio. Para Cali y Medellín se recolectaron en total 53 variables mensuales y para Popayán 51 variables (Ver anexo 1).

La fuente principal de los datos es el DANE, a través de la Gran Encuesta Integrada de Hogares (GEIH), que recoge información mensual a nivel nacional, sobre el mercado laboral (tamaño y estructura de la fuerza de trabajo), de la población del país, y de las características sociodemográficas de la población colombiana. A partir de esta información se construyen dos de las métricas más importantes para monitorear el comportamiento del mercado de trabajo, la tasa de desempleo que equivale a la relación porcentual entre el número de personas que están buscando trabajo (DS) y el número de personas que integran la fuerza laboral (PEA) y, la tasa de ocupación que es la relación porcentual entre la población ocupada (OC) y el número de personas que integran la población en edad de trabajar (PET).

Otra fuente relevante en este análisis es la plataforma de Google que reporta los índices de búsqueda de cualquier palabra en el explorador. Esta información se encuentra disponible en tiempo real para diferentes frecuencias, aunque en este proyecto de grado se utiliza su versión mensual. Los datos están disponibles a nivel nacional, departamental y para grandes ciudades, con acceso público sin restricción alguna.

Google Trends reporta una serie mensual con un índice que va de 0 a 100 en el periodo de tiempo seleccionado. “Los números reflejan el interés de búsqueda en relación con el mayor valor en una región y en un periodo determinados. Un valor de 100 indica la popularidad máxima de un

término, mientras que 50 y 0 indican una popularidad que es la mitad o inferior al 1%, respectivamente, en relación con el mayor valor" [45]. Existen varios artículos que predicen la tasa de desempleo a través de datos de Google Trends, véase por ejemplo [7], [6], [1]y [8].

En la Tabla 1 se presentan los términos de búsqueda de Google Trends (GT) incluidos en el catastro de datos inicial. Las 23 consultas que se incluyeron se asocian al proceso de búsqueda de empleo, a las plataformas, herramientas o entidades en línea que se dedican a la búsqueda de empleo y reclutamiento de personal y, a las ayudas del Gobierno para el desempleo. Se debe tener en cuenta que los términos fueron incluidos teniendo en cuenta el trabajo desarrollado por [10] y [11] en Colombia, el criterio de los investigadores y posteriormente ampliados con las sugerencias del motor de búsqueda Google para los términos más populares en cada ciudad. La información correspondiente a cada uno de estos términos fue descargada para el periodo comprendido entre abril 2007 y diciembre 2022, restringiendo el dominio de búsqueda al territorio nacional, departamental y municipal.

Tabla 1. Términos consultados en Google Trends para Cali, Medellín, Bogotá D.C. y Popayán

No.	Término	Acrónimo	No.	Término	Acrónimo
1	Empleo	GT_EMP	**13**	Computrabajo	GT_COMPUT
2	Ofertas de empleo	GT_OEMP	**14**	Ofertas de empleo	GT_OTRAB
3	Vacantes	GT_VAC	**15**	Olx empleo	GT_OLXE
4	Hoja de vida	GT_HOJAV	**16**	Trabajo Colombia	GT_TRABCOL
5	Trabajo sin experiencia	GT_AGEMP	**17**	Trabajo	GT_TRAB
6	Agencia de empleo	GT_TSEXP	**18**	Servicio de empleo	GT_SEREMP
7	Sena empleo	GT_SENA	**19**	Familias en acción	GT_FAMACC
8	El empleo	GT_EEMP	**20**	Subsidio desempleo	GT_SUBDES
9	Indeed	GT_INDEED	**21**	Prestamos	GT_PREST
10	Bolsa de empleo	GT_BOLEMP	**22**	Retiro cesantías	GT_RCES
11	Busco trabajo	GT_BTRA	**23**	Cesantías	GT_CES
12	Clasificados	GT_CLAS			

Fuente: Elaboración propia de los autores.

Los datos utilizados en el ACP para la construcción del indicador fueron los seleccionados por el método Lasso. Para Cali se seleccionaron 14 variables relacionadas con métricas del mercado laboral reportadas por el DANE o el Banco de la República y 5 variables relacionadas con términos de búsqueda en Google Trends, en total 19 variables. Para Medellín se seleccionaron 21 variables, 15 de mercado laboral y 6 de Google Trends. En Popayán, el método Lasso seleccionó 41 variables, 26 relacionadas directamente con mercado laboral y 15 de búsquedas en Google Trends (ver Anexo 2).

Las características que se consideran para la predicción de la tasa de ocupación (TO) y la tasa de desempleo (TD) en cada ciudad son series temporales mensuales y están disponibles desde abril 2007 hasta diciembre 2022 con un total de 189 observaciones. Así pues, los datos consisten en una matriz de 189×19, incluidas las dos variables objetivo, para Cali. En Medellín, es una matriz de 189×21, incluidas las dos variables objetivo. Para Popayán, una matriz de 189×41, incluidas las dos variables objetivo (ver Anexo 2). Se debe precisar que inicialmente, estas variables se utilizaron mediante ACP para construir indicadores del mercado laboral de las tres ciudades de análisis, principalmente, indicadores de la tasa de desempleo y de ocupación. Ahora incluyendo estos 2 nuevos indicadores como variables y el IMAE, una variable macroeconómica que mide el nivel de actividad económica regional para la ciudad de Cali y Popayán (véase [46] y [47] para mayor detalle de la metodología), se tiene una base de datos completa que contiene 22 variables para Cali; 23 para Medellín y 44 variables para Popayán (ver Tabla 2, 3 y 4).

Tabla 2. Variables utilizadas en la predicción de la TO y TD de Cali.

No.	Variables	Acrónimo	Descripción
1	Tasa de ocupación	TO	Es la relación porcentual entre la población ocupada (OC) y el número de personas que integran la población en edad de trabajar (PET)
2	Tasa de desempleo	TD	Es la relación porcentual entre el número de personas que están buscando trabajo (DS), y el número de personas que integran la fuerza laboral (PEA).
3	Tasa de subempleo	TS	Es la relación porcentual de la población ocupada que manifestó querer y poder trabajar más horas a la semana (PS) y el número de personas que integran la fuerza laboral (PEA).
4	Fuerza de trabajo	FT	Son las personas en edad de trabajar, que trabajan o están buscando empleo.
5	Población fuera de la fuerza laboral	INAC	Son las personas en edad de trabajar, que no trabajan porque están estudiando, en oficios del hogar, incapacitados permanentes para trabajar, rentista, pensionado o jubilados y personas que no les llama la atención o creen que no vale la pena trabajar.
6	Ocupados Alojamiento y servicios de comida	OCU_ALOJ	Personas en edad de trabajar vinculados al sector alojamiento y servicios de comida
7	Ocupados Transporte y almacenamiento	OCU_TRAN	Personas en edad de trabajar vinculados al sector transporte y almacenamiento
8	Ocupados Actividades profesionales	OCU_APRO	Personas en edad de trabajar vinculados al sector actividades, profesionales, científias y técnicas.
9	Ocupado como Empleado doméstico	OCU_ED	Personas en edad de trabajar vinculados como empleados domésticos
10	Tasa de desempleo de los jóvenes	TD_J	Es la relación porcentual entre el número de personas jpovenes (14-28 años) que están buscando trabajo (DS), y el número de personas jóvenes que integran la fuerza laboral (PEA).
11	Fuera de la fuerza laboral jóvenes	INAC_J	Son las personas entre 14 y 28 años en edad de trabajar, que no trabajan porque están estudiando, en oficios del hogar, incapacitados permanentes para trabajar, rentista, pensionado o jubilados y personas que no les llama la atención o creen que no vale la pena trabajar.
12	Tasa Global de Particpación Mujeres	TGP_M	Es la relación porcentual entre las mujeres económicamente activas y las mujeres en edad de trabajar Este indicador refleja la presión de la población de mujeres sobre el mercado laboral.
13	Tasa de desempleo Mujeres	TD_M	Es la relación porcentual entre el número de mujeres que están buscando trabajo (DS), y el número de mujeres que integran la fuerza laboral (PEA).
14	Índice de empleo industrial	IEMP_IND	Índice (0-100) que mide el empleo industrial en las diferentes ciudades de Colombia.
15	Búsqueda palabra “empleo” en GT	GT_EMP	Índice de búsquedas del término "empleo " en Google Trends (GT)
16	Búsqueda palabra “sena empleo” en GT	GT_SENA	Índice de búsquedas del término "sena empleo " en Google Trends (GT)
17	Búsqueda palabra “clasificados”en GT	GT_CLAS	Índice de búsquedas del término "clasificados " en Google Trends (GT)
18	Búsqueda palabra “ofertas de trabajo”en GT	GT_OTRAB	Índice de búsquedas del término "ofertas de trabajo " en Google Trends (GT)
19	Búsqueda palabra “olx empleo”en GT	GT_OLXE	Índice de búsquedas del término "olx empleo " en Google Trends (GT)
20	Indicador Mensual de Actividad Económica (IMAE)	IMAE	Índice sintético que da una medida de la evolución de la actividad real de la economía regional a corto plazo.
21	Indicador 1 del mercado laboral	Ind1	Indicadores construidos por los autores para monitorear la actividad laboral de la ciudad.
22	Indicador 2 del mercado laboral	Ind2	

Fuente: Elaboración propia de los autores.

Tabla 3. Variables utilizadas en la predicción de la TO y TD de Medellín.

No.	Variables	Acrónimo	Descripción
1	Tasa Global de Participación	TGP	Es la relación porcentual entre la población económicamente activa y la población en edad de trabajar. Este indicador refleja la presión de la población en edad de trabajar sobre el mercado laboral.
2	Tasa de ocupación	TO	Es la relación porcentual entre la población ocupada (OC) y el número de personas que integran la población en edad de trabajar (PET)
3	Tasa de desempleo	TD	Es la relación porcentual entre el número de personas que están buscando trabajo (DS), y el número de personas que integran la fuerza laboral (PEA).
4	Fuerza de trabajo	FT	Son las personas en edad de trabajar, que trabajan o están buscando empleo.
5	Población fuera de la fuerza laboral	INAC	Son las personas en edad de trabajar, que no trabajan porque están estudiando, en oficios del hogar, incapacitados permanentes para trabajar, rentista, pensionado o jubilados y personas que no les llama la atención o creen que no vale la pena trabajar.
6	Ocupados Construcción	OCU_CON	Personas en edad de trabajar vinculados al sector construcción
7	Ocupados Administración pública, educación y salud	OCU_ADM	Personas en edad de trabajar vinculados al sector administración pública y defensa
8	Tasa de ocupación jóvenes	TO_J	Es la relación porcentual entre la población ocupada jóven (14-28 años) y el número de personas que integran la población en edad de trabajar jóven (14-28 años)
9	Tasa de desempleo de los jóvenes	TD_J	Es la relación porcentual entre el número de personas jpovenes (14-28 años) que están buscando trabajo (DS), y el número de personas jóvenes que integran la fuerza laboral (PEA).
10	Fuerza de trabajo jóvenes	FT_J	Son las personas entre 14 y 28 años en edad de trabajar, que trabajan o están buscando empleo.
11	Fuera de la fuerza laboral jóvenes	INAC_J	Son las personas entre 14 y 28 años en edad de trabajar, que no trabajan porque están estudiando, en oficios del hogar, incapacitados permanentes para trabajar, rentista, pensionado o jubilados y personas que no les llama la atención o creen que no vale la pena trabajar.
12	Tasa Global de Particpación Mujeres	TGP_M	Es la relación porcentual entre las mujeres económicamente activas y las mujeres en edad de trabajar Este indicador refleja la presión de la población de mujeres sobre el mercado laboral.
13	Tasa de desempleo Mujeres	TD_M	Es la relación porcentual entre el número de mujeres que están buscando trabajo (DS), y el número de mujeres que integran la fuerza laboral (PEA).
14	Índice de empleo industrial	IEMP_IND	índice de 1-100 que mide el empleo industrial regional.
15	Vacantes u ofertas de empleo según anuncios de prensa ir	VAC	Este indicador mide la evolución del número de vacantes (ofertas laborales) utilizando los anuncios de empleo del principal diario impreso de cada ciudad.En Cali El País y en Medellín El Colombiano.
16	Búsqueda palabra “empleo” en GT	GT_EMP	Índice de búsquedas del término "empleo " en Google Trends (GT)
17	Búsqueda palabra “indeed”en GT	GT_INDEED	Índice de búsquedas del término "indeed " en Google Trends (GT)
18	Búsqueda palabra “ofertas de trabajo”en GT	GT_OTRAB	Índice de búsquedas del término "ofertas de trabajo " en Google Trends (GT)
19	Búsqueda palabra “trabajo colombia”en GT	GT_TRABCOL	Índice de búsquedas del término "trabajo Colombia " en Google Trends (GT)
20	Búsqueda palabra “trabajo”en GT	GT_TRAB	Índice de búsquedas del término "trabajo " en Google Trends (GT)
21	Búsqueda palabra “familias en accion”en GT	GT_FAMACC	Índice de búsquedas del término "familias en acción" en Google Trends (GT)
22	Indicador 1 del mercado laboral	Ind1	Indicadores construidos por los autores para monitorear la actividad laboral de la ciudad.
23	Indicador 2 del mercado laboral	Ind2	

Fuente: Elaboración propia de los autores.

Tabla 4. Variables utilizadas en la predicción de la TO y TD de Popayán.

No.	Variables	Acrónimo	Descripción
1	Tasa Global de Participación	TGP	Es la relación porcentual entre la población económicamente activa y la población en edad de trabajar. Este indicador refleja la presión de la población en edad de trabajar sobre el mercado laboral.
2	Tasa de ocupación	TO	Es la relación porcentual entre la población ocupada (OC) y el número de personas que integran la población en edad de trabajar (PET)
3	Tasa de desempleo	TD	Es la relación porcentual entre el número de personas que están buscando trabajo (DS), y el número de personas que integran la fuerza laboral (PEA).
4	Tasa de subempleo	TS	Es la relación porcentual de la población ocupada que manifestó querer y poder trabajar más horas a la semana (PS) y el número de personas que integran la fuerza laboral (PEA).
5	Fuerza de trabajo	FT	Son las personas en edad de trabajar, que trabajan o están buscando empleo.
6	Ocupados Agricultura	OCU_AGR	Personas en edad de trabajar vinculados al sector agricultura
7	Ocupados Industria	OCU_IND	Personas en edad de trabajar vinculados al sector industrial
8	Ocupados Construcción	OCU_CON	Personas en edad de trabajar vinculados al sector construcción
9	Ocupados Comercio y reparación de vehículos	OCU_COM	Personas en edad de trabajar vinculados al sector comercio y reparación de vehículos
10	Ocupados Alojamiento y servicios de comida	OCU_ALOJ	Personas en edad de trabajar vinculados al sector alojamiento y servicios de comida
11	Ocupados Transporte y almacenamiento	OCU_TRAN	Personas en edad de trabajar vinculados al sector transporte y almacenamiento
12	Ocupados Actividades financieras y de seguros	OCU_FIN	Personas en edad de trabajar vinculados al sector actividades financieras y de seguros
13	Ocupados Actividades profesionales	OCU_APRO	Personas en edad de trabajar vinculados al sector actividades, profesionales, científias y técnicas.
14	Ocupados Administración pública, educación y salud	OCU_ADM	Personas en edad de trabajar vinculados al sector administración pública y defensa
15	Ocupados Actividades artísticas, entretenimiento y recreación	OCUP_ART	Personas en edad de trabajr vinculados al sector activdiades artísticas y de entretenimiento
16	Ocupado como Empleado doméstico	OCU_ED	Personas en edad de trabajar vinculados como empleados domésticos
17	Ocupado como Trabajador por cuenta propia	OCU_TCP	Personas en edad de trabajar vinculados como trabajador por cuenta propia
18	Tasa de informalidad	TI	Relación porcentual entre el número de ocupados informales (según el tamaño de empresa y la afiliación al sistema de seguridad social en salud y pensiones) y el total de ocupados.
19	Tasa Global de Participación jóvenes	TGP_J	Es la relación porcentual entre la población económicamente activa joven (14-28 años) y la población en edad de trabajar jóven. Este indicador refleja la presión de la población sobre el mercado laboral.
20	Tasa de ocupación jóvenes	TO_J	Es la relación porcentual entre la población ocupada jóven (14-28 años) y el número de personas que integran la población en edad de trabajar jóven (14-28 años)
21	Tasa de desempleo de los jóvenes	TD_J	Es la relación porcentual entre el número de personas jpovenes (14-28 años) que están buscando trabajo (DS), y el número de personas jóvenes que integran la fuerza laboral (PEA).
22	Fuerza de trabajo jóvenes	FT_J	Son las personas entre 14 y 28 años en edad de trabajar, que trabajan o están buscando empleo.
23	Fuera de la fuerza laboral jóvenes	INAC_J	Son las personas entre 14 y 28 años en edad de trabajar, que no trabajan porque están estudiando, en oficios del hogar, incapacitados permanentes para trabajar, rentista, pensionado o jubilados y personas que no les llama la atención o creen que no vale la pena trabajar.
24	Tasa Global de Particpación Mujeres	TGP_M	Es la relación porcentual entre las mujeres económicamente activas y las mujeres en edad de trabajar Este indicador refleja la presión de la población de mujeres sobre el mercado laboral.
25	Tasa de Ocupación Mujeres	TO_M	Es la relación porcentual entre las mujeres ocupada (OCM) y el número de mujeres que integran la población en edad de trabajar (PET)
26	Tasa de desempleo Mujeres	TD_M	Es la relación porcentual entre el número de mujeres que están buscando trabajo (DS), y el número de mujeres que integran la fuerza laboral (PEA).
27	Búsqueda palabra “empleo” en GT	GT_EMP	Índice de búsquedas del término "empleo " en Google Trends (GT)
28	Búsqueda palabra"vacantes”en GT	GT_VAC	Índice de búsquedas del término "vacantes" en Google Trends (GT)
29	Búsqueda palabra “trabajo sin experiencia”en GT	GT_TSEXP	Índice de búsquedas del término "trabajo sin experiencia " en Google Trends (GT)
30	Búsqueda palabra “el empleo” en GT	GT_EEMP	Índice de búsquedas del término "el empleo" en Google Trends (GT)
31	Búsqueda palabra “indeed”en GT	GT_INDEED	Índice de búsquedas del término "indeed " en Google Trends (GT)
32	Búsqueda palabra “bolsa de empleo” en GT	GT_BOLEMP	Índice de búsquedas del término "bolsa de empleo " en Google Trends (GT)
33	Búsqueda palabra “busco trabajo” en GT	GT_BTRA	Índice de búsquedas del término "busco trabajo" en Google Trends (GT)
34	Búsqueda palabra “clasificados”en GT	GT_CLAS	Índice de búsquedas del término "clasificados " en Google Trends (GT)
35	Búsqueda palabra “ofertas de trabajo”en GT	GT_OTRAB	Índice de búsquedas del término "ofertas de trabajo " en Google Trends (GT)
36	Búsqueda palabra “olx empleo”en GT	GT_OLXE	Índice de búsquedas del término "olx empleo " en Google Trends (GT)
37	Búsqueda palabra “trabajo colombia”en GT	GT_TRABCOL	Índice de búsquedas del término "trabajo Colombia " en Google Trends (GT)
38	Búsqueda palabra “familias en accion”en GT	GT_FAMACC	Índice de búsquedas del término "familias en acción" en Google Trends (GT)
39	Búsqueda palabra “subsidio desempleo” en GT	GT_SUBDES	Índice de búsquedas del término "subsidio de desempleo " en Google Trends (GT)
40	Búsqueda palabra “prestamos”en GT	GT_PREST	Índice de búsquedas del término "prestamos" en Google Trends (GT)
41	Búsqueda palabra “ retiro cesantias”en GT	GT_RCES	Índice de búsquedas del término "retiro cesantias" en Google Trends (GT)
42	Indicador Mensual de Actividad Económica (IMAE)	IMAE	Índice sintético que da una medida de la evolución de la actividad real de la economía regional a corto plazo.
43	Indicador 1 del mercado laboral	Ind1	Indicadores construidos por los autores para monitorear la actividad laboral de la ciudad.
44	Indicador 3 del mercado laboral	Ind3	

Fuente: Elaboración propia de los autores.

6.2 FORMULACIÓN DEL MODELO

El conjunto de datos para la predicción se dividió en dos partes: el 70% de los datos se utilizó para el entrenamiento de los modelos (dentro de la muestra), en el que los parámetros se estimaron minimizando la función de pérdida cuando se conocen los valores objetivo. El 30% restante se utilizó para la evaluación (fuera de la muestra).

Se aplicó una metodología de ventana expansiva añadiendo un nuevo dato de fuera de muestra en cada iteración. Se estimó un modelo para el horizonte de previsión que se utilizó para predecir los valores con 1 paso de antelación y se probó mediante VC fuera de muestra para determinar su rendimiento. Por último, en la muestra externa se calcularon el RMSE, el MAE y la relación entre el RMSE de los modelos de machine learning propuestos. La Tabla 5 muestra los periodos para el conjunto de datos total, de entrenamiento y de validación.

Tabla 5. Periodos utilizados en los modelos de previsión la Tasas de ocupación y desempleo de Cali, Medellín y Popayán.

Periodos	Meses	Fecha de inicio	Fecha final
Conjunto de datos total	189	Abril 2007	Diciembre 2022
Conjunto de datos de formación (En muestra) 70%	132	Abril 2007	Marzo 2018
Conjunto de datos de validación cruzada (fuera de muestra) 30%	57	Abril 2018	Diciembre 2022

Fuente: Elaboración propia de los autores.

En concreto, para la formulación del modelo se siguen estos pasos:

1. Dividir la muestra en dos partes (datos de entrenamiento y datos de prueba) siguiendo la regla 70-30, como se indica en la Tabla 5.
2. Transformar los datos, escalando todas las variables a una misma unidad de medida o rango (estandarización con media 0 y varianza 1). Es importante mencionar que algunas

de las variables utilizadas en cada ciudad tienen valores atípicos debido a la pandemia del Covid-19 y/o el paro nacional (mayo 2021) que son necesarios mantener.

3. Imputar datos cuando faltan observaciones. La imputación se realiza según un método estadístico descrito en [44].
4. Ajustar todos los modelos descritos en la sección 5 para la tasa de desempleo y la tasa de ocupación en las tres ciudades de análisis.
5. Calcular los valores de pronóstico un mes adelante para cada modelo y para las variables objetivo (TO y TD). Se culmina con la muestra inicial de ajuste en 2018:04 y se evalúa el rendimiento dentro de la muestra.
6. Calcular la predicción futura con los modelos propuestos para enero 2023, para ambas variables objetivo (TO y TD) y para las tres ciudades.
7. Calcular la medida de precisión descrita en la sección 5.3 para todos los modelos y el horizonte de predicción con el fin de evaluar su rendimiento de pronóstico.

Como las variables objetivo son TD y TO, se requiere obtener la variable dependiente Y_{t+j}, donde j =1 mes por delante. Para el horizonte 1, se obtiene un conjunto de datos de diferente tamaño. Para obtener la serie temporal $\{Y_{t+j}\}$, se eliminan los primeros j valores de la serie temporal $\{Y_t\}$ y las últimas j filas del resto de variables. Para cada conjunto de datos se ajusta un modelo llamado $\mathbf{modelo}_j$.

6. ANÁLISIS Y RESULTADOS

6.1 INDICADORES DEL MERCADO LABORAL

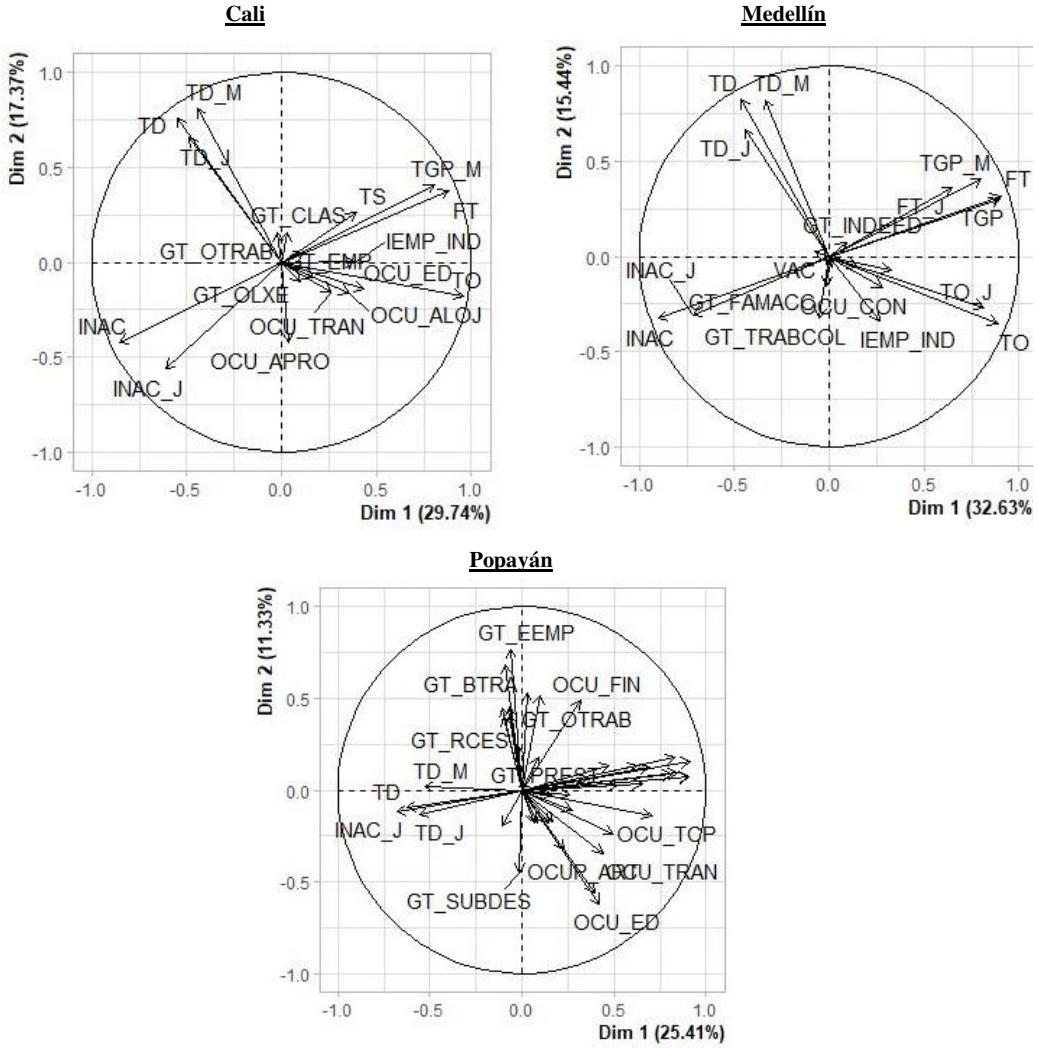
En este ejercicio, se hace la suposición que sólo se dispone de datos hasta diciembre de 2022 y se utiliza un marco para la previsión de la tasa de desempleo (TD) y la tasa de ocupación (TO) a corto plazo, para j=1 mes por delante. En este caso se tendrían la previsión para enero 2023.

Se realiza el ACP con las variables elegidas por la técnica Lasso en cada ciudad y se obtienen dos indicadores como primer y segundo componente principal que dan cuenta del 48% de la variabilidad total para Cali y Medellín respectivamente y del 37% de la variabilidad total para el caso de Popayán. Se demuestra que estos dos únicos factores explican una gran parte de la variación entre diferentes series del mercado laboral.

En la Figura 3, se muestran dos componentes principales para cada ciudad. Destaca que la tasa de ocupación y la tasa de desempleo se ajustan muy bien al primer y segundo componente principal. Esto es relevante porque muestra que estas dos nuevas series temporales obtenidas a partir del ACP resumen la dinámica de las demás variables del mercado laboral, mostrando que es posible predecir las tasas de ocupación y desempleo necesarias para establecer el estado del mercado laboral.

También, se observa que el peso de la tasa de ocupación es mayor en el primer componente que en el segundo, y a la inversa para la tasa de desempleo en cada ciudad. Las variables asociadas a las búsquedas en Google Trends son significativas en ambos componentes.

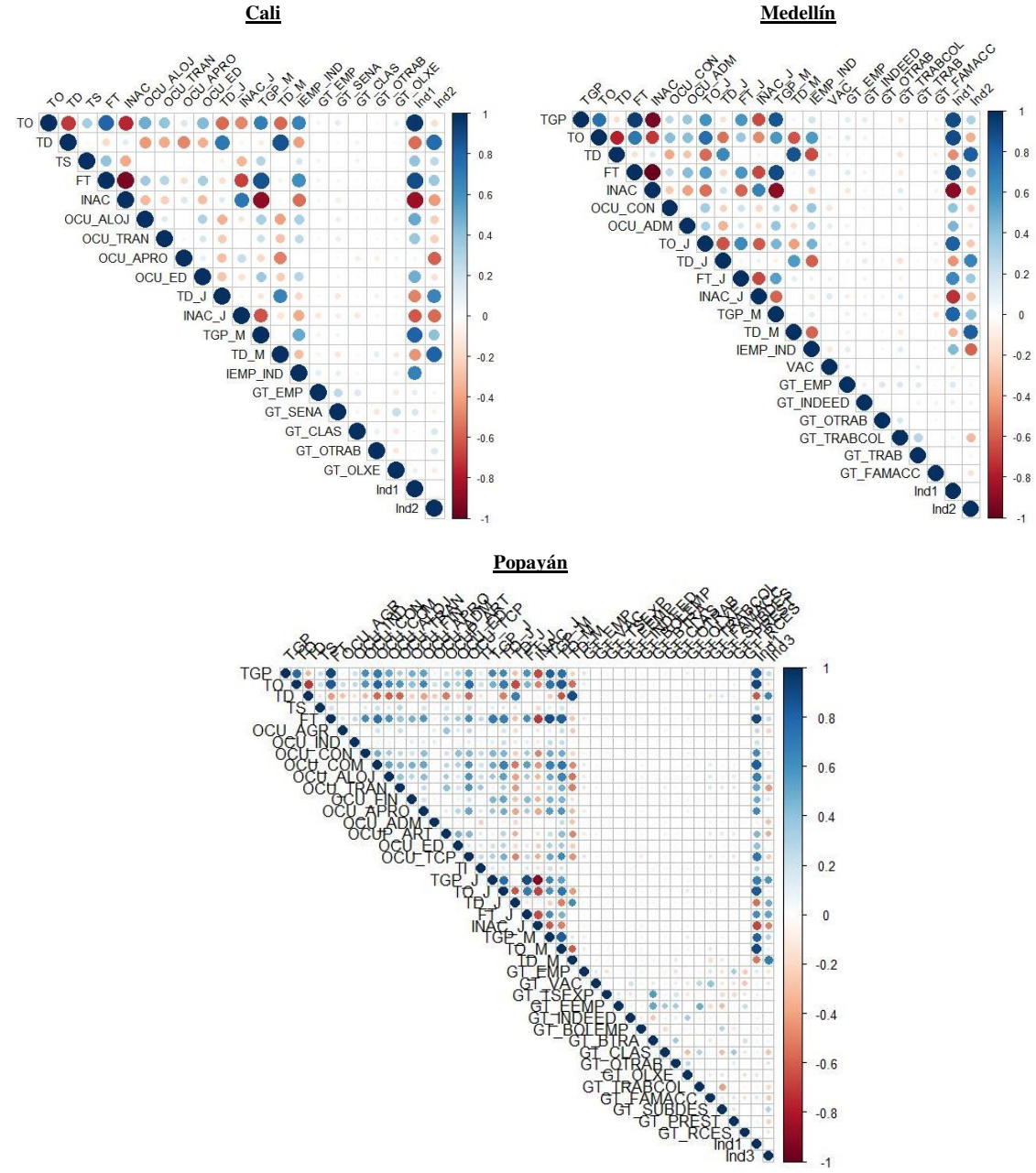
Figura 3. Representación de variables en las dos primeras componentes del indicador del mercado laboral en Cali, Medellín y Popayán.



Fuente: Elaboración propia de los autores.

La Figura 4 muestra la correlación gráfica entre las variables y demuestra que los distintos grupos de variables están correlacionados positiva y negativamente. Destaca que la correlación de las variables de Google Trends para las tres ciudades es positiva con la tasa de ocupación y el indicador del mercado laboral 1 (Ind1), lo que nos permite confiar en las variables de GT para rastrear la tasa de ocupación y desempleo más de cerca.

Figura 4. Correlación entre las variables y los indicadores del mercado laboral en Cali, Medellín y Popayán.



Fuente: Elaboración propia de los autores.

En las tres ciudades de estudio, la mayoría de las variables están bien representadas para el primer y el segundo o tercer componente principal (ver Tabla 6,7 y 8).

En la ciudad de Cali, el 70% del indicador 1 está representado por 5 variables: la tasa de ocupación, la fuerza de trabajo, la población inactiva, la TGP de las mujeres y los jóvenes inactivos. En el indicador 2 sobresalen la tasa de desempleo total, de las mujeres y jóvenes como las variables de mayor contribución. Con relación a las variables de Google Trends, destaca que su contribución es mayor en el indicador 2, relacionado en mayor medida con las condiciones de desempleo en la ciudad. La contribución conjunta de estas variables en el indicador 2 es del 2,5%.

Tabla 6. Contribución de las variables en las dos primeras componentes del Indicador de Monitoreo del mercado laboral de Cali.

No	Variables	Cod	Ind1	Ind2
1	Tasa de ocupación	TO	18.7	1.1
2	Tasa de desempleo	TD	6.2	20.1
3	Tasa de subempleo	TS	3.1	2.3
4	Fuerza de trabajo	FT	15.7	4.9
5	Población fuera de la fuerza laboral	INAC	14.9	6.2
6	Ocupados Alojamiento y servicios de comida	OCU_ALOJ	2.6	0.9
7	Ocupados Transporte y almacenamiento	OCU_TRAN	1.4	0.8
8	Ocupados Actividades profesionales	OCU_APRO	0.0	6.2
9	Ocupado como Empleado doméstico	OCU_ED	3.8	0.7
10	Tasa de desempleo de los jóvenes	TD_J	4.8	14.9
11	Fuera de la fuerza laboral jóvenes	INAC_J	7.5	10.9
12	Tasa Global de Particpación Mujeres	TGP_M	13.2	5.7
13	Tasa de desempleo Mujeres	TD_M	4.0	22.6
14	Índice de empleo industrial	IEMP_IND	3.1	0.0
15	Búsqueda palabra “empleo” en GT	GT_EMP	0.3	0.1
16	Búsqueda palabra “sena empleo” en GT	GT_SENA	0.6	0.2
17	Búsqueda palabra “clasificados” en GT	GT_CLAS	0.0	0.9
18	Búsqueda palabra “ofertas de trabajo” en GT	GT_OTRAB	0.0	0.9
19	Búsqueda palabra “olx empleo” en GT	GT_OLXE	0.2	0.4
Contribución total de las variables de Google Trends			**1.1**	**2.5**

Fuente: Elaboración propia de los autores.

En Medellín, la tasa de ocupación tiene un peso de 12,5% en el indicador 1 y la tasa de desempleo una contribución del 22,9% en el indicador 2. Destaca que las variables de Google trends en su conjunto contribuyen con 4,7% en el indicador 2 (ver Tabla 7).

Tabla 7. Contribución de las variables en las dos primeras componentes del Indicador de Monitoreo del mercado laboral de Medellín.

No	Variables	Cod	Ind1	Ind2
1	Tasa Global de Participación	TGP	12.8	3.0
2	Tasa de ocupación	TO	12.5	4.2
3	Tasa de desempleo	TD	3.4	22.9
4	Fuerza de trabajo	FT	13.0	3.4
5	Población fuera de la fuerza laboral	INAC	12.9	3.6
6	Ocupados Construcción	OCU_CON	1.2	0.9
7	Ocupados Administración pública, educación y salud	OCU_ADM	1.7	0.2
8	Tasa de ocupación jóvenes	TO_J	10.6	2.4
9	Tasa de desempleo de los jóvenes	TD_J	3.1	15.0
10	Fuerza de trabajo jóvenes	FT_J	6.7	4.4
11	Fuera de la fuerza laboral jóvenes	INAC_J	8.3	3.2
12	Tasa Global de Particpación Mujeres	TGP_M	10.2	5.5
13	Tasa de desempleo Mujeres	TD_M	1.8	22.7
14	Índice de empleo industrial	IEMP_IND	1.1	3.9
15	Vacantes según anuncios de prensa impresa	VAC	0.1	0.0
16	Búsqueda palabra “empleo” en GT	GT_EMP	0.4	0.1
17	Búsqueda palabra “indeed”en GT	GT_INDEED	0.1	0.2
18	Búsqueda palabra “ofertas de trabajo”en GT	GT_OTRAB	0.00	0.1
19	Búsqueda palabra “trabajo colombia”en GT	GT_TRABCOL	0.05	3.5
20	Búsqueda palabra “trabajo”en GT	GT_TRAB	0.00	0.1
21	Búsqueda palabra “familias en accion” en GT	GT_FAMACC	0.00	0.8
Contribución total de las variables de Google Trends			**0.5**	**4.8**

Fuente: Elaboración propia de los autores.

En Popayán, destaca que las variables de Google Trends en su conjunto representan el 15,3% del indicador 3, relacionado con el desempleo laboral. En este indicador 2 también destaca la contribución de la tasa de desempleo total, de las mujeres y de los jóvenes; la fuerza de trabajo juvenil y la TGP de los jóvenes (ver Tabla 8).

Tabla 8. Contribución de las variables en las dos primeras componentes del Indicador de Monitoreo del mercado laboral de Popayán

No.	Variables	Cod	Ind1	Ind3	No.	Variables	Cod	Ind1	Ind3
1	Tasa Global de Participación	TGP	7.5	2.2	22	Fuerza de trabajo jóvenes	FT_J	4.1	9.1
2	Tasa de ocupación	TO	8.8	1.0	23	Fuera de la fuerza laboral jóvenes	INAC_J	4.8	8.4
3	Tasa de desempleo	TD	4.1	14.1	24	Tasa Global de Particpación Mujeres	TGP_M	6.9	3.2
4	Tasa de subempleo	TS	0.7	0.0	25	Tasa de Ocupación Mujeres	TO_M	8.6	0.5
5	Fuerza de trabajo	FT	8.7	1.8	26	Tasa de desempleo Mujeres	TD_M	3.0	16.1
6	Ocupados Agricultura	OCU_AGR	0.5	0.3	27	Búsqueda palabra “empleo” en GT	GT_EMP	0.3	0.9
7	Ocupados Industria	OCU_IND	0.1	0.1	28	Búsqueda palabra "vacantes” en GT	GT_VAC	0.1	0.8
8	Ocupados Construcción	OCU_CON	2.6	0.5	29	Búsqueda palabra “trabajo sin experiencia” en GT	GT_TSEXP	0.1	0.0
9	Ocupados Comercio y reparación de vehículos	OCU_COM	4.5	0.1	30	Búsqueda palabra “el empleo” en GT	GT_EEMP	0.045	0.1
10	Ocupados Alojamiento y servicios de comida	OCU_ALOJ	2.8	0.8	31	Búsqueda palabra “indeed”en GT	GT_INDEED	0.049	0.0
11	Ocupados Transporte y almacenamiento	OCU_TRAN	2.0	2.2	32	Búsqueda palabra “bolsa de empleo”en GT	GT_BOLEMP	0.1	0.3
12	Ocupados Actividades financieras y de seguros	OCU_FIN	1.1	0.1	33	Búsqueda palabra “busco trabajo” en GT	GT_BTRA	0.1	0.3
13	Ocupados Actividades profesionales	OCU_APRO	2.3	0.0	34	Búsqueda palabra “clasificados” en GT	GT_CLAS	0.1	2.6
14	Ocupados Administración pública, educación y salud	OCU_ADM	0.0	2.0	35	Búsqueda palabra “ofertas de trabajo” en GT	GT_OTRAB	0.1	1.0
15	Ocupados Actividades artísticas, entretenimiento y recreación	OCUP_ART	1.7	1.2	36	Búsqueda palabra “olx empleo” en GT	GT_OLXE	0.2	0.3
16	Ocupado como Empleado doméstico	OCU_ED	1.9	0.0	37	Búsqueda palabra “trabajo colombia” en GT	GT_TRABCOL	0.1	1.3
17	Ocupado como Trabajador por cuenta propia	OCU_TCP	5.2	1.7	38	Búsqueda palabra “familias en accion” en GT	GT_FAMACC	0.3	2.4
18	Tasa de informalidad	TI	0.8	0.0	39	Búsqueda palabra “subsidio desempleo” en GT	GT_SUBDES	0.002	3.0
19	Tasa Global de Participación jóvenes	TGP_J	5.1	10.8	40	Búsqueda palabra “prestamos” en GT	GT_PREST	0.1	1.8
20	Tasa de ocupación jóvenes	TO_J	7.3	0.3	41	Búsqueda palabra “ retiro cesantias”en GT	GT_RCES	0.005	0.5
21	Tasa de desempleo de los jóvenes	TD_J	3.3	8.1	**Contribución total de las variables de Google Trends**			**1.6**	**15.3**

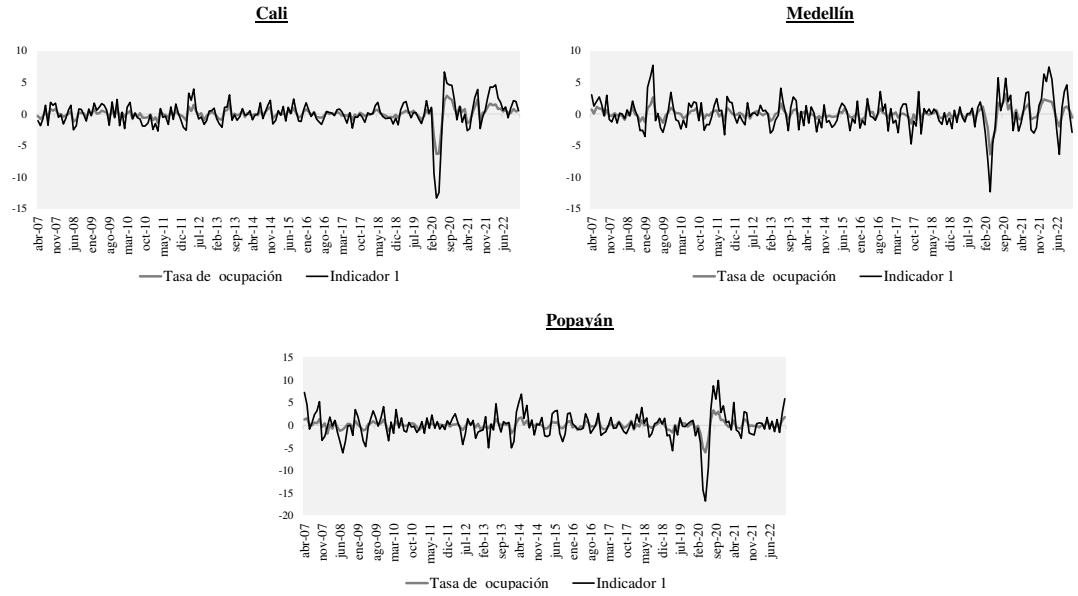
Fuente: Elaboración propia de los autores.

En la Figura 5 y 6 se muestran los indicadores obtenidos por ACP, tanto para la Tasa de Ocupación (TO) como para la Tasa de Desempleo (TD), en Cali, Medellín y Popayán. Se observa que ambos indicadores (1 y 2) representan una buena pista para las variables objetivo dado que capturan muy bien los puntos de quiebre del mercado laboral, sobre todo el cambio más reciente generado por la pandemia del covid-19. Además, permiten evidenciar el impacto diferenciado en cada ciudad y en cada indicador.

Esto resulta relevante, ya que demuestra que los indicadores construidos no sólo recogen la dinámica de la tasas de ocupación y desempleo, si no que contiene información adicional (preferencias, perspectivas, rigideces del mercado laboral, tendencias) que permiten una mejor

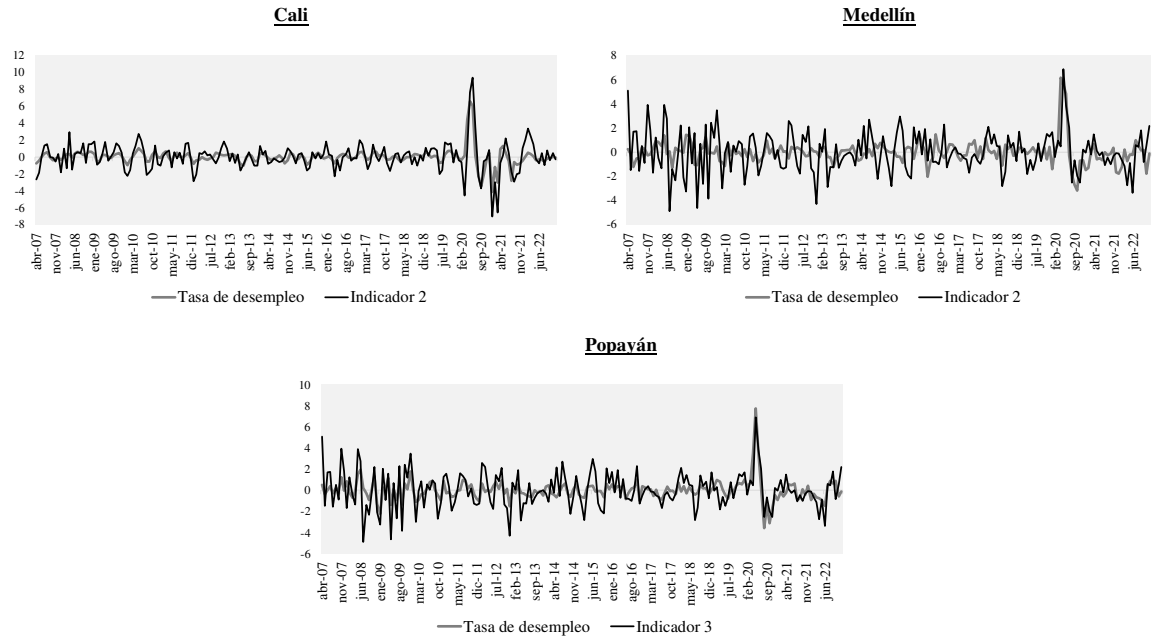
caracterización del ciclo laboral regional y contribuye a que la toma de decisiones sea más eficiente, puesto que los agentes económicos están más informados.

Figura 5. Indicador 1 del mercado laboral y la tasa de ocupación regional



Fuente: Elaboración propia de los autores.

Figura 6. Indicador 2 del mercado laboral y la tasa de desempleo regional



Fuente: Elaboración propia de los autores.

La trayectoria tanto del indicador 1 como del indicador 2 para las tres ciudades demuestran el gran impacto generado por el avance de la pandemia del covid-19 y las medidas de contención (cuarentenas estrictas) implementadas por el Gobierno nacional sobre todo en el segundo trimestre del año 2020 (ver Figura 7 y 8).

En junio 2020, el indicador 1, relacionado en mayor medida con la ocupación laboral, evidenció una contracción significativa. La ciudad de Popayán fue la ciudad que registró el mayor deterioro frente al nivel prepandemia: el indicador se ubicó un 33% por debajo de los niveles registrados en febrero 2020. Cali, es la segunda ciudad con mayor afectación, el indicador se ubicó un 27,1% por debajo de este nivel de comparación. Mientras que, en Medellín, el indicador se mantuvo sólo 14,2% por debajo de los niveles de febrero 2020 (ver Figura 7). El indicador 2 por su parte, demuestra en su trayectoria, un comportamiento similar. En julio 2020, Medellín es la ciudad que demuestra mayor afectación en su economía laboral. Durante este periodo el indicador 2 reflejó un incremento significativo que lo llevó a ubicarse un 99% por encima de los niveles registrados en febrero 2020 (nivel prepandemia). En Popayán y Cali, los niveles de julio 2020 se ubicaron

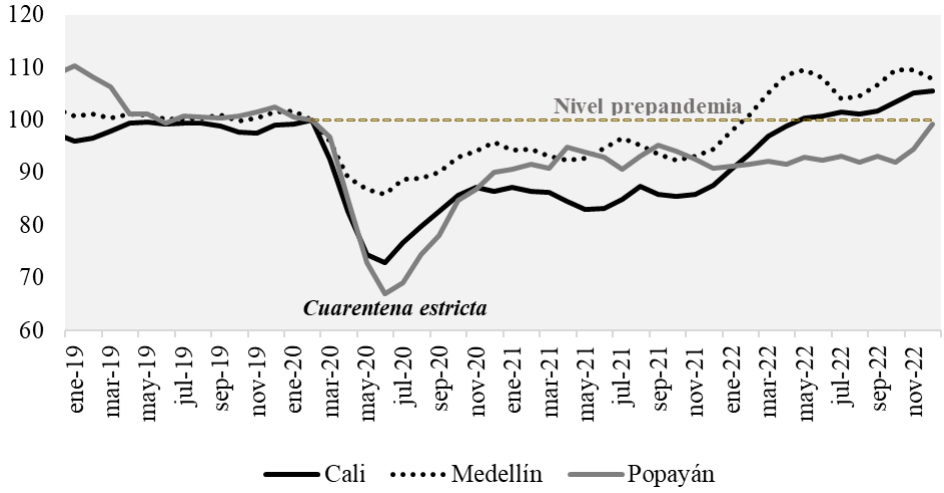
55,7% y 54,1% por encima de los niveles de febrero 2020 (ver Figura 8).

Con este resultado, se demuestra el impacto diferenciado de un choque sin precedentes en la economía regional y confirma que las restricciones a la movilidad tuvieron un efecto directo significativo en la destrucción de empleos durante los primeros meses de la pandemia, tal y como se afirma en [1].

Desde los últimos meses del año 2020, los indicadores del mercado laboral reportan para cada ciudad una tendencia de recuperación gradual, ver Figura 7 y 8. No obstante, en Cali y Popayán un nuevo choque transitorio afectó esta dinámica. El paro nacional y los bloqueos de las principales vías tanto internas como externas de la región del suroccidente del país llevadas a cabo durante el mes de mayo y junio 2021 provocaron una caída de la actividad económica de estas regiones hasta de un 20% respectivamente, una contracción que fue 3,2 y 3,4 veces mayor a la registrada en la economía de cada una de estas ciudades, durante el año 2020, tras el inicio de la pandemia. Esto sin duda, tuvo una repercusión en el mercado laboral, pero, de menor impacto a la registrada en 2020 para estas dos ciudades. El indicador 1 en Cali durante mayo 2021 se mantuvo 17,1% por debajo del nivel prepandemia y el indicador 2 en Popayán se mantuvo 25,7% por encima del nivel prepandemia. Mientras que, el indicador 1 y 2 del mercado laboral de Medellín se mantuvo solo 7,3% y 4,5% por debajo de los niveles de febrero 2020 respectivamente. Así, se afirma que el paro nacional tuvo un impacto transitorio pero sobresaliente sobre las dos ciudades del suroccidente del país principalmente (ver Figura 7 y 8).

En el año 2022, el indicador reporta la continuidad de la recuperación en la actividad productiva laboral regional. Medellín, destaca por ser la primera ciudad en superar los niveles prepandemia en relación con la ocupación laboral (indicador 1). En febrero de 2022, el indicador se ubicó por primera vez un 0,9% por encima de los niveles productivos de febrero 2020. El mercado de trabajo de Cali, desde la óptica del indicador 1, por su parte, superó los niveles prepandemia hasta el mes de mayo 2022, ubicándose 0,3% por encima del nivel prepandemia. En Popayán, por el contrario, hasta diciembre 2022 no se evidencia señales que muestren niveles del mercado laboral que superen el nivel prepandemia.

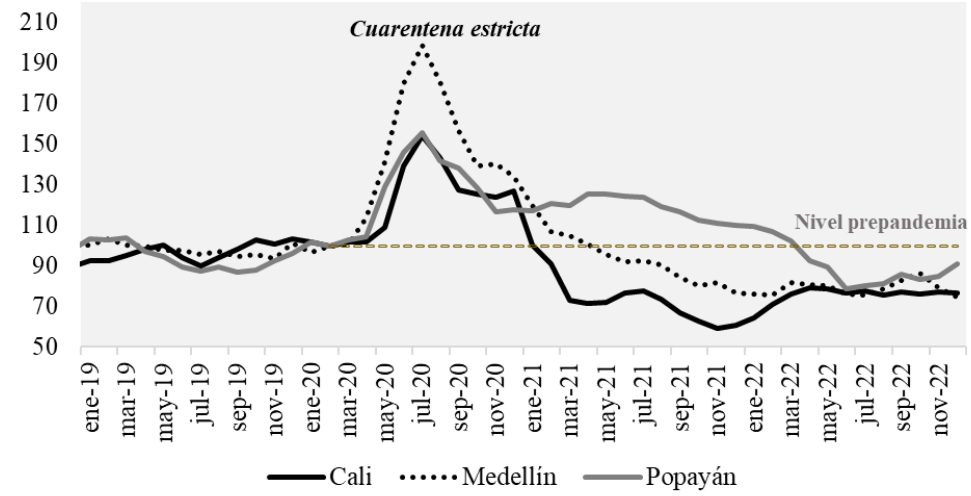
*Figura 7. Trayectoria del indicador 1 a nivel regional (2019-2022)**



*El indicador 1 se escala a índice, con base 100=febrero 2020 para evidenciar las condiciones del mercado laboral regional después de la pandemia (nivel comparativo).

Fuente: Elaboración propia de los autores.

*Figura 8. Trayectoria del indicador 2 a nivel regional (2019-2022) **



*El indicador 2 se escala a índice, con base 100=febrero 2020 para evidenciar las condiciones del mercado laboral regional después de la pandemia (nivel comparativo).

Fuente: Elaboración propia de los autores.

Bajo este panorama, se demuestra que Popayán hasta diciembre 2022 refleja unas condiciones del mercado de trabajo que en términos de ocupación laboral (indicador 1) aún no logra superar los niveles prepandemia (febrero 2020). Por el contrario, Medellín y Cali si reflejan desde los primeros meses del año 2022 una tendencia de recuperación más acelerada que ya supera los niveles prepandemia (ver Figura 7 y 8). En relación con el impacto del paro nacional en el mercado de trabajo de estas tres economías de análisis, se puede afirmar que el efecto no fue devastador como el evidenciado en 2020. Es decir, que el paro nacional fue un choque transitorio de menor duración y por tanto no tuvo un impacto tan notorio y significativo en el mercado de trabajo. Solo se observa un ligero cambio de tendencia que rápidamente vuelve a retornar hacia la recuperación, sobre todo en Cali y Popayán. En Medellín, este último choque no tuvo influencia en la dinámica de recuperación.

Po último, es importante resaltar que, en términos de desocupación laboral del mercado de trabajo (indicador 2), Medellín fue la ciudad donde la pandemia del Covid-19 reveló su mayor impacto. El incremento en el indicador de desocupación laboral es más significativo que en Cali y Popayán. También, se evidencia que la recuperación de empleos fue más acelerada en Medellín y Cali. Al cierre del año 2022, una noticia favorable es que el indicador 2 ya refleja niveles que se mantiene por debajo del nivel prepandemia en las tres ciudades de análisis.

6.2 RESULTADOS DEL PRONÓSTICO

En la tabla 9 y 10, se muestra el rendimiento los modelos predictivos de Machine Learning propuestos para la Tasa de Ocupación (TO) y la Tasa de Desempleo (TD) en un horizonte de previsión mensual un paso adelante. El rendimiento respectivo se evaluó mediante el RMSE de la muestra de entrenamiento y de prueba, el MAE y la relación o ratio entre el RMSE de los modelos propuestos y el modelo de referencia (ARIMA). La primera columna describe el modelo aplicado. La segunda columna contiene la previsión puntual para enero 2023.

La columna 3 muestra el RMSE para la muestra de entrenamiento. Las columnas 4 y 5 contienen el RMSE y MAE en la muestra de prueba total. Obsérvese que el RMSE para la muestra de entrenamiento y de prueba no tiene un tamaño similar. El RMSE de la muestra de prueba es mayor que el RMSE de la muestra de entrenamiento. Esto se debe al exceso de ajuste en el conjunto de entrenamiento, ya que la muestra de prueba contiene datos que el modelo no ha visto antes. La columna 6 muestra la relación entre el RMSE del modelo de machine learning y el modelo ARIMA. Si la relación es inferior a 1, las previsiones con el pronóstico de machine learning mejoran en comparación con el modelo de referencia.

En la tabla 9 y 10, se muestra que los modelos de machine learning utilizados tanto para TO como para TD mejoran la previsión de referencia para las tres ciudades de análisis excepto para Medellín en donde sólo el modelo de RN muestra una mejora respecto al modelo de referencia ARIMA tanto en TO como en TD (ver Tabla 10 y 11).

Sin embargo, en el caso de la previsión con RN, las mejoras respecto al modelo de referencia son mayores que las registradas por el modelo de SVR. Para la tasa de ocupación, la mejora en la previsión para Cali es del 29,9%, para Medellín es del 5,8% y para Popayán es del 14,1% (ver Tabla 9). De forma similar, para la tasa de desempleo, la mejora en la previsión en Cali es del 14,6%, para Medellín del 12,5% y en Popayán del 11,4% (ver Tabla 10).

Tabla 9. Medidas de precisión para la evaluación de los modelos de predicción de la tasa de ocupación en un horizonte de predicción de 1 mes por delante. Muestra de estimación inicial 2007:04-2018:03; muestra de previsión 2018:04-2022:12.

Cali					
Modelo	**Pronóstico**	**Muestra de entrenamiento**		**Muestra de prueba**	
		MAE	**RMSE**	**RMSE**	$Ratio = \frac{RMSE(modelo)}{RMSE\ (ARIMA)}$
Máquina de Soporte Vectorial para regresión (SVR)	59.6	0.483	0.483	0.673	0.9079
Redes Neuronales	58.7	0.189	0.312	0.519	0.7008
ARIMA	56.2	0.520	0.531	0.741	
Medellín					
Modelo	**Pronóstico**	**Muestra de entrenamiento**		**Muestra de prueba**	
		MAE	**RMSE**	**RMSE**	$Ratio = \frac{RMSE(modelo)}{RMSE\ (ARIMA)}$
Máquina de Soporte Vectorial para regresión (SVR)	56.1	0.586	0.726	0.820	1.1092
Redes Neuronales	57.7	0.315	0.476	0.571	0.9418
ARIMA	54.8	0.534	0.610	0.740	
Popayán					
Modelo	**Pronóstico**	**Muestra de entrenamiento**		**Muestra de prueba**	
		MAE	**RMSE**	**RMSE**	$Ratio = \frac{RMSE(modelo)}{RMSE\ (ARIMA)}$
Máquina de Soporte Vectorial para regresión (SVR)	49.5	0.491	0.813	0.724	0.9295
Redes Neuronales	52.1	0.193	0.425	0.669	0.8589
ARIMA	56.7	0.569	0.854	0.778	

Fuente: Elaboración propia de los autores.

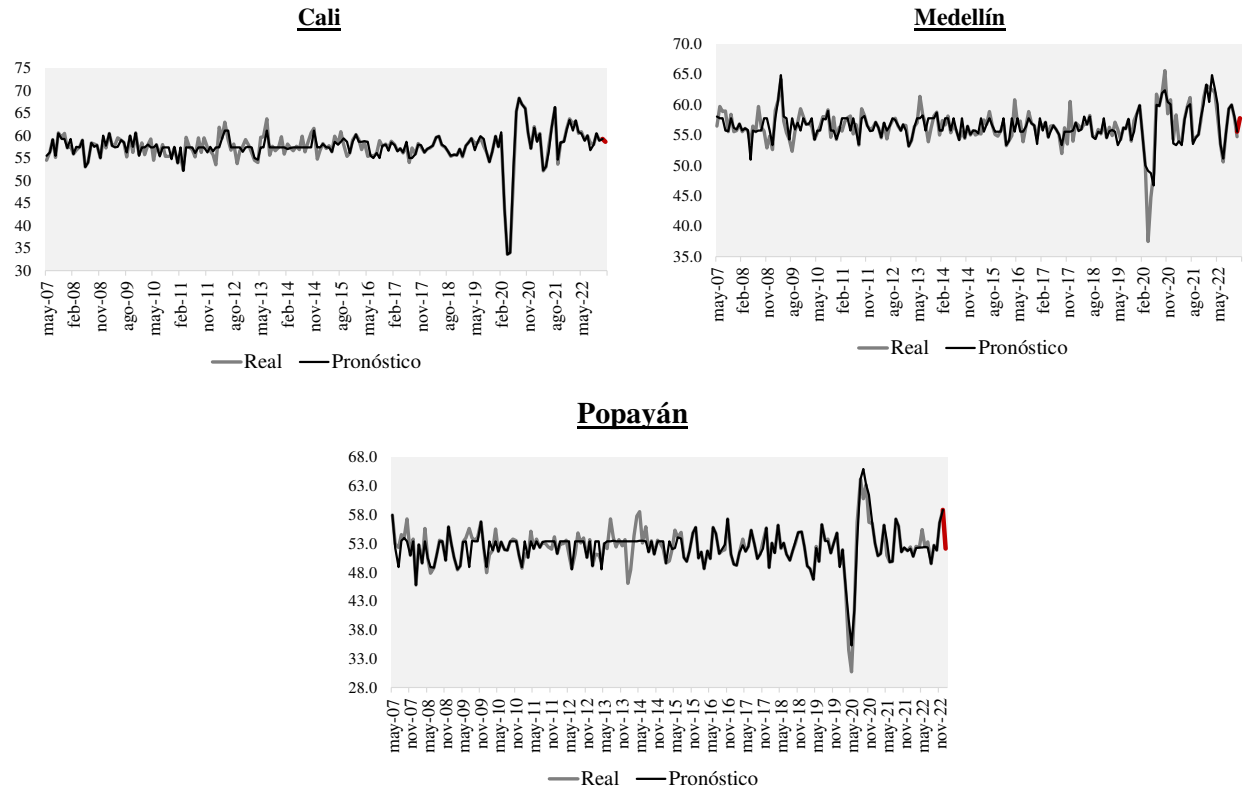
Tabla 10. Medida de precisión para la evaluación de los modelos de predicción de la tasa de desempleo en un horizonte de predicción de 1 mes por delante. Muestra de estimación inicial 2007:04-2018:03; muestra de previsión 2018:04-2022:12.

Cali					
Modelo	**Pronóstico**	**Muestra de entrenamiento**		**Muestra de prueba**	
		MAE	**RMSE**	**RMSE**	$Ratio = \frac{RMSE(modelo)}{RMSE\ (ARIMA)}$
Máquina de Soporte Vectorial para regresión (MSVR)	13.0	0.469	0.469	0.670	0.9565
Redes Neuronales	13.7	0.168	0.291	0.598	0.8539
ARIMA	13.6	0.375	0.569	0.701	
Medellín					
Modelo	**Pronóstico**	**Muestra de entrenamiento**		**Muestra de prueba**	
		MAE	**RMSE**	**RMSE**	$Ratio = \frac{RMSE(modelo)}{RMSE\ (ARIMA)}$
Máquina de Soporte Vectorial para regresión (MSVR)	13.4	0.510	0.585	0.738	1.0568
Redes Neuronales	12.1	0.203	0.315	0.611	0.8749
ARIMA	12.4	0.486	0.603	0.698	
Popayán					
Modelo	**Pronóstico**	**Muestra de entrenamiento**		**Muestra de prueba**	
		MAE	**RMSE**	**RMSE**	$Ratio = \frac{RMSE(modelo)}{RMSE\ (ARIMA)}$
Máquina de Soporte Vectorial para regresión (MSVR)	12.3	0.503	0.746	0.788	0.9613
Redes Neuronales	12.0	0.170	0.443	0.726	0.8858
ARIMA	13.3	0.557	0.789	0.820	

Fuente: Elaboración propia de los autores.

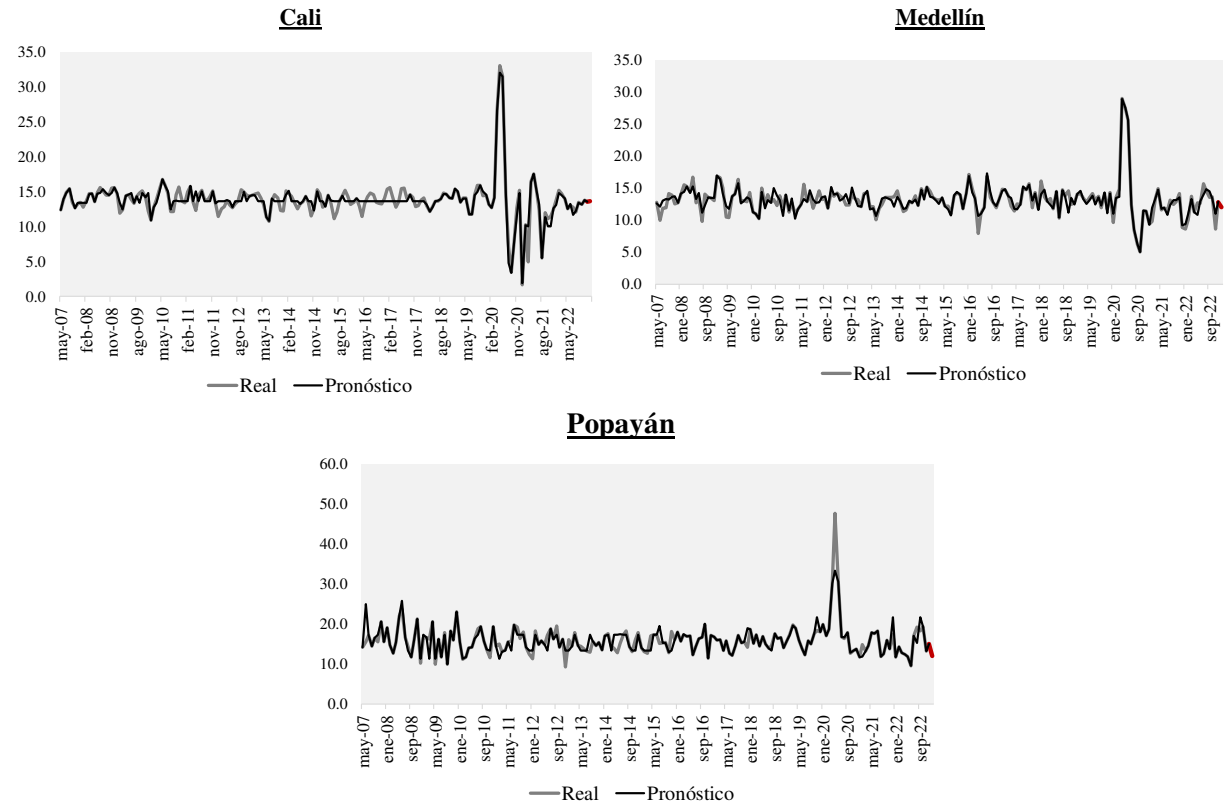
Las figuras 9 y 10 se muestran las series temporales de la TO y TD seguidas por el pronóstico con los modelos de RN, 1 mes adelante, con la previsión puntual para Cali, Medellín y Popayán, respectivamente. Se puede evidenciar que la predicción es muy ajustada dentro de muestra y fuera de ella. Además, llama la atención la gran capacidad que demuestran los modelos de machine learning para capturar los cambios estructurales del mercado de trabajo, sobre todo durante los últimos años, los cuales estuvieron permeados por choques económicos y sociales sin precedentes en la historia regional.

Figura 9. Tasa de ocupación y su pronóstico un mes adelante con Redes Neuronales (Perceptrón Multicapa).



Fuente: Elaboración propia de los autores.

Figura 10. Tasa de desempleo y su pronóstico un mes adelante con Redes Neuronales (Perceptrón Multicapa).



Fuente: Elaboración propia de los autores.

7. CONCLUSIONES

Con el ánimo de proveer en tiempo real información oportuna y contribuir a la toma de decisiones sobre el mercado laboral regional en Colombia, en este proyecto de grado se utilizaron por primera vez, métodos de Machine Learning e información de las búsquedas en Google Trends para construir un indicador de monitoreo del mercado laboral y pronosticar las tasas mensuales de desempleo y ocupación para tres ciudades de Colombia: Cali, Medellín y Popayán. Destaca que, el periodo de análisis elegido incluía las perturbaciones relacionadas con la pandemia del Covid-19 y en el caso del suroccidente del país, el paro nacional de 2021.

El método de selección Lasso, permitió identificar las variables más relevantes y significativas que se debían incluir en cada indicador del mercado laboral, permitiendo simplificar y mejorar la interpretación del indicador de acuerdo con las características y particularidades de mayor influencia dentro del mercado laboral de cada una de las ciudades.

En la información que se incluye de Google Trends en relación al mercado laboral, destaca que los términos clave que el método Lasso priorizó para las ciudades de análisis, se encuentran, "empleo", "ofertas de trabajo", "clasificados", "olx empleo", "trabajo Colombia", "familias en acción", vacantes", "agencia de empleo", "sena empleo", "indeed", "clasificados", "trabajo", "subsidio de desempleo", "trabajo sin experiencia", "el empleo", "busco trabajo", "prestamos" y "retiro de cesantías". Esto cobra relevancia, para el monitoreo de eventos y su impacto, ya que las búsquedas de estas palabras clave podrían indicar de forma directa o indirecta el impacto de cambios estructurales, crisis o eventos atípicos en el mercado laboral y ayudar a comprender las necesidades y/o intereses de los trabajadores, los hacedores de política y los agentes económicos tanto del sector público como privado.

Los indicadores de mercado laboral construidos evidenciaron el gran impacto diferenciado de la pandemia del covid-19 en cada una de las tres ciudades de análisis. Medellín y Cali, destacan por el acelerado y continuo ritmo de recuperación de su mercado laboral después del gran impacto del covid-19 en 2020 y en el caso del suroccidente, el paro nacional en 2021.Por el contrario, el mercado laboral de Popayán muestra una tendencia de recuperación más gradual y pausada, que

al cierre del 2022 no le permitió superar los niveles prepandemia (febrero 2020).

En la predicción con los modelos de machine learning, se logró evidenciar para las tres ciudades una mejora en la previsión de las tasas de desempleo y ocupación respecto al modelo de referencia tradicional-ARIMA. Esto, como consecuencia de la combinación de datos de la oficina nacional de estadística DANE con búsquedas en línea proporcionadas por Google Trends y una variable macroeconómica como el Indicador Mensual de Actividad Económica (IMAE). Se demostró que efectivamente la información proporcionada por Google Trends ayuda a predecir variables económicas, como la tasa de ocupación y la tasa de desempleo, esto es relevante porque esta es una fuente de información con una frecuencia más alta, que está disponible antes que la publicación de información de fuentes oficiales como el DANE y brinda información sobre las necesidades de las personas. En este sentido, en este proyecto de grado concluimos que Google Trends resultaría una herramienta importante sobre todo en épocas de crisis, en donde el seguimiento en tiempo real y predicción de variables económicas es relevante para la formulación de políticas oportunas y toma de decisiones eficientes.

Como referencia para las previsiones, se utilizó un modelo ARIMA para cada ciudad. La evaluación estadística de los modelos considera el periodo 2007- 2022, en el que las pruebas fuera de muestra abarcan los últimos cinco años. Los resultados muestran que las redes neuronales son más precisas que las predicciones con las máquinas de soporte vectorial para regresión y que el modelo de referencia ARIMA en cada ciudad para las tasas de desempleo y ocupación en un horizonte temporal de corto plazo (1 mes). Los términos de RMSE y MAE bajos en las muestras de entrenamiento y prueba sustentan este hecho. También, destaca que la predicción puntual a un mes vista es precisa y fiable en ambos modelos propuestos.

En este sentido, se espera que tanto los indicadores de mercado laboral y las previsiones con los modelos propuesto de machine learning contribuyan a la toma de decisiones por parte de los agentes económicos y que sirva de insumo para nuevas investigaciones en torno al mercado de trabajo a nivel regional, dado el poco avance en estos temas locales.

8. REFERENCIAS BIBLIOGRÁFICAS

[1] Morales, L. F., Mejía, L. B., Pulido, J., Flórez, L. A., Valderrama, F. L., Hermida, D., & Mahecha, K. L. P. (2022). *Efectos de la pandemia por Covid-19 en el mercado laboral colombiano. y desafíos en la economía colombiana*, 63.

[2] Cook, T., & Hall, A. (2017). *Macroeconomic Indicator Forecasting with Deep Neural Networks* (Working Paper RWP 17-11). Federal Reserve Bank of Kansas City. http://dx.doi.org/10.2139/ssrn.3046657

[3] Coulombe, P., Leroux, M., Stevanovic, D., & Surprenant, S. (2022). *How is machine learning useful for macroeconomic forecasting?*. Journal of Applied Econometrics, 37(5), 920-964.

[4] Medeiros, M. C., Vasconcelos, G. F., Veiga, Á., & Zilberman, E. (2021). *Forecasting inflation in a data-rich environment: the benefits of machine learning methods.* Journal of Business & Economic Statistics, 39(1), 98-119.

[5] D'Amuri, F., & Marcucci, J. (2017). The predictive power of Google searches in forecasting US unemployment. *International Journal of Forecasting*, *33*(4), 801-816. https://doi.org/10.1016/j.ijforecast.2017

[6] González-Fernández, M., & González-Velasco, C. (2018). Can Google econometrics predict unemployment? Evidence from Spain. *Economics Letters*, *170*(C), 42-45. https://doi.org/10.1016/j.econlet.2018.05

[7] Naccarato, A., Falorsi, S., Loriga, S., & Pierini, A. (2018). Combining official and Google Trends data to forecast the Italian youth unemployment rate. *Technological Forecasting and Social Change*, *130*(C), 114-122. https://doi.org/10.1016/j.techfore.2017.1

[8] Tuhkuri, J. (2016). *Forecasting Unemployment with Google Searches* (ETLA Working Papers N.º 35). The Research Institute of the Finnish Economy. https://ideas.repec.org/p/rif/wpaper/35.htm

[9] Xu, W., Li, Z., & Chen, Q. (2012). Forecasting the Unemployment Rate by Neural Networks Using Search Engine Query Data. *2012 45th Hawaii International Conference on System Sciences*, 3591-3599. https://doi.org/10.1109/HICSS.2012.284

[10] Rojas, L. F. C., & Aguilera, J. A. R. (2017). Pronósticos para la tasa de desempleo en Colombia a partir de Google Trends. Departamento Nacional de Planeación.

[11] Trespalacios Cárdenas, L. M. (2021). Modelo de nowcasting para pronosticar la tasa de desempleo de Colombia utilizando Google Trends.

[12] Ramos-Veloza, M. A., Cristiano-Botia, D. J., & Hernandez-Bejarano, M. D. (2021). Labor Market Indicator for Colombia. Latin American Economic Review, 30, 1-32.

[13] Gogas, P., Papadimitriou, T., & Sofianos, E. "Forecasting unemployment in the euro area with machine learning. *Journal of Forecasting*, *41*(3), 2022, pp. 551-566.

[14] Departamento Administrativo Nacional de Estadística. (DANE). Mercado laboral. Glosario.

[15] Chatzis, S. P., Siakoulis, V., Petropoulos, A., Stavroulakis, E., & Vlachogi-annakis, N. "Forecasting stock market crisis events using deep and statistical machine learning techniques". Expert Systems with Applications, 2018, pp. 353-371.

[16] Henrique, B. M., Sobreiro, V. A., & Kimura, H. "Literature review: Machine learning techniques applied to financial market prediction". Expert Systems with Applications, 2019, pp. 226-251.

[17] Hilbert, M. “Big Data for Development: A Review of Promises and Challenges”. Development Policy Review. 2016, PP. 135-174.

[18] Heinav, L. y J. Levin. “The Data Revolution and Economic Analysis, Innovation Policy and the Economy”. Pp. 1-24.

[19] Correa, A. (2021). Prediciendo la llegada de turistas a Colombia a partir de los criterios de Google Trends. Lecturas de Economía, (95), 105-134.

[20] Choi, H. y H. Varian. “Predicting the Present with Google Trends”. Economic Record, 2012, pp 29.

[21] Choi, H. “Predicting Initial Claims for Unemployment Benefits”, SSRN, 2010, https://ssrn.com/abstract=1659307

[22] Stephens-Davidowitz, S. “Everybody Lies: Big Data, New Data, and What the Internet Can Tell Us About Who We Really Are”. HarperCollins Publishers,2017, Nueva York.

[23] Baker, S. y A. Fradkin. “ The Impact of Unemployment Insurance on Job Search: Evidence from Google Search Data”, The Review of Economics and Statistics, 2017, pp. 756-768.

[24] Tuhkuri, J. “Big Data: Do Google Searches Predict Unemployment?”, tesis de maestría, Universidad de Helsinki, https://helda.helsinki.fi/handle/10138 /155258

[25] Goel, S., J. Hofman, S. Lahaie, D. Pennock y D. Watts. “ Predicting consumer behavior with Web search”, Proceedings of the National Academy of Sciences,2017, pp. 17486-17490.

[26] Pacheco-Bonrostro, J., Casado-Yusta, S., & Núñez Letamendía, L. (2007). Algoritmos meméticos para selección de variables en el análisis discriminante. Estadística española, 49(165), 333-347.

[27] Kim, H. H., & Swanson, N. R. (2014). Forecasting financial and macroeconomic variables using data reduction methods: New empirical evidence. Journal of Econometrics, 178, 352-367.

[28] Choi, J.-E., & Shin, D. W. (2019). The roles of differencing and dimension reduction in machine learning forecasting of employment level using the FRED big data. *Communications for Statistical Applications and Methods*, *26*(5), 497-506. https://doi.org/10.29220/CSAM.2019.26.5.497

[29] Katris, C. (2020). Prediction of Unemployment Rates with Time Series and Machine Learning Techniques. *Computational Economics*, *55*(2), 673-706.

[30] Kreiner, A., & Duca, J. V. (2019). Can machine learning on economic data better forecast the unemployment rate? *Applied Economics Letters*, 1-4. https://doi.org/10.1080/13504851.2019.1688237

[31] Sabino Parmezan, A. R., Souza, V. M. A., & Batista, G. E. A. P. A. (2019). Evaluation of statistical and machine learning models for time series prediction: Identifying the state-of-the-art and the best conditions for the use of each model. *Information Sciences*, *484*, 302-337. https://doi.org/10.1016/j.ins.2019.01.076

[32] Stasinakis, C., Sermpinis, G., Theofilatos, K., & Karathanasopoulos, A. (2016). Forecasting US Unemployment with Radial Basis Neural Networks, Kalman Filters and Support Vector Regressions. *Computational Economics*, *47*(4), 569-587. https://doi.org/10.1007/s10614-014-9479-y

[33] Sharma, S. & Singh, S. (2016). Unemployment rates forecasting using supervised neural networks. 2016 6th International Conference - Cloud System and Big Data Engineering (Confluence), 28-33. https://doi.org/10.1109/CONFLUENCE.2016.7508042

[34] Kim, H. H., & Swanson, N. R. (2018). Mining big data using parsimonious factor, machine learning, variable selection and shrinkage methods. *International Journal of Forecasting*, *34*(2), 339-354. https://doi.org/10.1016/j.ijforecast.2016.02.012

[35] Xu, W., Li, Z., Cheng, C., & Zheng, T. (2013). Data mining for unemployment rate prediction using search engine query data. Service Oriented Computing and Applications, 7, 33-42.

[36] Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing*, *14*(3), 199-222. https://doi.org/10.1023/B:STCO.0000035301.49549.88

[37] Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing*, *50*(0), 159-175. http://dx.doi.org/10.1016/S0925-2312(01)00702-0

[38] Varian, H. R. (2014). Big Data: New Tricks for Econometrics. *Journal of Economic Perspectives*, *28*(2), 3-28. https://doi.org/10.1257/jep.28.2.3

[39] Heinze, G., Wallisch, C., & Dunkler, D. (2018). Variable selection–a review and recommendations for the practicing statistician. Biometrical journal, 60(3), 431-449.

[40] Chung, H., Fallick, B. C., Nekarda, C. J., & Ratner, D. (2015). *Assessing the Change in Labor Market Conditions* (Working Papers (Old Series) N.º 1438). Federal Reserve Bank of Cleveland. https://ideas.repec.org/p/fip/fedcwp/1438.html

[41] Hakkio, C. S., & Willis, J. L. (2013). Assessing labor market conditions: The level of activity and the speed of improvement. *Macro Bulletin*, *july18*. https://ideas.repec.org/a/fip/fedkmb/y2013ijuly18.html

[42] Bergmeir, C., Hyndman, R. J., & Koo, B. (2018). A note on the validity of cross-validation for evaluating autoregressive time series prediction. *Computational Statistics & Data Analysis*, *120*, 70-83. https://doi.org/10.1016/j.csda.2017.11.003

[43] Vapnik, V. (1999). The nature of statistical learning theory. Springer science & business

media.

[44] Diebold, F. X., & Mariano, R. S. (1995). Comparing Predictive Accuracy. *Journal of Business & Economic Statistics*, *13*(3), 253-263.

[45] Google.(2022). Tendencias de Google trends. Recuperado de https://trends.google.es/trends/

[46] Vidal, P., Sierra, L. P., Sanabria, J., & Collazos, J. A. (2017). A monthly regional indicator of economic activity: An application for Latin America. Latin American Research Review, 52(4), 589-605.

[47] Sierra, L. P., Vidal, P., & Cerón, J. (2022). Capítulo 15. Una mirada regional al impacto económico del Covid-19 desde el indicador mensual de actividad económica (imae) para el Valle del Cauca. Capítulo 15. Una mirada regional al impacto económico del Covid-19 desde el indicador mensual de actividad económica (imae) para el Valle del Cauca. Pág.: 289-304.

9. ANEXOS

ANEXO 1. CATASTRO DE DATOS PARA CALI, MEDELLÍN Y POPAYÁN.

No.	Variables	Acrónimo	Unidad de medida	Fuente	Ciudad disponible
1	Tasa Global de Participación	TGP	Porcentaje-Tasa	DANE-GEIH	Cali, Medellín y Popayán
2	Tasa de ocupación	TO	Porcentaje-Tasa	DANE-GEIH	Cali, Medellín y Popayán
3	Tasa de desempleo	TD	Porcentaje-Tasa	DANE-GEIH	Cali, Medellín y Popayán
4	Tasa de subempleo	TS	Porcentaje-Tasa	DANE-GEIH	Cali, Medellín y Popayán
5	Fuerza de trabajo	FT	Miles de personas	DANE-GEIH	Cali, Medellín y Popayán
6	Población fuera de la fuerza laboral	INAC	Miles de personas	DANE-GEIH	Cali, Medellín y Popayán
7	Ocupados Agricultura	OCU_AGR	Miles de personas	DANE-GEIH	Cali, Medellín y Popayán
8	Ocupados Industria	OCU_IND	Miles de personas	DANE-GEIH	Cali, Medellín y Popayán
9	Ocupados Construcción	OCU_CON	Miles de personas	DANE-GEIH	Cali, Medellín y Popayán
10	Ocupados Comercio y reparación de vehículos	OCU_COM	Miles de personas	DANE-GEIH	Cali, Medellín y Popayán
11	Ocupados Alojamiento y servicios de comida	OCU_ALOJ	Miles de personas	DANE-GEIH	Cali, Medellín y Popayán
12	Ocupados Transporte y almacenamiento	OCU_TRAN	Miles de personas	DANE-GEIH	Cali, Medellín y Popayán
13	Ocupados Actividades financieras y de seguros	OCU_FIN	Miles de personas	DANE-GEIH	Cali, Medellín y Popayán
14	Ocupados Actividades profesionales	OCU_APRO	Miles de personas	DANE-GEIH	Cali, Medellín y Popayán
15	Ocupados Administración pública, educación y salud	OCU_ADM	Miles de personas	DANE-GEIH	Cali, Medellín y Popayán
16	Ocupados Actividades artísticas, entretenimiento y recreació	OCUP_ART	Miles de personas	DANE-GEIH	Cali, Medellín y Popayán
17	Ocupado como Empleado doméstico	OCU_ED	Miles de personas	DANE-GEIH	Cali, Medellín y Popayán
18	Ocupado como Trabajador por cuenta propia	OCU_TCP	Miles de personas	DANE-GEIH	Cali, Medellín y Popayán
19	Tasa de informalidad	TI	Porcentaje-Tasa	DANE-GEIH	Cali, Medellín y Popayán
20	Tasa Global de Participación jóvenes	TGP_J	Porcentaje-Tasa	DANE-GEIH	Cali, Medellín y Popayán
21	Tasa de ocupación jóvenes	TO_J	Porcentaje-Tasa	DANE-GEIH	Cali, Medellín y Popayán
22	Tasa de desempleo de los jóvenes	TD_J	Porcentaje-Tasa	DANE-GEIH	Cali, Medellín y Popayán
23	Fuerza de trabajo jóvenes	FT_J	Miles de personas	DANE-GEIH	Cali, Medellín y Popayán
24	Fuera de la fuerza laboral jóvenes	INAC_J	Miles de personas	DANE-GEIH	Cali, Medellín y Popayán
25	Tasa Global de Particpación Mujeres	TGP_M	Porcentaje-Tasa	DANE-GEIH	Cali, Medellín y Popayán
26	Tasa de Ocupación Mujeres	TO_M	Porcentaje-Tasa	DANE-GEIH	Cali, Medellín y Popayán
27	Tasa de desempleo Mujeres	TD_M	Porcentaje-Tasa	DANE-GEIH	Cali, Medellín y Popayán
28	Fuerza de trabajo Mujeres	FT_M	Miles de personas	DANE-GEIH	Cali, Medellín y Popayán
29	Índice de empleo industrial	IEMP_IND	Índice	DANE-GEIH	Cali y Medellín
30	Vacantes u ofertas de empleo según anuncios de prensa imp	VAC	No. de vacantes	Banco de la República	Cali y Medellín
31	Búsqueda palabra “empleo” en GT	GT_EMP	Índice de búsquedas	Google Trends	Cali, Medellín y Popayán
32	Búsqueda palabra “ofertas de empleo”en GT	GT_OEMP	Índice de búsquedas	Google Trends	Cali, Medellín y Popayán
33	Búsqueda palabra"vacantes”en GT	GT_VAC	Índice de búsquedas	Google Trends	Cali, Medellín y Popayán
34	Búsqueda palabra “hoja de vida”en GT	GT_HOJAV	Índice de búsquedas	Google Trends	Cali, Medellín y Popayán
35	Búsqueda palabra “agencia de empleo”en GT	GT_AGEMP	Índice de búsquedas	Google Trends	Cali, Medellín y Popayán
36	Búsqueda palabra “trabajo sin experiencia”en GT	GT_TSEXP	Índice de búsquedas	Google Trends	Cali, Medellín y Popayán
37	Búsqueda palabra “sena empleo” en GT	GT_SENA	Índice de búsquedas	Google Trends	Cali, Medellín y Popayán
38	Búsqueda palabra “el empleo” en GT	GT_EEMP	Índice de búsquedas	Google Trends	Cali, Medellín y Popayán
39	Búsqueda palabra “indeed”en GT	GT_INDEED	Índice de búsquedas	Google Trends	Cali, Medellín y Popayán
40	Búsqueda palabra “bolsa de empleo” en GT	GT_BOLEMP	Índice de búsquedas	Google Trends	Cali, Medellín y Popayán
41	Búsqueda palabra “busco trabajo” en GT	GT_BTRA	Índice de búsquedas	Google Trends	Cali, Medellín y Popayán
42	Búsqueda palabra “clasificados”en GT	GT_CLAS	Índice de búsquedas	Google Trends	Cali, Medellín y Popayán
43	Búsqueda palabra “computrabajo” en GT	GT_COMPUT	Índice de búsquedas	Google Trends	Cali, Medellín y Popayán
44	Búsqueda palabra “ofertas de trabajo”en GT	GT_OTRAB	Índice de búsquedas	Google Trends	Cali, Medellín y Popayán
45	Búsqueda palabra “olx empleo”en GT	GT_OLXE	Índice de búsquedas	Google Trends	Cali, Medellín y Popayán
46	Búsqueda palabra “trabajo colombia”en GT	GT_TRABCOL	Índice de búsquedas	Google Trends	Cali, Medellín y Popayán
47	Búsqueda palabra “trabajo”en GT	GT_TRAB	Índice de búsquedas	Google Trends	Cali, Medellín y Popayán
48	Búsqueda palabra “servicio de empleo”en GT	GT_SEREMP	Índice de búsquedas	Google Trends	Cali, Medellín y Popayán
49	Búsqueda palabra “familias en accion”en GT	GT_FAMACC	Índice de búsquedas	Google Trends	Cali, Medellín y Popayán
50	Búsqueda palabra “subsidio desempleo” en GT	GT_SUBDES	Índice de búsquedas	Google Trends	Cali, Medellín y Popayán
51	Búsqueda palabra “prestamos”en GT	GT_PREST	Índice de búsquedas	Google Trends	Cali, Medellín y Popayán
52	Búsqueda palabra “ retiro cesantias”en GT	GT_RCES	Índice de búsquedas	Google Trends	Cali, Medellín y Popayán
53	Búsqueda palabra “cesantias”en GT	GT_CES	Índice de búsquedas	Google Trends	Cali, Medellín y Popayán
Indicador Mensual de Actividad Económica (IMAE)* (Se utiliza para la predicción)		IMAE	Porcentaje-Tasa anual	Universidad Javeriana Cali	Cali y Popayán

ANEXO 2. SELECCIÓN DE VARIABLES CON EL MÉTODO BACKWARD Y LASSO PARA CALI, MEDELLÍN Y POPAYÁN.

No.	Variables	Acrónimo	Método Lasso			Método Backward		
			Cali	Medellín	Popayán	Cali	Medellín	Popayán
1	Tasa Global de Participación	TGP	✗	✓	✓	✓	✓	✓
2	Tasa de ocupación	TO	✓	✓	✓	✓	✓	✓
3	Tasa de desempleo	TD	✓	✓	✓	✓	✓	✓
4	Tasa de subempleo	TS	✓	✗	✓	✓	×	×
5	Fuerza de trabajo	FT	✓	✓	✓	✓	✓	✓
6	Población fuera de la fuerza laboral	INAC	✓	✓	✗	×	×	✓
7	Ocupados Agricultura	OCU_AGR	✗	✗	✓	✓	×	×
8	Ocupados Industria	OCU_IND	✗	✗	✓	✓	✓	✓
9	Ocupados Construcción	OCU_CON	✗	✓	✓	✓	✓	✓
10	Ocupados Comercio y reparación de vehículos	OCU_COM	✗	✗	✓	✓	✓	✓
11	Ocupados Alojamiento y servicios de comida	OCU_ALOJ	✓	✗	✓	✓	✓	✓
12	Ocupados Transporte y almacenamiento	OCU_TRAN	✓	✗	✓	✓	✗	✗
13	Ocupados Actividades financieras y de seguros	OCU_FIN	✗	✗	✓	✓	✗	✓
14	Ocupados Actividades profesionales	OCU_APRO	✓	✗	✓	✓	✗	✓
15	Ocupados Administración pública, educación y salud	OCU_ADM	✗	✓	✓	✓	✓	✓
16	Ocupados Actividades artísticas, entretenimiento y recreación	OCUP_ART	✗	✗	✓	✓	✗	✓
17	Ocupado como Empleado doméstico	OCU_ED	✓	✗	✓	✓	✗	✓
18	Ocupado como Trabajador por cuenta propia	OCU_TCP	✗	✗	✓	✓	✗	✗
19	Tasa de informalidad	TI	✗	✗	✓	✓	✗	✓
20	Tasa Global de Participación jóvenes	TGP_J	✗	✗	✓	✓	✗	✓
21	Tasa de ocupación jóvenes	TO_J	✗	✓	✓	✓	✓	✓
22	Tasa de desempleo de los jóvenes	TD_J	✓	✓	✓	✓	✓	✓
23	Fuerza de trabajo jóvenes	FT_J	✗	✓	✓	✓	✓	✓
24	Fuera de la fuerza laboral jóvenes	INAC_J	✓	✓	✓	✓	✓	✓
25	Tasa Global de Particpación Mujeres	TGP_M	✓	✓	✓	✓	✗	✓
26	Tasa de Ocupación Mujeres	TO_M	✗	✗	✓	✓	✓	✓
27	Tasa de desempleo Mujeres	TD_M	✓	✓	✓	✓	✓	✓
28	Fuerza de trabajo Mujeres	FT_M	✗	✗	✗	✗	✓	✓
29	Índice de empleo industrial	IEMP_IND	✓	✓		✓	✓	
30	Vacantes u ofertas de empleo según anuncios de prensa impre	VAC	✗	✓		✓	✓	
31	Búsqueda en Google de la palabra “empleo”	GT_EMP	✓	✓	✓	✗	✓	✓
32	Búsqueda en Google de la palabra “ofertas de empleo”	GT_OEMP	✗	✗	✗	✗	✗	✗
33	Búsqueda en Google de la palabra"vacantes”	GT_VAC	✗	✗	✓	✗	✗	✓
34	Búsqueda en Google de la palabra “hoja de vida”	GT_HOJAV	✗	✗	✗	✓	✗	✗
35	Búsqueda en Google de la palabra “agencia de empleo”	GT_AGEMP	✗	✗	✗	✗	✗	✓
36	Búsqueda en Google de la palabra “trabajo sin experiencia”	GT_TSEXP	✗	✗	✓	✗	✓	✓
37	Búsqueda en Google de la palabra “sena empleo”	GT_SENA	✓	✗	✗	✓	✗	✗
38	Búsqueda en Google de la palabra “el empleo”	GT_EEMP	✗	✗	✓	✗	✓	✗
39	Búsqueda en Google de la palabra “indeed”	GT_INDEED	✗	✓	✓	✗	✓	✗
40	Búsqueda en Google de la palabra “bolsa de empleo”	GT_BOLEMP	✗	✗	✓	✗	✓	✗
41	Búsqueda en Google de la palabra “busco trabajo”	GT_BTRA	✗	✗	✓	✓	✓	✗
42	Búsqueda en Google de la palabra “clasificados”	GT_CLAS	✓	✗	✓	✗	✓	✓
43	Búsqueda en Google de la palabra “computrabajo”	GT_COMPUT	✗	✗	✗	✗	✓	✗
44	Búsqueda en Google de la palabra “ofertas de trabajo”	GT_OTRAB	✓	✓	✓	✓	✓	✓
45	Búsqueda en Google de la palabra “olx empleo”	GT_OLXE	✓	✗	✓	✓	✗	✗
46	Búsqueda en Google de la palabra “trabajo colombia”	GT_TRABCOL	✗	✓	✓	✓	✗	✗
47	Búsqueda en Google de la palabra “trabajo”	GT_TRAB	✗	✓	✗	✗	✓	✗
48	Búsqueda en Google de la palabra “servicio de empleo”	GT_SEREMP	✗	✗	✗	✓	✗	✓
49	Búsqueda en Google de la palabra “familias en accion”	GT_FAMACC	✗	✓	✓	✗	✗	✓
50	Búsqueda en Google de la palabra “subsidio desempleo”	GT_SUBDES	✗	✗	✓	✓	✗	✓
51	Búsqueda en Google de la palabra “prestamos”	GT_PREST	✗	✗	✓	✓	✓	✓
52	Búsqueda en Google de la palabra “ retiro cesantias”	GT_RCES	✗	✗	✓	✗	✗	✓
53	Búsqueda en Google de la palabra “cesantias”	GT_CES	✗	✗	✗	✓	✗	✗
Total de variables de Google Trends seleccionadas para incluir en el indicador			**5**	**6**	**15**	**10**	**11**	**11**
Total de variables seleccionadas para incluir en el indicador			**19**	**21**	**41**	**38**	**28**	**35**