Nota de Aceptación

Aprobado por el Comité de Trabajo de Grado en cumplimiento de los requisitos exigidos por la Pontificia Universidad Javeriana para optar el título de Ingeniero de Sistemas y Computación.

Dr. HERNAN CAMILO ROCHA
Decano de la Facultad de Ingeniería

ING. GERARDO MAURICIO SARRIA
Director Carrera Ingeniería Sistemas y Computación.

ING. GLORIA INES ALVAREZ VARGAS
Director(a) Trabajo

ING. GERARDO MAURICIO SARRIA
Jurado 1

ING. ANDRES RODOLFO NAVARRO
Jurado 2

Pontificia Universidad JAVERIANA Cali

Facultad de Ingeniería y Ciencias

Acta de Correcciones al Proyecto de Grado
Ingeniería de Sistemas y Computación

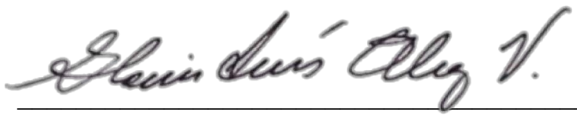
Fecha: 21 de febrero del 2023

Autores: Dicson Ferney Quimbayo Diaz

Nombre del Proyecto de Grado: Agente inteligente para la elección de carrera de estudiantes de bachiller

Director: Gloria Inés Álvarez Vargas

Como indica el artículo 2.27 de las Directrices de Trabajo de Grado, he verificado que los estudiantes indicados arriba han implementado todas las correcciones que los Jurados del Proyecto de Grado definieron que se efectuaran, como consta en el Acta de Calificación correspondiente.



Firma de Director(a) del Proyecto de Grado

PONTIFICIA UNIVERSIDAD JAVERIANA CALI
FACULTAD DE INGENIERÍA

AGENTE INTELIGENTE PARA PREDECIR LA ELECCIÓN DE CARRERA DE ESTUDIANTES DE BACHILLER

DICSON FERNEY QUIMBAYO DIAZ

FECHA: Febrero, 2023

Directora:

Dr. Gloria Inés Álvarez Vargas

Santiago de Cali, 21 de febrero de 2023

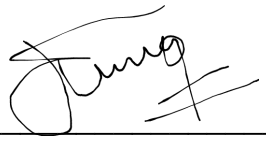
Señores
Pontificia Universidad Javeriana Cali.
Dr. Gerardo Sarria
Director Carrera de Ingeniería de Sistemas y Computación

Cordial saludo,

Me permito presentar a su consideración el proyecto de grado titulado "Agente Inteligente Para Predecir La Elección De Carrera De Estudiantes De Bachiller" con el fin de cumplir con los requisitos exigidos por la Universidad para optar al título de Ingeniero de Sistemas y Computación.

Al firmar aquí, doy fe de que el proyecto de grado se encuentra terminado, entendiendo y conociendo las directrices para la presentación de trabajos de grado de la Facultad de Ingeniería aprobadas el 26 de noviembre de 2009, donde se establecen los plazos, normas para el desarrollo del anteproyecto y trabajo de grado

Atentamente,



Dicson Ferney Quimbayo Diaz
Código: 8924459

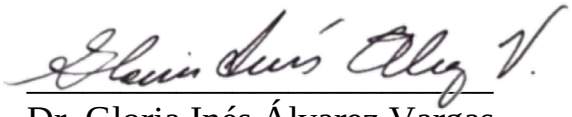
Santiago de Cali, 21 de febrero de 2023

Señores
Pontificia Universidad Javeriana Cali.
Dr. Gerardo Sarria
Director Carrera de Ingeniería de Sistemas y Computación

Cordial saludo,

Por medio de la presente me permito informarle que el proyecto de grado a cargo de mi dirección titulado “Agente Inteligente Para Predecir La Elección De Carrera De Estudiantes De Bachiller”, realizado por el estudiante Dicson Ferney Quimbayo Diaz (cód.: 8924459), se encuentra terminado y listo para sustentación

Atentamente,



Dr. Gloria Inés Álvarez Vargas

1 Descripción del Problema 11

1.1 Planteamiento del problema 11
1.1.1 Formulación 13
1.1.2 Objetivos 13
1.1.3 Objetivo General 13

1.1.1. Objetivos Específicos 13

1.2 Justificación 14

1.3 Delimitaciones y Alcances 14

2 Marco teórico 15

2.1 Orientación Vocacional 15

2.2 Aprendizaje Automático 16

2.3 Técnicas de aprendizaje automático 16
2.3.1 k-vecinos 17
2.3.2 Árboles de decisiones 17
2.3.3 Máquina de vectores de soporte 18
2.3.4 Naive bayes 19
2.3.5 Red neuronal 20
2.3.6 Random Forest 21
2.3.7 Regresión lineal 22

2.4 Trabajos Relacionados 22

3 Preparación de los datos 24

3.1 Recopilación de datos 24

3.2 Análisis exploratorio 26

3.3 Creación de clases 27

3.4 Limpieza de datos 29

3.5 Codificación de variables 30

4 Construcción del agente inteligente 33

4.1 Elección de técnicas 33
4.1.1 Criterios de selección 33
4.1.2 Técnicas seleccionadas 34

4.2 Estimación de parámetros 35

4.3 Entrenamiento de los modelos 36

5 Análisis de resultados 39

5.1 Comparación de los modelos 39

5.2 Consideraciones sobre la utilidad de los modelos desarrollados 41

5.3 Evaluación de la hipótesis de partida 42

6 Desarrollo del agente 43

7 Conclusiones 46

8 Trabajo futuro 47

9 Glosario 48

10 Bibliografía 49

Lista de figuras

Figura 1. Deserción universitaria de Latinoamérica en el 2016 11
Figura 2. Algoritmo K-vecinos…. 17
Figura 3. Árbol de decisión. 18
Figura 4. Máquina de vectores de soporte. 19
Figura 5. Naive Bayes. 20
Figura 6. Estructura de una red neuronal artificial 20
Figura 7. Random forest. 21
Figura 8. Estadísticas descriptivas de las variables numéricas 26
Figura 9. Diagrama de cajas y bigotes de los puntajes saber 11 27
Figura 10. Clases por núcleo de referencia 28
Figura 11. Clases por grupo de referencia 28
Figura 12. División del conjunto de datos en los data set's PRE Y TEC 36
Figura 13. Métrica Accuracy de las distintas técnicas sobre el dataset Pre 40
Figura 14. Métrica Accuracy de las distintas técnicas sobre el dataset TEC 40
Figura 15. Preguntas por el valor de cada área de evaluación de la prueba saber 11 43
Figura 16. Ingreso de información socio demográfico del estudiante. 44
Figura 17. Información de la ocupación y nivel académico de los padres. 44
Figura 18. Registro de las preferencias del estudiante. 45

Lista de tablas

Tabla 1. Diccionario de variables del conjunto de datos de las pruebas Saber 11. 25
Tabla 2 Carreras incluyendo las de carácter tecnólogo y técnicos 29
Tabla 3. Carrera solamente de Pregrado de carácter universitario 29
Tabla 4. Descripción de posibles valores para cada una de las variables del data set. 30
Tabla 5. Elección de técnicas de machine learning 33
Tabla 6. Selección de técnicas. 34
Tabla 7. Resultados de parámetros al seleccionar de las posibilidades con Grid Search 35
Tabla 8. Diseño de experimentos 37
Tabla 9. Resultados de Accuracy de los experimentos del data set TEC 39
Tabla 10. Resultados de la metrica Accuricy con el data set PRE 39

Resumen

El proceso orientación vocacional consiste en la recolección de información del estudiante relacionado con sus actitudes, talentos, personalidad e intereses profesionales, esta información se analiza de forma cuantitativa por medio pruebas vocacionales; además de entrevistas que ayudan a los profesionales en orientación vocacional a darle una guía al estudiante de las posibilidades que tiene en la elección de carrera universitaria.

En un país como Colombia donde solo el 39 % de los graduados de bachiller accede a educación superior, es vital la orientación vocacional. No obstante, brindar orientación vocacional a todos los bachilleres es una tarea qué requiere grandes recursos entre ellos: tiempo y capital humano que permita tener un gran alcance en las instituciones educativas de todo país. De aquí la importancia de poder aportar una forma de realizar una predicción automática de la información recolectada por el orientador vocacional desde una condición institucional a nivel nación para obtener el título de bachiller, como lo es la presentación de las pruebas saber 11.

En este proyecto de grado se llevó a cabo el diseño, implementación y evaluación de un agente inteligente para predecir las posibles carreras a estudiar de un estudiante de bachillerato con la evaluación de distintas técnicas de inteligencia artificial.

Al obtener los distintos modelos se realizó una evaluación del desempeño de los agentes. A partir de ello se analizaron los resultados y se obtuvo la técnica con mejor predicción de las carreras universitarias. Finalmente, se obtuvo una Accuracy de 0.62 utilizando la técnica k vecinos más cercanos sobre un dato set que corresponde solamente a las áreas de formación universitaria sin considerar los programas de nivel técnico o tecnológico.

Abstract

The vocational orientation process consists of gathering information about the student's attitudes, talents, personality, and professional interests. This information is analyzed quantitatively by means of vocational tests; in addition to interviews that help professionals in vocational orientation to give the student a guide to the possibilities he/she has in the choice of a university career.

In a country like Colombia, where only 39 % of high school graduates have access to higher education, vocational guidance is vital. However, providing vocational guidance to all high school graduates is a task that requires great resources including: time and human capital that allows to have a great scope in educational institutions throughout the country. Hence the importance of being able to provide a way to make an automatic prediction of the information collected by the vocational counselor from an institutional condition at the national level to obtain the title of baccalaureate as is the presentation of the saber 11 tests.

In this degree project was carried out the design, implementation, and evaluation of an intelligent agent to predict the possible careers to study of a high school student with the evaluation of different artificial intelligence techniques.

From the different models, an evaluation of the performance of the agents was carried out to analyze the results and obtain the technique that was best able to predict the university careers. Finally, an accuracy of 0.62 was obtained using the k nearest neighbors' technique on a data set that corresponds only to the areas of university education without considering technical or technological programs.

Introducción

El ingreso a la educación superior supone un cambio importante en la vida de los individuos, de allí que la elección de carrera permite una toma de decisión orientada y pensada en el desarrollo futuro. Sin embargo, la media de edad de ingreso a la universidad es de 17 años una edad en la que aún los estudiantes presentan grandes retos en la construcción de identidad[1]. Así, la orientación vocacional juega un papel crucial en este panorama porque su función es apoyar a los jóvenes a elegir su profesión a través de un proceso de guía y acompañamiento.

Hoy muchos orientadores vocacionales usan extensas pruebas psicotécnicas, entrevistas y sesiones de guía uno a uno que implican menor capacidad de atención y respuesta. Dado lo anterior, en la actualidad los psicólogos y orientadores vocacionales trabajan de la mano con la generación de soluciones tecnológicas tal es el caso de chatbots, y pruebas de auto reporte en línea. En este sentido cada vez es más necesaria la búsqueda de alternativas que permitan automatizar esta labor como, por ejemplo, la implementación de la inteligencia artificial para facilitar la toma de decisión en la orientación vocacional.

En el desarrollo de este proyecto se diseñará un agente inteligente para predecir la elección de carrera universitaria de estudiantes de bachillerato, utilizando técnicas de inteligencia artificial. Para ello se tiene en cuenta los datos reportados por el Instituto Colombiano para la Evaluación de la Educación (ICFES) en 2 momentos de la vida académica del estudiante durante la finalización de su bachillerato y al culminar su vida universitaria. Lo anterior, con el propósito de conocer las características sociodemográficas, pero además tener en cuenta el puntaje obtenido en la prueba saber 11.

En este sentido, la importancia del puntaje en las diferentes áreas del conocimiento se ampara en que múltiples partes interesadas (universidad, psicólogos, investigadores) están de acuerdo en que los estudiantes necesitan conocimientos y habilidades generales sólidos en áreas de conocimiento de su interés. Lo cual se sume a su competencia vocacional para que durante su vida universitaria puedan manejar las demandas personales y sociales[2], [3].

Capítulo 1

1 Descripción del Problema

1.1 Planteamiento del problema

La elección de carrera es una de las tareas más importantes en los proyectos de vida de los estudiantes graduados de la educación secundaria; esta decisión en algunas ocasiones se toma en una etapa en la cual no se cuenta con niveles de conocimiento u orientación suficientes [4]. En los adolescentes la toma de decisión sobre la carrera universitaria puede ser compleja, puesto que, tiene un impacto en la construcción de identidad de la persona, por la cual se asume la elección de un rol en la sociedad. Esta elección suele ser influenciada por los padres, maestros y personas cercanas, entre otras experiencias de manera positiva o negativa [5], [6].

Ahora bien, la escogencia incorrecta de profesión es uno de los múltiples factores para tener en cuenta en la deserción universitaria, uno de los problemas de la educación superior; de ahí que, el sistema educativo en Latinoamérica tenga altos índices de deserción académica[1] como lo muestra el estudio realizado por el Banco Mundial (2016), cabe resaltar que el análisis posiciona a Colombia como el segundo país con la tasa más alta de deserción universitaria con un 42 % (ver figura 1). El problema en la deserción universitaria influye de manera negativa no solamente la vida de los jóvenes desertores, sino también con el propósito de las instituciones educativas y el desarrollo de un país [7] de formar a los profesionales del futuro.

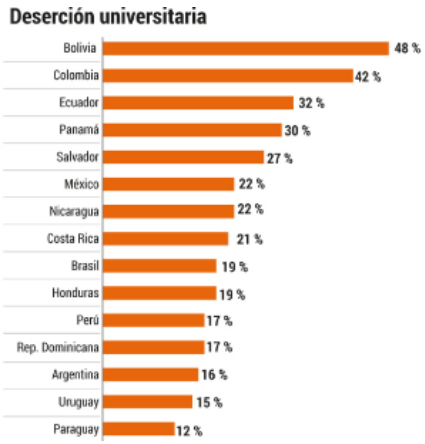


Figura 1. Deserción universitaria de Latinoamérica en el 2016. Tomado de: Informe sobre deserción del Banco Mundial

En Colombia, según el estudio de Urrego (2019) entre el 2006-2016 los principales componentes que ocasionaron la deserción se asociaron a factores: individuales, académicos, socioeconómicos e institucionales[8]. Adicionalmente, otro estudio realizado por la

[1]La deserción académica se entiende como la decisión que toma un estudiante cuando para sus estudios universitarios por un tiempo y por otro lado la finalización de su proyecto educativo se ve comprometido[8].

universidad San Buenaventura presenta dentro de las razones para interrumpir la carrera universitaria: la inseguridad de enfrentar la carrera; falta de motivación e inconformidad de la elección, es decir, la elección errónea de profesión resulta ser uno de los factores que aporten a la deserción universitaria, es importante observar de qué forma el proceso de orientación vocacional aporta a minimizar una escogencia de carrera equivoca.

De esta manera, la orientación vocacional se describe como el proceso que ayuda a determinar las capacidades del individuo y elegir una carrera acorde a la personalidad que tiene cada estudiante[6]. En Colombia se han implementado programas de orientación vocacional como "Buscando Carrera" o "Colombia Aprende" con el propósito de combatir la deserción universitaria en el país, sin embargo, es una estrategia que tiene un impacto reducido debido a que existe una brecha de conectividad a internet según informes del DANE [9], ocasionando dificultad para que los estudiantes pueden acceder a internet para hacer uso de los programas vocacionales brindados por el estado, en consecuencia, genera los resultados negativos de Colombia frente a la deserción mostrada en el estudio del Banco Mundial. Asimismo, en las zonas rurales la orientación vocacional es uno de los temas desafiantes en nuestro país. Dado lo anterior, es necesario contar con alternativas viables y replicables para apoyar a la labor de las instituciones educativas en orientar a los estudiantes, puesto que, la ayuda insuficiente para ayudar a los jóvenes a tomar decisiones coherentes con su personalidad, aptitudes, habilidades y valores[10].

El presente trabajo de grado se centra en el factor de elección de carrera individual, específicamente en la orientación vocacional que se brinda al estudiante para realizar una decisión acorde a sus aptitudes. La orientación tiene una relación directa con la deserción académica porque el estudiante al no contar con una guía oportuna se inclina a elegir una carrera desacorde a sus aptitudes, habilidades y conocimientos en diferentes áreas; la elección errónea es el efecto de una poca o nula orientación vocacional durante el paso de los bachilleres por la educación media[11].

Una forma de aportar a la elección correcta de carrera es mejorar los programas de orientación vocacional en el país por medio de un proceso que brinde las herramientas para hacer la inmersión de los jóvenes en la vida universitaria, en su transcurso y al finalizar la carrera profesional [8]. La Organización de las naciones Unidas (ONU) en el año 2017 realizo la primera "Cumbre de la inteligencia artificial" y propuso a la Inteligencia Artificial (IA) como una herramienta con el potencial de ayudar a las personas a resolver sus problemas, entre estos, la construcción de una educación con calidad [12]; es decir una reforma personalizada a los procesos de enseñanza y acompañamiento de los estudiantes.

El acompañamiento de orientación vocacional brindado por las instituciones educativas ha sido abordado desde la IA a través de sistemas vocacionales como lo son los chatbots y sistemas expertos. El propósito de integrar la IA es permitir al orientador vocacional, mediante un proceso rápido y eficiente, obtener información objetiva, estructurada y completa para así centrarse en la guía del estudiante y su toma de decisiones [13].

Partiendo de lo anterior, este trabajo de grado tiene como objetivo general: desarrollar un agente inteligente para brindar apoyo en la orientación vocacional de un estudiante por medio

de predicciones desde el cruce de pruebas de estado (saber 11, saber pro). En la construcción de un agente inteligente para la educación es necesario contar con la comprensión del entorno en el cual se aplican técnicas de la IA como lo son: el aprendizaje automático, la computación cognitiva, algoritmos genéticos, redes bayesianas, entre otras [14].

En esta medida, el agente inteligente deduce las posibles opciones de carrera universitaria a través del aprendizaje automático, y el uso de técnicas que reconocen y clasifican patrones en diferentes situaciones [15]. El reconocimiento de patrones se basará en el análisis de las pruebas del Instituto Colombiano para la Evaluación de la Educación (ICFES) que se realizan en el grado once y al finalizar la carrera universitaria, mediante el cruce de estas dos pruebas de estado, analizar el impacto de los resultados en las pruebas saber 11 en la elección de la carrera de los estudiantes.

1.1.1 Formulación

¿Cómo hacer la predicción de elección de carrera de un estudiante de bachillerato utilizando diferentes técnicas de inteligencia artificial?

1.1.2 Objetivos

1.1.3 Objetivo General

Construir un agente inteligente para predecir la elección de carrera universitaria de estudiantes de bachillerato, utilizando técnicas de inteligencia artificial.

1.1.1. Objetivos Específicos

- ✓ Diseñar un agente inteligente para la predicción de elección de carrera universitaria basada en pruebas de estado.
- ✓ Seleccionar técnicas de aprendizaje automático a emplear para predecir la elección de carrera.
- ✓ Implementar el agente inteligente para a predicción de elección de carrera universitaria.
- ✓ Evaluar el desempeño del agente inteligente para la predicción de carrera universitaria.

1.2 Justificación

Realizar la predicción de elección de carrera, desde el ámbito profesional se centra en la labor de un orientador vocacional, en esta labor se tienen en cuenta gustos, aptitudes e interés del estudiante. En consecuencia, un orientador vocacional recoge información cualitativa y cuantitativa con el fin sugerir carreras acordes con los elementos evaluados en el estudiante [11]. El orientador obtiene los datos para brindar una asesoría de extensas pruebas vocacionales, en consecuencia, surge la necesidad de contar con información estructurada y sintetizada para asesorar al estudiante en la toma de decisión de carrera universitaria[13].

La tecnología basada en inteligencia artificial tiene el propósito de proporcionar herramientas educativas para facilitar la obtención de conocimientos, así como la orientación de soluciones alternativas para procesos rutinarios y extensos [15]. Por lo anterior, la implementación de herramientas tecnológicas tiene un impacto relevante en las disciplinas que interactúan con la educación, como lo es la orientación vocacional.

El desarrollo del agente basado en la aptitud es el primer paso para apoyar los procesos de orientación vocacional inteligente para la elección de carrera según las aptitudes de estudiante evaluada en las pruebas saber 11. En esta medida, la importancia de la gestión y la construcción de recursos de inteligencia artificial permitirá acelerar la reforma de formación del personal y diversos métodos, entre ellos el proceso de elección de carrera que apoye y transforme al sistema educativo [12].

1.3 Delimitaciones y Alcances

- En el proceso de diseño del agente inteligente, se elegirá las técnicas de inteligencia artificial más acertadas para lograr hacer la predicción de la carrera universitaria mediante criterios de selección descritos en el capítulo 4.
- Para la etapa de implementación del agente inteligente se obtendrá la información necesaria para realizar el análisis y procesamiento de las bases de datos del gobierno colombiano ubicado en el sitio datos.gov.co.
- En la evaluación del desempeño del agente inteligente se utilizará la matriz de confusión y las métricas de rendimiento del aprendizaje automático.

Capítulo 2

2 Marco teórico

En el desarrollo de este proyecto de grado se utilizan técnicas de aprendizaje automático con técnicas de machine learning para abordar la solución de un agente inteligente que apoye el proceso vocacional de estudiantes de bachillerato y sea una herramienta que el profesional en orientación vocacional use para generar opciones de elección de carrera acorde a las aptitudes del estudiante. A continuación, se aborda las temáticas relacionadas con la orientación vocacional, aprendizaje automático y agentes inteligentes.

2.1 Orientación Vocacional

La orientación vocacional es un estudio que da apoyo y guía a los jóvenes en la toma de decisiones para la elección de carrera profesional y proporciona posibilidades basadas en las habilidades del estudiante para así escoger la carrera que se ajuste a su personalidad [6]. En este sentido, la orientación profesional se considera un proceso transversal en la vida del estudiante, debido a que surge desde temprana edad hasta luego de finalizar sus estudios académicos, puesto que es necesario otorgar desde el sistema educativo la guía de un profesional para que la calidad de vida se vea potencializada por las decisiones tomadas desde las estrategias que brinde la información de las opciones de desarrollo profesional[16].

El proceso de orientación vocacional desde la psicología se hace por medio de test o pruebas vocacionales que evalúan los aspectos de aptitudes, talentos, personalidad e intereses profesionales del estudiante para brindar las opciones oportunas de carrera que se adapten al contexto actual. Los aspectos evaluados profundizan en las siguientes características:

- Aptitudes: Habilidades que el estudiante posee y así permite vincularlo a áreas profesionales específicas.
- Talentos: habilidades del estudiante, se ha comprobado que quienes trabajan con sus habilidades son personas con un alto nivel de motivación y compromiso.
- Personalidad: mide el ajuste socioemocional que tiene el estudiante.
- Intereses profesionales: son test relacionados con la orientación vocacional y busca predecir la satisfacción del estudiante en una determinada carrera profesional.

2.2 Aprendizaje Automático

El impacto de herramientas basadas en la inteligencia artificial en la orientación vocacional permite a los profesionales en orientación vocacional contar con información de manera más sencilla y eficaz para poder dar las sugerencias de posibles elecciones de carrera profesional a los estudiantes que acompañan en su proceso vocacional. La IA brinda una manera eficiente de generar predicciones, específicamente el aprendizaje automático implica el uso de técnicas de clasificación de patrones con base a un conjunto de datos bajo distintos contextos [15]; adicionalmente el aprendizaje automático tiene las características de ser multidimensional, de procesamiento rápido, de fácil acceso entre otras [12].

El aprendizaje automático es la capacidad que tiene una computadora de aprender de datos recogidos del ambiente con la intención de mejorarse a sí misma por medio de experiencias previas y estímulos del ambiente donde se encuentre [17]. Para aprender o mejorar sus capacidades de deducción existen diferentes métodos de aprendizaje automático, entre los más comunes están aprendizaje supervisado, aprendizaje no supervisado y aprendizaje forzado.

Ahora se da una breve de cada método de aprendizaje automático:

- **Aprendizaje supervisado:** El agente inteligente aprende por medio parejas de datos (entrada y salida), por los cuales se hace una comparación de las salidas del modelo respecto a las salidas esperadas para realizar los arreglos necesarios para mejorar la predicción.

- **Aprendizaje no supervisado:** este método no requiere datos previos (conocimiento), por medio de un algoritmo genera la estructura de los datos, requiere solo de datos de entrada para dar respuesta.

- **Aprendizaje por refuerzo:** el aprendizaje se hace por la capacidad de medir acciones como recompensas o castigos, por los cuales se genera una política de toma de decisiones que maximice las recompensas futuras y minimice los castigos.

2.3 Técnicas de aprendizaje automático

A partir de revisión de la literatura, se investigaron técnicas de aprendizaje automático que se aplican para la predicción de programas de estudio en educación superior, a continuación, se presentan las más usadas:

2.3.1 k-vecinos

K-vecinos es una técnica de análisis no paramétrica, supervisada de conglomerados y es comúnmente usada en estudios de segmentación. A diferencia de los métodos paramétricos, K-vecinos no asume una distribución subyacente de los datos. Además, con K-vecinos, no existe una fase de entrenamiento explícita. Los cálculos se basan en el conjunto de datos completo en contraste con otras técnicas de aprendizaje automático[18]; en la figura 2 se puede observar el paso a paso del algoritmo para la clasificación de las clases.

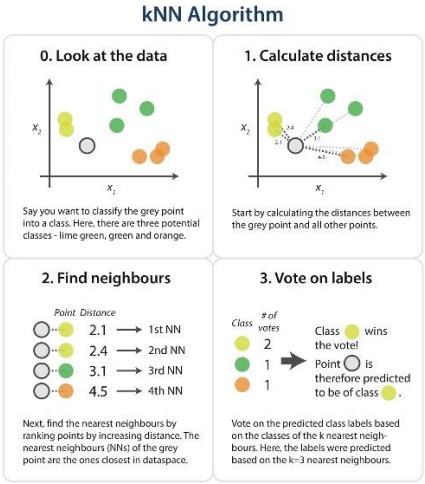


Figura 2. Algoritmo K-vecinos. Tomado de: *https://www.kdnuggets.com/2016/01/implementing-your-own-knn-using-python.html*

Ventajas: Una de las principales ventajas de K-vecinos es que no necesita la especificación de ningún modelo predictivo; al ser no paramétricos toma relevancia para problemas del mundo real donde los datos no suelen seguir los supuestos teóricos [19]. Además, tiene una implementación sencilla y suele ofrecer un buen rendimiento en comparación con otros métodos.

Limitaciones: En el peor de los casos, podría utilizar todos los puntos de datos para tomar una decisión, o tal vez se necesite un gran bloque de memoria para almacenar todos los datos de entrenamiento[18].

2.3.2 Árboles de decisiones

En este algoritmo de carácter aprendizaje supervisado se crea una estructura similar a un árbol basado en el peso de la entropía. Cuanto mayor sea la profundidad del árbol, mayor será la precisión del resultado. La poda del árbol de decisión se refiere a podar o recortar las ramas de un árbol para evitar el sobre ajuste[20].

Un árbol de decisión se representa como una función que toma como entrada una secuencia de valores de atributo sobre un elemento y devuelve una "decisión" a partir de ella, los valores de entrada y salida pueden ser discretos o continuos; para ver un ejemplo del proceso de decisión ver figura 3. Los modelos de árbol en los que la variable de destino puede tomar un conjunto discreto de valores se denominan árboles de clasificación; en estas estructuras de árbol, las hojas representan etiquetas de clase y las ramas representan combinaciones de características que conducen a esas etiquetas de clase. Los árboles de decisión donde la variable objetivo puede tomar valores continuos (generalmente números reales) se denominan árboles de regresión[21].

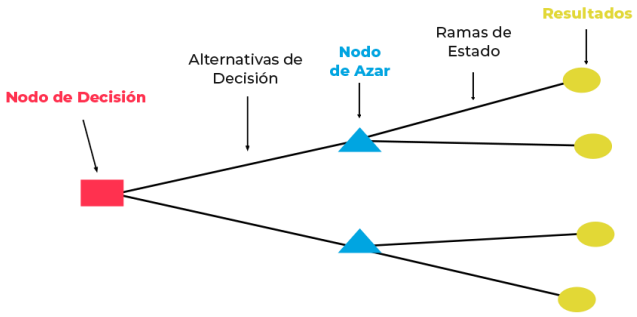


Figura 3. Árbol de decisión. Tomado de: https://www.masterdatascienceucm.com/que-es-machine-learning/

Ventajas: los modelos de árbol de decisión pueden ser fáciles de entender e interpretar y requieren un bajo preprocesamiento de datos. Además, poseen métodos de validación basados en test estadísticos. Tienen buen desempeño con grandes conjuntos de datos. Soportan datos numéricos y categóricos.

Limitaciones: un árbol de decisión puede ser poco robusto; además, un cambio ligero en los datos de entrenamiento puede afectar el modelo y por ende sus predicciones. Sumado a lo anterior, cuando los modelos de árboles de decisiones son entrenados a con extensos conjuntos de datos, se necesitan técnicas de poda.

2.3.3 Máquina de vectores de soporte

Máquina de vectores de soporte es un tipo de algoritmo de aprendizaje supervisado que se utiliza principalmente para problemas de clasificación y regresión. Funciona en el procedimiento en el que cada elemento de datos se traza en la forma de la dimensión *n*, donde *n* es el número de características y la coordenada muestra el valor de una coordenada particular. El siguiente paso es colocarlo en el hiperplano que separa estas dos clases como se representa en la figura 4. Este algoritmo se implementa principalmente mediante kernels.

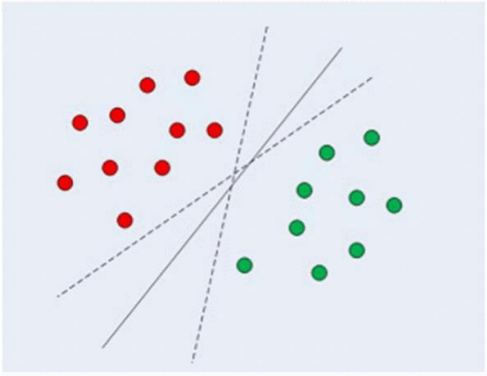


Figura 4. Máquina de vectores de soporte. Tomado de 10.1109/ICSSS49621.2020.9202024

Un algoritmo tipo Máquina de vectores de soporte utiliza líneas discontinuas y sólidas para clasificar los datos. Entonces, de esta manera, SVM es diferente de otros algoritmos de clasificación, puesto que utiliza un límite de decisión que maximiza la distancia desde los puntos de datos cercanos de todas las clases. Por lo tanto, encuentra el límite de decisión óptimo [22].

Ventajas: Una máquina de vectores de soporte simple que utiliza un algoritmo de aprendizaje automático está tratando de encontrar un límite que separe los datos de tal manera que se puedan minimizar los errores de clasificación.

Limitaciones: sus principales desventajas se asocian con la memoria y la complejidad de cálculo. Se han desarrollado muchas técnicas para superar estas limitaciones, que se clasifican en algoritmos basados en la descomposición y en algoritmos basados en la variante. Sumado a lo anterior, las máquinas de vectores fueron desarrolladas para clasificar de forma binaria. Sin embargo, en su mayoría, los problemas de clasificación incluyen más de dos clases[23].

2.3.4 Naive bayes

Este algoritmo supervisado se basa en una red bayesiana que es un grafo dirigido acíclico que indica la distribución de probabilidad de forma comprimida. Un nodo en este grafo muestra una variable aleatoria, X^i. Un borde dirigido entre dos nodos indica una posible interdependencia entre una variable mostrada por el nodo principal y otra variable mostrada por un nodo secundario. La estructura de esta red asume que un nodo X^i es condicionalmente independiente de otros nodos vectoriales y no padres[24]. En la figura 5 se determina la fórmula de cálculo de probabilidades para cada clase.

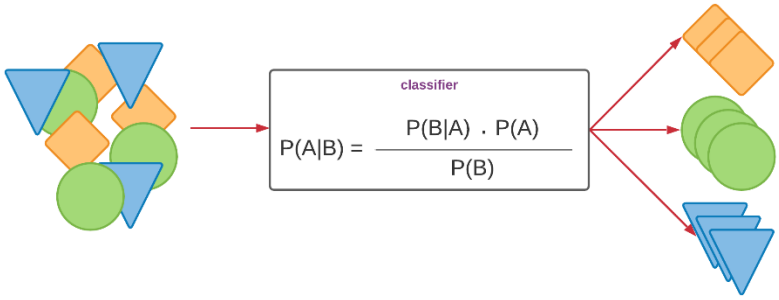


Figura 5. Naive Bayes. Tomado de https://hands-on.cloud/implementing-naive-bayes-classification-using-python/

Ventajas: Estos clasificadores son relativamente fáciles de entender y construir. Son fáciles de entrenar y no requieren grandes conjuntos de datos para producir resultados efectivos. [25].

Limitaciones: su limitación está relacionada con la suposición de la independencia de la característica, que no es válida para la mayoría de las situaciones de la vida real.

2.3.5 Red neuronal

Una red neuronal imita un cerebro humano con neuronas interconectadas en capas; así, las redes neuronales simulan la actividad eléctrica del cerebro y el sistema nervioso. Los elementos de procesamiento (también conocidos como neurodo o perceptrón) están conectados a otros elementos de procesamiento. La red normalmente incluye un número de entradas y salidas, varias capas ocultas y una capa de salida[26], en la figura 6 se establece la estructura base de una red neuronal.

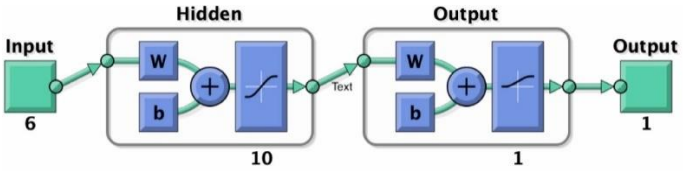


Figura 6. Estructura de una red neuronal artificial. Tomado de 10.1016/J.EGYAI.2022.100163

Las neuronas (o neurodos) generalmente se organizan en capas o vectores, y la salida de una capa se usa como entrada para la siguiente capa, y posiblemente también para otras capas. Una neurona puede conectarse a todas o algunas de las neuronas en capas posteriores, y estas conexiones imitan las conexiones sinápticas del cerebro [27].

Ventajas: es a través del ajuste de las fuerzas o pesos de conexión que se emula el aprendizaje en las redes neuronales [27]. Además, Las redes neuronales ofrecen muchas ventajas, incluida la necesidad de un entrenamiento estadístico menos formal, la capacidad de detectar implícitamente relaciones no lineales complejas entre variables dependientes e independientes, la capacidad de detectar todas las posibles interacciones entre predictores y una variedad de algoritmos de entrenamiento[28].

Limitaciones: las redes neuronales tienen naturaleza de "caja negra", mayor carga computacional, propensión al sobre ajuste y la naturaleza empírica del desarrollo del modelo, esto se convierte en una desventaja a la hora del aprendizaje automático.

2.3.6 Random Forest

El Random forest es un algoritmo de clasificación supervisado. En este algoritmo se generan una serie de árboles de decisión para formar el bosque como se observa en la figura 7. Por lo tanto, también implementa el algoritmo del árbol de decisión. Cuanto mayor sea el número de árboles en el bosque, más precisos serán los resultados. El algoritmo es un proceso de dos etapas. En la primera etapa, primero se crea un bosque aleatorio. En la segunda etapa, se realizan predicciones basadas en los clasificadores creados en la primera etapa.[29].

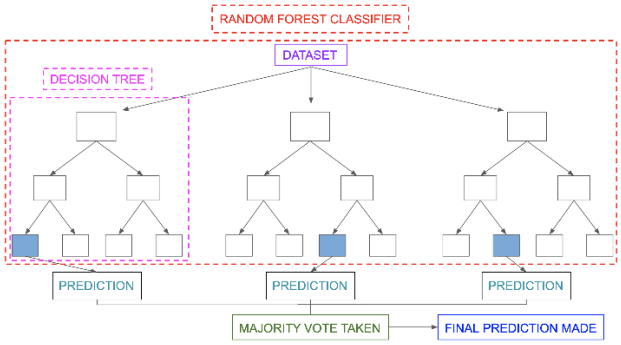


Figura 7. Random forest. Tomado de https://www.section.io/engineering-education/introduction-to-random-forest-in-machine-learning/

Ventajas: el bosque aleatorio se usa más comúnmente, ya que ayuda a superar los problemas causados por el sobre ajuste. También ayuda a identificar los atributos y características más importantes en el conjunto de datos[29].

Limitaciones: La falta de interpretabilidad limita su uso en algunos campos específicos como la salud y la economía, puesto que no se pueden determinar relaciones de clasificación y podría comportarse como una caja negra. En suma, la literatura señala que un árbol de decisión no es competitivo en términos de precisión con respecto a otros enfoques de regresión y clasificación. Sin embargo, al agregar muchos árboles de decisión, se puede mejorar sustancialmente el rendimiento predictivo[30].

2.3.7 Regresión lineal

Este es un método estadístico para analizar un conjunto de datos en los casos en que más de una variable independiente puede determinar el resultado. La función logística [21] se inventó para comparar la tasa de crecimiento de una cantidad con la población y encontrar el nivel de saturación superior de la tasa de crecimiento. Hoy en día, se utiliza para ajustar parámetros a una función ajustando sus pesos para que se ajusten al modelo. [29]

2.4 Trabajos Relacionados

- ***Sistema Experto vocacional:*** el propósito de este trabajo fue contar con una herramienta que le brindara una guía al estudiante para definir sus verdaderos intereses, gustos y habilidades para poder recolectar la información necesaria para la sugerencia de una carrera profesional y ser de apoyo en el rol del orientador vocacional de seguimiento [13]. La sugerencia de carrera profesional se hizo por medio de un sistema experto para ejecutar la aplicación de conocimientos específicos y procedimientos de inferencia.

- ***Robot virtual en orientación vocacional:*** robot que simula a un humano experto en orientación vocacional, por medio de una conversación con el estudiante obtiene información de sus capacidades e interés profesionales con el propósito de generarle un informe al orientador vocacional y así poder brindarle apoyo en la elección de carrera al estudiante [5]; el proceso se realizó por medio de un agente inteligente con el cual tenía un procesamiento de lenguaje natural. Las limitaciones de este trabajo son que la información recogida no da una sugerencia de carrera, sino que proporciona información al orientador; el profesional debe realizar el proceso, analizar la información recogida por el chat y hacer el proceso de guía de elección de carrera para el estudiante.

- ***An intelligent career guidance system using machine learning:*** Para creación del modelo de predicción se entrevistaron estudiantes de grados 11-12 y estudiantes universitarios inscritos en el departamento de ingeniería, más allá de las notas el conjunto de predicciones estuvo dado por las habilidades. En este sentido, el modelo analizó el conjunto de habilidades de un candidato en contraste con las requeridas por un departamento específico, de esta manera según la habilidad principal y un papeo de habilidades secundarias del estudiante se recomendó un departamento y opciones secundarias o terciarias, al estar basado en habilidades permite al estudiante tener una visión objetiva de las habilidades a desarrollar para cursar una carrera específica [31].

- **Modelo de predicción de carrera usando minería de datos y clasificación lineal:** el modelo se basa en la aptitud y personalidad del estudiante. La muestra usada para la construcción del modelo fueron 200 estudiantes, a los cuales se les aplicó un cuestionario para evaluar la aptitud; en suma, se determinó el tipo de personalidad usando las publicaciones del perfil de Facebook del estudiante. La precisión promedio para la aptitud fue del 77 % y para la personalidad del 75 % [29].
- ***Online career counsellor system based on artificial intelligence: An approach***: se propone que los estudiantes entren a un sitio web, el cual está compuesto por el componente de Chatbot, y un motor inteligente que ofrece las mejores opciones de carrera al estudiante en las siguientes áreas: Ingeniería y Tecnología, Artes, Comercio, Derecho, Humanidades, Gestión Hotelera, Ciencias Sociales. Para generar la predicción, el sistema toma las respuestas de los estudiantes a los cuestionarios [22].

- ***Intelligent Decision Support System Using Decision Tree Method for Student Career:*** para el Sistema inteligente se usaron varios conjuntos de datos, dentro de ellos se contaba con: variables de registro académico, prueba de aptitud y detalles como pasatiempos, calificaciones, gustos entre otros. Como resultado, la aplicación web proporcionaba una sugerencia de carrera [20].

La diferencia de los anteriores trabajos con respecto al propuesto en este documento es el uso de diferentes técnicas de inteligencia artificial supervisadas para encontrar el modelo predictivo más acertado para la predicción de una carrera universitaria, además se hace uso de datos del gobierno colombiano proporcionado por la entidad del ICFES.

Capítulo 3

Esta sección se centra en el conjunto de datos para conocer a detalle las variables que contiene, cómo se crearon las clases a predecir, la limpieza de los datos y categorización de estas que fue usada en la construcción de los modelos predictivos.

3 Preparación de los datos

Para la etapa de preparación de los datos se hace un proceso que consta de los siguientes pasos:

- Primero se dio la recopilación de datos
- Segundo se hizo un análisis exploratorio de los datos obtenidos.
- Tercero se crearon las clases para la predicción de la carrera
- Cuarto se procedió a hacer la limpieza de los datos (eliminar datos nulos, reemplazar datos faltantes)
- Quinto se categorizaron las variables socio demográficas.

Estos pasos se realizaron para poder obtener una base de datos organizada y limpia para proceder a la construcción de los modelos con las técnicas de machine learning seleccionadas.

3.1 Recopilación de datos

El conjunto de datos utilizado para la creación del agente inteligente se construyó de la unión de las bases de datos correspondientes a las pruebas del saber 11 y saber Pro. Los datos exportados desde el Instituto Colombiano para la Evaluación de la Educación (en adelante ICFES) contienen la información de los estudiantes en 2 momentos: al salir del bachillerato y al cumplir mínimo el 70 % del plan de estudios de un programa de educación superior. El rango de años en que se enmarca el conjunto de datos es el siguiente:

- Saber 11: 2006 al 2013
- Saber Pro: 2012 al 2017

La base de datos contenía información de las condiciones socioeconómicas, la educación de los padres, las expectativas del estudiante, la información del colegio y los resultados de las

pruebas (ver tabla 1). En 2014, el ICFES cambió la estructura de sus pruebas, lo que a su vez supuso cambios en los campos de sus bases de datos y la eliminación de otros. Las áreas de conocimiento evaluadas dentro del conjunto de datos seleccionado son:

- Lenguaje
- Ciencias sociales
- Matemáticas
- Filosofía
- Biología
- Química
- Física
- Inglés

Se tuvo en cuenta el criterio de expertos para la selección de variables en la construcción de la base de datos del modelo predictivo, considerando los aspectos relevantes para la orientación vocacional. La tabla 1 muestra las variables seleccionadas[2].

Tabla 1. Diccionario de variables del conjunto de datos de las pruebas Saber 11.

Diccionario de Variables	Variables
Información de contacto	edad_11', 'genero', 'reside_mcpio_11', 'reside_depto_11', DEPTO_PRO', 'MCPIO_PRO
Socioeconómica	FAMI_EDUCACIONPADRE, FAMI_EDUCACIONMADRE, FAMI_OCUPACIONPADRE, FAMI_OCUPACIOMADRE, FAMI_ESTRATOVIVIENDA, 'ESTRATO_VIVIENDA', 'TIENE_INTERNET', 'TIENE_COMPUTADOR', ESTU_TRABAJAACTUALMENTE, 'CABEZA_FAMILIA', 'NUM_PERSONAS_ACARGO'.
Datos académicos saber pro	'VALOR_UNIVERSIDAD', 'BECADO', 'CREDITO_ESTUDIO', 'PAGO_PADRES', 'PAGO_PROPIO', "NOMBRE_UNIVERSIDAD', 'PRGM_ACADEMICO', 'GRUPO_REFERENCIA', , 'NIVEL_PRGM_ACADEMICO', 'TIPO_PRGM', 'NUCLEO_PREGRADO', 'SEMESTRE',

[2] entre ellos psicopedagogos (Hesam Camilo Sadeghian), psicólogos (Diana Arce, Yurdey Herran) y la directora del Centro de Enseñanza y Aprendizaje de la universidad Javeriana Cali (Carolina Duque)

Resultados	'punt_lenguaje', 'punt_matematicas', 'punt_c_sociales', 'punt_filosofia', 'punt_biologia', 'punt_quimica', 'punt_fisica', 'punt_ingles', 'puesto_saber_11'
Expectativas	'ESTU_TIPOCARRERADESEADA', 'ESTU_RAZON_PROG_DESEADO'
Información Colegio	'cole_bilingue_cod', 'cole_caracter_cod', 'cole_jornada_cod', 'COLE_NATURALEZA'

3.2 Análisis exploratorio

La librería sklearn proporcionó estadísticas descriptivas para las variables numéricas, incluyendo medidas como la moda y la distribución de los datos, lo cual da una idea de las características y el comportamiento de las variables numéricas en el conjunto de datos (ver figura 8).

```
[ ] #Medidas de centralidad y desviación para atributos numéricos:
    df.describe()
```

	edad_11	punt_lenguaje	punt_matematicas	punt_c_sociales	punt_filosofia	punt_biologia	punt_quimica	punt_fisica	punt_ingles
count	911055.000000	911055.000000	911055.000000	911055.000000	911055.000000	911055.000000	911025.000000	911055.000000	911055.00000
mean	16.755690	51.196813	50.679413	50.473581	47.461614	50.362136	49.923432	48.298922	50.58015
std	1.833661	8.422459	10.938572	9.313465	9.467601	8.694212	8.955390	9.120024	13.24142
min	11.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.00000
25%	16.000000	46.000000	43.800000	44.000000	41.270000	44.420000	44.000000	42.690000	41.93000
50%	16.000000	50.600000	49.840000	50.260000	46.810000	49.870000	48.970000	47.950000	47.58000
75%	17.000000	56.410000	56.590000	56.070000	53.590000	55.530000	54.290000	54.120000	55.84000
max	79.000000	113.190000	127.000000	100.490000	97.350000	123.000000	118.800000	121.790000	117.29000

Figura 8. Estadísticas descriptivas de las variables numéricas del conjunto de datos inicial.

En la base de datos inicial se observan incoherencias en el análisis estadístico de los atributos numéricos, dado que existen registros con puntajes superiores a 100 puntos en las áreas del saber 11, cuando el puntaje máximo a obtener era 100 puntos. Para corroborar esta falta de coherencia, se elaboró un diagrama de cajas y bigotes que ofrece la distribución de los datos y garantiza la existencia de datos atípicos (consultar la figura 9).

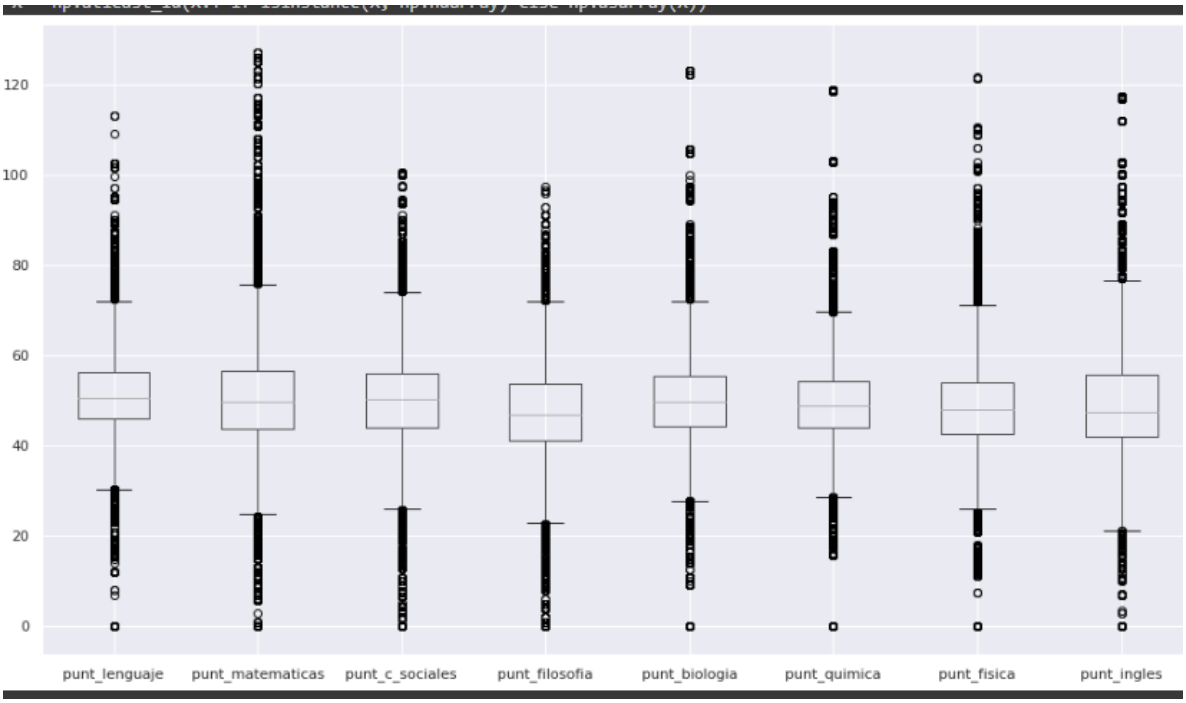


Figura 9. Diagrama de cajas y bigotes de los puntajes saber 11

Una vez comprobados los datos atípicos, en la etapa de limpieza se eliminaron todos los casos atípicos, con el objetivo de que los datos fueran coherentes. Es importante señalar que el ICFES llevó a cabo el proceso de lectura y recopilación de datos para las bases de datos originales.

Se realizó un cruce entre la base de datos de saber 11 y saber pro con el objetivo de disponer de un conjunto de datos que contenga la variable objetivo a predecir por las técnicas de inteligencia artificial, que es la carrera universitaria. Se realizó este cruce por un diccionario de clave-valor facilitado por el ICFES, utilizando la librería de pandas enfocada en el manejo de conjuntos de datos en el lenguaje de Python. El resultado es un conjunto de datos con todas las variables necesarias (ver tabla 1) para el entrenamiento de un agente inteligente.

3.3 Creación de clases

La variable objetivo a predecir debe ser una indicación del área disciplinar que el estudiante pueda elegir de un programa de pregrado; de la data obtenida con el cruce del saber pro y saber 11 se identifica dos variables que pueden aportar a la creación de esta variable, ‘Grupo de Referencia’ y ‘Núcleo de pregrado’.

La elección de la variable objetivo se realizó con el análisis de los histogramas de ambas variables (ver figura 10 y figura 11,), se observa que la variable grupo de referencia tiene 34 valores y núcleo de pregrado 57 valores, esta diferencia es a causa de la especificación de la variable núcleo que contiene las carreras de los estudiantes que presentaron el Saber Pro. En

este sentido, un problema de clasificación debe priorizar el menor número de clases posibles para facilitar el entrenamiento y predicción del modelo predictivo, por esa razón se eligió la variable grupo de referencia como variable objetivo.

Figura 10. Clases por núcleo de referencia

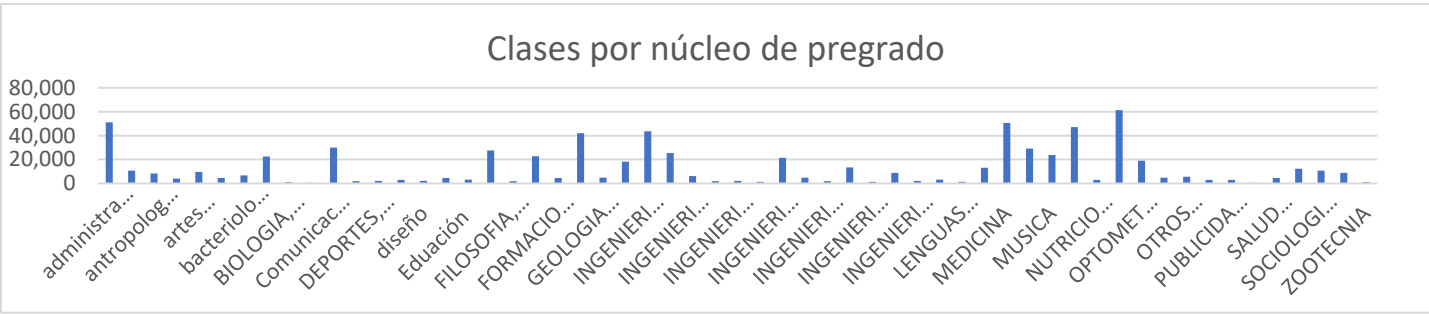


Figura 11. Clases por grupo de referencia

Con la elección de la variable objetivo se procedió a la reducción de la cantidad de clases que esta posee; se agruparon las carreras en clases representativas por áreas de conocimiento. Por ejemplo: las diferentes ingenierías en una sola clase llamada ingeniería, es preciso señalar que los programas técnicos y tecnológicos se agruparon separadamente de los programas de pregrado.

Las etiquetas creadas (ver tabla 4) siguen lo propuesto por el SNIES (Sistema Nacional de Información de la Educación Superior)[32] y se definieron con la asesoría de psicólogas con enfoque educativo de la universidad Pontificia Javeriana Cali. Las tablas 2 y 3 presentan las clases que se usarán en el entrenamiento del agente inteligente; en los datos, se les identificará con el nombre de “target”.

Tabla 2 Carreras incluyendo las de carácter tecnólogo y técnicos

Grupo con Tecnólogos y técnicos	Cantidad
Ciencias agropecuarias	35.423
ciencias económicas	48.201
Administración y afines	98.645
Ingeniería y tecnología	153.588
Industria y minas	63.387
Carrera militar	30.751
Artes	56.758
Salud	80.656
Educación	69.247
Derecho	46.030
Turismo y afines	151.243
Total de clases	**11**

Tabla 3. Carrera solamente de Pregrado de carácter universitario

Grupo Sin Tecnólogos y técnicos (Pregrado)	Cantidad
Humanidades	67.072
Ciencias agropecuarias	20.214
ciencias económicas	48.201
Administración y afines	98.645
Ingeniería	123.483
Artes	31.020
Ciencias de la Salud	57.873
Educación	69.247
Derecho	44.266
Total de clases	**9**

3.4 Limpieza de datos

La limpieza de los datos tiene como propósito resolver anomalías en un data set para mejorar la calidad de los datos, obteniendo al finalizar el proceso de la limpieza datos coherentes y libres de errores. Entre las anomalías que presentan los datos comúnmente son registros incompletos, duplicados, incorrectos; se debe eliminar o corregir porque es información que genera ruido en la predicción de la variable objetivo.

En el proceso de limpieza se realizó la eliminación de los registros de estudiantes que tenían datos atípicos en los puntajes del saber 11, la unificación de los valores de la variable objetivo dado que se encontraban duplicados por mala escritura o lectura. Además, se identificó y

reemplazó los registros con datos faltantes por el dato de moda correspondiente a la columna que corresponda.

3.5 Codificación de variables

La mayoría de las columnas de la data set son de tipo texto, es decir variables que necesitan una reinterpretación de su valor para facilitar el entrenamiento por el agente, se procedió hacer una codificación de estas para obtener un valor número correspondiente a cada valor único de las variables, con la intención de facilitar el entrenamiento del agente inteligente.

La categorización se realizó con la librería scikit learn con su módulo de preprocesamiento enfocado en la codificación de variables categóricas, el resultado de este proceso es la tabla 4:

Tabla 4. Descripción de posibles valores para cada una de las variables del data set.

Variable	Descripción	Codificación
Target	Carreras por predecir por el agente inteligente	Administracion_afines (0) artes (1) carrera militar (2) ciencias agropecuarias (3) ciencias económicas (4) derecho (5) educación (6) humanidades (7) industria y minas (8) ingeniería (9) salud (10) sin categoría (11) turismo y afines (12)
Genero	Genero del estudiante	Femenino (0), Masculino (1)
Estrato vivienda	Estrato socioeconómico del estudiante al momento de presentar la prueba saber 11	Estrato 0 (0) Estrato 1 (1) Estrato 2 (2) Estrato 3 (3) Estrato 4 (4) Estrato 5 (5) Estrato 6 (6) Sin estrato o zona rural (7)
Cabeza de familia	El estudiante es cabeza de familia	No (0), Sí (1)
Colegio Bilingüe	El colegio del estudiante se considera bilingüe	No (0), Sí (1)

Carácter del colegio	Identifica el tipo de planteamiento académico del colegio	Académico (0), Normalista (1), Técnico (2), Técnico/Académico (3)
Jornada del colegio	Tiempo diario en que el colegio presta el servicio educativo a los estudiantes	Completa (0), Mañana (1), Noche (2), Sabatina (3), Tarde (4), Única (5).
Genero del colegio	Genero preferente en el colegio para sus estudiantes	Femenino (0), Masculino (1), Mixto (2)
Naturaleza del colegio	El colegio es de sentido oficial o no	No oficial (0), Oficial (1)
Tiene internet	El estudiante cuenta con internet en su vivienda	No (0), Sí (1)
Tiene computador	El estudiante posee un computador	No (0), Sí (1)
Becado	El estudiante salió becado de su colegio	No (0), Sí (1)
Razón de programa deseado	El estudiante podía dar una razón por la cual le gustaría estudiar una carrera de educación superior	Servir a la comunidad (1), Seguir inclinaciones vocacionales (2), Tener éxito y prestigio (3), Mejor posición social (4), Profundizar conocimiento (5), Responder expectativas familiares (6), No definido (7).
Estudiante trabaja	Si el estudiante recibe un tipo de remuneración por algún trabajo realizado	Si (1), No (2)
Educación Padres	Nivel educativo al que llegaron los padres del estudiante, el padre y madre compartes la misma codificación	Primaria incompleta (1), Primaria completa (2), Secundaria incompleta (3), Secundaria completa (4), Educación profesional completa (5), Técnica o Tecnológica completa (6), Postgrado (7), Educación profesional incompleta (8), Técnica o Tecnológica incompleta (9), No sabe o ninguno (9).
Ocupaciones padres	Rol que desempeña el padre o madre	Empresario (1), Pequeño empresario (2), Empleado cargo directivo (3), Empleado nivel técnico o profesional (4), Empleado auxiliar o administrativo (5), Obrero u operario (6), Independiente (7), Pensionado (8), No sabe (9), Hogar (10).

Tipo carrera deseada	Nivel de la carrera deseada por el estudiante	Técnica (1), Tecnológica (2), Profesional (3), ninguna (4)

El data set inicial contaba con 911.056 registros, al realizar la limpieza de los datos el data set final contiene 857.008 registros con una eliminación de 54.048 filas por datos incorrectos (4903), vacíos o nulos (49.145) para así poder contar con un conjunto de datos limpio para el entrenamiento de los distintos modelos predictivos basados en las técnicas de clasificación.

Capítulo 4

4 Construcción del agente inteligente

En esta sección se presenta el proceso de selección de técnicas de machine learning, la estimación de parámetros de los modelos iniciales a entrenar, configuración de los distintos modelos construidos, entrenamiento y evaluación de los modelos.

4.1 Elección de técnicas

Por medio de la revisión de literatura de trabajos de grado relacionados con la misma problemática de diseñar un modelo predictivo de la carrera profesional se realizará un análisis de cada una de las técnicas según criterios explicados en la sección, para esto vemos la frecuencia de uso de las técnicas de machine learning en la revisión del estado del arte (ver tabla 5).

Tabla 5. Elección de técnicas de machine learning

Técnicas de Machine Learning	**Artículos referenciados**
K-vecinos	[29], [31]
Árbol de decisión	[20], [22]
Máquina de vectores de soporte (SVM)	[33], [31], [22], [15]
Naive Bayes	[33], [31]
Red Neuronal	[15]
Random Forest	[29]
Regresión Lineal	[29]

4.1.1 Criterios de selección

Se presentan los criterios de selección para las técnicas a utilizadas en la creación del agente inteligente y experimentación de los diferentes modelos a evaluar, estos fueron:

- Desempeño en trabajos previos: resultados de métricas del desempeño de las técnicas al utilizarlas en la predicción de la carrera universitaria.

- Facilidad de implementación: nivel de conocimiento previo y disponibilidad de librerías para implementar la técnica.

- Capacidad de clasificación multiclase: naturalidad con que las técnicas realizan clasificación multiclase desde su base teórica y práctica

- Manejo de datos discretos: se analiza el uso de frecuencia de datos discretos en las técnicas, debido a que en el dato set este tipo de dato se presenta con mayor frecuencia.

4.1.2 Técnicas seleccionadas

Para la selección de las técnicas se realizó la siguiente tabla (ver tabla 6) con el análisis de cada criterio, bajo la escala de puntación bajo (1 y 2), medio (3), alto (4 y 5):

Tabla 6. Selección de técnicas.

Técnica	*Desempeño en trabajos previos*	*Facilidad de implementación*	*Capacidad de clasificación multiclase*	*Manejo de datos discretos*	*Puntaje*
K - vecinos	**5**	**5**	**5**	**5**	**20**
Árbol de decisión	**4**	**5**	**5**	**3**	**17**
SVM	**4**	**3**	**4**	**4**	**15**
Naive bayes	**2**	**4**	**3**	**3**	**12**
Red neuronal	**4**	**4**	**4**	**4**	**16**
Random Forest	**5**	**5**	**5**	**4**	**19**
Regresión lineal	**1**	**4**	**2**	**2**	**9**

Se observa que las técnicas con mayor puntaje y las cuales fueron utilizadas en la creación de los modelos predictivos son k-vecinos, árbol de decisión, máquina de vectores de soporte (SVM) y red neuronal; se omite random forest en la tabla de experimentación descrita en la tabla 8 porque es una técnica que implementa árbol de decisión en su algoritmo, se podría utilizar en algún escenario en particular para ver una variación del árbol de decisión.

4.2 Estimación de parámetros

Para lograr la creación de modelos es necesario considerar el valor de los parámetros de cada una de las técnicas seleccionadas para maximizar el rendimiento y precisión de estos [34].

La estimación se realizó por medio de la técnica de Grid Search, la cual proporciona la configuración de parámetros mejor posicionada con la métrica de Accuracy como medida de evaluación porque mide el porcentaje de casos que el modelo ha acertado la variable objetivo. Las técnicas a hacer estimación de parámetros fueron árbol de decisión, máquina de vectores de soporte, k vecinos y red neuronal con los siguientes parámetros a estimar:

- Árbol de decisión
  - Max_depth: La profundidad máxima del árbol
  - Min_samples_split: El número mínimo de muestras requeridas para dividir un nodo interno
  - Min_samples_leaf: número mínimo de muestras requeridas para estar en un nodo hoja.

- Máquina de vectores de soporte
  - Kernel: la manera en que se hará la división por medio de una recta, plano o un hiperplano n-dimensional.
- K- vecinos
  - N_neighbors: número de vecinos más cercano al punto a evaluar
  - Weights: función de peso utilizada en la predicción
  - P: función de cálculo de los n vecinos más cercanos del punto a evaluar
- Red neuronal
  - Hidden_layers_sizes: estructura las capas ocultas, definiendo cantidad de neuronas y capas.

Al hacer diferentes procesos de estimación de parámetros, los resultados finales fueron la tabla 7

Tabla 7. Resultados de parámetros al seleccionar de las posibilidades con Grid Search

Técnica	Posibilidades	Parámetros
Árbol de decisión	Max_depth: rango de 9 a 21 (saltos de 1 unidad) Min_samples_split: rango de 50 a 1500 (saltos de 5 unidades)	Max_depth: 9 Min_samples_split: 1200 Min_samples_leaf: 57

	Min_samples_leaf: 50 a 1500 (salto de 7 unidades)	
Máquina de vectores de soporte	Kernel: poly, lineal, rbf, sigmoid	Kernel: poly
K vecinos	N_neighbors: 5 a 100 (salto de 1 unidad) Weights: uniform, distance P: 1, 2,5	N_neighbors: 9 Weights: uniform P: 1
Red neuronal	Hidden_layer sizes: [(50,100,150), (50,50,50), (200,150,50)]	Hidden_layer sizes: (200, 150, 50)

4.3 Entrenamiento de los modelos

Para el entrenamiento de los modelos se debe contar con la claridad de cómo se dividirá el data set y configuración de cada experimento a realizar. En la figura 12, se muestra da división del conjunto de datos en su parte de entrenamiento y test

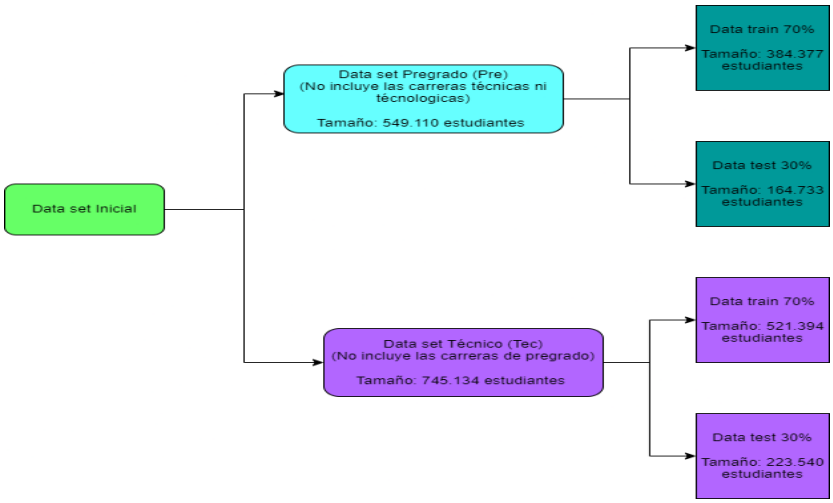


Figura 12. División del conjunto de datos en los data sets PRE Y TEC

El diseño de los experimentos se basó en las recomendaciones de profesionales en orientación vocacional de la Universidad Pontificia Javeriana[3] junto a una asesoría del CEA (Centro de

[3] Diana Arce (Magister en Educación) | Hesam Sadeghian (Magister en educación)
Marcela Valencia (Doctorado en Psicología)

Enseñanza y Aprendizaje) para la elección de distintos atributos que son comunes en pruebas vocacionales.

Se presentan en la tabla 8 los diferentes experimentos con su respectiva explicación. Se ejecutaron en los dos conjuntos de datos creados, Pre (carreras solamente de pregrado) y Tec (carreras técnicas y tecnológicas), mencionados en la sección 3.3.

Tabla 8. Diseño de experimentos

# Experimento	Resumen
#1 Puntajes Saber 11	Como data de entrada del agente inteligente se eligen solamente los puntajes de las áreas de conocimiento de la prueba Saber 11.
#2 Socio	Se añaden las variables de género, estrato y cabeza de familia.
#3 Colegio	Se añaden las variables relacionadas con el colegio (carácter, bilingüe, jornada) a los atributos del experimento #2, se reduce la cantidad de clases a clasificar de 11 a 9.
#4 Normalización	Se realiza la normalización de los datos de entrada de los puntajes saber 11 y se incluyen las variables del experimento anterior.
#5 Balanceo de clases	Se balancea las clases de la variable a predecir realizando un sobre muestreo y sub muestreo de las clases para obtener cada clase en un rango de 14000 a 16000 registros por clase.
#6 CEA y puntajes 11	Se añaden las variables: Genero del colegio, Naturaleza del colegio, Tiene computador, Tiene internet, Razón de programa deseado, Estudiante trabaja, Educación Padres, Ocupación Padres, Tipo de Carrera.
#7 CEA	Se elimina de las variables de entrada los puntajes de las áreas de conocimiento
#8 Dummies	Transforma las variables categóricas en variables dummies.
#9 Reducción dimensionalidad	Selección de variables con métodos cuantitativos como selection features, PCA y feature importance proporcionados por la librería Scikit-learn

Al contar con el diseño de los experimentos se procedió al entrenamiento del agente inteligente y así capturar los resultados a analizar en el siguiente capítulo los resultados obtenidos con los diferentes experimentos y técnicas de clasificación.

Capítulo 5

5 Análisis de resultados

En esta sección se expone el análisis realizado de los resultados obtenidos de los diferentes experimentos entre las técnicas de clasificación, adicional se presenta una comparación del desempeño del resultado de la experimentación para predecir la carrera universitaria.

En las tablas 8 y 9 se consolida los resultados de la métrica Accuracy de cada experimento junto a las técnicas de clasificación escogidas.

Tabla 9. Resultados de Accuracy de los experimentos del data set TEC

Técnica/Experimento	Exp1	Exp2	Exp3	Exp4	Exp5	Exp6	Exp7	Exp8	Exp9
K vecinos	0,14	0,15	0,19	0,23	0,46	0,40	0,30	0,54	0,39
Árbol de decisión	0,23	0,25	0,29	0,28	0,24	0,30	0,29	0,37	0,30
Red Neuronal	0,17	0,1	0,27	0,16	0,23	x	x	x	x
Máquina de vectores de soporte	0,24	0,24	0,28	0,28	0,24	0,30	0,29	0,30	0,25

Tabla 10. Resultados de la métrica Accuracy con el data set PRE

Técnica/Experimento	Exp1	Exp2	Exp3	Exp4	Exp5	Exp6	Exp7	Exp8	Exp9
K vecinos	0,19	0,20	0,31	0,33	0,41	0,47	0,46	0,62	0,49
Árbol de decisión	0,25	0,28	0,38	0,37	0,29	0,38	0,45	0,44	0,38
Red Neuronal	0,22	0,17	0,37	0,25	0,27	x	x	x	x
Máquina de vectores de soporte	0,22	0,28	0,38	0,38	0,28	0,38	0,37	0,39	0,22

5.1 Comparación de los modelos

La ejecución de los experimentos se utilizó los data sets Pre y Tec descritos en la figura 12, Test Accuracy fue la métrica para evaluar el desempeño de los modelos creados, cabe resaltar que a partir del experimento 5, la técnica de máquinas de soporte fue sacada de la experimentación porque su entrenamiento tomaba mucho tiempo y no presentaba mejoras significativas respecto a las otras técnicas.

De acuerdo con los resultados obtenidos de los experimentos, la técnica de k vecinos presenta el mejor resultado de clasificación de las carreras para ambos data sets como se observa en la figura 13 y 14; en el data set Pre obtuvo un Accuracy de 62 % mientras que en Tec obtuvo 54 %.

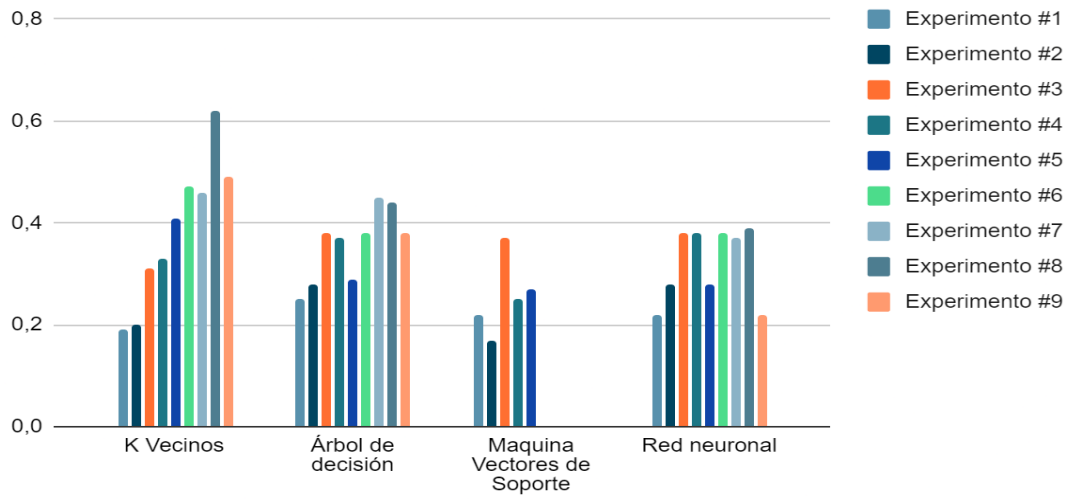


Figura 13. Métrica Accuracy de las distintas técnicas sobre el data set Pre

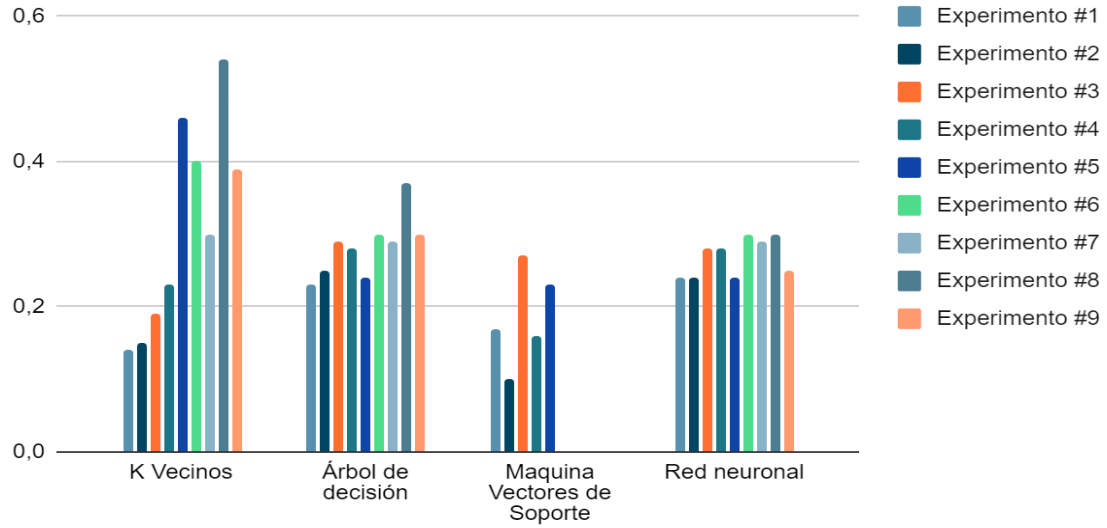


Figura 14. Métrica Accuracy de las distintas técnicas sobre el data set TEC

Las variaciones de los resultados con el pasar de cada experimento se debe a factores como la cantidad de clases a predecir, balanceo de los datos de entrenamiento, transformación de variables dummies, proporcionando una mejora secuencial en la tarea de la clasificación correcta; ahora una explicación a detalle de la ejecución de estos modelos.

- Disminución de clase: inicialmente la cantidad de clases a predecir eran 11 al realizar el paso de los experimentos, se disminuyó el número de las clases con el objetivo de observar alguna variación en el resultado de la métrica Accuracy; para la reducción en el número de clases se hicieron agrupaciones por área de conocimiento entre los programas de educación superior.

- Balanceo de clases: las variables obtenidas del data set inicial del ICFES tenían una distribución no paramétrica, algunas clases contaban con datos mayoritarios que superaban miles a otras clases, que contaban con menor número de casos; en consiguiente, no llegaban a establecer una distribución normal para facilitar la técnica la clasificación de la de la variable objetiva predecir. Aun cuando la base datos recopilaba 549.000 (PRE) y 745.134 (TEC).

- Variables dummies: el data set en su mayoría contaba con variables categóricas estás, tienen la particularidad que desde las técnicas de machine learning se pueden transformar a variables dummy. Para ello, se hace un arreglo por las n clases que tenga esta variable en la que se representa con un 1 en la casilla donde esté presente el dato, esto se realizó para poder aumentar la diversidad de los datos para las técnicas o predecir.

5.2 Consideraciones sobre la utilidad de los modelos desarrollados

Los modelos obtenidos con la ejecución de los experimentos 8 y 7 en las técnicas de K-vecinos y árbol de decisión son los que más se adecuan al objetivo de este trabajo de grado, el cual es brindar opciones de carreras técnicas, tecnológicas o profesionales para apoyar el proceso de orientación vocacional de un estudiante recién graduado de bachiller y su ingreso a la educación superior.

Para facilitar la noción de las carreras a elegir no se recogieron datos, sino que se realizó un análisis con las bases de datos ya existentes recolectadas por el ICFES, esto hace que el agente sea replicable a nivel nacional en todos los bachilleres. Puesto que la presentación del examen saber 11, se constituye en un requisito de grado de los estudiantes. En suma, es importante considerar que el uso de las características socio demográficas, socioeconómicas y los puntajes del saber 11, marcan una tendencia en la elección de carrera adecuada. Es pertinente resaltar que las variables con mayor poder de predicción estuvieron relacionadas con los gustos, el contexto y las personas de referencia para el estudiante, como se observa en las figuras 13 y 14, donde se evalúa los diferentes experimentos con la integración o eliminación de las variables mencionadas.

5.3 Evaluación de la hipótesis de partida

La hipótesis inicial se basa en que los puntajes del saber 11 relacionados con las áreas del conocimiento eran los suficientemente significativos por sí solos para llegar a predecir la carrera universitaria de un estudiante de bachillerato. Sin embargo, en la ejecución de los diferentes experimentos se observó que al aumentar variables relacionadas con el contexto del estudiante se obtuvieron resultados mejores en comparación con la hipótesis inicial de la cual se evaluaban solamente los puntajes del saber 11.

Desde la teoría de la orientación vocacional existen 3 ejes para la vocación de un estudiante: los gustos, las aptitudes y el contexto, sin embargo, en este trabajo de grado se enfocó solamente en las actitudes y datos del contexto recolectados por el ICFES para lograr la predicción de la carrera obteniendo un Accuracy de 0.62 utilizando la técnica k vecinos más cercanos sobre el data set de programas universitarios.

Capítulo 6

6 Desarrollo del agente

En esta sección se presenta el diseño y el desarrollo del agente inteligente reportado en los capítulos anteriores hace referencia a la construcción de un modelo predictivo con las distintas técnicas de clasificación, luego de la comparación de los modelos se determina k vecinos es la técnica más adecuada para el desarrollo del agente inteligente el cual consiste en una aplicación sencilla para ingresar los atributos establecidos para un estudiante de bachillerato.

El agente inteligente se compone de una interfaz web desarrollada con streamlit un framework para aplicaciones de machine learning, donde se presenta al usuario los campos a rellenar estilo formulario para ingresar los atributos necesarios para que el modelo proporcione una predicción de la carrera universitaria a estudiar por el recién graduado del bachillerato.

A continuación, se muestra la interfaz del agente inteligente

Figura 15. Preguntas por el valor de cada área de evaluación de la prueba saber 11

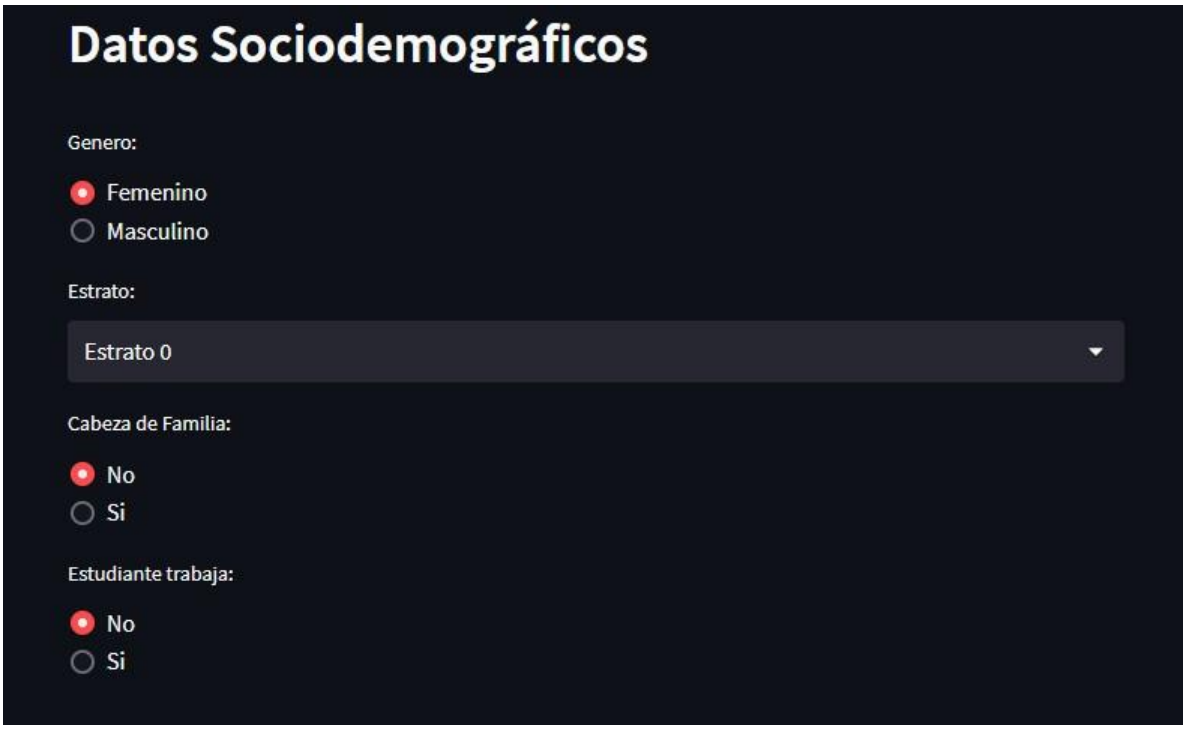


Figura 16. Ingreso de información socio demográfico del estudiante.

Figura 17. Información de la ocupación y nivel académico de los padres.

Figura 18. Registro de las preferencias del estudiante.

El objetivo es que el orientador vocacional use esta interfaz web para recolectar la información del estudiante para que el agente inteligente pueda realizar la predicción de la carrera universitaria que se adecuada a los datos suministrados.

Capítulo 7

7 Conclusiones

- Se construyo el agente inteligente cumpliendo el objetivo de su implementación y evaluación de este con diferentes técnicas de clasificación con escenarios de experimentación variados.

- La técnica de clasificación que mejor funcionó al construir un modelo predictivo fue k-vecinos, y su efectividad fue incrementada en cada experimento.

- Durante los experimentos se observó que disminuir la cantidad de clases a predecir facilitó el entrenamiento de los modelos a través de las técnicas de clasificación.

- El trabajo interdisciplinario con profesionales de psicología y expertos en orientación vocacional brindó una visión diversa al trabajo de grado, lo que permitió obtener una variedad de experimentos ejecutados y contrastados.

Capítulo 8

8 Trabajo futuro

- Realizar predicción de carreras sobre un data set con data que incluya datos relacionados con su contexto, gustos particulares y referente de su núcleo familiar, de forma más robusta y profunda.

- Implementar técnicas de clasificación no supervisada como K-means para evaluar su desempeño y compararlo con el resultado obtenido.

- Entrenar el agente inteligente con datos recolectados propios, con un enfoque en los 3 ejes de la orientación vocacional (gusto, contexto y aptitudes), para complementar los expuestos en repositorios del gobierno.

- Añadir una prueba de orientación vocacional o prueba psicotécnica como insumo de los datos de entrada del agente inteligente para su entrenamiento.

- Desarrollar una aplicación móvil o web para que los orientadores vocacionales a nivel nacional hagan uso de esta herramienta.

Capítulo 9

9 Glosario

- Data set: conocido también como conjunto de datos, hace referencia a la colección de datos, organizado por columnas.

- Variable Dummie: Es la representación de una variable categórica que toma el valor de 0 o 1 para indicar la ausencia o presencia.

- Chatbot: Es un programa informático que simula y procesa conversaciones humanas (ya sean escritas o habladas), permitiendo a los humanos interactuar con dispositivos digitales como si se estuvieran comunicando con una persona real.

- Accuracy: métrica que representa la cantidad de predicciones que el modelo realizo correctamente

- Precisión: métrica que evalúa la calidad del modelo, determina cuantos aciertos fueron correctos

- Recall: representa la cantidad de verdaderas predicciones se realizaron.

- F1 Score: El valor F1 se utiliza para combinar las medidas de precisión y recall en un solo valor. Esto es práctico porque hace más fácil el poder comparar el rendimiento combinado de la precisión y la exhaustividad entre varias decisiones.

- Matriz de confusión: herramienta que permite la visualización del desempeño de un algoritmo de clasificación.

10 Bibliografía

[1] Ministerio de Educación Nacional, “La Revolución Educativa: Plan Sectorial 2006- 2010.” 2010. Accessed: Jul. 24, 2022. [Online]. Available: https://www.mineducacion.gov.co/1621/articles-156179_recurso_7.unknown

[2] L. Brewer and P. Comyn, “Integrating core work skills into TVET systems: Six country case studies,” 2015, Accessed: Jul. 24, 2022. [Online]. Available: www.ilo.org/publns

[3] K. G. Skarpaas and G. O. Hellekjær, “Vocational orientation – A supportive approach to teaching L2 English in upper secondary school vocational programmes,” *Int. J. Educ. Res. Open*, vol. 2, p. 100064, Jan. 2021, doi: 10.1016/J.IJEDRO.2021.100064.

[4] A. Alfaro-Barquero and S. Chinchilla-Brenes, “Construcción y validación de un instrumento de evaluación de preferencias y habilidades vocacionales para carreras científico-tecnológicas,” *Rev. Tecnol. en Marcha*, vol. 30, no. 4, p. 138, Dec. 2017, doi: 10.18845/tm.v30i4.3418.

[5] A. Claudia and R. Tade, “Robot Virtual en Orientación Vocacional,” no. 02, pp. 1–18, 2007.

[6] A. Claudia and R. Tadeo, “Sistema Inteligente Conversacional para la Orientación Vocacional,” Universidad de Colima, 2009.

[7] H. Alberto and B. Peñaloza, “INCIDENCIA DE LOS PROGRAMAS DE ORIENTACIÓN VOCACIONAL EN COLOMBIA,” Educación Media.

[8] M. R. Urrego, “La investigación sobre deserción universitaria en Colombia 2006-2016. Tendencias y resultados,” 2019.

[9] DANE, “Convocatoria 2022 - Índice de Capacidad Estadística Territorial - ICET.” https://www.dane.gov.co/index.php/convocatorias-y-contratacion/informacion-laboral/convocatorias-roles-operativos-para-operaciones-estadisticas/convocatoria-2022-indice-de-capacidad-estadistica-territorial-icet (accessed Jul. 24, 2022).

[10] L. Myaka, “IMPACT OF FORMAL CAREER GUIDANCE AND COUNSELLING DURING HIGH SCHOOL AT UNIZULU.”

[11] L. López Villafaña, A. B. Solache, and M. Antonio Pérez Chávez, “Deserción escolar en universitarios del centro universitario UAEM Temascaltepec, México: estudio de caso de la licenciatura de Psicología School dropout in university students from UAEM Temascaltepec center, Mexico: a case study of Psychology degree,” *Rev. Iberoam. Evaluación Educ.*, vol. 7, no. 1, 2014, Accessed: Oct. 06, 2020. [Online]. Available: www.rinace.net/riee/

[12] L. Shuguang, L. Zheng, and B. Lin, “Impact of Artificial Intelligence 2.0 on Teaching and Learning,” Feb. 2020, pp. 128–133. doi: 10.1145/3383923.3383928.

[13] J. TAPIA, “Sistema Experto Para El Apoyo Del Proceso De Orientación Vocacional Para Las Carreras De Ingeniería En La Pontificia Universidad Católica Del Perú.,” *Test*, pp. 1–125, 2006.

[14] IEEE Staff, *Construction of a Basic Intelligent Agent*. IEEE, 2017.

[15] O. Zahour, E. H. Benlahmar, A. Eddaoui, H. Ouchra, and O. Hourrane, “A system for educational and vocational guidance in Morocco: Chatbot E-Orientation,” *Procedia Comput. Sci.*, vol. 175, pp. 554–559, Jan. 2020, doi: 10.1016/j.procs.2020.07.079.

[16] V. González Maura, “AUTODETERMINACIÓN Y CONDUCTA EXPLORATORIA. ELEMENTOS ESENCIALES EN LA COMPETENCIA PARA LA ELECCIÓN PROFESIONAL RESPONSABLE,” 2009.

[17] V. N. Drozdov, V. A. Kim, and L. B. Lazebnik, *[Modern approach to the prevention and treatment of NSAID-gastropathy].*, no. 2. 2011.

[18] A. Jhamtani, R. Mehta, and S. Singh, “Size of wallet estimation: Application of K-nearest neighbour and quantile regression,” *IIMB Manag. Rev.*, vol. 33, no. 3, pp. 184–190, Sep.

2021, doi: 10.1016/j.iimb.2021.09.001.
[19] M. Cubillos, S. Wøhlk, and J. N. Wulff, “A bi-objective k -nearest-neighbors-based imputation method for multilevel data,” *Expert Syst. Appl.*, vol. 204, p. 117298, Oct. 2022, doi: 10.1016/j.eswa.2022.117298.
[20] A. Shankhdhar, A. Agrawal, D. Sharma, S. Chaturvedi, and M. Pushkarna, “Intelligent Decision Support System Using Decision Tree Method for Student Career,” *2020 Int. Conf. Power Electron. IoT Appl. Renew. Energy its Control. PARC 2020*, pp. 140–142, Feb. 2020, doi: 10.1109/PARC49193.2020.246974.
[21] M. Rezapour, A. Mehrara Molan, and K. Ksaibati, “Analyzing injury severity of motorcycle at-fault crashes using machine learning techniques, decision tree and logistic regression models,” *Int. J. Transp. Sci. Technol.*, vol. 9, no. 2, pp. 89–99, Jun. 2020, doi: 10.1016/J.IJTST.2019.10.002.
[22] K. Joshi, A. K. Goel, and T. Kumar, “Online career counsellor system based on artificial intelligence: An approach,” Jul. 2020. doi: 10.1109/ICSSS49621.2020.9202024.
[23] A. Ganatra and H. Bhavsar, “Variations of Support Vector Machine classification Technique: A survey,” *Int. J. Adv. Comput. Res.*, vol. 2, no. 4, pp. 223–227, 2013, Accessed: Jul. 20, 2022. [Online]. Available: https://www.researchgate.net/publication/305073135_Variations_of_Support_Vector_Machine_classification_Technique_A_survey
[24] A. S. Aski and N. K. Sourati, “Proposed efficient algorithm to filter spam using machine learning techniques,” *Pacific Sci. Rev. A Nat. Sci. Eng.*, vol. 18, no. 2, pp. 145–149, Jul. 2016, doi: 10.1016/j.psra.2016.09.017.
[25] E. Afacan, N. Lourenço, R. Martins, and G. Dündar, “Review: Machine learning techniques in analog/RF integrated circuit design, synthesis, layout, and test,” *Integration*, vol. 77, pp. 113–130, Mar. 2021, doi: 10.1016/J.VLSI.2020.11.006.
[26] J. I. Ryu, “Applying machine learning techniques to predict detonation initiation from hot spots,” *Energy AI*, vol. 9, p. 100163, Aug. 2022, doi: 10.1016/J.EGYAI.2022.100163.
[27] S. Walczak and N. Cerpa, “Artificial Neural Networks,” *Encycl. Phys. Sci. Technol.*, pp. 631–645, Jan. 2003, doi: 10.1016/B0-12-227410-5/00837-1.
[28] J. V. Tu, “Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes,” *J. Clin. Epidemiol.*, vol. 49, no. 11, pp. 1225–1231, Nov. 1996, doi: 10.1016/S0895-4356(96)00002-9.
[29] R. H. Rangnekar, K. P. Suratwala, S. Krishna, and S. Dhage, “Career Prediction Model Using Data Mining and Linear Classification,” *Proc. - 2018 4th Int. Conf. Comput. Commun. Control Autom. ICCUBEA 2018*, Jul. 2018, doi: 10.1109/ICCUBEA.2018.8697689.
[30] M. Aria, C. Cuccurullo, and A. Gnasso, “A comparison among interpretative proposals for Random Forests,” *Mach. Learn. with Appl.*, vol. 6, p. 100094, Dec. 2021, doi: 10.1016/J.MLWA.2021.100094.
[31] S. Vignesh, C. Shivani Priyanka, H. Shree Manju, and K. Mythili, “An Intelligent Career Guidance System using Machine Learning,” in *2021 7th International Conference on Advanced Computing and Communication Systems, ICACCS 2021*, Mar. 2021, pp. 987–990. doi: 10.1109/ICACCS51430.2021.9441978.
[32] “Inicio - Sistema Nacional de Información de la Educación Superior.” https://snies.mineducacion.gov.co/portal/ (accessed Aug. 12, 2022).
[33] R. Ade and P. R. Deshmukh, “An incremental ensemble of classifiers as a technique for prediction of student’s career choice,” *1st Int. Conf. Networks Soft Comput. ICNSC 2014 - Proc.*, pp. 384–387, Sep. 2014, doi: 10.1109/CNSC.2014.6906655.
[34] “Ajuste de modelos de aprendizaje automático - RiskSpan.” https://riskspan.com/tuning-machine-learning-models/ (accessed Aug. 09, 2022).