INSTITUCIÓN UNIVERSITARIA POLITÉCNICO GRAN COLOMBIANO

FACULTAD DE INGENIERÍA Y CIENCIAS BÁSICAS

MAESTRÍA EN INGENIERÍA DE SISTEMAS

GRUPO DE INVESTIGACIÓN FICB-PG

LINEA DE PROFUNDIZACIÓN: CONSTRUCCIÓN DE SOFTWARE Y DESCUBRIMIENTO DE CONOCIMIENTO EN BASES DE DATOS

SISTEMA DE RECOMENDACIÓN DE CARRERAS PROFESIONALES PARA ESTUDIANTES DE GRADO 11 EN COLOMBIA

PRESENTA:

JEISSON RAUL LEIVA TELLEZ

1810020078

ASESOR TEMÁTICO:

ISABEL ANDREA MAHECHA NIETO

MAGISTER EN INGENIERÍA DE SISTEMAS Y COMPUTACIÓN

DICIEMBRE DE 2021

Resumen

El incremento de la cantidad de datos que se recolectan en cada instante de tiempo hace que las técnicas estadísticas tradicionales sean menos eficientes y difíciles de mantener, sin embargo, durante los últimos 30 años, se han venido madurando técnicas que permiten no solo describir una situación si no también incluso predecir tendencias de consumo entre las personas.

Estas técnicas son el resultado de un largo proceso de investigación y desarrollo de productos. La minería de datos es un intento de buscarle sentido a la explosión de información que actualmente puede ser almacenada [1]. Es por esto por lo que en la era de la tecnología y la información los datos se constituyen en el mayor valor para cualquier compañía.

El siguiente trabajo pretende explicar el proceso de realización de una propuesta metodológica usando la minería de datos para la orientación vocacional, con el fin de clasificar a los aspirantes a una carrera de pregrado, segmentarlos y con esto poder establecer posibles tendencias de selección satisfactoria de la carrera a seguir y así tener una retención y fidelización dentro del sistema educativo.

Palabras clave: Minería de datos, orientación vocacional, retención, fidelización, clasificación

Abstract

The increase in the amount of data collected at each instant of time makes traditional statistical techniques less efficient and difficult to maintain, however, during the last 30 years, techniques have been maturing that allow not only describing a situation if not also predicting consumption trends among people.

These techniques are the result of a long process of research and product development. Data mining is an attempt to make sense of the explosion of information that can currently be stored

[1]. Therefore, in the age of technology and information data is the greatest value for any company.

The following work tries to explain the process of realization of a methodological proposal using data mining for vocational orientation, in order to classify applicants for an undergraduate career, segment them and with this be able to establish possible trends of satisfactory selection of the career to follow and thus have retention and loyalty within the educational system

Key words: Data mining, Vocational orientation, retention, loyalty

TABLA DE CONTENIDO

Contenido

ÍNDICE DE ILUSTRACIONES

ÍNDICE DE TABLAS

ÍNDICE DE ECUACIONES

ÍNDICE DE ANEXOS

1. INTRODUCCIÓN

El problema de la deserción estudiantil universitaria se expresa en todos los países del mundo. Países como España, Francia y Austria, la tasa oscila entre el 30 y 50% de deserción. En otros como Finlandia, Alemania, Países bajos y Suiza oscila entre el 7% y el 30% [2].

Durante la transición entre la educación media y la educación superior se presenta uno de los momentos críticos de la deserción, en los tres primeros semestres de formación superior es cuándo se presenta el mayor porcentaje de este fenómeno, argumentando problemas académicos y de orientación profesional [3].

Teniendo en cuenta esto, y en pro de aprovechar los datos que aportan las universidades en Colombia existe una iniciativa del Ministerio de Educación denominada SPADIES[1][4], la cual es un sistema que permite realizar seguimiento a las condiciones académicas y socioeconómicas de los estudiantes que han ingresado a la educación superior en el país [5]. Fue diseñada junto al Centro de Estudios Económicos de la Universidad de los Andes para dar seguimiento al problema de la deserción en la educación superior, calcular el riesgo de deserción de cada estudiante y clasificarlo por grupos[4].

Este conocimiento es usado en las universidades para tomar decisiones y prevenir la deserción estudiantil, la cual se debe a factores individuales, socioeconómicos, apoyos financieros y académicos e institucionales en donde el estudiante se desvincula de la IES[2] o del sistema educativo del país [6].

Dentro de los factores individuales se encuentra el aspecto vocacional, que es el factor motivacional que lleva a un individuo a elegir una carrera universitaria y la identificación de saber por qué se toma una decisión es relevante por al menos tres razones: primero, es el reconocimiento de que la elección es una acción compleja que involucra a todo lo que rodea al

[1] Sistema de Prevención y Análisis de la Deserción de Instituciones de Educación Superior
[2] Institución de Educación Superior

individuo; segundo, porque esta decisión afecta el encuentro pedagógico entre estudiantes y docentes; y tercero, porque esos aspectos motivacionales demandarán del estudiante un compromiso que puede obstaculizar o facilitar su rendimiento académico [7].

Con estos antecedentes y teniendo en cuenta que la minería de datos es definida como el proceso de descubrimiento de información útil en repositorios grandes de datos (Tan & Steinbach, n.d.), Tsai (2013) citado por [8] agrega que la minería de datos es un campo interdisciplinario que combina inteligencia artificial (IA), gestión de bases de datos, visualización de datos, aprendizaje automático, algoritmos matemáticos y estadísticos para la toma de decisiones, análisis, diagnóstico, planificación, resolución de problemas, la prevención, el aprendizaje y la innovación; en el presente trabajo se explica el proceso que se tuvo en cuenta desde el momento del entendimiento del problema hasta el momento de diseñar el modelo para el sistema de recomendación que se propone como solución. En primera instancia se presenta el planteamiento del problema donde se muestran las cifras y posibles causas del estudiante que lo impulsan a abandonar un programa de educación superior, se definen los objetivos del proyecto, metodología usada para su realización, para finalmente describir el trabajo desde la documentación técnica, presentando hallazgos y resultados obtenidos, enunciando conclusiones y recomendaciones respecto a la problemática de la deserción y de la solución presentada.

2. Planteamiento del problema

2.1. Justificación

La educación es uno de los instrumentos fundamentales con los que cuenta un país para asegurar su desarrollo humano y social, y la deserción universitaria incide negativamente de forma significativa la realización de esto [9]. Existen varias razones por las cuales un estudiante abandona la universidad, sin embargo hay un grupo de causas que son comunes en la mayoría de grupos de estudiantes, como económicas, familiares o de una mala profesión a

seguir [10], esto no solo afecta a la institución, ya que la tasa de abandono es uno de los indicadores de baja calidad, sino que también al estudiante, su entorno y la sociedad, como lo menciona Velasco en su tesis sobre el Análisis de las causas de deserción universitaria, al ser desertor de la educación superior retrasa los avances socioeconómicos y tecnológicos del país [10].

Aunque el factor vocacional es una de las causas por la cual un estudiante abandona una carrera universitaria, para el 2016 este porcentaje era de un 16,7% de los estudiantes [9], en la mayoría de los estudios se tienen en cuenta factores socioeconómicos y demográficos dejando de lado las causas vocacionales particulares de cada joven.

Con los avances en la tecnología con respecto al desarrollo de aplicaciones, recolección de datos y el análisis de estos y, teniendo en cuenta la gran cantidad de datos que se almacenan cada segundo surge la pregunta de ¿Cómo a partir de procesos de minería de datos se puede disminuir la tasa de deserción estudiantil?

2.2. Problema

La deserción estudiantil ha sido estudiada durante las últimas 2 décadas, en primera instancia, para conocer los motivos que generan este fenómeno y, segundo para intentar combatir el aumento del número de estudiantes que abandonan la educación superior. Esta reducción de la tasa de abandono se ha convertido en uno de los objetivos de las universidades del mundo [11] debido a que la educación es uno de los instrumentos fundamentales con los que cuenta un país para asegurar su desarrollo humano y social [9], además que representa no solo pérdidas económicas para las instituciones, ya que la tasa de abandono es un indicador de baja calidad debido a que se entiende que la universidad no proporcionó los medios necesarios para que los estudiantes no obtuvieran el titulo esperado [12], el estudiante y su entorno familiar sino que también representa pérdida de tiempo que a

mediano plazo se convierte en retraso a la hora de iniciar una vida laboral profesional, impactando en la calidad de vida que representa esto.

En Estados Unidos los estudiantes de programas como Ciencia, Tecnología, Ingeniería y Matemáticas son especialmente vulnerables en los años iniciales de sus programas académicos; ya que más del 60% de los abandonos ocurren durante los dos primeros años [13].

En Colombia el fenómeno de la deserción universitaria no es diferente, y a pesar de que el índice ha sido estable con el tiempo, sigue siendo bastante alto con cifras preocupantes siendo este, con un 50% de deserción en promedio por semestre, esto quiere decir que por cada 2 estudiantes matriculados, solo 1 culmina el programa académico elegido [14].

A pesar del esfuerzo realizado por las universidades y el gobierno nacional para evitar la deserción universitaria, esta sigue manteniéndose durante el tiempo, en [12] se afirma que en Colombia las políticas gubernamentales están siendo dirigidas a la inclusión educativa, con incrementos del 48% de la tasa neta de matrículas, esto para el año 2013, y donde para el año 2019 fue aprobado un presupuesto de $41.4 Billones de pesos siendo el sector educativo con el presupuesto más alto [15].

No obstante, estimaciones de costos de deserción arrojan que para el año 2009 fueron de $778 mil millones de pesos [12], lo que es un desaprovechamiento de los recursos públicos, esto sin contar los costos asumidos por los estudiantes y sus familias[16].

Según [17], la condición de abandono de los estudios se asoció con el poco interés a estudiar la carrera en la cual fue admitido y a la matricula a una carrera no deseada. En esta misma investigación se encontró que el 46.6% de los estudiantes desertores lo hicieron porque no les gustaba la carrera y no llenaba sus expectativas.

Lo anterior nos lleva a que, aunque el abandono estudiantil universitario se ha estudiado desde perspectivas sociales y económicas[10][18][19], y aunque también, existen modelos de predicción de abandono [20][21] y sistemas de recomendación [3] [22] este no ha sido estudiado teniendo en cuenta los datos Saber11 y Saber Pro, usando la minería de datos como herramienta de prevención de una posible deserción a futuro por la mala elección de una carrera profesional a estudiar.

3. OBJETIVOS

3.1. Objetivo General

Desarrollar un sistema de recomendación que ayude a estudiantes de grado 11 a la elección de una carrera profesional a partir de un modelo de minería de datos y así contribuir con una estrategia que disminuya los porcentajes de deserción universitaria causada por la mala elección de esta.

3.2. Objetivos Específicos

- Caracterizar los elementos asociados al modelo, incluyendo los datos, procesos de minería de datos y exámenes aplicados.
- Realizar la recolección de los datos asociados a los estudiantes a los cuales se les va a realizar el proceso de minería de datos usando una bodega de datos; esto con el fin de tener la información centralizada y uniforme.
- Proponer un modelo de minería de datos con la información recolectada a través de las diferentes fases que van desde la extracción de datos, hasta la entrega de los resultados encontrados.
- Validar el modelo propuesto basado en minería de datos aplicándolo en un caso de estudio.

4. Alcance

El presente trabajo busca segmentar y clasificar los estudiantes de grado 11 y su elección vocacional mediante la aplicación de un algoritmo de minería de datos, usando la información recolectada desde la página de datos abiertos publicados por el ICFES.

Estos datos dependen de una estrategia basada en la explotación de los datos basada en un proceso de minería de datos la cual incluye un preprocesamiento de datos,

entrenamiento del sistema y posterior verificación, debido a esto no se pueden tener en cuenta todos los algoritmos.

Finalmente, el resultado de los datos explotados en el presente trabajo, tienen un propósito analítico y propositivo, es decir, el resultado será un modelo aplicado a un prototipo de herramienta que permita la explotación de datos sistemático para la toma de decisiones.

5. MARCO REFERENCIAL

5.1. MARCO CONTEXTUAL

La deserción estudiantil universitaria es un asunto que se evidencia en todos los países del mundo, dando lugar a estudios que pretenden encontrar las posibles causas internas y externas que influyen en el estudiante al momento de tomar esta decisión.

Existe un alto porcentaje de abandono universitario alrededor del mundo, según un estudio presentado en el informe Education at a Glance en el año 2016, en los países pertenecientes a la OCDE (Organización para la Cooperación y el Desarrollo Económicos) alcanzan una taza de abandono del 31%, en el caso del Espacio Europeo de Educación Superior (EEES), la tasa de deserción varia de un 20% a un 55% mientras que en Latinoamérica las tasas de deserción van de un 8% a un 48% y finalmente en Colombia durante el año 2016 la tasa de deserción es del 48.8% [23].

Para realizar un seguimiento a las cifras de deserción universitaria en Colombia, el Ministerio de educación nacional en el año 2002 puso en marcha el Sistema para la Prevención de la Deserción en las Instituciones de Educación Superior (i) que es una herramienta para medir y monitorear los factores determinantes de la deserción, conocer su evolución en el tiempo y ver cómo se comportan las diferentes instituciones y regiones.[24]

Existen trabajos relacionados a la deserción universitaria que tratan sobre sistemas de recomendación en educación [22], uso de Deep learning [21], minería de datos [20], entre otros. En [22] se muestra un sistema de recomendación donde tienen en cuenta los datos demográficos recolectados en el momento de la admisión a la universidad, los datos del estudiante, los datos de la institución educativa donde realizó el estudiante el bachillerato y la información de los exámenes de estado disponibles en el Instituto Colombiano para el Fomento de la Educación Superior (ICFES). En [25] donde también tienen en cuenta los datos recolectados provenientes del ICFES, con la diferencia que utilizan datos históricos académicos y utilizan Procesamiento

de Lenguaje Natural (PLN) así como algoritmos de redes neuronales, este proyecto contiene 3 requerimientos funcionales, recomendación basado en las características, recomendación basada en texto abierto y generar una explicación del texto abierto. Para [3] el sistema planteado tiene un enfoque de desarrollo de software, sin embargo en el proceso de minería de datos solamente tienen en cuenta los recolectados al interior de la institución educativa a la cual hace referencia el documento, además de, tener en cuenta 2 carreras como son ingeniería de sistemas y Psicología. Otro campo abordado en los trabajos de investigación recolectados es el de Deep Learning mostrado en [21], donde muestra como a partir de datos recolectados a través de la misma institución usan herramientas y plataformas en la nube para realizar este proceso.

En estas revisiones se evidencian las diferentes formas en que es posible abordar el tema en cuestión, predominando el análisis técnico, algoritmos y las técnicas. Si bien estos temas imperan, el tener en cuenta los intereses vocacionales, fortalezas académicas que el individuo pueda tener y que se evidencian en el examen de estado Saber11 y poder establecer similitudes con estudiantes que están por finalizar sus estudios universitarios usando los resultados de los exámenes de estado Saber Pro para lograr una orientación vocacional como lo menciona [26] en su trabajo de búsqueda temática digital sobre orientación vocacional, que es el enfoque del presente trabajo.

5.2. MARCO CONCEPTUAL

5.2.1. Deserción

La deserción estudiantil ha sido estudiada durante décadas con el objetivo de disminuir su porcentaje y elevar la calidad en la educación, ya que esto quiere decir que la institución educativa proporcionó los medios necesarios para que los estudiantes, que lo intentaron, obtuvieron el título esperado.

Para Himmel, citado en el trabajo de [27]. Define deserción como el abandono prematuro de un programa de estudios antes de alcanzar el titulo o grado, y considera un tiempo suficientemente largo como para descartar la posibilidad de que el estudiante se reincorpore.

Dentro de las múltiples definiciones y discusiones que existen sobre deserción se encuentra un consenso en precisarla como un abandono que puede ser explicado por diferentes categorías de variables socioeconómicas, individuales, institucionales y académicas [9].

Dentro de las variables socioeconómicas se encuentran bajos ingresos personales y familiares, cambios sociodemográficos, periferia de la universidad, ausencia de actividades recreativas y de interacción.

Como factores individuales se encuentra la baja escolaridad de los padres, motivacionales, emocionales, problemas de salud, edad, ausencia de disciplina académica, influencia de la familia u otros grupos primarios, rebeldía hacia las figuras de autoridad, metas inciertas, tendencia a la depresión, apatía, conflictos familiares, etc.

Institucionales: métodos pedagógicos deficientes, falta de apoyos didácticos, cambio de institución educativa, vivienda ubicada lejos de la universidad, influencias ejercidas por profesores y por otros centros educativos.

Académicas: baja aptitud intelectual, deficiente formación previa y deficiente orientación vocacional. [10]

5.2.2. Orientación Vocacional

Según la Real Academia de la Lengua Española (RAE por sus siglas en español) la vocación es la inclinación a un estado, una profesión o una carrera [28].

Teniendo esto como base puede entenderse como un proceso que dé ayuda a la elección de una profesión, la preparación para ella, el acceso al ejercicio de esta y la evolución y progreso posterior. El objetivo de este proceso es el conocimiento de sí mismo, de las ofertas capacitantes y académicas, de las propuestas de trabajo, de las competencias que debe desarrollar para alcanzar un buen desempeño en estas propuestas, esto permitirá tomar las decisiones que considere de acuerdo con sus capacidades y aptitudes para ubicarse en el contexto social-laboral [26].

Dentro de este ámbito la teoría de los tipos de personalidad vocacional y ambientes laborales de Holland, es considerada como una de las más influyentes en la psicología vocacional [29].

Holland plantea que, en la cultura estadounidense, es posible categorizar a las personas en uno o en varios de los siguientes tipos de personalidad (R) Realista, (I) Investigador, (A) Artístico, (S)Social, (E)Emprendedor y (C) Convencional. Asimismo, los ambientes de trabajo también se pueden describir, por sus semejanzas, en uno o en una combinación de los tipos RIASEC. Según este autor la personalidad y el ambiente de trabajo interaccionan entre sí influyendo de manera bidireccional [29].

Para la orientación vocacional es importante tener en cuenta la edad, esto obedece a la teoría de Ginzerg, Ginsburg, Axelard y Herma mencionada en el trabajo [30], donde marcan que el proceso vocacional es un proceso irreversible que ocurre en periodos marcados donde están implicados 4 factores significativos: La Realidad, la influencia del proceso educativo, los factores emocionales del sujeto y los valores que posee.

Los periodos de este proceso son:

- Periodo de fantasía: hasta los 11 años, donde los niños ignoran sus habilidades
- Periodo tentativo: de los 12 hasta los 18 años, donde conocen sus intereses capacidades y valores

Periodo realista: desde los 18 a los 24 años de edad, que es donde se selecciona un camino que permita seguir con los intereses del sujeto [30]

5.2.3. Minería de datos

La minería de datos es un paso particular en el proceso de KDD (Knowledge Discovery in Databases) el cual consiste en la aplicación de algoritmos específicos para extraer patrones de los datos [31].

Dentro de los modelos de la minería de datos se encuentran los descriptivos y los predictivos. Entre los modelos predictivos se encuentran los árboles de decisión, la regresión lineal, redes neuronales, el algoritmo de Naive Bayes, Random Forest, K-NN entre otros; mientras que en los modelos descriptivos se encuentran el Agrupamiento (clustering), Reglas de asociación, Reglas de asociación secuenciales y las correlaciones [32].

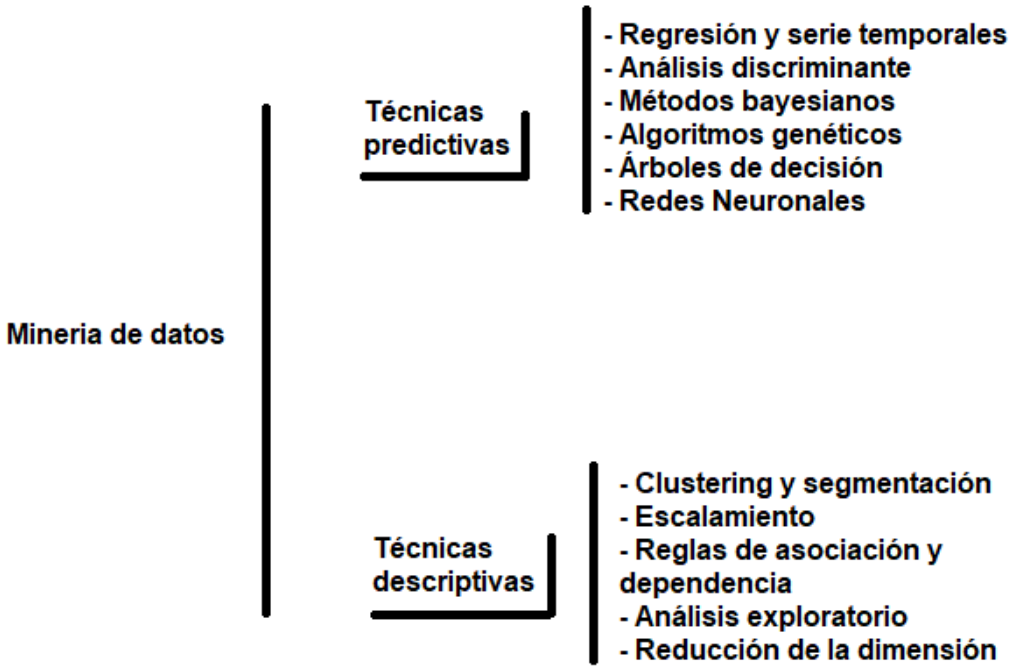


Ilustración 1: Tareas de la minería de datos, elaboración propia

Random Forest: Es un algoritmo de clasificación que consiste en la combinación de árboles predictores, en la que cada árbol depende de los valores de un vector aleatorio probado independientemente y con la misma distribución para cada uno [33].

K-NN: Es un algoritmo que admite el intercambio de la función de proximidad, esta función de proximidad puede decidir la clasificación de un nuevo ejemplo atendiendo a la clasificación del ejemplo o de la mayoría de los k ejemplos más cercanos [34]. En otras palabras, el algoritmo K-NN clasifica un nuevo dato en un grupo dependiendo que tan cerca esté de sus k vecinos más cercanos, donde k es un valor entero definido.

Naive Bayes: Este algoritmo clasifica un nuevo ejemplo de acuerdo con el valor más probable dados los valores de sus atributos, además asumen que el efecto de un valor del atributo en una clase dada es independiente de los valores de los otros atributos [34].

Árboles de decisión: Este modelo aprende una serie de preguntas para deducir las etiquetas de clase de las muestras. En otras palabras, es una descomposición de los datos mediante la toma de decisiones basada en la formulación de una serie de preguntas. La estructura de un árbol de decisión está compuesta por una raíz y los datos se van dividiendo en la característica que resulta en la mayor Ganancia de la información (IG) y esto se repite de manera iterativa hasta que las hojas sean puras [35].

5.2.4. Fases de la minería de datos

Además de los algoritmos existen unas tareas previas a realizar antes de hacer el modelo de minería de datos, estas tareas dependen de la metodología a seguir, sin embargo, se pueden agrupar en las siguientes fases:

- Entendimiento del negocio: Esta agrupa las tareas de comprensión de los objetivos y requisitos del proyecto desde la perspectiva empresarial o institucional, con el fin de convertirlos en objetivos técnicos y en un plan de proyecto.
- Comprensión de los datos: Esta fase comprende la recolección inicial de los datos, donde se establece un primer contacto con el problema, identificar la calidad de los datos y establecer las relaciones más evidentes que permitan definir las hipótesis iniciales.
- Preparación de los datos: En esta fase, luego de haber recolectado los datos, se preparan para adaptarlos a las técnicas de minería de datos que se utilicen posteriormente. Esta preparación incluye tareas de selección de datos, limpieza de datos, generación de variables adicionales, integración entre diferentes orígenes de datos y cambios de formato.
- Modelado: En la generación del modelo depende de las características de los datos y de la precisión que se quiera lograr con el modelo. Para esto se debe seleccionar la técnica de modelado, donde se debe elegir la más apropiada para resolver el problema

y se debe considerar el objetivo principal del proyecto y la relación con las herramientas de minería de datos existentes. Es decir, aquí es donde se debe tener en cuenta si es un problema de clasificación, de predicción o de segmentación.

- Evaluación: En esta fase se evalúa el modelo, teniendo en cuenta el cumplimiento de los criterios de éxito del problema. Para lograr esto existen diferentes herramientas para la interpretación de resultados. [36]

5.2.5. Investigaciones previas

Debido a la necesidad de las instituciones educativas y gubernamentales en Colombia de conocer la tasa de deserción de los estudiantes y disminuir su porcentaje, se han realizado investigaciones y elaborado herramientas que les permiten tomar decisiones para retener a las personas dentro del sistema educativo.

5.2.5.1. Antecedente sobre el seguimiento de la deserción universitaria en Colombia

Es el caso del Centro de Estudios Económicos (CEDE[3]) y citado en la tesis de Maestría en Educación de Guzman Puentes (2009), donde por solicitud del Ministerio de educación Nacional, desarrolló en el 2004 una herramienta informática llamada SPADIES[4], la cual permite hacer el seguimiento a los estudiantes desde que ingresan a la educación superior hasta el momento de su deserción o graduación, teniendo en cuenta los diferentes factores (Institucionales, académicos, personales, financieros) relacionados con la deserción y su evolución en el tiempo [37].

Esta investigación brinda un aporte significativo porque muestra la importancia de comprender el negocio además de comprender la problemática donde, establecer estrategias para evitar la deserción universitaria en los diferentes tipos de instituciones (Técnicas, tecnológicas y universitarias) implicaría obtener un mayor reconocimiento y aceptación,

[3] El CEDE pertenece a la Universidad de los Andes
[4] Sistema de Prevención y Análisis de la Deserción en las Instituciones de Educación Superior

además de ser vista como una institución de alta calidad ya que ofrecer al estudiante una posibilidad mayor de alcanzar su objetivo y ofrecer los medios necesarios para esto, se considera a nivel general como una medida de calidad. [12]

5.2.5.2. Causas de la deserción universitaria, comparación entre el sistema educativo estadounidense y el latinoamericano

Otro de los estudios realizados con respecto a la deserción estudiantil se encuentra el de [27], donde menciona diferencias entre el sistema educativo estadounidense y el latinoamericano, y menciona que los estudios suelen hacer énfasis en las difíciles condiciones socioeconómicas de la juventud. Sin embargo, en este estudio existe una mención a la ausencia de orientación vocacional entre una de las causas de la deserción estudiantil [38].

Este trabajo nos muestra como factores internos y externos al estudiante pueden determinar el abandono de este a nivel superior, el aporte de este se centra en establecer posibles variables a la hora de realizar la investigación, centrándose en factores sociales y económicos que rodean al estudiante, también como su familia y amigos pueden influir en esta decisión, así como su rendimiento académico en el bachillerato.

5.2.5.3. Antecedentes de la orientación vocacional

En el trabajo de De la Ossa Farias (2015), presenta la orientación vocacional como uno de los principales problemas que tiene la educación superior y que se manifiesta en el abandono de los estudiantes el cual se debe atacar desde el fortalecimiento de la orientación profesional y vocacional.[9]

Este trabajo tiene un aporte importante en la investigación ya que se encuentra que uno de los factores más importantes al momento de tomar la decisión de abandonar los estudios de educación superior es la falta de orientación vocacional, el cual es la base de este trabajo.

5.2.5.4. Antecedentes de un sistema de recomendación vocacional

Uno de los más recientes es el de Orozco (2019) donde presenta un sistema de recomendación híbrido donde combina 2 estrategias, de filtrado colaborativo y la otra de filtrado por contenido. Esto significa que la primera estrategia está enfocada a recomendar programas en los que otros estudiantes han tenido un buen rendimiento académico y tienen características similares al usuario mientras que la segunda, está enfocada al procesamiento de lenguaje natural y obteniendo los conceptos asociados al programa y a través de vectores semánticos[5].[22]

Se establece la importancia de este trabajo en la evidencia del uso de la minería de datos teniendo en cuenta los datos del examen de estado Saber 11 y el procesamiento de lenguaje natural (PLN), además de brindar el primer acercamiento a las metodologías de minería de datos, en este caso ASUM-DM de IBM.

5.2.5.5. Sistemas de recomendación basado en Deep Learning y Procesamiento de lenguaje natural (PLN)

El trabajo de Gomez (2019) muestra el proceso de diseño y desarrollo de un sistema de recomendación basado en Deep Learning y Procesamiento de Lenguaje Natural, además de realizar un proceso de minería de datos a los datos del examen de estado Saber 11 recolectados de los estudiantes de la universidad de los Andes. [25]

Este trabajo es de gran valor debido a que utilizan como fuente de datos los recolectados al interior de la institución universitaria y de allí se puede extraer una metodología mixta para el desarrollo del trabajo, como es la metodología para realizar el proceso de minería de datos y el desarrollo del aplicativo final.

Con base en estas investigaciones y en la búsqueda de información se encontró que no existe hasta el momento una investigación sobre recomendación vocacional que relacione los

[5] Almacenamiento de palabras según su similitud, coocurrencia estadística[60]

datos obtenidos del examen Saber11 y los obtenidos en el examen Saber Pro realizados por el ICFES.

6. ESTRATEGIA METODOLÓGICA

En la estructuración de este trabajo de investigación, se analizaron diferentes metodologías orientadas a la minería de datos, entre las analizadas están KDD (Knowledge Discovery in Databases), CRISP-DM (Cross-Industry Standard Process for Data Mining) y SEMMA (Sample, Explore, Modify, Model and Access) las cuales son las más utilizadas para el desarrollo de un proceso de minería de datos; es por esto por lo que a partir de la revisión del trabajo [39] Titulado “A comparative study of data mining process models (KDD, CRISP-DM and SEMMA)” donde se describen los diferentes pasos de cada una de las metodologías así como las ventajas y desventajas de cada una, se elegirá una para el desarrollo de este trabajo.

En una encuesta realizada en el año 2014 por la firma Kdnuggets a 200 votantes, donde se preguntó cuál es la metodología que usaba para el desarrollo de proyectos de análisis, minería de datos y ciencia de datos, se encontró que el favorito es CRISP-DM con un 43% de preferencia, mientras que las metodologías propias tienen un 27.5% de preferencia, seguida de SEMMA con un 8.5% y en 5to lugar se encuentra KDD con un 7.5% de preferencia. Otro aspecto que llama la atención de esta encuesta es que los resultados con respecto a la preferencia de CRISP-DM y KDD no han sufrido grandes cambios con respecto a la anterior realizada en 2007, mientras que SEMMA ha perdido preferencia en un 4.5% con respecto a esta misma [40]. Estos resultados se pueden ver en la Ilustración 2.

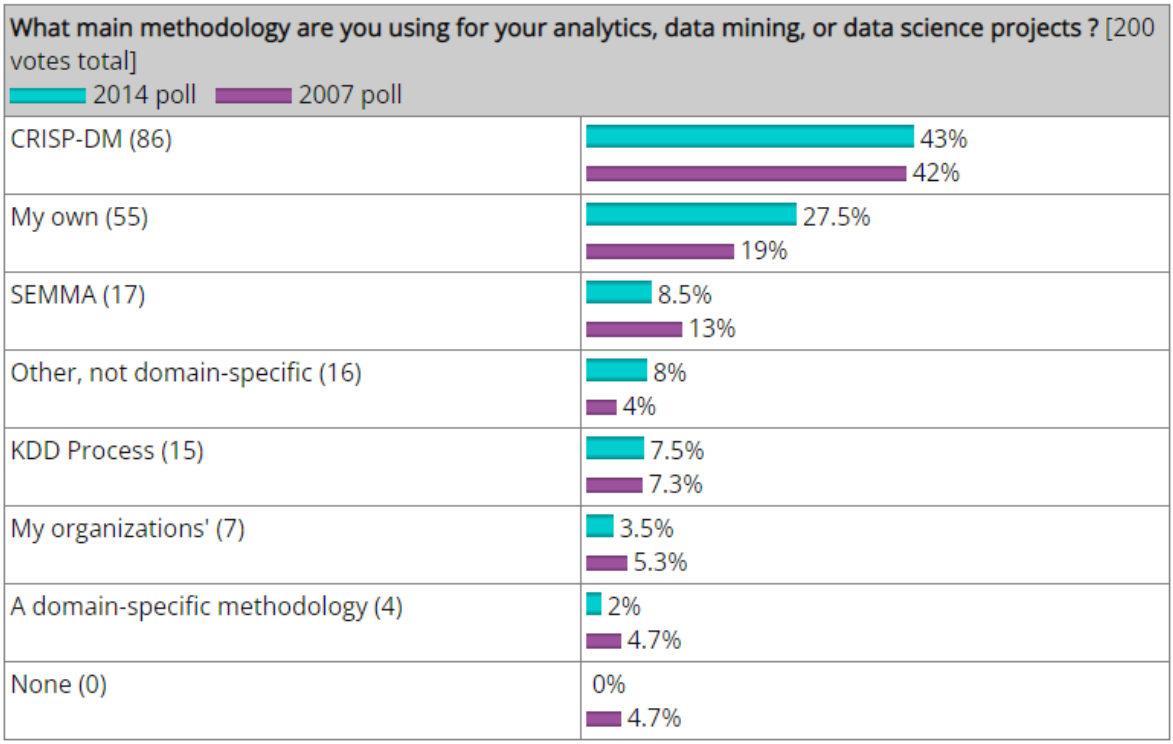


Ilustración 2: Resultados a la encuesta realizada por KdNuggets, Tomada de la página de KdNuggets

Un estudio comparativo entre KDD, SEMMA y CRISP-DM muestra las diferencias entre estas tres metodologías incluyendo sus ventajas, desventajas, fases y número de fases entre ellas. A continuación se presenta un resumen de cada una y un cuadro comparativo de estas [39].

6.1. KDD

Es el proceso de extraer el conocimiento escondido a partir de los datos. KDD requiere el conocimiento y el entendimiento del dominio de la aplicación y sus objetivos, es iterativo e interactivo de manera natural. Contiene 9 diferentes pasos y son los siguientes:

- Entendimiento del dominio de la aplicación
- Creación de un conjunto de datos objetivo
- Limpieza y preprocesamiento de datos

- Transformación de los datos
- Elección adecuada de la tarea de minería de datos
- Elección adecuada del algoritmo de minería de datos
- Implementación del algoritmo de minería de datos
- Interpretación de los patrones minados
- Uso del conocimiento encontrado

6.2. CRISP-DM

Cross-Industry Standard Process for Data Mining (CRISP-DM) ofrece un marco de trabajo y una guía para los mineros de datos. Consiste en las siguientes 6 fases:

- Entendimiento del negocio
- Entendimiento de los datos
- Preparación de los datos
- Modelamiento
- Evaluación
- Despliegue

6.3. SEMMA

Sample, Explore, Modify, Model and Access (SEMMA) es un método de minería de datos desarrollado por el instituto SAS quien ofrece el entendimiento, organización, desarrollo y mantenimiento de los proyectos de minería de datos. SEMMA está ligado a la empresa SAS de minería y básicamente una organización lógica de las herramientas para ellos y tiene 5 fases o pasos:

- Ejemplificar (Sample)
- Explorar
- Modificar
- Modelar

- Acceso

Modelos de procesos de minería de datos	KDD	CRISP-DM	SEMMA
No. De pasos	9	6	5
Nombre de los pasos	Desarrollo y entendimiento de la aplicación	Entendimiento del negocio	--
	Creación de un conjunto de datos objetivo	Entendimiento de los datos	Muestra
	Limpieza de datos y preprocesamiento		Exploración
	Transformación de datos	Preparacion de los datos	Modificación
	Elegir la tarea de minería de datos adecuada	Modelamiento	Modelado
	Empleo del algoritmo de mineria de datos		
	Interpretación de los patrones minados	Evaluación	Evaluación
	Usar el conocimiento descubierto	Despliegue	--

Ilustración 3: Cuadro comparativo de las metodologías para minería de datos, Elaboración propia

Luego de realizar la revisión y análisis de las diferentes metodologías, identificando ventajas, desventajas y características y, teniendo en cuenta la cantidad de datos recolectados, se ha decidido usar la metodología CRISP-DM (Cross Industry Standard Process For Data Mining)[41] por ser la más completa debido a que incluye el entendimiento del negocio, el cual nos permite establecer el objetivo desde el inicio, también por su filosofía iterativa, además de contar con la parte documental, donde para este trabajo es de gran importancia. Además, esta metodología, la cual se describe en términos de un proceso jerárquico, que consiste en un conjunto de tareas que describen 4 niveles de abstracción, adicional puede ser usada independiente de la herramienta tecnológica a utilizar en la exploración de datos. Esta metodología consiste en la realización de 6 pasos Ilustración 4:

- Entendimiento del negocio

Consiste en entender los objetivos y requerimientos del proyecto desde el punto de vista del negocio. Para realizar esto se debe realizar una investigación sobre los objetivos de la educación, cuáles son las variables que inciden en la medición de calidad de las instituciones y como la deserción estudiantil impacta negativamente esto.

- Entendimiento de los datos

Esta fase consiste en la recolección y familiarización de los datos. En este punto se tomarán los datos recolectados de las pruebas SABER 11 que el ICFES publica a través del portal de datos abiertos, además realizar una serie de datos simulados con respecto a los diferentes enfoques que una prueba pueda presentar analizando que datos se tendrán en cuenta para su almacenamiento y preprocesamiento y análisis.

- Preparación de datos

Son las actividades necesarias para construir el subconjunto de datos (data set) desde los datos originales, sin procesar. Con los datos recolectados se realiza el preprocesamiento de estos, significa que se hará un proceso de ETL (Extraer, Transformar y Limpiar) los datos para su posterior modelado.

- Modelado

Selección y aplicación de la técnica de minería de datos que servirá para obtener un modelo para presentar el conocimiento. En esta fase se evaluarán los diferentes algoritmos tales como Redes neuronales, Naive Bayes, árboles de decisión, regresión lineal, etc. y se elegirá el que de acuerdo con las necesidades se usará para el desarrollo del proyecto.

- Evaluación

Consiste en la revisión de los pasos ejecutados en la construcción del modelo asegurándose que este alcanza los objetivos del negocio para la toma de decisiones. Una vez hecho el modelado y presentación de los datos, se procederá a la realización de la propuesta de un recomendador vocacional realizando una prueba con una serie de datos elegidos del set de datos original.

- Desarrollo

Es aquí donde se define si tan solo se genera un reporte o si se implementa un proceso de minería de datos repetible en un centro de educación media. [42]

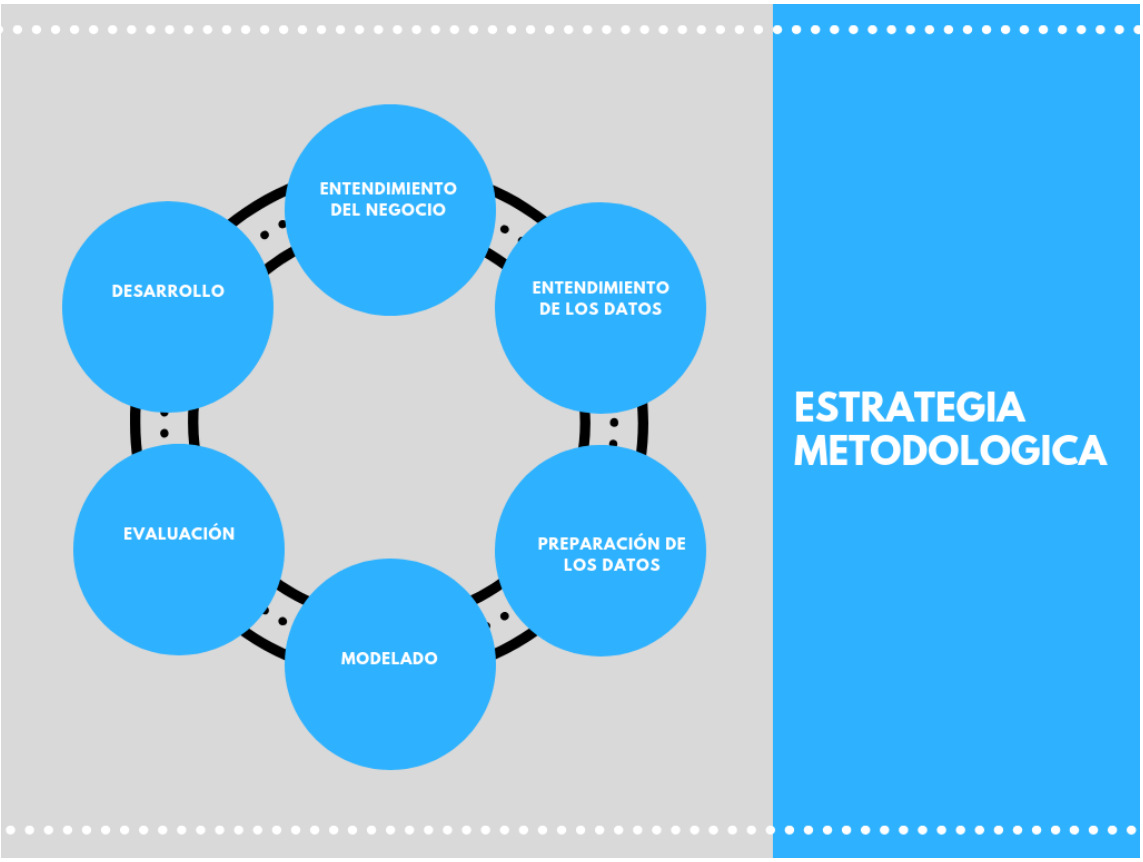


Ilustración 4: Estrategia Metodológica Elaboración propia [43]

7. PRESENTACIÓN Y ANÁLISIS DE RESULTADOS: APLICACIÓN DE LA METODOLOGÍA CRISP-DM

En la guía de referencia “Step-by-step data mining guide”[44] mencionan de forma detallada las tareas de cada fase y como llevar a cabo un proyecto de minería de datos. A continuación, se presenta cada una de ellas con sus respectivos resultados, iniciando con la comprensión del negocio y hasta el modelado del proyecto.

7.1. COMPRENSION DEL NEGOCIO

Esta primera fase lo que se busca es conocer lo que el cliente quiere lograr, para esto, se deben identificar las variables “objetivo” las cuales son las que determinan el éxito del proyecto, adicional se identifican los orígenes de los datos obtenidos, como son los puntajes históricos de las pruebas Saber11 y Saber Pro de todo el país, así como también obtener los recursos de hardware y de software necesarios para iniciar la elaboración de este trabajo de tesis.

7.1.1. Objetivos del negocio

Así como menciona [43] el objetivo de esta fase es entender lo que el cliente quiere alcanzar desde una perspectiva de negocio. Al final de esta fase se debe obtener la situación actual del problema, los objetivos del negocio y cuáles son los criterios del éxito del proyecto.

El objetivo del negocio se definió a partir del análisis del estado del arte de los sistemas de recomendación para estudiantes, donde se evidencia la necesidad de guiar a los estudiantes en esta importante elección; con esto dicho el objetivo principal es guiar a estudiantes de educación media en la selección de una carrera universitaria o que quieran continuar con su educación superior, a través de un modelo de clasificación y predicción de sus posibles intereses según los puntajes obtenidos en el examen de estado Saber11.

7.1.2. Valoración de la situación

Actualmente un 16% de los estudiantes abandona la carrera universitaria por factores vocacionales, La mayoría de los estudios tienen en cuenta aspectos socioeconómicos y demográficos, una de estas herramientas llamada SPADIES, la cual permite realizar el seguimiento de las condiciones académicas y socioeconómicas de los estudiantes que han ingresado a la educación superior, según las cifras de esta herramienta actualmente el 50% de los estudiantes por semestre desertan, esto quiere decir, que de cada 2 estudiantes 1 finaliza su proceso educativo.

Para evaluar la situación es necesario encontrar los recursos, limitaciones, supuestos y factores que pueden considerarse determinantes al momento del desarrollo del proyecto. En primera instancia están los recursos de software, entre los cuales están aquellos programas diseñados para el almacenamiento de la información como el motor de base de datos MySql, el cual es un motor de base de datos relacionales de código abierto y gratuito; además de ser usado para el almacenamiento de datos, también se utilizó para la administración, proceso de limpieza y normalización de los datos. La selección de este software va orientada a la facilidad de integración con lenguajes de programación orientados a web, además de ofrecer una documentación amplia alimentada por la propia comunidad, experiencia por parte del autor de la investigación en el uso de este y disminuir los tiempos de realización de la investigación y así evitar incurrir en tiempo adicional en la curva de aprendizaje en cualquier otro motor de base de datos. Dentro del software utilizado se encuentra la aplicación Rapid Miner, el cual es un programa de código abierto para los sistemas operativos Windows, Linux y MacOs donde se pueden realizar procesos de minería de texto y de datos. Este software fue usado para realizar la analítica descriptiva de las fuentes de datos recopiladas y todo lo concerniente a la experimentación y modelado del proceso de minería de datos. La razón para ser seleccionado es la facilidad de realizar estos procesos, además que su curva de aprendizaje no es pronunciada y con su licencia educativa permite tener durante 1 año una licencia que permite

filas ilimitadas a la hora de usar los datos recopilados en la base de datos. Para este propósito también se usó el programa Weka, que consiste en una colección de algoritmos de machine learning de código abierto bajo licencia Publica General (GNU) para llevar a cabo tareas de minería de datos. Contiene herramientas de preparación de datos, clasificación, regresión, agrupamiento, reglas de asociación y visualización. Este software fue usado para realizar experimentación inicial, así como la analítica descriptiva de las fuentes de datos recopiladas. La razón de haberlo seleccionado es la facilidad de uso, además de ofrecer uso ilimitado en los datos que se pueden usar para estos procesos. Por último, tenemos los recursos de hardware, donde se usó un equipo de cómputo de escritorio con sistema operativo Windows 10, procesador Intel Core i7 de 4 núcleos, 16GB de memoria RAM, disco de estado Solido de 222GB y un disco mecánico de 1TB. Este computador fue usado para el almacenamiento, preprocesamiento, limpieza de los datos, además fue usado para realizar la visualización descriptiva de los datos, modelamiento, entrenamiento y validación de modelos.

En cuanto a los datos disponibles, se encuentra la información de los estudiantes en diferentes fuentes, la fuente principal para este trabajo se encuentra de manera pública en la página del Instituto Colombiano para el Fomento de la Educación Superior (ICFES), dichos datos se encuentran protegidos por la ley 1581 de 2012 de protección de datos personales, donde según el artículo 10, no es necesaria su autorización debido a que son datos de naturaleza pública al encontrarse almacenados en una entidad pública como lo es el ICFES [45].

7.1.3. Determinación de los objetivos de la minería de datos

El objetivo principal de la minería de datos para este proyecto es predecir el top 3 de los programas académicos donde el estudiante puede desempeñarse, a través de la clasificación de estudiantes que ya están finalizando su ciclo de pregrado y que ya cuentan con un registro

en el examen de estado Saber Pro que tuvieron un rendimiento similar en las pruebas Saber 11.

8. COMPRENSION DE LOS DATOS

Para tener un contexto del alcance del presente trabajo, es importante mostrar las características de los datos recopilados en cuanto a integridad y calidad, así como los hallazgos importantes.

Entender los datos, para la metodología CRISP-DM, está destinado para registrar los problemas asociados a la adquisición de estos y como fue solucionado esto, esto ayudará a futuras posibles réplicas del proyecto o con una nueva ejecución de un proyecto futuro similar.

Las tareas aplicadas son:

- Recolección de los datos: De esta tarea se pretende conseguir el listado de los diferentes sets de datos adquiridos, junto con sus localizaciones, los métodos usados y los problemas encontrados. También el de registrar los problemas encontrados y sus soluciones.
- Descripción de los datos: Examina lo superficial de los datos adquiridos y reportarlo en los resultados, aquí se describe los datos incluyendo el formato, la cantidad de datos, la identidad de los campos y cualquier otra característica que haya sido descubierta. Evalúa que los datos adquiridos satisfagan los requerimientos relevantes.
- Exploración de los datos: Esta tarea direcciona las preguntas de minería de datos usando consultas, visualizaciones y técnicas de reporteo. Se describen los resultados incluyendo los primeros hallazgos o la hipótesis inicial y su impacto en el proyecto y, si es apropiado, se incluyen gráficos para indicar las características que sugieren un examen más profundo de subconjuntos de datos.

- Verificación de la calidad de los datos: Se examina la calidad de los datos, direccionando hacia preguntas como: los datos están completos, contienen errores, que tan comunes son los errores, hay valores incompletos en los datos, como están representados, donde ocurren y que tan común son. Su objetivo es listar los resultados de la verificación de la calidad de los datos, si hay problemas cuáles son sus posibles soluciones. En este punto se pueden encontrar tipos de problemas como que los datos faltantes incluyen valores que están en blanco, codificados con el número cero (0) o como no-response (null, “?” o 9999), también se pueden encontrar errores tipográficos, errores de medición donde los datos están ingresados correctamente, pero con un esquema de medición incorrecto. Incoherencias de codificación o unificación en unidades de medidas (uso de F y de femenino para el género).

8.1. Recolección de datos iniciales

La fuente de datos en este caso es a través del portal del ICFES, donde reposa la información de manera pública, donde a través de la creación de una cuenta se tiene acceso a estos. Y se dividen de acuerdo con el año de presentación del examen, es decir, los archivos se encuentran separados por año y semestre, siendo que para el año 2010 existen los archivos SB11_20101 y SB11_20102. Además, se encuentran divididos por el tipo de prueba presentada, donde SB11 es Saber11 y SaberPro_Genericas_20101 son los resultados de las pruebas Saber pro Genéricas, donde genéricas son las materias comunes evaluadas en el examen. Se tiene un 3er set de datos llamado llaves, donde se encuentra almacenada la información de los códigos de estudiante almacenado en el examen Saber 11 y su par en el examen Saber Pro, es con este archivo se logra emparejar los resultados del estudiante que presento la prueba Saber 11 con la prueba Saber pro (si ya la presentó) [46].

8.2. Descripción de los datos

En cada uno de los datos descargados se encuentra un diccionario de datos donde se puede ver el significado de cada columna. En cuanto a la cantidad de datos en el archivo de las pruebas saber 11 tenemos 1.229.724 registros, los cuales tienen 82 atributos descritos en el Anexo 1.

Por parte de los datos Saber Pro se cuenta con 2.927.405 datos y 169 columnas tal como se muestra en el Anexo 2.

Por parte de los datos de las llaves se cuenta con 1.910.470 datos y 2 columnas como se muestra en la Tabla 1:

Estu_Consecutivo_11	Código del estudiante que se asignó en el momento de presentar la prueba Saber 11
Estu_Consecutivo_PRO	Código del estudiante que se asignó en el momento de presentar la prueba Saber PRO

Tabla 1: Diccionario de datos tabla llaves, tomado del sitio de Sharepoint del ICFES

Para realizar estos cruces el ICFES, a partir del año 2016 incluyó en el cuestionario de la prueba Saber PRO, una pregunta relacionada con el tipo de documento de identidad y el número que usó al momento de presentar la prueba Saber 11, con esto se realizó lo siguiente:

- Cruce entre variable de Saber Pro que indica el documento que usó el estudiante al momento de presentar Saber 11°" y la variable documento registrada en Saber 11.
- Cruce entre las variables de documentos de identidad reportadas en el momento de registrarse en Saber Pro y Saber 11.
- Cruce fonético entre los estudiantes de Saber Pro a los que no se les encontró su registro de Saber 11 con los pasos anteriores y los estudiantes registrados en Saber

11°. Este cruce fonético tiene en cuenta nombres, apellidos, género y fecha de nacimiento. [47]

8.3. Exploración de los datos

El objetivo de esta fase es encontrar una estructura general para los datos. Se aplican pruebas estadísticas básicas donde se revelen propiedades de los datos adquiridos.

Inicialmente se cuenta con un total de 1.229.724 de registros de exámenes Saber11 que son el histórico de las personas que han presentado el examen del ICFES o Saber 11, también se cuenta con 2.927.405 registros de exámenes Saber Pro correspondientes a estudiantes que están finalizando su carrera universitaria.

Para el caso de las pruebas Saber 11 la distribución entre los diferentes niveles de desempeño se encuentra descrita en la Tabla 2 y Tabla 3 para la materia inglés:

Escala Desempeño	Biología	Lectura Critica	Matemáticas	Ciencias sociales y ciudadanas
0	0	0	102	0
1	259006	167115	123041	423404
2	657082	492805	555666	483837
3	277414	472229	485271	286386
4	36222	97575	65644	36097

Tabla 2: Distribución de los datos por desempeño, Elaboración propia

Nivel Inglés	Cantidad
A-	610423
A1	370321
A2	139188
B+	41195
B1	68597

Tabla 3: Distribución de los datos por desempeño inglés, Elaboración propia

El primer hallazgo es en la prueba de Matemáticas que hay 102 registros que tienen como desempeño cero en Matemáticas, por lo tanto, estos registros pueden ser candidatos para no ser tenidos en cuenta o imputarles un valor de desempeño de acuerdo con la media.

Con la clasificación de desempeño de cada una de las materias se obtienen resultados como los que se ven en la Ilustración 5 la cuál vemos como la mayoría de los resultados se encuentran en el nivel de desempeño 2, mientras que la menor cantidad está en el nivel de desempeño 4.

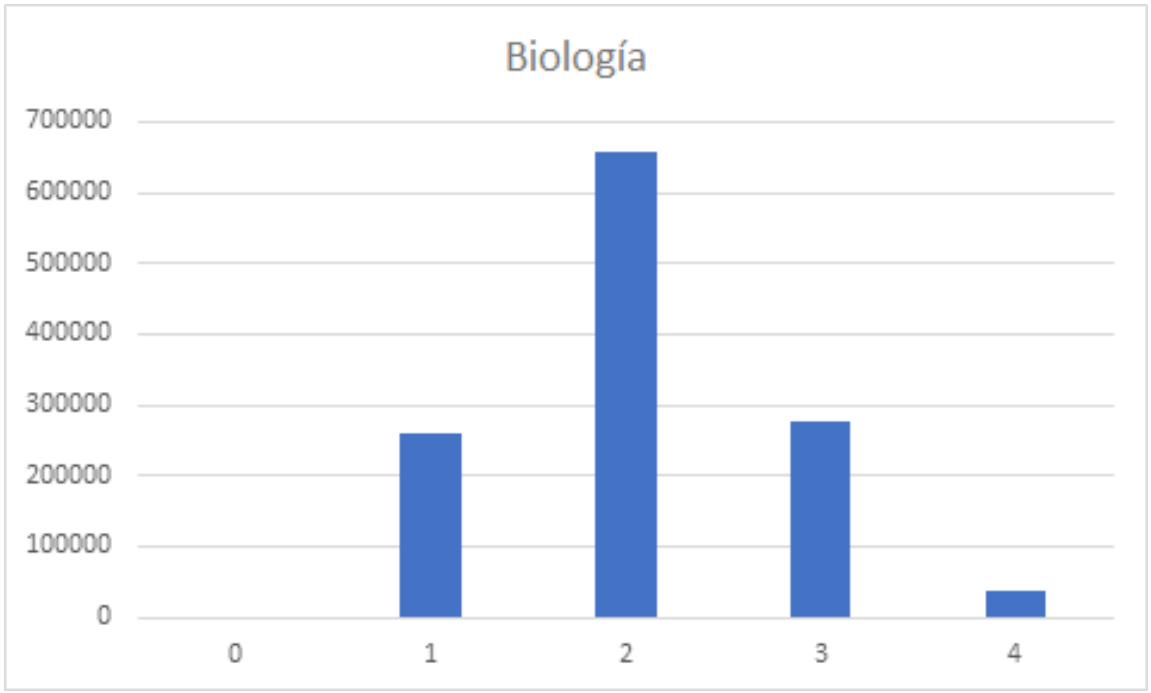


Ilustración 5: Grafico de distribución de desempeño Biología Elaboración propia

En la Ilustración 6 se evidencia como el nivel de desempeño 2 y 3 son similares, mientras que en el nivel de desempeño 1 hay menor cantidad con respecto a Biología, pero aumenta en el nivel 4 con respecto a la misma materia.

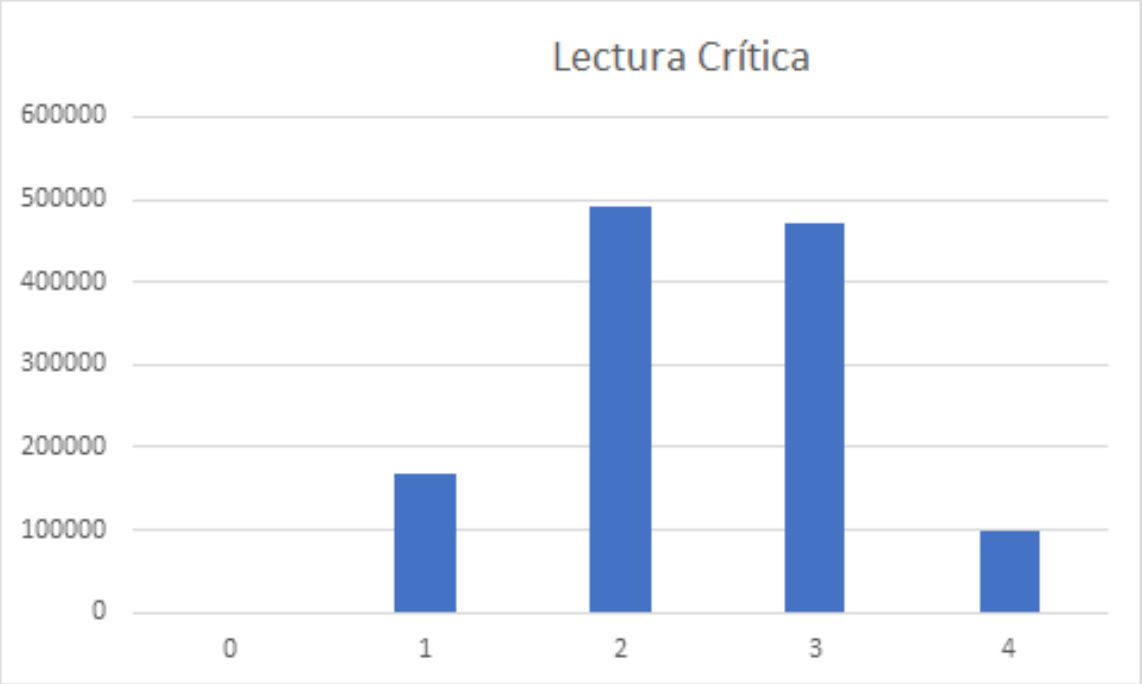


Ilustración 6: Grafico de distribución de desempeño Lectura Crítica Elaboración propia

En la Ilustración 7 se evidencia como aparecen datos en el nivel cero (0) el cual lo muestra debido que existen estudiantes que no tienen este puntaje registrado. También se evidencia que al igual que las dos materias anteriores el mayor número de estudiantes se ubica en el nivel 2, y la minoría se ubica en el nivel 4.

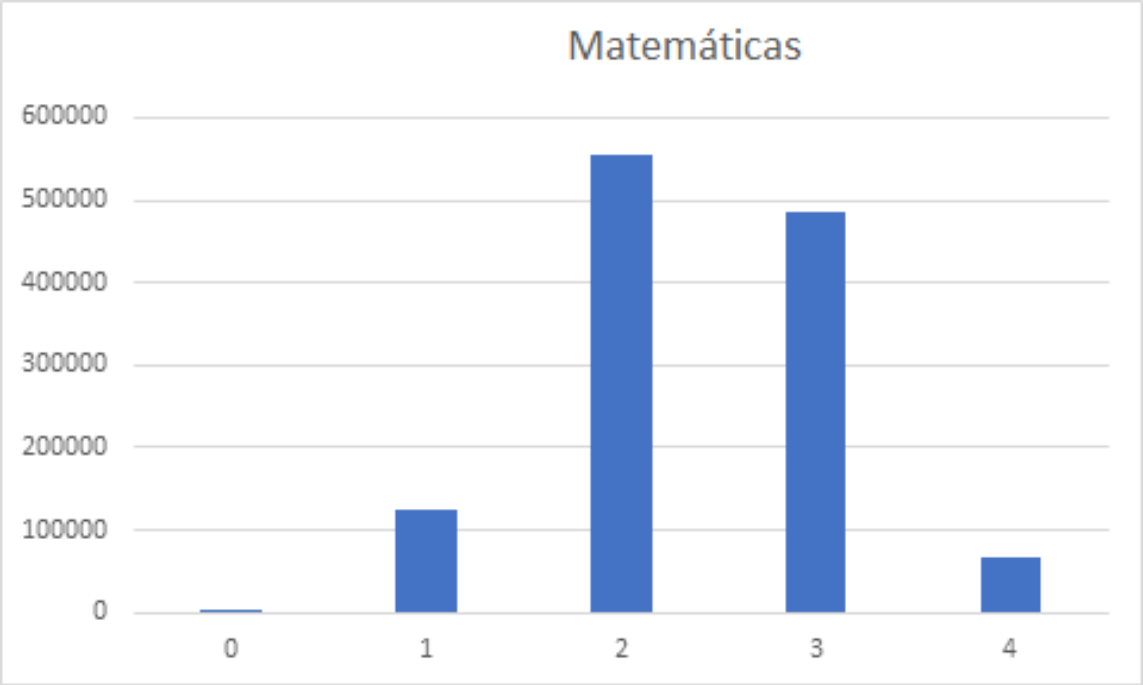


Ilustración 7: Grafico de distribución de desempeño Matemáticas Elaboración propia

En la Ilustración 8 hay un cambio de comportamiento en cuanto al nivel 1, ya que se acerca bastante al nivel 2, sin embargo, este último sigue siendo el que tiene más concentración de estudiantes.

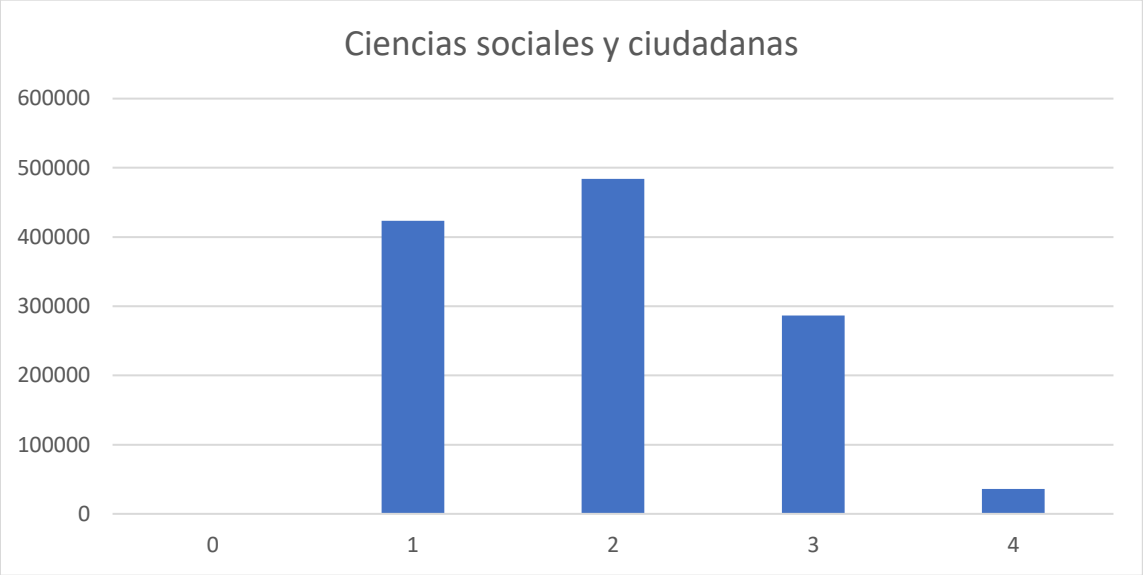


Ilustración 8: Grafico de distribución de desempeño Ciencias sociales y ciudadanas Elaboración propia

La Ilustración 9 muestra como en los diferentes niveles de inglés el que tiene mayor número de estudiantes es el nivel A- mientras que el que representa una minoría es el nivel B+ el cual es considerado un nivel muy superior para la edad escolar.

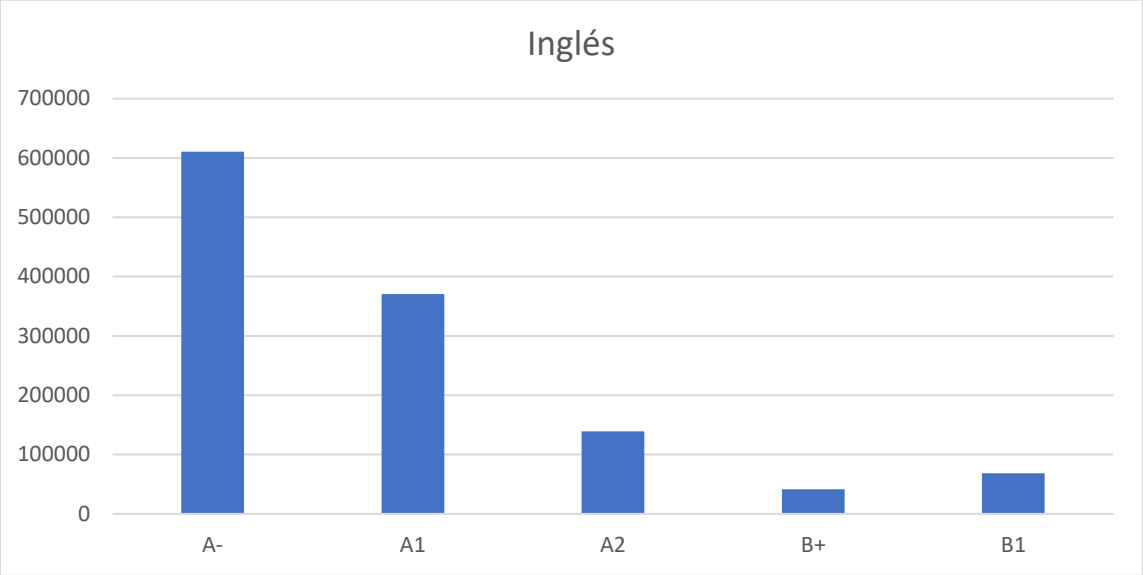


Ilustración 9: Grafico de distribución de desempeño Inglés Elaboración propia

Para este apartado entre los hallazgos más sobresalientes se encuentran el observar que el nivel predominante en todas las pruebas es el número 2, mientras que los niveles más altos presentan un número de estudiantes muy inferior.

Con respecto a los datos individuales y puntajes se encontró que las materias por si solas tienen datos esperados, los cuales se muestran en los Anexo 3, Anexo 4, Anexo 5, Anexo 6, Anexo 7, Anexo 8, Anexo 9, Anexo 10, Anexo 11 y Anexo 12.

Genero

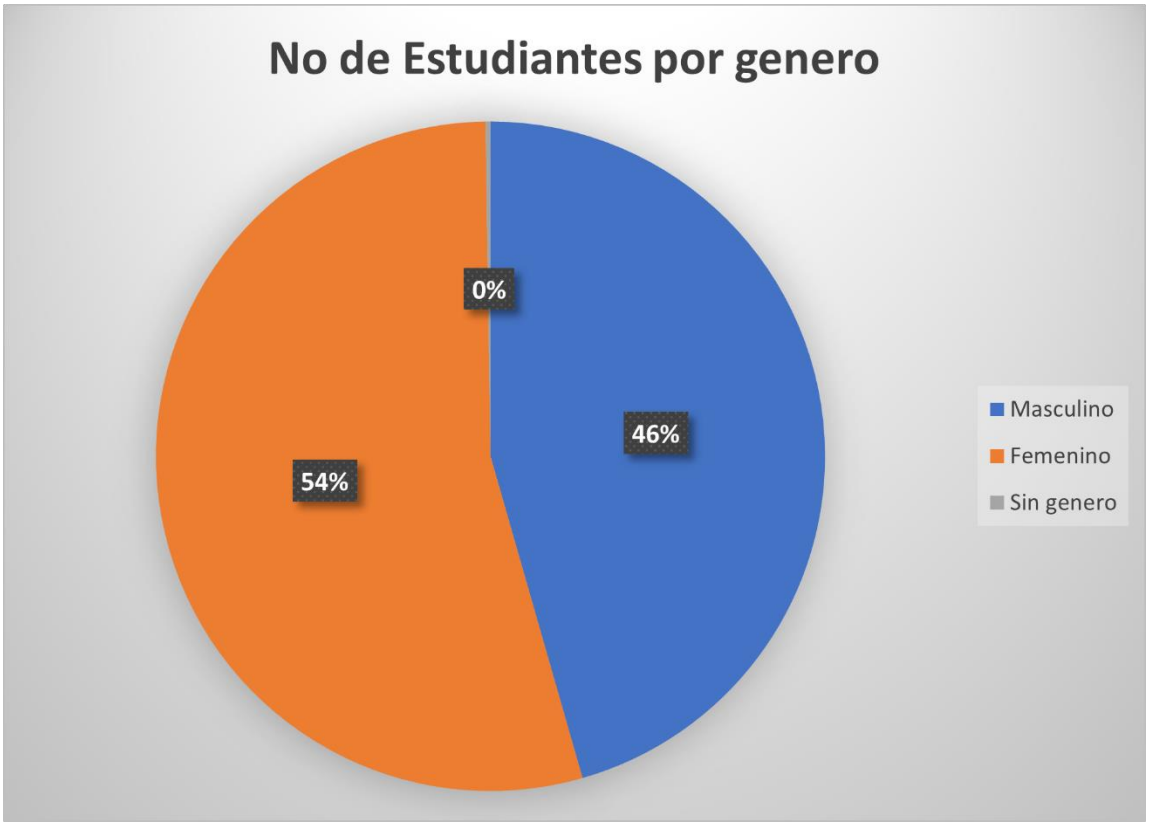


Ilustración 10: Distribución de estudiantes por género Elaboración propia

En las gráficas concernientes al género (Ilustración 10 e Ilustración 11) se puede ver que la distribución entre los dos sexos es similar, sin embargo, el número de mujeres es superior que al de los hombres. También se puede ver que existe un valor que no contiene ninguno de los dos géneros, esto es porque para estos estudiantes el valor de género es inexistente.

No.	Label	Count	Weight
1	M	560294	560294.0
2	F	666483	666483.0
3	-	2947	2947.0

Ilustración 11: Tabla de distribución de estudiantes por género Elaboración propia

Datos Saber Pro

En el caso de los datos del examen Saber Pro se tiene que grupo de referencia que más registros tiene es Ingeniería con 6.678 registros, le sigue Administración y Afines con 4.381, Derecho Psicología y Educación. Los grupos de referencia con menos registros son los Técnicos y tecnológicos. Aquí también se puede ver que los datos se encuentran desbalanceados, donde de los 37 grupos de referencia, Ingeniería tiene mayor cantidad de registros, en la presente investigación se quiere evitar recomendaciones del programa más popular o con más registros.

Grupo Referencia	Cantidad de Registros
INGENIERIA	6678
ADMINISTRACION Y AFINES	4381
DERECHO	2722
PSICOLOGIA	2276
EDUCACION	1755
CONTADURIA Y AFINES	1254
SALUD	1147
COMUNICACION, PERIODISMO Y PUBLICIDAD	1001
ARQUITECTURA Y URBANISMO	847
BELLAS ARTES Y DISEÑO	835

Tabla 4: Grupos de Referencia con más registros Elaboración propia

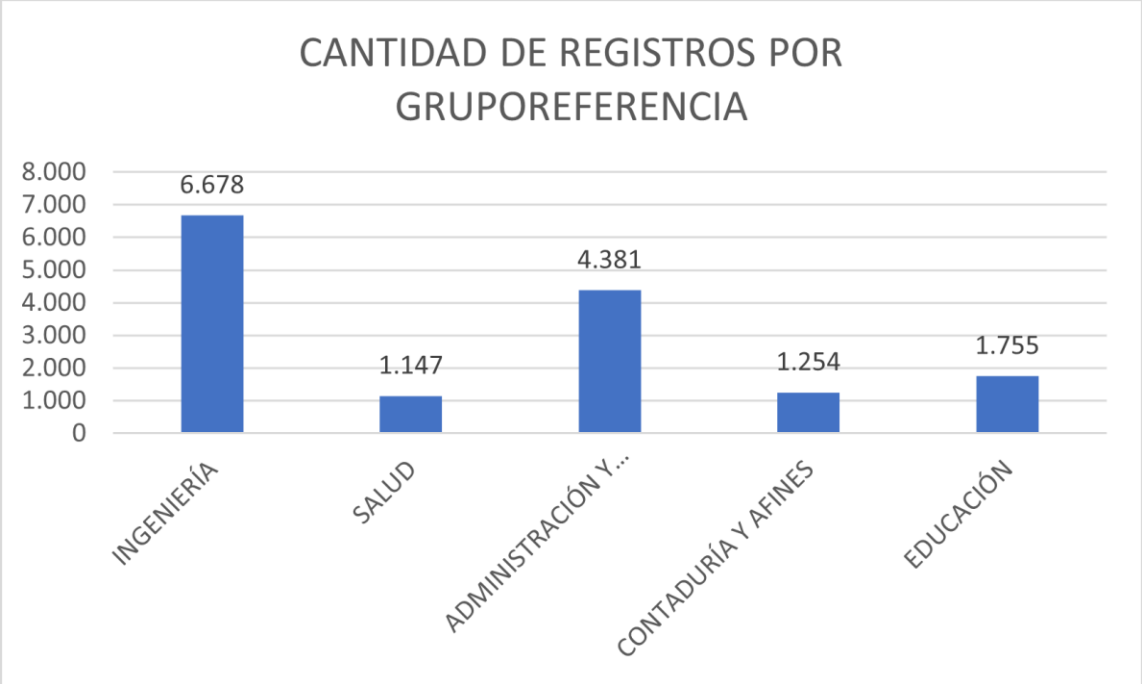


Ilustración 12: Registros de estudiantes por grupo de referencia elaboración propia

En la Ilustración 12 se muestra como el grupo de referencia que más agrupa estudiantes es el de ingeniería, mientras que las Ciencias militares y navales es la que menos concentración de estudiantes tiene.

En estos mismos datos y luego de realizar una asociación entre los puntajes obtenidos en el examen Saber 11 y Saber Pro, se visualiza que los puntajes de lectura crítica y Sociales y ciudadanas con respecto a estudiantes de Medicina son bajos, donde se tiene que el puntaje obtenido en promedio es de 14 puntos en ambas materias como se muestra en la Ilustración 13 y Ilustración 14.

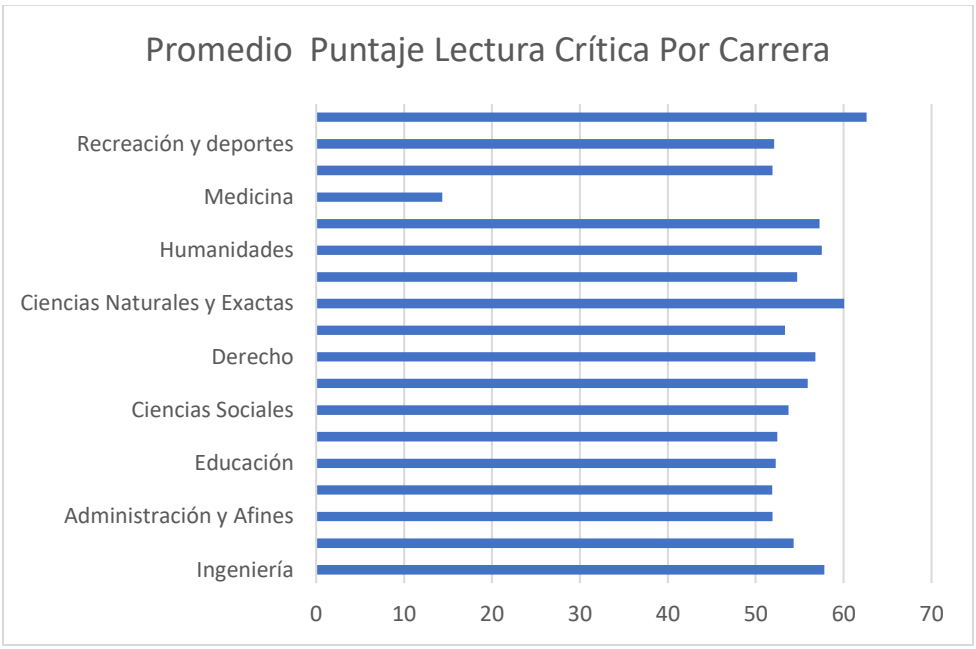


Ilustración 13 Puntaje lectura crítica obtenida en Examen Saber 11 por carrera. Elaboración propia

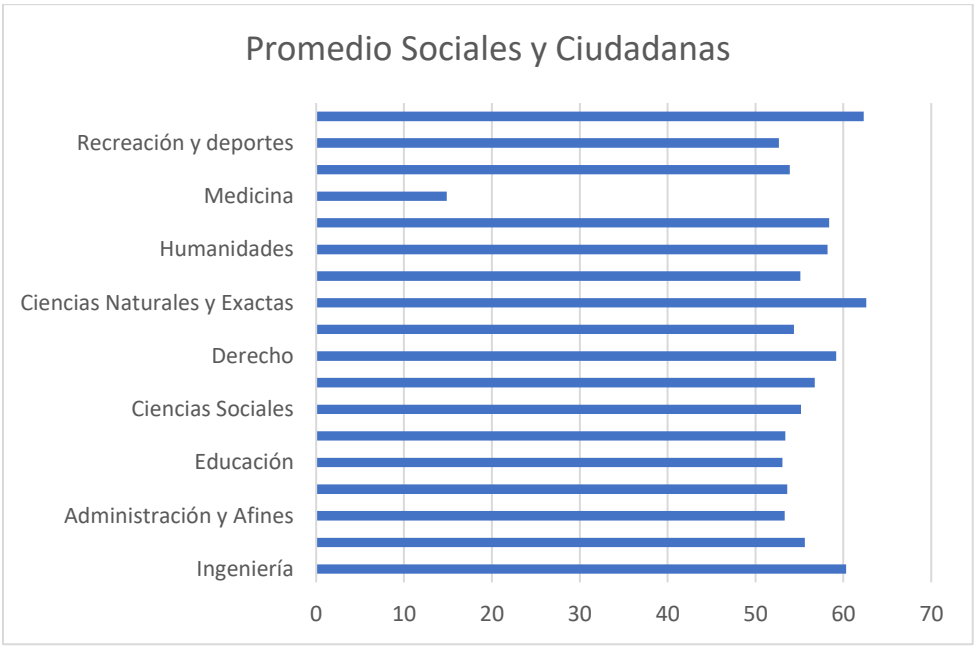


Ilustración 14 Puntaje obtenido en Sociales y ciudadanas Saber 11 por Carrera. Elaboración propia

También se encontró que en materias como matemáticas, los estudiantes de Bellas artes y humanidades obtuvieron un promedio de puntaje similar a estudiantes de Ingeniería, Ciencias naturales y exactas y Medicina, tal como se muestra en la

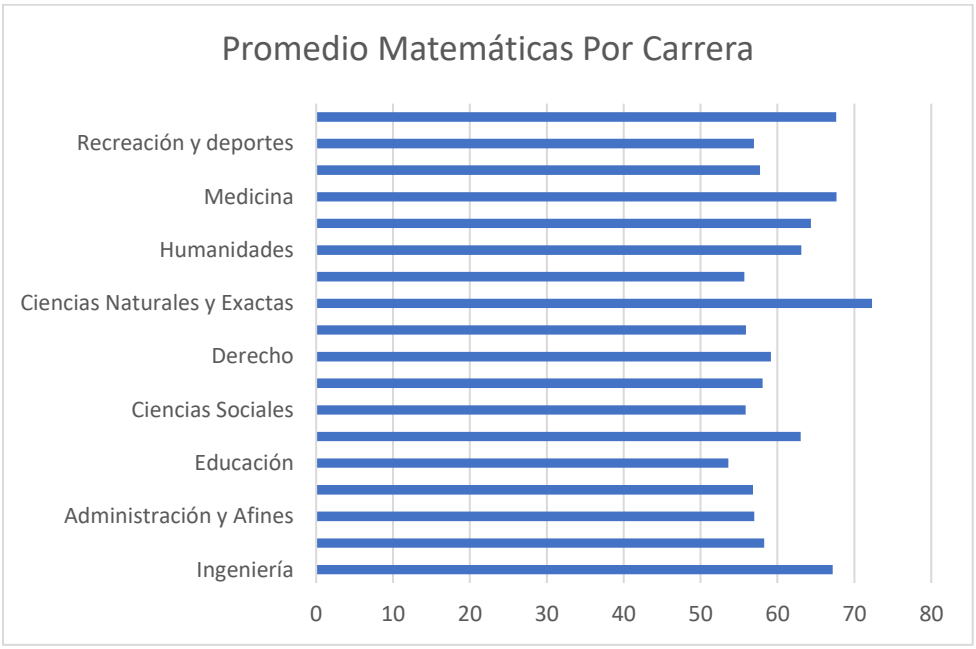


Ilustración 15 Puntaje promedio obtenido en Matemáticas Saber 11 por carrera. Elaboración propia

En la Ilustración 16 se visualiza como los estudiantes de Economía tienen un promedio de puntaje similar a los estudiantes de Ciencias Naturales y exactas, siendo estas dos las de mayor promedio en esta materia.

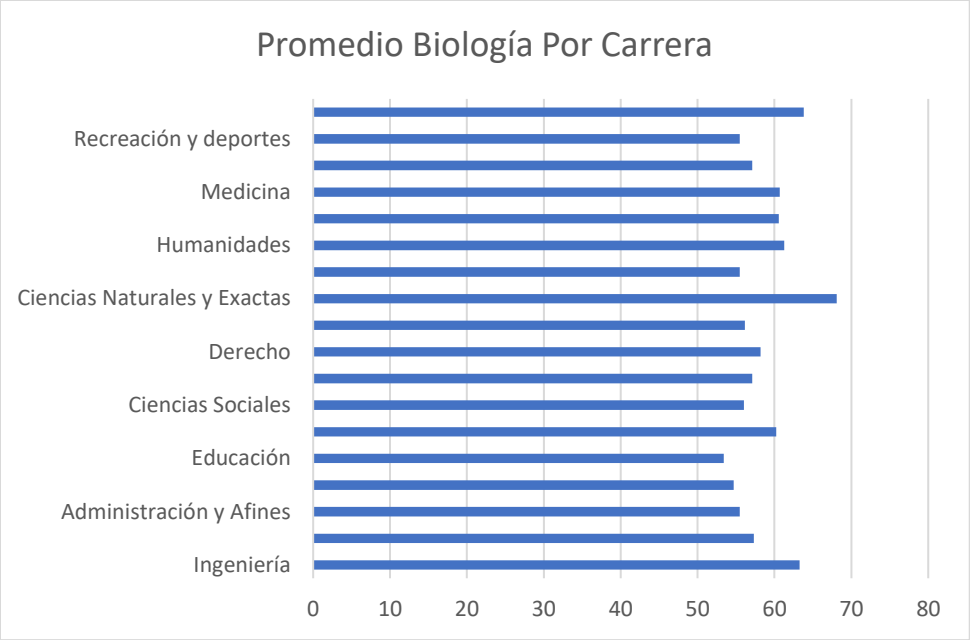


Ilustración 16 Promedio de puntaje obtenido en Biología Saber 11 por carrera. Elaboración propia

Se encontró en la Ilustración 17 que los estudiantes de Medicina obtuvieron puntajes superiores en inglés sobre otros estudiantes, mientras que los estudiantes de contaduría tienen el promedio más bajo en esta materia.

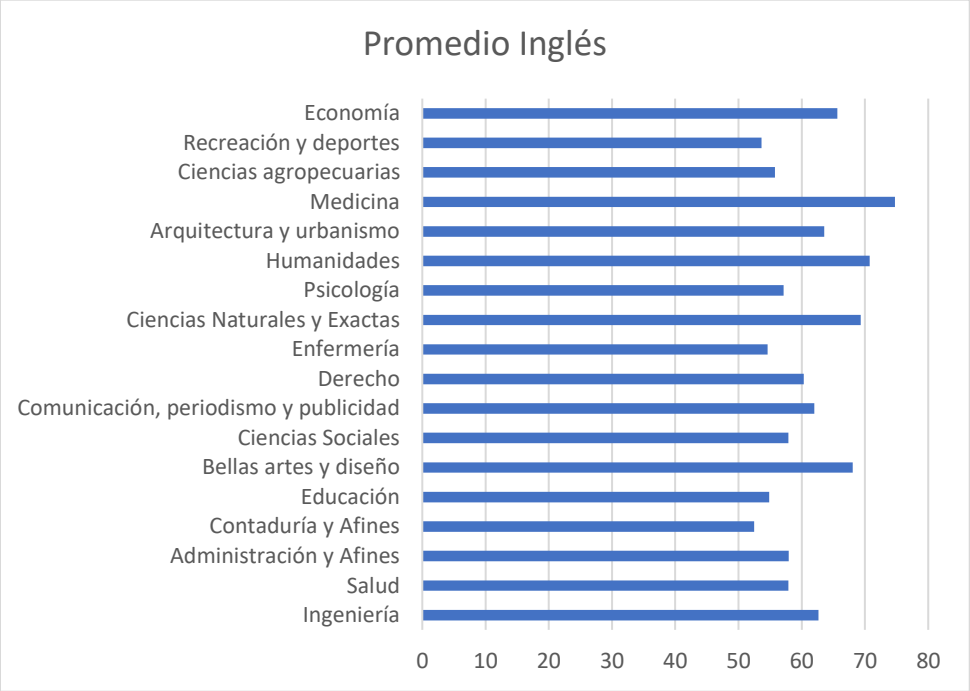


Ilustración 17 Promedio de puntaje obtenido en Inglés Saber 11 por carrera. Elaboración propia

8.4. Verificación de la calidad de los datos

El objetivo de esta etapa es determinar la consistencia de los valores individuales de los campos, la cantidad y distribución de los valores nulos, encontrar valores fuera de rango, valores atípicos (Outliers).

Lo primero que se hace es encontrar los valores nulos o datos perdidos, para esto en la herramienta WEKA se realiza estadística descriptiva que muestra el comportamiento de las diferentes variables.

La primera variable que encontramos es de naturaleza categórica y es la variable GENERO tal como se describe en la Ilustración 10 y su tabla de distribución Ilustración 11

En esta variable encontramos que solo existen 2 posibles valores M (masculino) y F(Femenino) sin embargo aparece un tercer valor que son aquellos que no tienen el dato y se consideran datos perdidos o missing data, para poder tratar este tipo de datos existen 2 diferentes procedimientos o técnicas; la eliminación de los casos que los contienen o la imputación de un valor estimado a la variable[49], en esta última existen sub técnicas como son: sustitución por la media[50], sustitución por la moda[51] y K-vecinos más próximos[52]. Para el uso de cada una de estas técnicas es necesario tener en cuenta la naturaleza de la variable que se va a tratar, si es categórica o cuantitativa.

En este caso la variable Género es una variable categórica por lo tanto hay que analizar si se eliminan estos casos o se sustituyen. En el caso que se eliminaran debemos cuantificar cual es la cantidad de estos y, como se expone en la Ilustración 10 son 2947 casos que equivalen al 0.23% del total, es decir si se eliminan solo corresponden a menos de 1% de los datos, donde según Mesa y Useche la cantidad de pérdidas pueden estar entre el 1% y el 20%, esto depende de la exactitud del estudio y el área en el que se aplica [53], sin embargo hay que tener en cuenta que se busca que se tenga el mayor número de datos posibles, por lo tanto se procede a probar la técnica de la imputación de un valor estimado, también expuesto en el trabajo de Mesa y Useche en su artículo “Una introducción a la imputación de valores perdidos” [53] presentan la imputación a través de la media como una técnica que se usa para variables cuantitativas y la moda para variables cualitativas, además de técnicas como un K-vecino más cercano como estimación de valores perdidos.

Un comportamiento similar se encuentra con las variables día de nacimiento, mes de nacimiento y año de nacimiento, donde se debe tener en cuenta el año de presentación del examen de las personas y las posibles edades que estas presentan, ya que una persona con una edad inferior a 10 años o mayor a 65 años tiene una baja probabilidad de haber presentado

las pruebas de estado SABER11 por estar por fuera del promedio, estos datos son considerados como outliers que pueden agregar ruido al estudio que se está realizando.

Se identifican como datos numéricos aquellos relacionados con los puntajes por materia, como son Biología, Matemáticas, Filosofía, Física, Historia, Química, Lenguaje, Geografía, Lectura crítica, Ciencias sociales ciudadanas, Razonamiento cuantitativo, Competencias ciudadanas, Interdisciplinar e Idioma, esto debido a que las demás variables pueden ser consideradas Nominales ya que tienen valores que representan categorías que no obedecen a una clasificación intrínseca.

El otro set de datos que se va a utilizar es el de las pruebas Saber Pro hallado a través del portal del ICFES. En este set de datos solo nos interesa tener el código del estudiante, el grupo de referencia y el núcleo de pregrado, ya que estos son los datos que se van a usar para realizar el cruce entre los 2 sets de datos (Saber 11 y Saber Pro) y así conseguir un etiquetado de los datos para el proceso de minería de datos.

La primera variable para verificar es GrupoReferencia, uno de los hallazgos para esta variable fue la forma en que sus valores se encuentran escritos, ya que contiene caracteres como tildes o virgulillas o incluso letras con diptongo (Ä, ö), esto puede generar agrupaciones incorrectas al momento de realizar posteriores análisis, para resolver esto se presenta como estrategia, la corrección desde la base de datos.

8.5. PREPARACION DE LOS DATOS

Para esta fase se revisan los datos según la calidad de estos expuesto en el punto anterior y se inicia el filtrado de estos siempre teniendo en cuenta los objetivos del proyecto.

8.5.1. Selección de datos

Para esta fase se debe seleccionar los datos de acuerdo con los objetivos que se quieren alcanzar. Para el primer set de datos (SABER11) se van a seleccionar los datos que son numéricos, es el caso de los puntajes (Punt_Biologia, Punt_matematicas, Punt_filosofia,

punt_fisica, punt_historia, punt_quimica, punt_lenguaje, punt_geografia, punt_lecturacritica, punt_sociales_ciudadanas, punt_razona_cuantitativo, punt_comp_ciudadana, punt_interdisciplinar, punt_idioma, punt_global) también se dejan las variables que brindan información demográfica, esto con el fin de establecer si estas inciden en las capacidades y virtudes del estudiante.

Para el segundo set de datos (Saber PRO) se seleccionan las columnas Estu_consecutivo, Estu_Nucleo_pregrado, Esto_Pgrm_academico, Estu_SNIES_PrgmAcademico, GrupoReferencia, Inst_cod_institucion y inst_nombre_institucion, esto para posteriormente con los datos de las llaves poder realizar el cruce y marcar la carrera estudiada.

8.5.2. Limpieza de los datos

El objetivo de esta fase es aumentar la calidad de los datos requerido, para esto se aplica una selección de subconjuntos de datos, se insertan valores adecuados o se aplican técnicas como la estimación de los datos mediante el modelado.

Para el caso de las variables del examen Saber11, se utilizaron varias estrategias, una de ellas fue por medio de métodos estadísticos, como la moda o la media. Para establecer estos valores de puntajes que están como nulos se procede a la técnica de imputación de valores mediante la técnica de la media, que constituye una de las técnicas más sencillas de usar, y debido a la cantidad de datos a modificar no constituyen un riesgo para el estudio.

También para aumentar la calidad de los datos se modificaron aquellos caracteres extraños de los nombres de los Grupos de Referencia de los resultados de las pruebas Saber PRO.

Otra de las técnicas usadas para la limpieza de datos es la de eliminar datos duplicados, esto se realiza mediante la exploración de los códigos de estudiante que debe ser único, dejando solamente los datos del último examen presentado.

8.5.3. Estructuración de los datos

Esta fase del proceso incluye la creación de atributos derivados, completar nuevos registros o transformar valores basados en atributos existentes.

La primera actividad que se llevó a cabo fue una estandarización de los datos de los puntajes del examen Saber 11, creando nuevas variables que agruparan estos valores teniendo en cuenta los niveles de desempeño de cada prueba teniendo en cuenta la información del ICFES. A saber, en las materias evaluadas como Lectura Critica, Matemáticas, Sociales y ciudadanas y Ciencias Naturales son 4 niveles, mientras que en Ingles son 5 formas de agrupación. [54]

Cada una de las pruebas tiene una agrupación de puntajes diferentes, esto quiere decir, que para Lectura crítica en el nivel 1 están agrupados aquellos que tienen entre 0 y 35 puntos en la prueba, mientras que para Sociales y Ciudadanas el nivel 1 va desde 0 a 40 puntos, la prueba de inglés tiene 5 niveles los cuales son los siguientes A-, A1, A2, B1 y B+. Esta clasificación está resumida en la Tabla 5.

Otra de las actividades realizadas fue la conversión de las variables que agrupan los desempeños de los estudiantes de variables numéricas a variables categóricas, esto para que el árbol pueda clasificar de manera correcta cada una de estas características, las cuales a pesar de que son números, son agrupaciones de los puntajes. Tabla 5

Prueba	Niveles de desempeño – Agrupación de puntajes
Lectura Critica	1 – 0 a35 2 – 36 a 50 3 – 51 a 65 4 – 66 a 100
Matemáticas	1 – 0 a 35 2 – 36 a 50 3 – 51 a 70 4 – 71 a 100
Sociales y Ciudadanas	1 – 0 a 40 2 – 41 a 55 3 – 56 a 70 4 – 71 a 100
Ciencias Naturales	1 – 0 a 40 2 – 41 a 55 3 – 56 a 70 4 – 71 a 100
Inglés	A- - 0 a 47 A1 – 48 a 57 A2 – 58 a 67 B1 – 68 a 78 B+ - 79 a 100

Tabla 5: Niveles de desempeño por prueba elaboración propia

8.5.4. Integración de los datos

La integración de los datos tiene como objetivo combinar información de otros recursos de datos para crear nuevos registros o valores, para el presente trabajo se realizó una combinación de datos entre los estudiantes y puntajes del examen Saber 11 con los resultados de las pruebas Saber Pro, esto con el objetivo de obtener las carreras que habrían estudiado en su ciclo profesional, para este cruce de datos se utilizó los datos de las pruebas Saber Pro, las llaves que contienen el número consecutivo del estudiante tanto del examen Saber Pro como del examen Saber 11 y también los datos del examen Saber 11, estas llaves asocian los dos exámenes y, por medio de consultas realizadas en un motor base de datos MySql se realizó este cruce. Para obtener los datos de las llaves y realizar este procedimiento, el ICFES a partir del año 2016, en el cuestionario de la prueba Saber Pro incluyó una pregunta relacionada con el documento usado en la prueba Saber 11 y, donde estos datos pasan por 3 etapas, cruce entre variable de documento que se usó en Saber 11 obtenida en el cuestionario del examen Saber Pro, cruce entre variables de documentos de identidad reportadas en el registro de Saber Pro y Saber 11 y, finalmente el cruce fonético entre los estudiantes de Saber Pro a los que no se les encontró el registro Saber 11, este procedimiento se encuentra detallado en la Documentación y diccionarios del portal [47].

9. MODELADO

Los sistemas de recomendación están diseñados para sugerir elementos nuevos a quien lo está usando, basándose en elecciones anteriores y elecciones de otros usuarios con historial de votaciones o valoraciones similares. Para este proceso de valoración existen dos posibles formas de recolectar las valoraciones, explícita, cuando el usuario da una puntuación a cada elemento y esta puntuación es un valor discreto entre un mínimo y un máximo. La segunda es de forma implícita, donde la valoración se extrae de información dejada por el usuario de acuerdo con sus acciones. Un ejemplo de esto es el número de veces que reproduce una canción, el tipo de información que consume usando un buscador determinado o el tiempo que pasa leyendo una página web.[55]

Basado en esto, existen diferentes tipos de sistemas de recomendación aquellos que recomiendan ítems basados en idioma, edad, ubicación geográfica o cualquier característica del usuario, los cuales son conocidos como Demográficos, según la utilidad que pueda tener para el usuario, los basados en conocimiento, los cuales evalúan los requerimientos iniciales, reparan inconsistencias de los requisitos y explican los resultados de las recomendaciones. Para finalizar están los híbridos, que combinan técnicas donde se usan las ventajas de uno para solucionar las desventajas del otro.[22]

Teniendo lo anterior en cuenta, el modelo de recomendación propuesto está basado en conocimiento con valoración explicita, ya que se realizará la recomendación de acuerdo con unos puntajes entregados por parte del estudiante donde existe un mínimo y un máximo y, realiza la clasificación de acuerdo con estos.

9.1. Selección de características

Para lograr el objetivo propuesto en este trabajo, solo se tuvo en cuenta los puntajes obtenidos en el examen de estado Saber 11 y se realizó un cruce con la base de datos Saber

Pro para obtener la carrera estudiada en el ciclo profesional, además de esto se excluyen los datos sociodemográficos para el análisis de esta tesis. Tabla 6

Característica	Descripción
Estu_Genero	Género del estudiante
DesemLectCritica	Grupo de desempeño en el que se encuentra el estudiante en la materia Lectura Crítica
DesemMatematicas	Grupo de desempeño en el que se encuentra el estudiante en la materia Matemáticas
DesemBiologia	Grupo de desempeño en el que se encuentra el estudiante en la materia Biología
DesemSocyCompCiu	Grupo de desempeño en el que se encuentra el estudiante en la materia Sociedad y competencias ciudadanas
DesemIngles	Grupo de desempeño en el que se encuentra el estudiante en la materia inglés
GrupoReferencia	Nombre del grupo de referencia al que pertenece el programa académico del estudiante
Estu_nucleo_pregrado	Nombre del programa académico que estudia

Tabla 6: Espacio característico seleccionado a partir de los objetivos elaboración propia

9.2. Entrenamiento y validación del modelo

En este proyecto de tesis se ha planteado hacer el uso de los algoritmos de clasificación para construir un modelo de recomendación, el cual busca sugerir a un estudiante las posibles carreras que debería estudiar en su ciclo profesional en función de los puntajes obtenidos en las cinco materias núcleo en el examen Saber 11. De esta manera del total de registros (27.343) obtenidos después del cruce, se reemplazaron los valores perdidos (missing values) realizando la imputación de datos utilizando el promedio para realizar esto. Así la cantidad de

datos a utilizar para el entrenamiento se utilizará el 70% de estos, quedando 19.140 datos, partiendo de esto se compararon varios algoritmos de clasificación con el objetivo de obtener el mejor rendimiento. Tabla 7

Conjunto de datos	Cantidad de registros	Porcentaje %
Entrenamiento	19.140	70%
Validación	8.203	30%

Tabla 7: Distribución conjuntos de datos elaboración propia

9.3. Balanceo de datos

En un proceso de minería de datos ocurre cuando el número de registros o instancias de cada clase es muy diferente entre sí,[56] esto representa un problema para los clasificadores ya que tienden a realizar la clasificación hacia la clase mayoritaria.

En la Tabla 8, se evidencia que los datos para cada grupo de referencia se encuentran desbalanceados, por ejemplo, en Ingeniería tiene 4670 registros mientras que Economía, Administración, contaduría y afines solo tiene 48. Para esto se aplicó una estrategia llamada *SAMPLING* que permite mejorar estos indicadores. *SAMPLING* es una estrategia que permite realizar el balanceo de manera automática tanto hacia arriba como hacia abajo.

Este método funciona de la misma manera que funcionaría la combinación de multiplicación y muestreo (absoluto). Para un conjunto de datos de entrada dado para cada etiqueta incluida, se selecciona un número fijo de ejemplos. En caso de que no se incluyan suficientes ejemplos, los ejemplos existentes se multiplican. Si se marca la opción, las etiquetas, que están representadas por más ejemplos de los ingresados, se elimina el número de ejemplos. De lo contrario, se deja como está[6]. Tabla 8

[6] Tomado de la documentación de RapidMiner – Sampling (balance)

GRUPO REFERENCIA	Cantidad Registros
INGENIERÍA	4670
ADMINISTRACIÓN Y AFINES	3063
DERECHO	1896
PSICOLOGIA	1580
EDUCACIÓN	1230
CONTADURIA Y AFINES	887
SALUD	822
COMUNICACIÓN, PERIODISMO Y PUBLICIDAD	709
ARQUITECTURA Y URBANISMO	600
BELLAS ARTES Y DISEÑO	584
CIENCIAS SOCIALES	563
CIENCIAS NATURALES Y EXACTAS	316
ENFERMERÍA	300
ECONOMÍA	288
MEDICINA	220
CIENCIAS AGROPECUARIAS	193
HUMANIDADES	176
ECONOMÍA, ADMINISTRACIÓN, CONTADURÍA Y AFINES - UNIVERSITARIA	48

Tabla 8: Cantidad de registros por programa elaboración propia

Luego del balanceo de datos, en la Ilustración 18 se puede comprobar que la cantidad de registros para cada programa es el mismo.

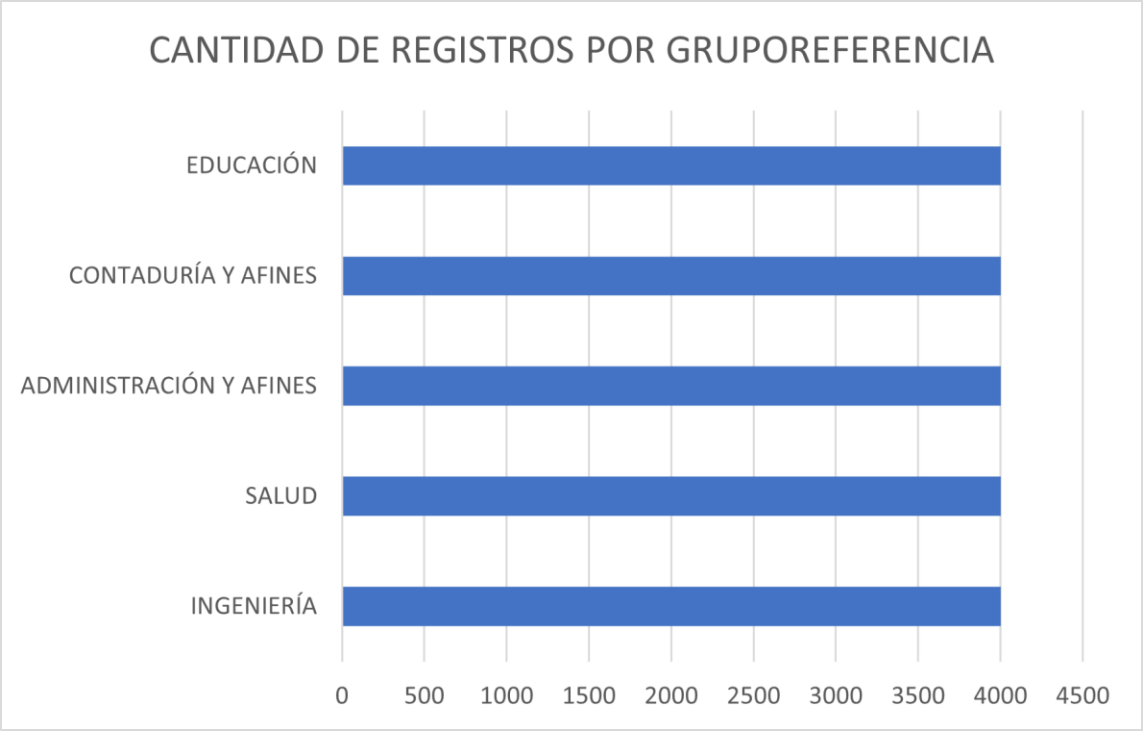


Ilustración 18: Datos Balanceados por programa elaboración propia

10. EVALUACION

En los problemas de clasificación, existen 9 posibles métricas de desempeño que se pueden usar:

- Matriz de confusión o error: Esta métrica permite visualizar mediante una tabla de contingencia la distribución de errores cometidos por un clasificador. La matriz de confusión tiene como elementos descriptores los Verdaderos Positivos (VP), son instancias correctamente reconocidas por el sistema; Falsos Negativos (FN), son instancias que son positivas y el sistema dice que no lo son; Falsos Positivos (FP), instancias que son negativas pero el sistema dice que no lo son y los Verdaderos Negativos (VN), son instancias que son negativas y el sistema las reconoce de manera correcta como tal. En la Ilustración 19 se describe una matriz que contiene dos clases. La suma de estos elementos debe tener como total el número del conjunto de datos de entrenamiento [32] Ecuación 1.

	Clase Predicha	
Clase real	**Play**	**No Play**
Play	Verdaderos positivos (VP)	Falsos negativos(FN)
No Play	Falsos positivos (FP)	Verdaderos negativos (VN)

Ilustración 19: Matriz de confusión tomado de [32]

$$N = VP + FN + FP + VN$$

Ecuación 1: Número total de datos tomado de [32]

- Exactitud (accuracy): Porcentaje de data clasificada correctamente. Ecuación 2

$$Accuracy = \frac{VP + VN}{Total}$$

Ecuación 2: Exactitud tomado de [32]

- Recall o sensibilidad o TPR (Tasa Positiva Real): Calcula el porcentaje de clasificación cuándo la clase es positiva. Mide la proporción de términos correctamente reconocidos con respecto al total de términos reales, mide en que grado están todos los que son. Ecuación 3

$$Sensibilidad = \frac{VP}{VP + VN}$$

Ecuación 3: Sensibilidad tomado de [30]

- Precisión: Calcula el porcentaje de clasificación correcta cuando se predice positivos. Es decir, mide el número de términos correctamente reconocidos respecto al total de términos predichos, sin importar si son verdaderos o falsos Ecuación 4

$$Precision = \frac{VP}{VP + FP}$$

Ecuación 4: Precisión tomado de [32]

- Especificidad o TNR (Tasa Negativa Real): Calcula el porcentaje de clasificación cuando la clase es negativa, mide la probabilidad de que un elemento negativo del conjunto sea rechazado por el filtro. Ecuación 5

$$Especificidad = \frac{VN}{TotalNegativos}$$

Ecuación 5: Especificidad tomado de [32]

- F1-Score: Promedio armónico entre la precisión y sensibilidad. Una calificación F1 alta indica que la precisión y la sensibilidad son altas y está dada por la Ecuación 6.[57]

$$F1 = \frac{2 * (presición * sensibilidad)}{presisión + sensibilidad}$$

Ecuación 6: f1-Score tomado de [32]

Para realizar la medición de los algoritmos se utiliza RapidMiner, donde se realizó la ejecución y comparación de los algoritmos de clasificación. En la se puede ver la comparación de los diferentes algoritmos usando como métricas la exactitud, sensibilidad y precisión.

Algoritmo	**Exactitud (Accuracy)**	**Sensibilidad (Recall)**	**Precisión**
Árboles de decisión	27.78%	4.63%	5.01%
Naive Bayes	23.45%	7.22%	5.62%
K-NN	23.82%	4.62%	4.95%
Random Forest	29.86%	5.19%	5.21%

Tabla 9: Comparación de métricas de medición antes de balanceo de datos elaboración propia

Algoritmo	Exactitud (Accuracy)	Sensibilidad (Recall)	Precisión
Árboles de decisión	44.69%	44.63%	42%
Naive Bayes	94.84%	94.84%	94.91%
K-NN	28.50%	28.39%	25.99%
Random Forest	81.20%	81.23%	79.13%

Tabla 10: Comparación de métricas de medición después de balanceo de datos elaboración propia

Después de realizar los procesos antes y después de realizar balanceo de datos, se observa que el algoritmo que presenta un mayor porcentaje de exactitud es Random Forest antes de balanceo de datos, sin embargo, después del balanceo de datos con los datos de entrenamiento se observa que el algoritmo que presenta mayores estadísticas es Naive Bayes. Tabla 9 y Tabla 10

Ahora bien, después del balanceo de datos los algoritmos presentan un incremento de estas métricas sustancialmente, excepto K-NN, no obstante, el algoritmo Naive Bayes presenta unas estadísticas que podrían hacernos sospechar de un sobre ajuste, para comprobar esto validamos el modelo con el set de datos de validación. Tabla 11

La exactitud con la que el algoritmo Random forest clasifica es del 82.90%, mientras que Naive Bayes lo hace con el 27.94%, árboles de decisión con un 44.93% y K-NN con un 75.98% de precisión.

Algoritmo	Exactitud (Accuracy)
Árboles de decisión	44.93%
Naive Bayes	27.94%
K-NN	75.98%
Random Forest	82.90%

Tabla 11: Comparación de exactitud con datos de validación elaboración propia

De acuerdo con estas estadísticas se decide usar el algoritmo Random Forest ya que, entre las investigaciones realizadas y tal como se comprueba en las tablas mostradas anterior mente es el algoritmo que presenta mejor rendimiento, además de su escalabilidad y facilidad de uso[35]. Además, estos son excelentes a la hora de aprender relaciones complejas, generalmente no es necesario podar el bosque aleatorio, puesto que el modelo es bastante robusto ante el ruido de los árboles de decisión individuales.

11. Resultados

Para la comprobación del modelo se realiza un prototipo, donde se capturarán los datos del estudiante y como respuesta se recibirá una sugerencia de la carrera a estudiar. Para esto se realiza un diagrama de casos de uso y de componentes.

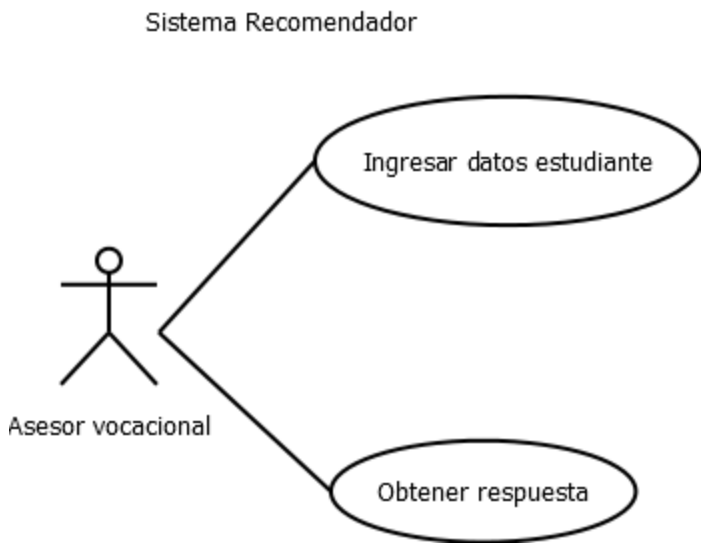


Ilustración 20: Diagrama casos de uso Asesor Vocacional Elaboración propia

Teniendo en cuenta que es un recomendador vocacional y el marco legal colombiano, donde según El Decreto 2277 de 14 septiembre 1979 [58] en el artículo 2. “Considera a quienes se encargan de la consejería y orientación educativa como educadores. “Profesión Docente. Las personas que ejercen la profesión docente se denominan genéricamente educadores. Se entiende por profesión docente el ejercicio de la enseñanza en planteles oficiales y no oficiales de educación en los distintos niveles de que trata este decreto. Igualmente incluye esta definición a los docentes que ejercen funciones de dirección y coordinación de los planteles educativos de supervisión e inspección escolar, de programación y capacitación educativa, de consejería y orientación del educando, de educación especial, de alfabetización de adultos y demás actividades de educación formal autorizadas por el Ministerio de Educación Nacional en los términos que determine el reglamento ejecutivo”, y la Resolución 12712 del 21 de julio, “reglamenta la orientación escolar para los niveles de educación básica y media vocacional y se asignan las funciones de los docentes especialistas en esta área” [58], el recomendador no puede ser accedido por un estudiante, además de estar regulado por un marco legal, también obedece a que se encuentran en una edad aproximada entre los 15 y 18

años llamada periodo tentativo, donde están conociendo sus intereses, capacidades y valores [30].

En el diagrama de casos de uso se muestra cómo el Asesor vocacional tiene la posibilidad de ingresar los datos del estudiante para obtener un resultado de la clasificación del recomendador. Existen dos casos de uso que incluyen el ingreso de los datos del estudiante, Preparar los datos ingresados, el cual se encarga de transformar los datos a otros que el algoritmo pueda procesar para la posterior clasificación y, el último caso de uso, Hacer la predicción, es aquel encargado de procesar los datos y entregar un resultado final al Asesor Vocacional.

En la Ilustración 21 se muestra como el científico de datos está relacionado con los diferentes casos de uso para el desarrollo del modelo de minería de datos.

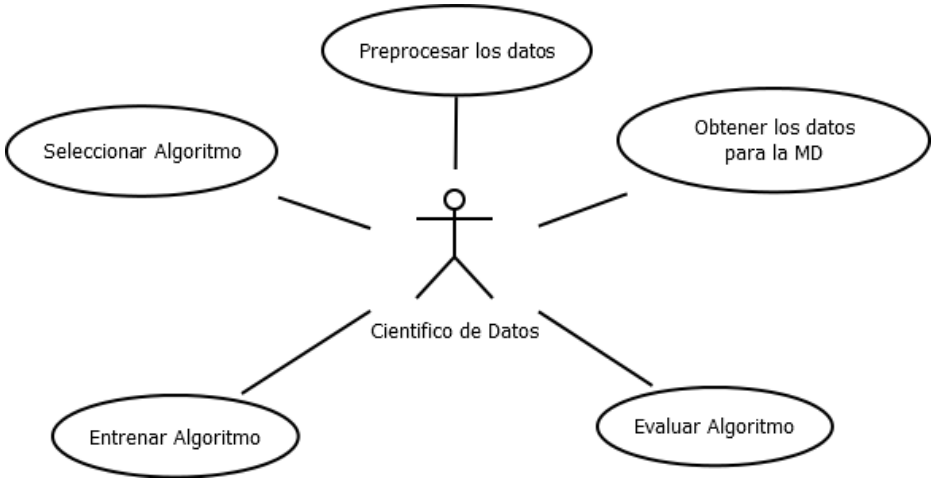


Ilustración 21: Diagrama casos de uso Científico de datos Elaboración propia

En la Ilustración 22 se muestran los diferentes componentes que hacen parte del prototipo, por un lado, está el interfaz, que es la interfaz que comunica al usuario con el sistema, tenemos un componente llamado Flask el cual es un microentorno de trabajo el cual permite la creación de aplicaciones web escrito en Python rápidamente y con un mínimo de

líneas de código, permite a partir del motor de plantillas Jinja2 se pueda intercambiar información entre el programa y la interfaz de usuario. Flask incluye un servidor web de desarrollo para que se pueda probar e ir viendo los resultados sin necesidad de instalar otro servidor web [59]. Por otro lado está el Recomendador, que es el componente que realiza la clasificación de los datos y recorrer los diferentes árboles generados y regresar un resultado al usuario final, este componente tiene relación con 2 bibliotecas de software que ayudan a realizar la tarea de clasificación de los datos, Pandas es usado para manipular datos y entregarlos en el formato adecuado al clasificador que se encuentra en la segunda biblioteca de software llamada Sklearn, la cual es usada para el aprendizaje automático y en este caso realizar la predicción con respecto a los datos entregados por el usuario.

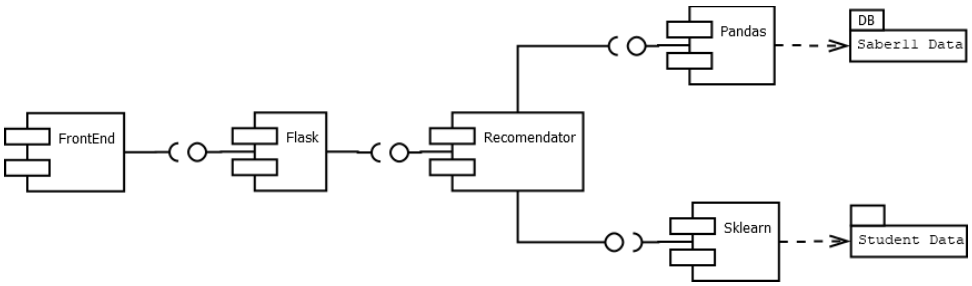


Ilustración 22: Diagrama de componentes Elaboración propia

En la Ilustración 23 se encuentra la interfaz de la solicitud de datos del estudiante

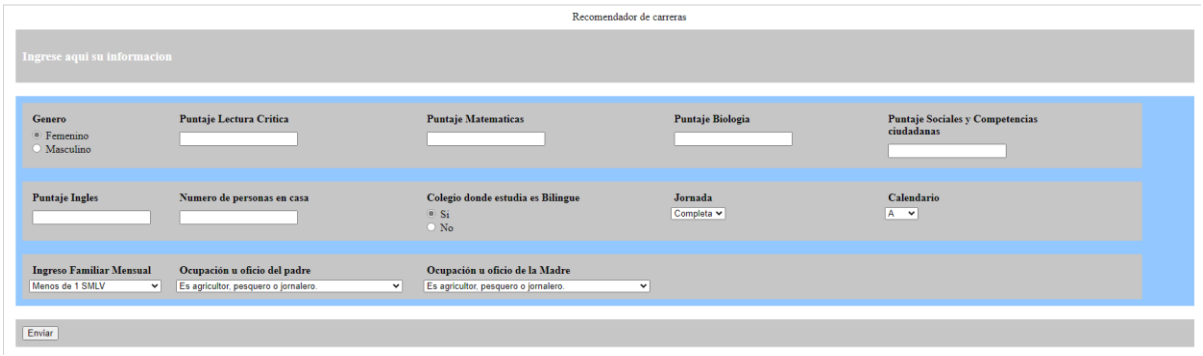


Ilustración 23: Interfaz de usuario elaboración propia

En la Ilustración 24 se encuentra la página de retorno de resultados.

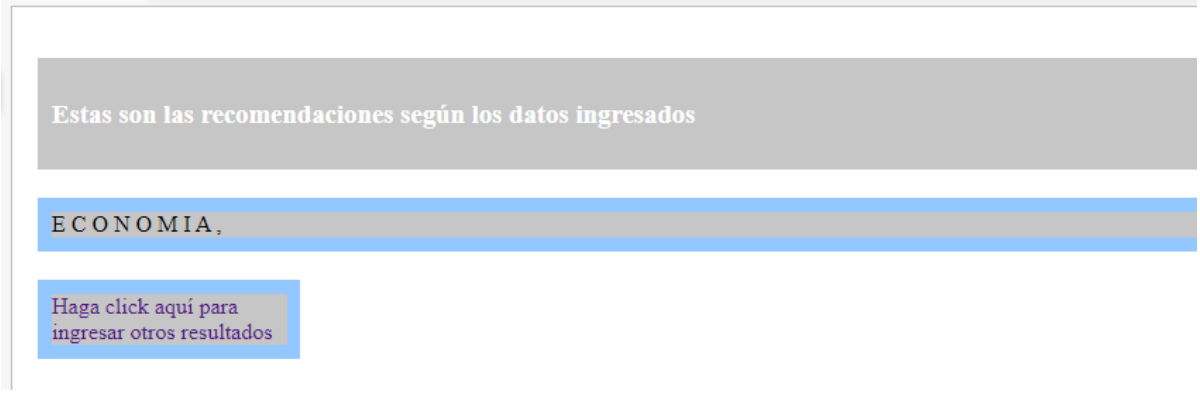


Ilustración 24: Retorno de resultados elaboración propia

Una vez se tiene el modelo implementado, se realiza una ejecución de este y se obtiene una precisión del 75% con respecto al grupo de datos de prueba. Ilustración 25.

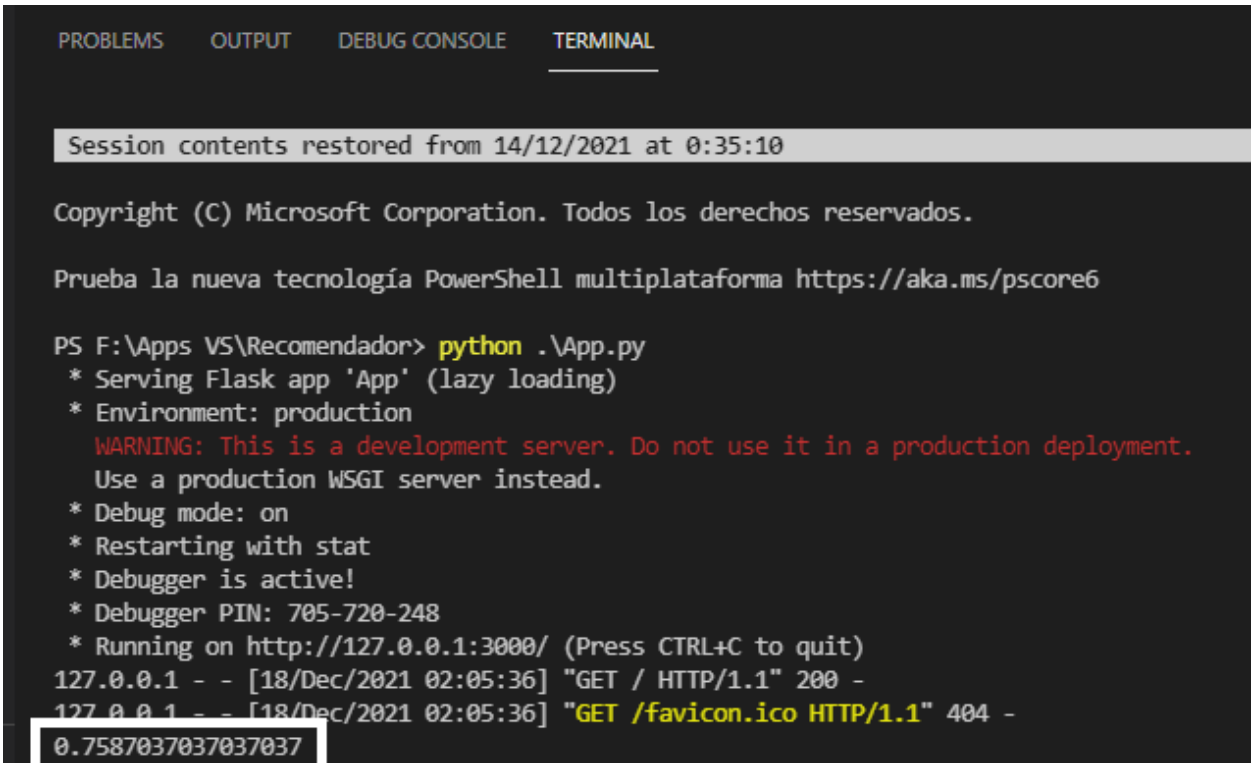


Ilustración 25 Precisión grupo de datos de prueba, elaboración propia

Ahora bien, también se requiere verificar si es capaz de clasificar un estudiante que no esté en ninguno de los 2 conjuntos de datos, para esto previamente se separaron 3 estudiantes y sobre estos se realizó la prueba.

Los datos de prueba del estudiante fueron:

Género: Masculino

Puntaje Lectura crítica: 46

Puntaje Matemáticas: 65

Puntaje Biología: 55

Puntaje Sociales y ciudadanas: 55

Puntaje Ingles: 54

Número de personas en el hogar: 3

El colegio es bilingüe: No

Jornada del colegio: Mañana

Calendario: A

Ingreso Familiar Mensual: Entre 1 y menos de 2 SMLV

Ocupación del padre: Trabajador por cuenta propia

Ocupación de la madre: Hogar

Con estos datos y según el recomendador la carrera que debería haber estudiado esta persona es Enfermería Ilustración 26, y según los datos separados debería estudiar algo relacionado con la Salud como se muestra en la Ilustración 27.

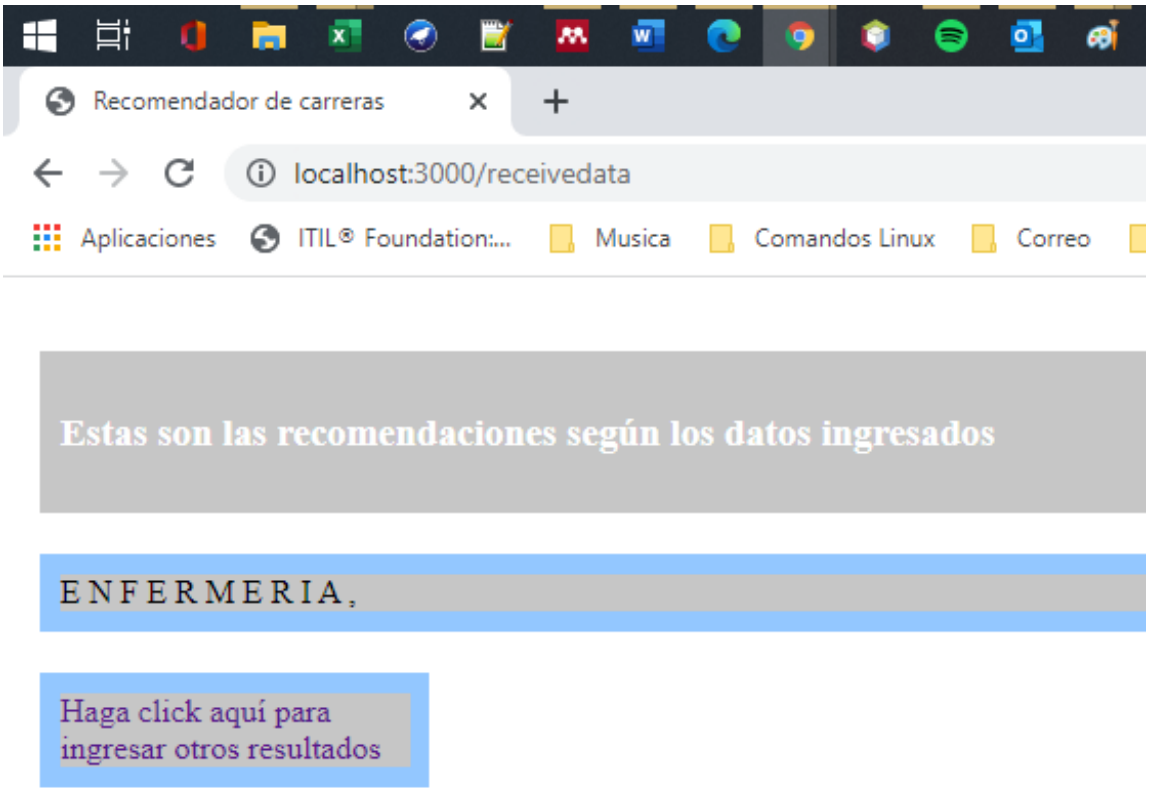


Ilustración 26: Clasificación realizada por el recomendador, elaboración propia

IO	FAMI_INGRESO_FMILIAR_MENSUAL	FAMI_OCUPA_PADRE	FAMI_OCUPA_MADRE	GRUPOREFERENCIA
	1	12	7	SALUD

Ilustración 27: La carrera que estudió según datos consultados, elaboración propia

Con estos datos lo primero a realizar es su clasificación de rendimiento, donde tenemos que en matemáticas se encuentra en un nivel de desempeño 3, donde según [54] un estudiante en este nivel puede enfrentarse a problemas que involucran el uso de conceptos de proporcionalidad, factores de conversión, áreas y desarrollos planos, en contextos laborales u ocupacionales, matemáticos o científicos, y comunitarios o sociales.

También se encuentra que en biología se encuentra en un nivel de desempeño 2, que se describe como, el estudiante que se ubica en este nivel reconoce información suministrada en tablas, gráficas y esquemas de una sola variable independiente, y la asocia con nociones de los conceptos básicos de las ciencias naturales (tiempo, posición, velocidad, imantación y filtración).

En Sociales y ciudadanas también, se encuentra en un nivel de desempeño 2 donde, en este nivel, se presentan contextos cuya descripción es corta, con pocos actores, enunciados directos y posturas o posiciones explícitas, sencillas y claras. Además, se presentan situaciones cercanas a la cotidianidad del estudiante o de conocimiento y amplia discusión pública.

Tomando en cuenta cada uno de los descriptores de los niveles de desempeño, se puede decir que el estudiante puede enfrentarse a problemas que involucran lo social y comunitario, además también es capaz de reconocer información estadística básica y tiene la capacidad de presentar un contexto de descripción corta, si comparamos esto con una descripción de carrera de enfermería, donde es una carrera dirigida a personas con sensibilidad social, motivada hacia las necesidades humanas, dispuesta a conocer y comprender las diferentes situaciones de salud y vida en diferentes escenarios[7], coincidiendo con el perfil descrito en el estudiante mencionado.

Es decir, el clasificador toma estos niveles de desempeño y aspectos sociales del estudiante de su entorno cercano y navega a través del bosque generado, arrojando un resultado aceptable que permite al asesor vocacional tener en cuenta este para realizar una asesoría vocacional más acertada al estudiante.

[7] Descripción tomada de la página de la Universidad Nacional de Colombia Requisitos: Universidad Nacional de Colombia (unal.edu.co)

12. CONCLUSIONES

El objetivo del proyecto, como se ha mencionado, es apoyar a estudiantes de grado 11 a la elección de una carrera profesional, buscando similitudes con otros estudiantes que ya han pasado por este proceso y están a punto de finalizar su ciclo profesional, esto beneficia tanto al estudiante, como a las familias y a las instituciones, al estudiante por tener el convencimiento de haber seleccionado la carrera adecuada para seguir y tener la satisfacción de haberlo conseguido, para las familias porque el riesgo del costo asociado a un probable abandono se habrá disminuido y a las instituciones porque con esto tendrán un crecimiento en su reputación ofreciendo a los estudiantes lo que realmente esperan al momento de elegir una institución para continuar con sus estudios de educación superior.

Para caracterizar los elementos asociados al modelo, se realizó una implementación de la metodología CRISP-DM, siguiendo la guía, donde partiendo desde el entendimiento del negocio, donde se indagó sobre la problemática del abandono de estudiantes en la educación superior, sus estadísticas, así como también entender qué objetivos persigue la educación y, articularlos con los de los estudiantes, así como también tener en cuenta el punto de vista de la orientación vocacional y como a través de los años ha sido estudiada y ha sido implementada en los distintos escenarios en las que se da; una vez entendido esto, se realizó la recolección de los datos a través de la descarga de los archivos de manera individual de cada uno de los años que se tenían los datos de los exámenes de estado Saber11 y SaberPro, para luego a través de los diccionarios de datos entenderlos para su caracterización.

Para la bodega de datos se implementó una aplicación de entorno local web en lenguaje PHP y almacenamiento en un motor de base de datos MySql. El aplicativo se diseñó para leer los archivos separados por comas (CSV) de acuerdo con el tipo de examen que se fuera a consultar y almacenar (Saber11, SaberPro o Llaves), separar cada una de las columnas

y una vez hecho esto almacenarlo, esto se realizó por cada archivo obtenido, en total se procesaron 22 archivos del examen Saber11, 26 del examen SaberPro y 1 de llaves. En este paso el algoritmo fue modificado varias veces para disminuir la cantidad de errores generados por las diferencias entre cada uno de los archivos, ya que cada uno cuenta con formato y un orden en las columnas diferente, también se realizó cambios en el tiempo de espera de ejecución del programa ya que el lenguaje PHP al ser un lenguaje orientado a la web, tiene la limitante de su tiempo de ejecución del lado del servidor, sin embargo al ser una aplicación local este tiempo fue extendido para evitar estos errores de tiempo terminado, fue necesario realizar estos cálculos teniendo en cuenta el tiempo que se tardaba en ingresar 1000 datos y sobre esto modificar el valor de tiempo de espera en el servidor llegando en ocasiones a tiempos que superaban los 6 días de ejecución del programa por cada archivo. Con esto se logró el objetivo de tener los datos en una bodega de datos donde se pudiera consultar de manera fácil, además de estandarizar los tipos de datos de cada columna.

Para el modelo de minería de datos, se usa la bodega de datos realizada, teniendo separados y unificados los datos de los exámenes de estado Saber11 y Saber Pro y, teniendo una tabla con las llaves que unen a los 2, se realizó la consulta que uniera estos datos y así poder etiquetar los datos del examen Saber11 con la variable a clasificar, GRUPOREFERENCIA. Con esta consulta en el programa Weka se realizó la conexión a la base de datos mediante el controlador JDBC de java para disminuir el tiempo de consulta de los datos, con esto se obtuvo la estadística descriptiva de cada una de las columnas, y así poder identificar los valores perdidos, outliers, columnas sin datos y posibles ajustes a realizar para realizar un buen entrenamiento del algoritmo que se usaría posteriormente. Luego de esto se usó el programa RapidMiner con una licencia de educación, para realizar el preprocesamiento de los datos, tal como identificar valores perdidos o nulos, remplazarlos con valores usando la estrategia de imposición de valores usando la media, filtrado de aquellos

registros que por diversos temas no se encontraban en condiciones de ser usados ya que tenían valores que no correspondían y balanceo de datos a través de la estrategia oversampling y downsampling. Con esto se realizó el modelo de minería de datos y por medio de la comparación de métricas entre los diferentes algoritmos de clasificación como K-NN, Árboles de decisión, Naïve Bayes y Random Forest, se encontró que el clasificador que mejor resultado arroja es el de Bosques aleatorios (Random Forest) el cual obtuvo un 82.9% de precisión de clasificación, a pesar que en las métricas obtenidas con el set de entrenamiento el que mejores estadísticas presentaba era K-NN, sin embargo al ejecutarlo con el conjunto de pruebas su rendimiento bajó considerablemente, además si se tiene en cuenta el tiempo de ejecución, al ser un algoritmo “vago”, no era el mejor.

El método de validación fue tomar datos de varios estudiantes que estuvieran fuera del conjunto de datos de entrenamiento y pruebas y comprobar la capacidad de clasificación.

13. TRABAJOS FUTUROS

En este trabajo de tesis se utilizaron algoritmos de clasificación tales como arboles de decisión, K-NN, Naïve Bayes y Random Forest, con esto se logró comprender como cada variable puede afectar la decisión de un estudiante al seguir una carrera profesional. También se realizó la combinación de los datos de las pruebas Saber11 y SaberPro, encontrando a otros estudiantes en la finalización de su ciclo de estudios profesionales y lograr etiquetar cada uno de los datos para lograr el objetivo planteado. De esta manera como trabajo futuro se plantea poder realizar la automatización de la recolección de estos datos y tener unificada esta información, así como establecer procesos de ETL para simplificar la tarea de almacenamiento y preprocesamiento de los datos, así como el etiquetado de las carreras que los estudiantes han ido finalizando y así tener cada vez más posibilidades de entrenamiento del modelo.

REFERENCIAS

[1] J. C. Riquelme, R. Ruiz, and K. Gilbert, “Minería de Datos: Conceptos y Tendencias,” 2006. Accessed: Oct. 01, 2018. [Online]. Available: http://www.aepia.org.

[2] A. V. D. López, “Estrategias para vencer la deserción universitaria,” *Educ. y Educ.*, vol. 7, no. 0, pp. 177–203, Aug. 2009, Accessed: Oct. 29, 2021. [Online]. Available: https://educacionyeducadores.unisabana.edu.co/index.php/eye/article/view/555.

[3] R. A. Leyva Osorio and K. A. Medina Arango, “SISTEMA DE RECOMENDACIÓN PARA VOCACIÓN PROFESIONAL, APLICADO A LA CARRERA DE INGENIERÍA DE SISTEMAS DE LA UNIVERSIDAD DE CUNDINAMARCA, EXTENSIÓN FACATATIVA,” 2019. http://repositorio.ucundinamarca.edu.co/bitstream/handle/20.500.12558/3092/SISTEMA DE RECOMENDACIÓN PARA VOCACIÓN PROFESIONAL%2C APLICADO A LA CARRERA DE INGENIERIA DE SISTEM.pdf?sequence=1&isAllowed=y (accessed Sep. 23, 2020).

[4] “Vista de Comparación de técnicas de minería de datos para identificar indicios de deserción estudiantil, a partir del desempeño académico.” https://revistas.uis.edu.co/index.php/revistauisingenierias/article/view/9834/10291 (accessed Sep. 01, 2020).

[5] “¿Qué es el SPADIES? - Sistemas información.” https://www.mineducacion.gov.co/sistemasinfo/spadies/Informacion-Institucional/254648:Que-es-el-SPADIES (accessed Sep. 01, 2020).

[6] I. L. Fontecha Barbosa and D. M. Pinzon Plazas, “Incidencia de la Financiación en la Deserción Universitaria en Bogotá,” 2018. https://repository.ucatolica.edu.co/bitstream/10983/22658/1/Trabajo de investigación Barbosa y Pinzon.pdf (accessed Sep. 01, 2020).

[7] D. Ramírez, O. Alberto Tapasco Alzate, F. Javier Ruiz Ortega, and D. Osorio García, “Deserción estudiantil: incidencia de factores institucionales relacionados con los procesos de admisión Diógenes Ramírez Ramírez,” doi: 10.5294/edu.2019.22.1.5.

[8] C. Marulanda, M. López, and M. (2017) Mejía, “Minería de datos en gestión del conocimiento de pymes de Colombia,” Manizales, 2017. Accessed: Sep. 13, 2018. [Online]. Available: http://revistavirtual.ucn.edu.co/index.php/RevistaUCN/article/view/821/1339.

[9] C. A. De La Ossa Farias, “Aportes a la disminución de la deserción universitaria causada por la mala elección de carrera en el Programa de Administración de Empresas de la Facultad de Ciencias Administrativas y Contables de la Universidad de La Salle, en el período 2015,” 2016. Accessed: Aug. 28, 2020. [Online]. Available: https://ciencia.lasalle.edu.co/administracion_de_empresas/871.

[10] I. VELASCO QUINTERO, “ANÁLISIS DE LAS CAUSAS DE DESERCIÓN

UNIVERSITARIA.”

[11] S. Ameri, M. J. Fard, R. B. Chinnam, and C. K. Reddy, “Survival Analysis based Framework for Early Prediction of Student Dropouts,” doi: 10.1145/2983323.2983351.

[12] O. A. Tapasco Alzate, F. J. Ruiz Ortega, D. Osorio García, and D. Ramírez Ramírez, “cidencia de factores institucionales relacionados con los procesos de admisión,” *Educ. y Educ.*, vol. 22, no. 1, pp. 81–100, May 2019, doi: 10.5294/edu.2019.22.1.5.

[13] Y. Chen, A. Johri, and H. Rangwala, “Running out of STEM: A Comparative Study across STEM Majors of College Students At-Risk of Dropping Out Early,” p. 10, doi: 10.1145/3170358.3170410.

[14] “SPADIES - Sistema para la Prevención y Análisis de la Deserción en las Instituciones de Educación Superior.” https://spadies3.mineducacion.gov.co/spadiesWeb/#/page/basicas (accessed Sep. 01, 2020).

[15] “Aprobado Presupuesto General de la Nación 2019, enfocado en una mayor equidad.” https://id.presidencia.gov.co/Paginas/prensa/2018/181018-Aprobado-Presupuesto-General-de-la-Nacion-2019-enfocado-en-una-mayor-equidad.aspx (accessed Sep. 01, 2020).

[16] M. Esteban García, A. B. Bernardo Gutiérrez, E. Tuero Herrero, R. Cerezo Menéndez, and J. C. Núñez Pérez, “El contexto sí importa: identificación de relaciones entre el abandono de titulación y variables contextuales,” *Eur. J. Educ. Psychol.*, vol. 9, no. 2, pp. 79–88, Dec. 2016, doi: 10.1016/j.ejeps.2015.06.001.

[17] M. T. Hernández-Jiménez, T. E. Moreira-Mora, M. Solís-Salazar, and T. Fernández-Martín, “Estudio descriptivo de variables sociodemográficas y motivacionales asociadas a la deserción: la perspectiva de personas universitarias de primer ingreso.” https://www.scielo.sa.cr/pdf/edu/v44n1/2215-2644-edu-44-01-00210.pdf (accessed Sep. 11, 2020).

[18] M. De Grado, “Estudio sobre las posibles causas de deserción estudiantil de las facultades de Ingenierías y FACEACO en la UCC campus Villavicencio.”

[19] P. Morán, J. Carlos, T. Chavez, R. Vargas, and A. Alejandra, “ANÁLISIS DEL ABANDONO, DEL PROCESO DE ELECCIÓN Y DEL CAMBIO DE CARRERA EN ESTUDIANTES UNIVERSITARIOS Línea 1. Factores asociados al abandono. Tipos y perfiles de abandono.”

[20] E. J. Polo, S. Sebastian, and R. Bauer, “IMPLEMENTACIÓN DE TÉCNICAS DE DATA MINING, PARA LA PREDICCIÓN DE LA DESERCIÓN DE LOS ESTUDIANTES DEL PROGRAMA DE INGENIERIA INDUSTRIAL DE LA UNIVERSIDAD ICESI,” 2016.

[21] J. I. Álvarez García and 79691279, “Uso de Deep Learning en sistemas de recomendación para reducir la deserción en Educación Superior Colombiana,” 2020, Accessed: Oct. 30, 2021. [Online]. Available: https://repository.ean.edu.co/handle/10882/10196.

[22] J. A. Orozco Cacique, “Sistema de recomendación de programas universitarios para la orientación profesional de estudiantes de educación media,” *instnameUniversidad los Andes*, 2019, Accessed: Oct. 30, 2021. [Online]. Available: http://hdl.handle.net/1992/44058.

[23] “Vista de Indicadores de deserción universitaria y factores asociados.” https://hemeroteca.unad.edu.co/index.php/educat/article/view/4738/4538 (accessed Oct. 29, 2021).

[24] “Conozca qué información brinda el SPADIES - Sistemas información.” https://www.mineducacion.gov.co/sistemasinfo/spadies/Informacion-Institucional/254651:Conozca-que-informacion-brinda-el-SPADIES (accessed Oct. 29, 2021).

[25] J. Sebastian Gómez, C. Carlos, A. Mendoza Patalagua, A. Darío, and M. Barbosa, “Sistemas de Recomendación de Programas Universitarios Basados en Deep Learning y Procesamiento de Lenguaje Natural,” 2019.

[26] M. Vidal, B. Fernández, and O. Ii, “BÚSQUEDA TEMÁTICA DIGITAL Orientación vocacional Vocational guiding,” 2009. Accessed: Sep. 22, 2020. [Online]. Available: http://scielo.sld.cu.

[27] M. Saldaña Villa and O. A. Barriga, “Adaptación del modelo de deserción universitaria de Tinto a la Universidad Católica de la Santísima Concepción, Chile,” 2010. http://ve.scielo.org/scielo.php?pid=S1315-95182010000400005&script=sci_arttext&tlng=en (accessed Sep. 21, 2020).

[28] “vocación | Definición | Diccionario de la lengua española | RAE - ASALE.” https://dle.rae.es/vocación (accessed Sep. 16, 2020).

[29] M. Teresa, F. Nistal, J. Keven, M. Soto, F. Aglaé, and P. Zaragoza, “La validez estructural de los modelos de Holland y Gati sobre los intereses vocacionales RIASEC en estudiantes mexicanos,” Dec. 2019. Accessed: Sep. 23, 2020. [Online]. Available: http://ojs.ual.es/ojs/index.php/EJREP/article/view/2634.

[30] A. Calderón, S. Dámaris, V. Rebaza, and J. Carlos, “Desarrollo de un sistema experto de orientacion vocacional para la clasificación de carreras profesionales basado en la teoría de Holland,” 2007.

[31] J. Benalcázar, *ANÁLISIS COMPARATIVO DE METODOLOGÍAS DE MINERÍA DE DATOS Y SU APLICABILIDAD A LA INDUSTRIA DE SERVICIOS.*, vol. 1. 2017.

[32] I. Corso and C. Lorena, “Aplicación de algoritmos de clasificación supervisada usando Weka.”

[33] S. Labañino Urbina and O. G. Valencia Zayas, Hugo Alberto Toledano López, “Vista de Algoritmo Random Forest para la detección de fallos en redes de computadoras,” 2019. https://publicaciones.uci.cu/index.php/serie/article/view/445/361 (accessed Dec. 10, 2021).

[34] J. García, J. M. Molina, A. Berlanga, M. Á. Patricio, Á. L. Bustamante, and W. R. Padilla, *Ciencia de datos. Técnicas analíticas y aprendizaje estadístico en un*

enfoque práctico. Alfaomega, 2018.

[35] S. Raschka and V. Mirjalili, *PYTHON MACHINE LEARNING: APRENDIZAJE AUTOMÁTICO Y APRENDIZAJE PROFUNDO*, Segunda Ed. Alfaomega, 2019.

[36] J. A. Gallardo Arancibia, “Modelos de proceso para proyectos de Data Mining (DM) CRISP-DM (Cross Industry Standard Process for Data Mining).”

[37] S. P. Guzman Puentes, “Deserción universitaria, causas de deserción estudiantil,” 2009. https://repository.javeriana.edu.co/bitstream/handle/10554/425/edu54.pdf;jsessionid=C253EF1A0A62A5D47D3E22CB9724F6E2?sequence=1 (accessed Aug. 27, 2020).

[38] W. De Vries, P. León Arenas, J. F. Romero Muñoz, and I. Hernández Saldaña, “¿Desertores o decepcionados? Distintas causas para abandonar los estudios universitarios,” 2011. http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S0185-27602011000400002 (accessed Aug. 27, 2020).

[39] U. Shafique and H. Qaiser, “A Comparative Study of Data Mining Process Models (KDD, CRISP-DM and SEMMA),” *Int. J. Innov. Sci. Res.*, vol. 12, no. 1, pp. 217–222, 2014, Accessed: Nov. 04, 2021. [Online]. Available: http://www.ijisr.issr-journals.org/.

[40] “CRISP-DM, still the top methodology for analytics, data mining, or data science projects - KDnuggets.” https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html (accessed Nov. 04, 2021).

[41] D. M. Pabon Contreras and J. O. Bautista Maldonado, “El papel de la minería de datos en la inteligencia de negocios, una revisión literaria,” 2010. Accessed: Nov. 04, 2018. [Online]. Available: http://ciinatic2017.ufps.edu.co/wordpress/wp-content/uploads/2010/08/CIINATIC_PONENCIA_Mineria-de-datos.pdf.

[42] L. Contreras Chinchilla and K. R. Ferreira, “Análisis del comportamiento de los clientes en las redes sociales mediante técnicas de Minería de Datos.” Accessed: Oct. 15, 2018. [Online]. Available: http://repositorio.uigv.edu.pe/bitstream/handle/20.500.11818/661/COMTEL 2016 - Paper18.pdf?sequence=1&isAllowed=y.

[43] P. Chapman *et al.*, “Step-by-step data mining guide,” 2000.

[44] (Ncr and J. Clinton, “Step-by-step data mining guide,” DaimlerChrysler, 2000. Accessed: Nov. 04, 2018. [Online]. Available: https://www.the-modeling-agency.com/crisp-dm.pdf.

[45] “Ley 1581 de 2012 - EVA - Función Pública.” https://www.funcionpublica.gov.co/eva/gestornormativo/norma.php?i=49981 (accessed Nov. 08, 2021).

[46] ICFES, “DataIcfes.” https://icfesgovco-my.sharepoint.com/:f:/g/personal/dataicfes_icfes_gov_co/EkLXei-dqdlFuRb9hlf1b8IBhZHmwkhRJtY3wNNjttNCoA?e=9PROAP&CT=15892967714

89&OR=OWA-NT&CID=5cc96871-447f-0e87-9de5-3893e123b5ba.

[47] “8. Documentación y Diccionarios - OneDrive.” https://icfesgovco-my.sharepoint.com/personal/dataicfes_icfes_gov_co/_layouts/15/onedrive.aspx?ct=1589296771489&or=OWA-NT&cid=5cc96871-447f-0e87-9de5-3893e123b5ba&id=%2Fpersonal%2Fdataicfes_icfes_gov_co%2FDocuments%2FDataIcfes%2F8. Documentación y Diccionarios%2FCruce_Documentación SaberPro - Saber11.pdf&parent=%2Fpersonal%2Fdataicfes_icfes_gov_co%2FDocuments%2FDataIcfes%2F8. Documentación y Diccionarios (accessed Nov. 09, 2021).

[48] Ministerio de educación Nacional, “Consolidación de un sistema nacional de evaluación estandarizada,” 2013. https://view.officeapps.live.com/op/view.aspx?src=https%3A%2F%2Fwww.mineducacion.gov.co%2F1621%2Farticles-342919_Nov27_alineacion_pruebas_saber.pptx&wdOrigin=BROWSELINK (accessed Dec. 11, 2021).

[49] “Revista Cubana de Ciencia Agrícola,” Accessed: Apr. 12, 2021. [Online]. Available: http://www.redalyc.org/articulo.oa?id=193033033004.

[50] F. J. Muñoz Rosas and E. Álvarez Verdejo, “Métodos de imputación para el tratamiento de datos faltantes: Aplicación mediante R/Splus,” *Rev. Metod. Cuantitativos para la Econ. y la Empres.*, vol. 7, no. 7, pp. 3–30, 2009, [Online]. Available: https://fcecon.unr.edu.ar/web-nueva/sites/default/files/u16/Decimocuartas/bussi_hernandez_mari_mendez_mitas.pdf.

[51] R. O. Calafati, “Memoria del Trabajo Directora: Nuria Pérez Álvarez Profesor Responsable de la Asignatura: Alexandre Sánchez Pla 6 de junio de 2017.”

[52] J. Bussi, L. Hernández, G. Marí, F. Méndez, and G. Mitas, “VISUALIZACIÓN Y MÉTODOS DE IMPUTACIÓN DE DATOS FALTANTES EN LA ENCUESTA DE GASTO DE LOS HOGARES,” 2018. https://fcecon.unr.edu.ar/web-nueva/sites/default/files/u16/Decimocuartas/bussi_hernandez_mari_mendez_mitas.pdf (accessed Apr. 12, 2021).

[53] L. Useche and D. Mesa, “UNA INTRODUCCIÓN A LA IMPUTACIÓN DE VALORES PERDIDOS AN INTRODUCTION TO THE IMPUTATION OF LOST VALUES,” 2006.

[54] J. M. Santos, C. Ministra De Educación, N. Yaneth, G. Tovar, and P. J. Quintero, “Guía de orientación Saber 11.° para instituciones educativas,” Accessed: Nov. 18, 2021. [Online]. Available: www.icfes.gov.co.

[55] S. Manuel and G. Nieto, “Filtrado Colaborativo y Sistemas de Recomendación,” 2007.

[56] D. María José Ramírez Quintana, “APLICACIÓN DE TÉCNICAS DE MINERÍA DE DATOS EN ACCIDENTES DE TRÁFICO,” 2014.

[57] X. Ma, Y. Yang, and Z. Zhou, “Using Machine Learning Algorithm to Predict Student Pass Rates In Online Education,” 2018, doi: 10.1145/3220162.3220188.

[58] M. De and E. Nacional, “PLAN NACIONAL DE ORIENTACIÓN ESCOLAR,” 2021.

[59] D. López, M. Tutores, I. Roberto, V. Rodríguez, L. Claudia, and R. Rodríguez, “API REST PARA EL RECONOCIMIENTO FACIAL DE EMOCIONES (FER REST API).”

[60] S. Vázquez Pérez, “Resolución de la ambigüedad semántica mediante métodos basados en conocimiento y su aportación a tareas de PLN,” 2009.

ANEXOS

Anexo 1: Diccionario de datos Saber 11 Tomado del Sharepoint del ICFES

ESTU_CONSECUTIVO STRING	Código consecutivo Estudiante
COLE_COD_ICFES INTEGER	Código del colegio registrado en el ICFES
COLE_NOMBRE_SEDE STRING	Nombre de la institución
IND_ANNO_TERMINO_BACHILLERATO INTEGER	Año en que terminó el bachillerato
COLE_JORNADA {M, T, C, S, N, U}	Jornada en la que estudió
COLE_VALOR_PENSION {1,2,3,4,5,6,7,8,9,10,11,12,0}	Valor pensión mensual del colegio
ESTU_GENERO {M, F, -}	Genero del estudiante
ESTU_NACIMIENTO_DIA INTEGER	Dia de nacimiento estudiante
ESTU_NACIMIENTO_MES INTEGER	Mes de nacimiento del estudiante
ESTU_NACIMIENTO_ANNO INTEGER	Año de nacimiento del estudiante
ESTU_DEPTO_PRESENTACION STRING	Departamento de presentación del examen
ESTU_MCPIO_PRESENTACION STRING	Municipio de presentación del examen
ESTU_LIMITA_INVIDENTE {-, x}	Tiene limitación de invidente
ESTU_LIMITA_SORDOINTERPRETE {-, x}	Tiene limitación de sordo y necesita interprete
ESTU_LIMITA_SORDONOINTERPRETE {-, x}	Tiene limitación de sordo y no necesita interprete
ESTU_LIMITA_MOTRIZ {-, x}	Tiene limitación motriz
PUNT_BIOLOGIA INTEGER	Puntaje obtenido en Biología
PUNT_MATEMATICAS INTEGER	Puntaje obtenido en Matemáticas
PUNT_FILOSOFIA INTEGER	Puntaje obtenido en Filosofía
PUNT_FISICA INTEGER	Puntaje obtenido en Física
PUNT_HISTORIA INTEGER	Puntaje obtenido en Historia
PUNT_QUIMICA INTEGER	Puntaje obtenido en Química
PUNT_LENGUAJE INTEGER	Puntaje obtenido en Lenguaje

PUNT_GEOGRAFIA INTEGER	Puntaje obtenido en Geografía
PUNT_LECTURA_CRITICA INTEGER	Puntaje obtenido en Lectura Critica
PUNT_SOCIALES_CIUDADANAS INTEGER	Puntaje obtenido en Sociales Ciudadanas
PUNT_RAZONA_CUANTITATIVO INTEGER	Puntaje obtenido en Razonamiento cuantitativo
PUNT_COMP_CIUDADANA INTEGER	Puntaje obtenido en Competencias ciudadanas
COD_INTERDISCIPLINAR INTEGER	Código de Interdisciplinar
PUNT_INTERDISCIPLINAR INTEGER	Puntaje obtenido en Interdisciplinar
COD_IDIOMA INTEGER	Código de idioma seleccionado
PUNT_IDIOMA INTEGER	Puntaje de idioma
ESTU_IES_COD_DESEADA INTEGER	Código de la institución en la que se desea estudiar
ESTU_RAZONINSTITUTO {1,2,3,4,5,6,7,8,-}	Razón para elegir la institución en la que se desea estudiar
ESTU_CARRDESEADA_COD INTEGER	Carrera que se desea estudiar
ESTU_CARRDESEADA_RAZON {1,2,3,4,5,6,-}	Razón por la cual se desea estudiar la carrera
FAMI_PERSONAS_HOGAR INTEGER	Número de personas en el hogar
FAMI_VIVIENDA_PROPIA {S, N, -}	Vivienda propia
FAMI_DEUDA_VIVIENDA {S, N, -}	La vivienda se encuentra en deuda o no
FAMI_APORTANTES INTEGER	Número de personas que aportan en el hogar
FAMI_INGRESO_FMILIAR_MENSUAL {0,1,2,3,4,5,6,7,8,9,-}	Ingreso familiar mensual
FAMI_LEE_ESCRIBE_PADRE {0,1,'99', S, N, *, -}	El padre lee y escribe
FAMI_LEE_ESCRIBE_MADRE {0,1,'99', S, N, *, -}	la madre lee y escribe
FAMI_EDUCA_PADRE {0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,99,-}	Educación del padre

FAMI_EDUCA_MADRE {0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,99,-}	Educación de la madre
FAMI_OCUPA_PADRE {0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,99,-}	ocupación del padre
FAMI_OCUPA_MADRE {0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,99,-}	ocupación de la madre
FAMI_NUM_HERMANOS INTEGER	Numero de hermanos
FAMI_NUM_HERMANOS_EDUSUPERIOR INTEGER	Numero de hermanos con educación superior
FAMI_POSICION_HERMANOS INTEGER	posición entre los hermanos
FAMI_SOST_PERSONAL {N, P, T, S, -}	Sostenimiento personal
ESTU_TRABAJA {0,1,2,3,4,5,6,7, N, S, -}	El estudiante trabaja
ESTU_ACT_PROX_ANNO {1,2,3,4,-}	Actividad del estudiante el siguiente año
COLE_CALENDARIO {A, B, F, OTRO, -}	Calendario del colegio
COLE_BILINGUE {-, N, S}	El colegio es bilingüe
COLE_CARACTER {ACADEMICO, TECNICO,"TECNICO/ACADEMICO", NORMALISTA, DESCONOCIDO, NO_APLICA, -}	Carácter del colegio
COLE_DEPTO_UBICACION STRING	Departamento de Ubicación de la institución donde se estudia
COLE_MCPIO_UBICACION STRING	Municipio de ubicación de la institución donde estudia
ESTU_RESIDE_DEPTO STRING	Departamento de residencia del estudiante
ESTU_RESIDE_MCPIO STRING	Municipio de residencia del estudiante
ESTU_TIPO_DOCUMENTO {T, C, E, R, Q, P, N, V, -}	Tipo de documento del estudiante
ESTU_PILOPAGA {S, N, -}	Pertenece al programa ser pilo paga
fami_estratovivienda STRING	Estrato de la vivienda donde vive el estudiante
fami_personashogar_1 STRING	Personas en el hogar

fami_cuartoshogar {-,0,1,2,3,4,5,6}	Numero de cuartos en el hogar
fami_tieneinternet {-, S, N}	Tiene internet
fami_tienecomputador {-, S, N}	Tiene computador
fami_tienelavadora {-, S, N}	Tiene Lavadora
fami_tienehornomicroogas {-, S, N}	Tiene horno microondas o gas
fami_tieneserviciotv {-, S, N}	Tiene servicio de Tv
fami_tieneautomovil {-, S, N}	Tiene Automovil
fami_tienemotocicleta {-, S, N}	Tiene Motocicleta
fami_tieneconsolavideojuegos {-, S, N}	Tiene Consola de videojuegos
fami_comelechederivados {1,2,3,4,-}	Frecuencia Come leche o derivados
fami_comecarnepescadohuevo {1,2,3,4,-}	Frecuencia Come Carne pescado o huevo
fami_comecerealfrutoslegumbre {1,2,3,4,-}	Frecuencia come cereal frutos legumbre
fami_situacioneconomica {0,1,2,3,-}	situación económica actual
estu_dedicacionlecturadiaria {0,1,2,3,4,5,-}	Número de horas que dedica a la lectura diaria
estu_dedicacioninternet {0,1,2,3,4,5,-}	Número de horas que dedica a navegar por internet
estu_horassemanatrabaja {0,1,2,3,4,-}	Número de horas a la semana en las que el estudiante trabaja
estu_tiporemuneracion {0,1,2,3,-}	Tipo de remuneración
punt_global INTEGER	Puntaje global prueba Saber11

Anexo 2: Diccionario de datos pruebas Saber Pro

Campo	Descripción
ESTU_TIPODOCUMENTO	Tipo de Documento
ESTU_NACIONALIDAD	Nacionalidad
ESTU_GENERO	Género
ESTU_FECHANACIMIENTO	Fecha de Nacimiento
ESTU_EXTERIOR	Indica si el estudiante presenta la prueba Saber Pro-Exterior
PERIODO	Periodo de aplicación de la prueba
ESTU_CONSECUTIVO	Id público del estudiante
ESTU_ESTADOCIVIL	Estado civil
ESTU_ESTUDIANTE	Indica si realizó la inscripción por medio de la Institución de Educación Superior (estudiante) o fue individual
ESTU_PAIS_RESIDE	Código del país donde reside actualmente el estudiante
ESTU_TIENEETNIA	¿Pertenece usted a un grupo étnico minoritario?
ESTU_ETNIA	¿Cuál es el grupo étnico minoritario al que pertenece?
ESTU_DISC_FISICA	Se inscribió indicando que tiene discapacidad -Física movilidad
ESTU_DISC_SISTEM	Se inscribió indicando que tiene discapacidad -Sistémica
ESTU_DISC_AUDITCASTELL	Se inscribió indicando que tiene discapacidad -Auditiva Usuario del Castellano
ESTU_DISC_AUDITLSC	Se inscribió indicando que tiene discapacidad -Auditiva Usuario Lengua de Señas Colombiana
ESTU_DISC_VISUALCEGUE	Se inscribió indicando que tiene discapacidad -Visual Ceguera
ESTU_DISC_VISUALBAJA	Se inscribió indicando que tiene discapacidad -Visual Baja Visión Irreversible
ESTU_DISC_VOZHABLA	Se inscribió indicando que tiene discapacidad -Trastorno permanente de la voz y el habla
ESTU_DISC_INTELEC	Se inscribió indicando que tiene discapacidad -Intelectual
ESTU_DISC_SORDOCEGUERA	Se inscribió indicando que tiene discapacidad -Sordo ceguera
ESTU_DISC_ASPERGER	Se inscribió indicando que tiene discapacidad -Asperger
ESTU_DISC_PSICOSOCIAL	Se inscribió indicando que tiene discapacidad -Psicosocial

ESTU_LIMITA_MOTRIZ	Se inscribió indicando que tiene una discapacidad - Motriz
ESTU_LIMITA_INVIDENTE	Se inscribió indicando que tiene una discapacidad - Invidente
ESTU_LIMITA_CONDICIONESPECIAL	Se inscribió indicando que tiene una discapacidad - Condición especial
ESTU_LIMITA_SORDO	Se inscribió indicando que tiene una discapacidad - Sordo
ESTU_LIMITA_SDOWN	Se inscribió indicando que tiene una discapacidad – Síndrome de Down
ESTU_LIMITA_AUTISMO	Se inscribió indicando que tiene una discapacidad - Autismo
ESTU_LIMITA_SORDOCONINTERPRETE	Se inscribió indicando que tiene una discapacidad – Sordo con intérprete
ESTU_LIMITA_SORDOSININTERPRETE	Se inscribió indicando que tiene una discapacidad – Sordo sin intérprete
ESTU_LIMITA_SORDOCEGUERA	Se inscribió indicando que tiene una discapacidad – Sordo ceguera
ESTU_APO_DESPLAZASITIO	Solicitó apoyo para la discapacidad seleccionada
ESTU_APO_ACOMPAÑAMIENTO	Solicitó apoyo para la discapacidad seleccionada
ESTU_APO_JEFESALON	Solicitó apoyo para la discapacidad seleccionada
ESTU_APO_LECTORAPOYO	Solicitó apoyo para la discapacidad seleccionada
ESTU_APO_MANIOBRARMAT	Solicitó apoyo para la discapacidad seleccionada
ESTU_APO_INTERPSEÑAS	Solicitó apoyo para la discapacidad seleccionada
ESTU_APO_INTERPSORDOCIEGO	Solicitó apoyo para la discapacidad seleccionada
ESTU_APO_NOREQUIERE	Seleccionó que no necesita apoyo para la discapacidad seleccionada
ESTU_DEPTO_RESIDE	Departamento de residencia
ESTU_COD_RESIDE_DEPTO	Código Dane del departamento de residencia
ESTU_MCPIO_RESIDE	Municipio de residencia
ESTU_COD_RESIDE_MCPIO	Código Dane del municipio de residencia
ESTU_AREARESIDE	Área de residencia
ESTU_GRADUO_BACHILLER	¿Se graduó usted de bachiller?
ESTU_FECHAGRADOBACHILLER	Fecha de grado de bachiller
ESTU_MES_TERMINOBACHILLER	Mes en que terminó el bachillerato
ESTU_ANO_TERMINOBACHILLER	Año en que terminó el bachillerato
ESTU_ANO_EGRESO	Año en que salió de bachiller
ESTU_CODICFESCOLE_TERMINO	Código Icfes del colegio donde terminó bachillerato
ESTU_COLE_TERMINO	Nombre del colegio donde terminó bachillerato
ESTU_CODDANE_COLE_TERMINO	Código DANE del colegio donde terminó bachillerato

ESTU_COD_COLE_MCPIO_TERMINO	Código DANE del municipio donde está ubicado el colegio donde terminó bachillerato
ESTU_OTROCOLE_TERMINO	Escriba el nombre completo del colegio donde terminó
ESTU_TITULOOBTENIDOBACHILLER	Título de bachiller obtenido
ESTU_ANO_EXAMENESTADO_SB11	Año en el que presentó el examen SB11
ESTU_SEMESTRE_EXAMENESTADOSB11	Semestre en el que presentó el examen SB11
ESTU_PORCENTAJECREDITOSAPROB	¿Qué porcentaje de créditos necesarios para optar el título, ha cursado y aprobado hoy?
ESTU_VALORMATRICULAUEXT	¿Cuál es el valor de la matrícula del último semestre cursado (sin considerar descuentos o becas)? SOLO PARA PRO-EXTERIOR
ESTU_VALORMATRICULAUNIVERSIDAD	¿Cuál es el valor de la matrícula del último semestre cursado (sin considerar descuentos o becas)?
ESTU_PAGOMATRICULAEXT	Los recursos con que usted canceló los semestres del año pasado eran en su mayoría: (SOLO PARA PRO-EXTERIOR)
ESTU_PAGOMATRICULABECA	Variable que define si el pago de matrícula es por beca
ESTU_PAGOMATRICULACREDITO	Variable que define si el pago de matrícula es mediante crédito
ESTU_PAGOMATRICULAPADRES	Variable que define si el pago de matrícula lo realizan los padres del estudiante
ESTU_PAGOMATRICULAPROPIO	Variable que define si el pago de matrícula es por recursos propios
ESTU_COMOCAPACITOEXAMENSB11	¿Cómo se preparó para el examen SABER 11?
ESTU_TOMO_CURSOPREPARACION	¿Cómo se preparó para el examen SABER PRO?
ESTU_CURSODOCENTESIES	Se preparó para el examen Saber Pro en su IES con docentes de la institución (número de horas)
ESTU_CURSOIESAPOYOEXTERNO	Se preparó para el examen Saber Pro en un curso organizado por la institución con apoyo de un instituto de preparación de exámenes externos (número de horas)
ESTU_CURSOIESEXTERNA	Se preparó para el examen Saber Pro en un instituto de preparación de exámenes (número de horas)
ESTU_SIMULACROTIPOICFES	¿Qué actividades desarrolló en el curso de preparación para el examen Saber Pro?: Simulacros con preguntas tipo ICFES

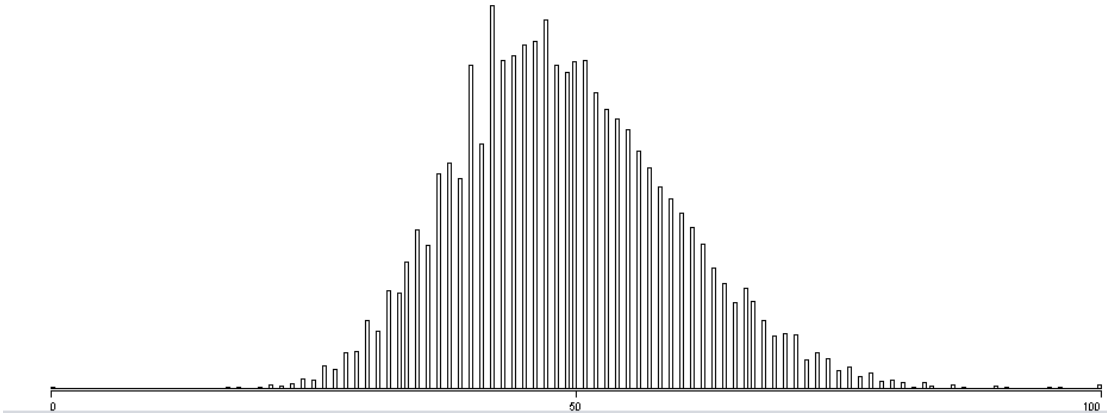
ESTU_ACTIVIDADREFUERZOAREAS	¿Qué actividades desarrolló en el curso de preparación para el examen Saber Pro?: Clases de refuerzo en algunas áreas
ESTU_ACTIVIDADREFUERZOGENERIC	¿Qué actividades desarrolló en el curso de preparación para el examen Saber Pro?: Refuerzo en desarrollo de competencias genéricas
ESTU_PAISDOCUMENTOSB11	País de origen del documento de identidad con el cual presentó la prueba SABER 11
ESTU_TIPODOCUMENTOSB11	Tipo de documento de identidad con el cual presentó la prueba SABER 11
ESTU_SEMESTRECURSA	Semestre que cursa actualmente el estudiante
ESTU_ULTIMOGRADOAPROBO	¿Cuál fue el último grado a probado en una institución de educación formal?
ESTU_ANOTERMINOULTIMOGRADO	¿En qué año aprobó el último grado a probado en una institución de educación formal?
ESTU_VECESPRESENTOEXAMENESTAD	¿Cuántas veces ha presentado el examen SBPRO? Para INDIVIDUALES
ESTU_VECESPRESENTOEXAMEN	¿Cuántas veces ha presentado el examen SBPRO? Para ESTUDIANTES
FAMI_HOGARACTUAL	Variable que indica si el hogar actual donde vive es permanente o temporal
FAMI_CABEZAFAMILIA	¿Es usted jefe de hogar o cabeza de familia?
FAMI_NUMPERSONASACARGO	¿Cuántas personas dependen económicamente de usted? (Incluya parientes, no parientes y servicio doméstico que viven permanentemente en su hogar)
FAMI_EDUCACIONPADRE	Nivel educativo más alto alcanzado por el padre
FAMI_EDUCACIONMADRE	Nivel educativo más alto alcanzado por la madre
FAMI_OCUPACIONPADRE	Ocupación u oficio del padre
FAMI_OCUPACIONMADRE	Ocupación u oficio de la madre
FAMI_TRABAJOLABORPADRE	Señale aquella labor que sea más similar al trabajo que realizó su padre durante la mayor parte del último año:
FAMI_TRABAJOLABORMADRE	Señale aquella labor que sea más similar al trabajo que realizó su madre durante la mayor parte del último año:
FAMI_ESTRATOVIVIENDA	Estrato socioeconómico de su vivienda según recibo de energía eléctrica
FAMI_NIVEL_SISBEN	Puntaje de SISBEN en el que está clasificada su familia

FAMI_PERSONASHOGAR	¿Cuántas personas conforman el hogar donde vive actualmente, incluido usted?
FAMI_CUARTOSHOGAR	En total, ¿en cuántos cuartos duermen las personas de su hogar?
FAMI_PISOS_HOGAR	¿Cuál es el material de los pisos que predomina en su vivienda?
FAMI_TIENEINTERNET	¿Su hogar cuenta con servicio o conexión a internet?
FAMI_TIENESERVICIOTV	¿Su hogar cuenta con servicio cerrado de televisión?
FAMI_TIENECOMPUTADOR	¿Cuáles de los siguientes bienes posee su hogar?: Computador
FAMI_TIENELAVADORA	¿Cuáles de los siguientes bienes posee su hogar?: Máquina lavadora de ropa
FAMI_TIENEHORNOMICROOGAS	¿Cuáles de los siguientes bienes posee su hogar?: Horno Microondas u Horno eléctrico o a gas
FAMI_TIENE_MICROONDAS	¿Cuáles de los siguientes bienes posee su hogar?: Horno microondas
FAMI_TIENE_HORNO	¿Cuáles de los siguientes bienes posee su hogar?: Horno eléctrico o a gas
FAMI_TIENEAUTOMOVIL	¿Cuáles de los siguientes bienes posee su hogar?: Automóvil particular
FAMI_TIENEMOTOCICLETA	¿Cuáles de los siguientes bienes posee su hogar?: Motocicleta
FAMI_TIENE_NEVERA	¿Cuáles de los siguientes bienes posee su hogar?: Nevera
FAMI_TIENE_CELULAR	¿Cuáles de los siguientes bienes posee su hogar?: Servicio de teléfono móvil
FAMI_TIENECONSOLAVIDEOJUEGOS	¿Cuáles de los siguientes bienes posee su hogar?: Consola para juegos electrónicos (PlayStation, Xbox, Nintendo, etc.) *En 2017 solo se preguntó para Pro exterior
FAMI_CUANTOSCOMPARTEBAÑO	¿Con cuántas personas comparte usted el baño en su hogar? *En 2017 solo se preguntó para Pro exterior
FAMI_NUMLIBROS	¿Cuántos libros físicos o electrónicos hay en su hogar excluyendo periódicos, revistas, directorios telefónicos y libros del colegio?
FAMI_TELEFONO	¿Con cuáles de los siguientes servicios públicos, privados o comunales cuenta su hogar?: Teléfono (fijo)

FAMI_INGRESO_FMILIAR_MENSUAL	¿Cuál es el total de ingresos mensuales de su hogar, en términos de salarios mínimos (SM)?
ESTU_DEDICACIONLECTURADIARIA	Usualmente, ¿cuánto tiempo al día dedica a leer por entretenimiento?
ESTU_DEDICACIONINTERNET	Usualmente, ¿cuánto tiempo al día dedica a navegar en internet? Excluya actividades académicas
ESTU_TRABAJA_ACTUALMENTE	¿Trabaja usted actualmente?
ESTU_HORASSEMANATRABAJA	¿Cuántas horas trabajó usted durante la semana pasada?
ESTU_TIPOREMUNERACION	¿Usted recibe algún tipo de remuneración por trabajar?
INST_COD_INSTITUCION	Código de la Institución de Educación Superior
INST_NOMBRE_INSTITUCION	Nombre de la Institución de Educación Superior
ESTU_PRGM_ACADEMICO	Nombre del programa académico que estudia
ESTU_SNIES_PRGMACADEMICO	Código SNIES del programa académico que estudia
GRUPOREFERENCIA	Nombre del Grupo de Referencia al que pertenece el programa académico del estudiante
ESTU_PRGM_CODMUNICIPIO	Código del municipio donde se ofrece el programa académico
ESTU_PRGM_MUNICIPIO	Nombre del municipio donde se ofrece programa académico
ESTU_PRGM_DEPARTAMENTO	Nombre del departamento donde se ofrece el programa académico
ESTU_NIVEL_PRGM_ACADEMICO	Nivel del programa académico
ESTU_METODO_PRGM	Metodología del programa académico
ESTU_NUCLEO_PREGRADO	Nombre del núcleo de pregrado al que pertenece el programa académico
ESTU_INST_CODMUNICIPIO	Código del municipio donde está ubicada la IES
ESTU_INST_MUNICIPIO	Nombre del municipio donde está ubicada la IES
ESTU_INST_DEPARTAMENTO	Nombre del departamento donde está ubicada la IES
INST_CARACTER_ACADEMICO	Carácter académico de la IES
INST_ORIGEN	Naturaleza u origen de la IES
ESTU_PRIVADO_LIBERTAD	Respuesta del evaluado si actualmente se encuentra privado de la libertad
ESTU_COD_MCPIO_PRESENTACION	Código Dane del municipio presentación del examen
ESTU_MCPIO_PRESENTACION	Municipio de presentación del examen

ESTU_DEPTO_PRESENTACION	Código Dane del departamento del municipio de presentación del examen
ESTU_COD_DEPTO_PRESENTACION	Departamento del municipio de presentación del examen
MOD_RAZONA_CUANTITAT_PUNT	Puntaje razonamiento cuantitativo
MOD_RAZONA_CUANTITAT_DESEM	Nivel de desempeño razonamiento cuantitativo
MOD_RAZONA_CUANTITATIVO_PNAL	Percentil nacional razonamiento cuantitativo
MOD_RAZONA_CUANTITATIVO_PGREF	Percentil por grupo de referencia razonamiento cuantitativo
MOD_RAZONA_CUANTITATIVO_PNBC	Percentil por NBC razonamiento cuantitativo
MOD_LECTURA_CRITICA_PUNT	Puntaje lectura crítica
MOD_LECTURA_CRITICA_DESEM	Nivel de desempeño lectura crítica
MOD_LECTURA_CRITICA_PNAL	Percentil nacional lectura crítica
MOD_LECTURA_CRITICA_PGREF	Percentil por grupo de referencia lectura crítica
MOD_LECTURA_CRITICA_PNBC	Percentil por NBC lectura crítica
MOD_COMPETEN_CIUDADA_PUNT	Puntaje competencias ciudadanas
MOD_COMPETEN_CIUDADA_DESEM	Nivel de desempeño competencias ciudadanas
MOD_COMPETEN_CIUDADA_PNAL	Percentil nacional competencias ciudadanas
MOD_COMPETEN_CIUDADA_PGREF	Percentil por grupo de referencia competencias ciudadanas
MOD_COMPETEN_CIUDADA_PNBC	Percentil por NBC competencias ciudadanas
MOD_INGLES_PUNT	Puntaje inglés
MOD_INGLES_DESEM	Nivel de desempeño inglés
MOD_INGLES_PNAL	Percentil nacional inglés
MOD_INGLES_PGREF	Percentil por grupo de referencia inglés
MOD_INGLES_PNBC	Percentil por NBC inglés
MOD_COMUNI_ESCRITA_PUNT	Puntaje comunicación escrita
MOD_COMUNI_ESCRITA_DESEM	Nivel de desempeño comunicación escrita
MOD_COMUNI_ESCRITA_PNAL	Percentil nacional comunicación escrita
MOD_COMUNI_ESCRITA_PGREF	Percentil por grupo de referencia comunicación escrita
MOD_COMUNI_ESCRITA_PNBC	Percentil por NBC comunicación escrita
PUNT_GLOBAL	Puntaje total obtenido
PERCENTIL_GLOBAL	Percentil global en que se encuentra el evaluado
PERCENTIL_NBC	Percentil nacional por NBC
ESTU_INSE_INDIVIDUAL	Índice socioeconómico a nivel de estudiante
ESTU_NSE_INDIVIDUAL	Nivel socioeconómico a nivel de estudiante
ESTU_NSE_IES	Nivel Socioeconómico del Establecimiento
ESTU_ESTADOINVESTIGACION	Identifica los usuarios que están en proceso de investigación en el Icfes

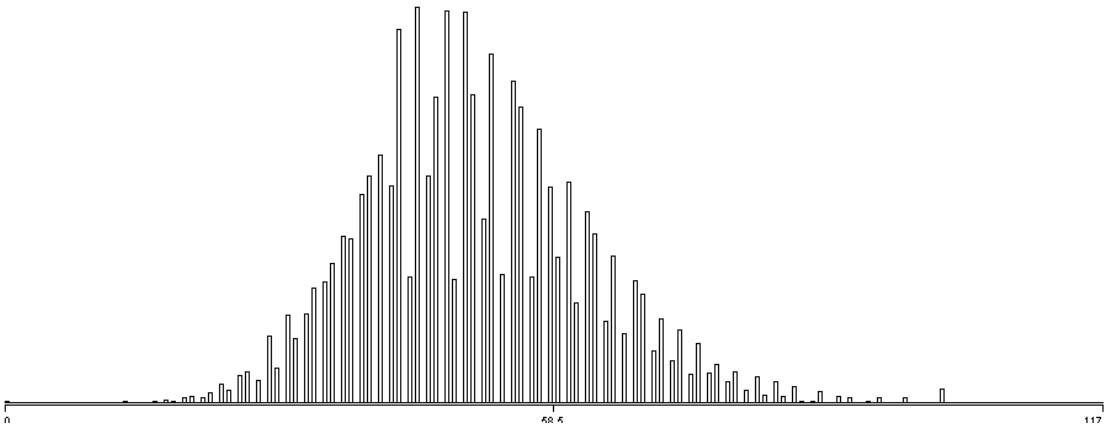
Anexo 3 Distribución puntajes Biología Elaboración propia



Anexo 4 Tabla de distribución de puntajes Biología Elaboración propia

Statistic	Value
Minimum	0
Maximum	100
Mean	49.042
StdDev	10.444

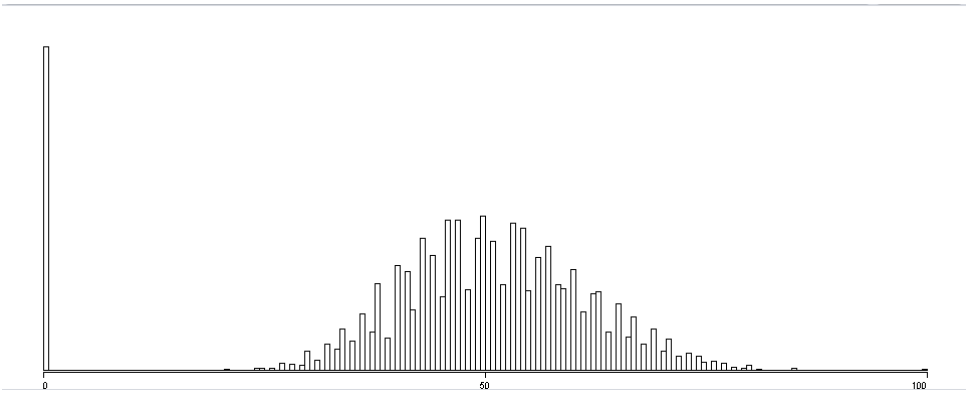
Anexo 5 Distribución puntajes Matemáticas Elaboración propia



Anexo 6 Tabla de distribución de puntajes Matemáticas Elaboración propia

Statistic	Value
Minimum	0
Maximum	117
Mean	50.056
StdDev	12.085

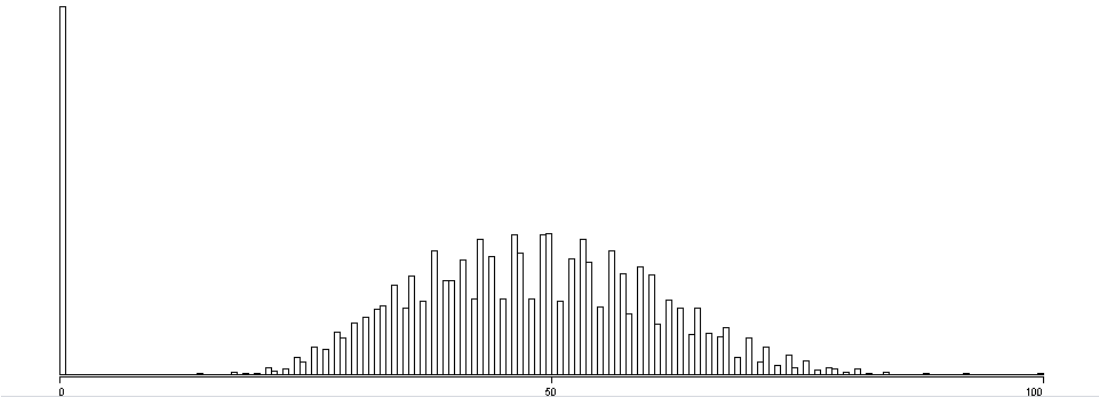
Anexo 7 Distribución puntajes Lectura Crítica Elaboración propia



Anexo 8 Tabla distribución de puntajes lectura crítica Elaboración propia

Statistic	Value
Minimum	0
Maximum	100
Mean	46.711
StdDev	17.476

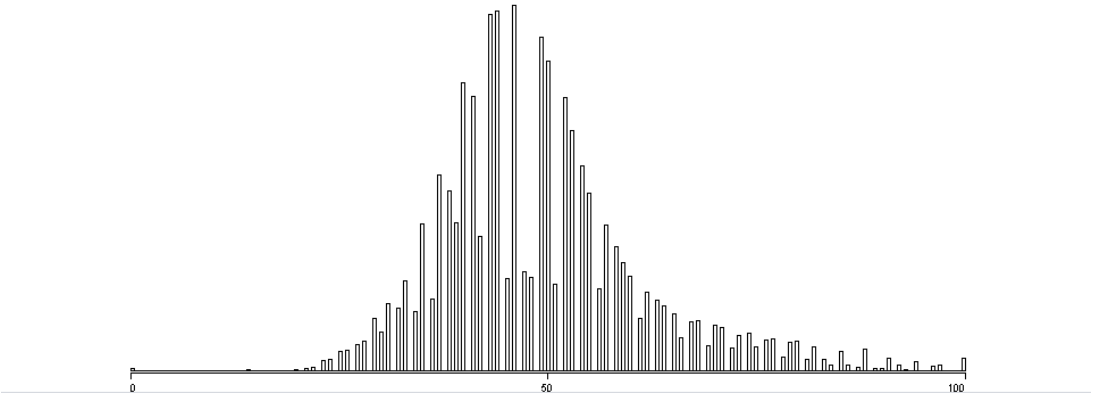
Anexo 9 Distribución puntajes Sociales Ciudadanas Elaboración propia



Anexo 10 Tabla distribución puntajes Sociales Ciudadanas Elaboración propia

Statistic	Value
Minimum	0
Maximum	100
Mean	44.06
StdDev	17.832

Anexo 11 Distribución puntajes Idioma Elaboración propia



Anexo 12 Tabla de distribución puntajes Idioma Elaboración propia

Statistic	Value
Minimum	0
Maximum	100
Mean	49.312
StdDev	12.552