Santiago de Cali, 30 de 07 del 2019

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

Cumplido los requisitos establecidos en los artículos 5.6 y 5.7 de las Directrices para Trabajo de Grado de Maestría, solicitamos se autorice la sustentación del Trabajo de Grado denominado "**Uso de Minería de datos para predecir el rendimiento académico de estudiantes de la Institución Educativa Libardo Madrid Valderrama"**, realizado por el (la) estudiante **Andrés Racines Bolaños** con código **8924860** perteneciente al énfasis en Ingeniería Sistemas y Computación, bajo la dirección del profesor **María Constanza Pabón, PhD.**

El suscrito director del Trabajo de Grado autoriza para que se proceda a hacer su sustentación ante el Tribunal que para el efecto se designe, toda vez que ha revisado meticulosamente el documento y avala que el Trabajo de Grado ya se encuentra listo para ser evaluado oficialmente.

Atentamente,

_____	_____
Andrés Racines Bolaños	María Constanza Pabón, PhD.
C.C. 94.507.629 de Cali	C.C. 34.559.226

Documentación anexa:
Dos copias anilladas del documento de Trabajo de Grado, con impresión por lado y lado y paginación completa.
El resumen del Trabajo de Grado en formato electrónico (máximo 1 página).

Datos del Estudiante

Nombres y Apellidos: Andrés Racines Bolaños

Dirección: Calle 12 Oeste # 10 -50 Apto 403 Torre 2

Teléfono: 3059275

Celular: 3216447063

Correo electrónico: andracin@gmail.com

Profesión: Licenciado en Informática

Universidad: Fundación Universitaria Católica Lumen Gentium Cali

Empresa: Servicio Nacional de Aprendizaje /Regional Valle/Centro de Electricidad y Automatización Industrial C.E.A.I

Cargo: Instructor

Experiencia laboral y profesional como Coordinador de Sistemas y Telecomunicaciones, Cargo desempeñado en la Corporación Observatorio Sismológico del Sur Occidente de la Universidad del valle hasta Julio 2012, Actualmente soy Docente instructor de SENA Centro de Electricidad y Automatización Industrial Cali.

Me considero una persona comprometida, hábil, proactiva y creativa, con grandes capacidades para trabajar bajo presión y grandes cualidades de adaptación a diferentes contextos laborales y profesionales, con una constante motivación a la superación personal y de excelentes relaciones humanas.

FICHA RESUMEN

PROYECTO DE TRABAJO DE GRADO DE MAESTRÍA

TITULO: "**Uso de Minería de datos para predecir el rendimiento académico de estudiantes de la Institución Educativa Libardo Madrid Valderrama"**

1. ÉNFASIS: Ingeniería **Sistemas y Computación**
2. ÁREA DE INVESTIGACIÓN: **Minería de Datos**
3. ESTUDIANTE: **Andrés Racines Bolaños**
4. CORREO ELECTRÓNICO: **andracin@gmail.com**
5. DIRECTOR: María Constanza Pabón, PhD
6. CO-DIRECTOR(ES):
7. GRUPO QUE LO AVALA:
8. OTROS GRUPOS:

9. PALABRAS CLAVE: Minería de datos, demográfico, CRISP-DM.
10. CÓDIGOS UNESCO CIENCIA Y TECNOLOGÍA: **1203.04** , **1203.99**
11. FECHA DE INICIO: 16 de 03 de 2018 DURACIÓN ESTIMADA: 13 Meses

12. RESUMEN (máximo una página).

Resumen

El presente es el informe final del trabajo realizado para optar por el título de Maestría en Ingeniería con énfasis en sistemas y computación de la Pontificia Universidad Javeriana Cali. Se describen las actividades realizadas en la creación de un modelo de minería de datos para predecir el rendimiento académico en matemáticas de los estudiantes de la Institución Educativa Libardo Madrid Valderrama para el siguiente año lectivo, a partir de los datos socioeconómicos, socioculturales, y de las notas de los períodos académicos del año anterior. La Institución educativa es de carácter oficial y pertenece a la secretaria de educación de Cali.

Para la creación del modelo de minería de datos se implementó la metodología CRISP-DM, los datos fueron suministrados por la Institución Educativa y contienen las notas finales de matemáticas de los cuatro periodos académicos de grados 8°, 9° y 10 cursados por los estudiantes de secundaria en el año lectivo 2016, la nota final de matemáticas del año lectivo 2017 es la clase objetivo, adicionalmente se recopiló la información socioeconómica y sociocultural de dichos estudiantes a través de una encuesta web.

Palabras Claves: Minería de datos, CRISP-DM, Modelado, Socioeconómica, Sociocultural.

Abstract

The present is the final report of the graded work done to opt for the Master's degree in Engineering with emphasis in systems and computing from the Pontificia Universidad Javeriana Cali, It describes the activities carried out for the creation of a model using data mining to predict the academic performance in mathematics of the students of the Libardo Madrid Valderrama Educational Institution for the following academic year based on socio-economic, socio-cultural data and the grades of other academic periods. The educational institution is of an official nature and belongs to the secretary of education of Cali.
.

For the creation of the data mining model, the CRISP-DM methodology was implemented. The data were provided by the Educational Institution and contain the final math grades of the four academic periods of grades 8, 9 and 10 in the 2016 school year, the final grade of mathematics for the 2017 school year is the target class, in addition, the socioeconomic and sociocultural information was collected through a web survey.

Keywords: Data Mining, CRISP-DM, Predictive Modeling, Socioeconomic, Sociocultural

Uso de Minería de datos para predecir el rendimiento académico de estudiantes de la Institución Educativa Libardo Madrid Valderrama.

Por: Andrés Racines Bolaños

Director de Proyecto:

María Constanza Pabón, PhD

Pontificia Universidad Javeriana Cali

Maestría en Ingeniería con Énfasis en Sistemas y Computación

Julio-2019

CONTENIDO

INTRODUCCIÓN

Predecir el rendimiento académico de los estudiantes es un estrategia prometedora para mitigar el riesgo académico en Colombia [1], el riesgo académico se define como la probabilidad que tiene un estudiante en desertar del sistema educativo, el fenómeno de la deserción no se puede atribuir a un único factor, algunos estudios señalan que el rendimiento académico bajo y básico en la mayoría de casos es un factor común entre los estudiantes que tienden a desertar del sistema educativo [1].

En el presente trabajo se describen las actividades realizadas para creación de un prototipo de aplicación Web que incluye un modelo de minería de datos, el cual permite cargar las notas finales de periodos académicos cursados por los estudiantes más los datos socioeconómicos y socioculturales, con el objetivo predecir mediante el uso de técnicas de minería de datos el rendimiento académico de los estudiantes para el siguiente año lectivo. Dadas las características del problema y de los datos disponibles en el diseño del prototipo Web de minería de datos, se evaluaron técnicas de minería de datos orientadas a la predicción de tipo supervisado.

Los datos para la creación del prototipo Web de minería de datos, fueron suministrados por la Institución educativa Libardo Madrid Valderrama de Santiago de Cali y contienen las notas finales de matemáticas de los cuatro periodos académicos de los grados 8°, 9° y 10° cursados por los estudiantes de secundaria en el año lectivo 2016. La nota final de matemáticas del año lectivo 2017 es la clase objetivo. Adicionalmente se recopiló la información socioeconómica y sociocultural de dichos estudiantes a través de una encuesta realizada cuando cursaban los año lectivos 9°, 10° y 11° en 2017.

1. PLANTEAMIENTO DEL PROBLEMA

El Ministerio de Educación Nacional realiza el seguimiento a los niños y jóvenes que abandonan el sistema educativo con el fin de conocer las causas y buscar soluciones [1]. Sin embargo, los estudios se han enfocado la deserción estudiantil, subestimando la importancia de generar estrategias preventivas que den alcance a sus causas; particularmente estrategias que se ocupen del riesgo académico bajo y básico, como una de las principales causas de la deserción estudiantil.

A partir del año 2008 el Ministerio de Educación Nacional inicia la aplicación de Encuesta Nacional de Deserción Escolar (ENDE) [2], que son aplicadas en todas las secretarias de educación del país y que contiene información del sistema de matrículas estudiantil (SIMAT) y de los Directivos Docentes de cada colegio público adscrito a una Entidad Territorial Certificada (ETC) de Colombia [2], Los resultados de la (ENDE) publicados en 2011, permiten identificar los factores asociados a la deserción por departamento y Entidad Territorial Certificada (ETC), se destacan 21 factores asociados a la deserción, dichos factores se dividen en dos grupos: (1) gestión y administración de la educación con 24% de impacto sobre la deserción estudiantil, (2) las dificultades académicas con 43% de impacto sobre la deserción estudiantil.

La Institución Educativa Libardo Madrid Valderrama, carece de un mecanismo para prevenir el riesgo académico, que le permitiera identificar las causas y actuar sobre estas causas antes de que la deserción ocurra. Un primer acercamiento a este mecanismo fue proponer un prototipo Web de minería de datos que permita brindar información útil a directivos y docentes para diseñar mejores estrategias para prevenir el riesgo académico, disminuyendo la probabilidad de la deserción.

1.1. OBJETIVOS DEL PROYECTO

1.1.1 OBJETIVO GENERAL

- Desarrollar un prototipo de una aplicación Web que incluya un modelo entrenado de minería de datos que pueda ser usado por la Institución Educativa Libardo Madrid Valderrama, para hacer la predicción del rendimiento académico de los estudiantes para el siguiente año lectivo.

1.1.2 OBJETIVOS ESPECÍFICOS

- Aplicar el descubrimiento de conocimiento con el uso de la metodología CRISP-DM.
- Evaluar varios modelos de clasificación de minería de datos para encontrar el que mejor prediga el rendimiento de los estudiantes.
- Desarrollar un prototipo Web de un modelo entrenado de minería de datos que, dadas las características socioculturales, socioeconómicas y las notas finales de periodos anteriores cursados por los estudiantes, prediga el rendimiento académico para el siguiente año lectivo.

2. ALCANCES DEL TRABAJO DE GRADO

Creación de un prototipo de aplicación Web que incluye un modelo entrenado de minería de datos para predecir el rendimiento académico de los estudiantes de la Institución Educativa Libardo Madrid Valderrama a partir de: (1) Las notas de otros periodos académicos cursados, (2) datos socioeconómicos y socioculturales de los estudiantes.

El prototipo de aplicación Web no incluye la autenticación ni seguridad. Los requerimientos de hardware y software de la aplicación Web se describen en el anexo 4. El funcionamiento de la aplicación se describe en el anexo 5.

3. JUSTIFICACIÓN DEL TRABAJO DE GRADO

Existen muchas circunstancias y contextos válidos para explicar el por qué se presenta el fenómeno de deserción académica. Las explicaciones pueden ser tan subjetivas y complejas que no existiría una sola respuesta que abarque todas las circunstancias y contextos posibles [3].

Los factores que determinan la deserción académica no son expresables como una suma de características, adicionalmente, la complejidad del problema no permite una única solución. Por lo tanto es necesario explorar en los datos socioeconómicos, socioculturales y académicos de los estudiantes para descubrir patrones y tendencias que permitan soportar nuevas hipótesis sobre el porqué se presenta dicho fenómeno [4].

Existen trabajos similares a este realizados en otras instituciones educativas que no son aplicables a este trabajo, dado que un modelo de minería describe el comportamiento de un grupo particular de estudiantes. Por lo cual fue necesario generar un modelo a partir de los datos propios de la Institución Educativa Libardo Madrid Valderrama que describa el comportamiento en el rendimiento académico de los estudiantes de esta institución en particular.

Prevenir el riesgo académico es importante para la Institución Educativa Libardo Madrid Valderrama de Santiago de Cali. Porque permite mitigar la deserción estudiantil [1] permitiendo a directores, docentes y psicólogos tomar decisiones oportunas para diseñar estrategias que actúen sobre los procesos de formación de los estudiantes.

4. MARCO DE REFERENCIA

4.1. MARCO TEÓRICO

En esta sección se presentan los fundamentos teóricos que soportan el trabajo realizado, teorías como: minería de datos (DM), descubrimiento de conocimiento (KDD) y la metodología CRISP-DM, fueron implementadas en el desarrollo del trabajo.

4.1.1 MINERÍA DE DATOS (DM)

En la última década la aplicación de técnicas de minería de datos ha tomado una importante relevancia para modelar distintos tipos de problemas en diferentes ramas de la ciencia, principalmente en estudios relacionados con

la predicción [7]. Los procesos de minería datos se clasifican de acuerdo al tipo de tareas que se pueden realizar, en dos tipos:

a) Predicción, la cual usa algunas variables para construir un modelo que permita predecir el valor desconocido o futuro, de otra variable.
b) Descripción, permite encontrar patrones interpretables por los humanos. Dichas tareas a su vez pueden ser divididas en dos tipos: Supervisadas en la cual la categoría o clase están predefinidas y las tareas no supervisadas en la cual la categoría o clases del conjunto de datos no está predefinido [7].

Las técnicas de minería de datos más usadas de acuerdo con el tipo de tarea que puede realizar son [16]:

- Clasificación (descripción o predicción). Que permite asignar objetos a una o varias categorías predefinidas (variable discreta). Dentro de las técnicas de clasificación más usadas son:
  - Métodos basados en árboles de decisión.
  - Métodos basados en reglas.
  - Vecino más próximo.
  - Redes neuronales artificiales.
  - Clasificadores bayesianos.
  - Máquinas de soporte de vectores.
- Regresión (predicción) Que permite evaluar variables de tipo continúo.
- Clustering (descripción). Que permite dividir los datos en grupos (clusters) que tienen significado (entender los datos) y/o son útiles.
- Descubrir reglas de asociación (descripción). Que permite descubrir relaciones escondidas en los datos.
- Descubrir patrones (descripción). Que permite descubrir patrones escondidos en los datos.

El análisis de regresión busca la relación entre variables independientes, el procedimiento de clasificación es procedimiento de variables individuales que están contenidas en los grupos basados en la información cuantitativa respecto a una o muchas características inherentes a las variables y basados en un entrenamiento de conjuntos previos con etiquetas por cada variable [13].

4.1.2 DESCUBRIMIENTO DE CONOCIMIENTO (KDD)

La extracción de conocimiento, conocido como Knowledge Discovery in Databases (KDD), es el proceso de descubrir conocimiento e información potencialmente útil a partir de los datos No es un proceso automático, es un proceso iterativo que explora grandes volúmenes de datos para encontrar relaciones. [15].

Una de las metodologías más usadas para orientar los procesos de descubrimiento de conocimiento es la metodología CRISP-DM, por sus siglas en inglés: Cross-Industry Standard Process for Data Mining [8].
La metodología incluye descripciones de las fases del proyecto de minería de datos, las tareas necesarias en cada fase y una explicación de las relaciones entre las tareas. Como modelo de proceso, CRISP-DM ofrece un resumen del ciclo vital de minería de datos.

El ciclo vital del modelo contiene seis fases que se muestran en la Figura 1, las flechas indican las dependencias más importantes y frecuentes entre las fases. La secuencia de las fases no es estricta, la mayoría de los proyectos avanzan y retroceden entre fases de ser necesario [5].

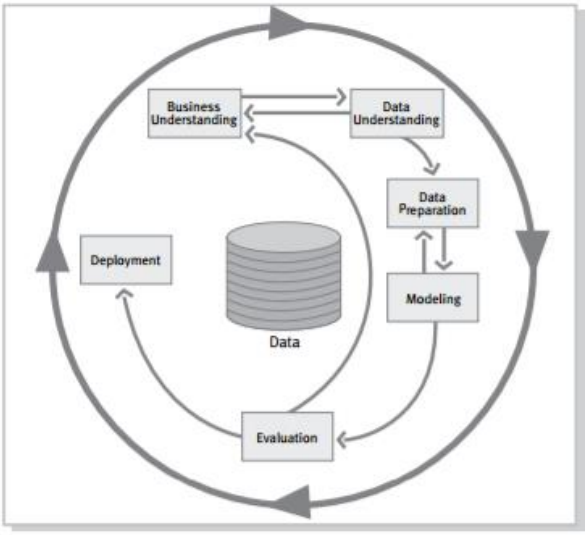


Figura 1: fases de Modelo CRISP-DM [19]

La aplicación de la metodología CRISP-DM además de evaluar el proyecto de minería de datos en cada fase permite evaluar la técnica de minería de datos a implementar.

Las fases de la metodología CRISP-DM son:

- **Comprensión del negocio**

  Determinar los objetivos del negocio. Esta es la primera tarea para desarrollar y tiene como metas, determinar cuál es el problema que se desea resolver, por qué existe la necesidad de utilizar minería de datos y definir los criterios éxito.

- **Comprensión de los datos**

  La fase de comprensión de los datos comprende la recolección inicial de datos, con el objetivo de establecer un primer contacto con el problema, familiarizándose con estos, identificar la calidad de los datos y establecer las relaciones más evidentes que permitan definir las primeras hipótesis.

- **Preparación de datos**

  En esta fase las tareas son: Selección de datos y Limpieza de los datos, se selecciona un subconjunto de los datos adquiridos en la fase anterior, apoyándose en criterios previamente establecidos en las fases anteriores, la calidad de los datos en cuanto a completitud y corrección de los datos y las limitaciones en el volumen o en los tipos de datos que están relacionadas con las técnicas de minería de datos seleccionadas.

  Eliminación de valores atípicos en los datos. Esta tarea complementa a la anterior, y es una de las que más tiempo y esfuerzo consume, debido a la diversidad de técnicas que pueden aplicarse para optimizar la calidad de los datos con el objetivo de prepararlos para la fase de modelación. Algunas de las técnicas a utilizar para este propósito son: normalización de los datos, discretización de campos numéricos, tratamiento de valores ausentes, reducción del volumen de datos, etc.

- **Modelado**

  La descripción de las principales tareas de esta fase son las siguientes:

  - Selección de la técnica de modelado, esta tarea consiste en la selección de la técnica de minería de datos más apropiada al tipo de problema a resolver.

  - Generación del plan de prueba. Una vez construido un modelo, se debe generar un procedimiento destinado a probar la calidad y validez de este.
  - Construcción del Modelo. Después de seleccionada la técnica, esta se ejecuta sobre los datos previamente preparados para generar uno o más modelos. Todas las técnicas de modelado tienen un conjunto de parámetros que determinan las características del modelo a generar. La selección de los

mejores parámetros es un proceso iterativo y se basa exclusivamente en los resultados generados.

Se seleccionaron varias técnicas para hacer una comparación entre las técnicas y elegir la técnica que obtenga el mejor rendimiento. Las técnicas seleccionadas para la realización del proyecto fueron:

- k- vecinos más cercanos (KNN).

El algoritmo clasifica cada dato nuevo en el grupo que corresponda, según tenga k vecinos más cerca de un grupo o de otro. Es decir, calcula la distancia (distancia Euclidiana) del elemento nuevo a cada uno de los existentes, y ordena dichas distancias de menor a mayor para ir seleccionando el grupo al que pertenece [6]. Este grupo será, por tanto, el de mayor frecuencia con menores distancias.
El K-NN es un algoritmo de aprendizaje supervisado, su objetivo será el de clasificar correctamente todas las instancias nuevas. El conjunto de datos típico de este tipo de algoritmos está formado por varios atributos descriptivos y un solo atributo objetivo o clase [6].

- Máquinas de vectores de soporte (SVM).

Los algoritmos SVM pertenecen a la familia de los clasificadores lineales. En estos clasificadores se tiene la característica de que, a priori, se conocen las clases a las que pertenecen los individuos, no se trata de una agrupación por similitudes, sino que se tiene las clases bien definidas [6]. Dado un conjunto de ejemplos de entrenamiento (muestras) podemos etiquetar las clases y entrenar una SVM para construir un modelo que prediga la clase de una nueva muestra. Intuitivamente, una SVM es un modelo que representa a los puntos de

muestra en el espacio, separando las clases por un espacio lo más amplio posible. Cuando las nuevas muestras se ponen en correspondencia con dicho modelo, en función de su proximidad pueden ser clasificadas a una u otra clase, dependiendo de la proximidad a cada una. Más formalmente, una SVM construye un hiperplano o conjunto de hiperplanos en un espacio de dimensionalidad muy alta (o incluso infinita) que puede ser utilizado en problemas de clasificación o regresión. Una buena separación entre las clases permitirá una clasificación más correcta

- Perceptrón multicapa (MLP)

Las redes neuronales son una técnica de aproximación paramétrica útil para construir modelos de densidad. Esta red presenta una capa de entrada con (n) neuronas y una capa de salida con (m) neuronas y al menos una capa de neuronas ocultas internas. Cada neurona (menos en la capa de entrada) recibe entrada de todas las neuronas de la capa previa y genera salida hacia todas las neuronas de la capa siguiente (salvo las de salida) [6]. No hay conexiones hacia atrás (feedback) ni laterales o autor recurrentes. El funcionamiento de la red consiste en un aprendizaje de un conjunto predefinido de pares de entradas-salidas dados como ejemplo, empleando un ciclo propagación y adaptación de dos fases: Primero se aplica un patrón de entrada como estímulo para la primera capa de las neuronas de la red, se va propagando a través de todas las capas superiores hasta generar una salida, se compara el resultado obtenido en las neuronas de salida con la salida que se desea obtener y se calcula un valor del

error para cada neurona de salida. Estos errores se transmiten hacia atrás, partiendo de la capa de salida, hacia todas las neuronas de la capa intermedia que contribuyan directamente a la salida, recibiendo el porcentaje de error aproximado a la participación de la neurona intermedia en la salida original. Este proceso se repite, capa por capa, hasta que todas las neuronas de la red hayan recibido un error que describa su aportación relativa al error total. Basándose en el valor del error recibido, se reajustan los pesos de conexión de cada neurona, de manera que en la siguiente vez que se presente el mismo patrón, la salida este más cercana a la deseada, es decir, que el error disminuya.

- Otras consideraciones a la etapa de modelado

En la mayoría de los casos los modelos requieren que se realicen ajustes al conjunto de datos, problemas como la alta dimensionalidad y/o el desbalance de las clases en el conjunto de datos, se deben abordar necesariamente para mejorar los resultados en los modelos de minería de datos, los métodos más usadas son: selección de atributos y balanceo de clases.

- Selección de atributos. Es un método de evaluación que determina la calidad del conjunto de atributos para discriminar las clases. Se pueden distinguir dos categorías de métodos de evaluación, en la primera se utiliza directamente un clasificador para medir la calidad del subconjunto de atributos a través de la tasa de error del clasificador. Estos métodos necesitan un proceso completo de entrenamiento y evaluación en cada caso

de búsqueda, lo que resulta de un elevado costo computacional. Otra alternativa es la utilización de métodos que no utilizan un clasificador especifico, en el cual se busca calcular la correlación de las clases con cada atributo, y eliminar los atributos que tienen una correlación muy alta con otros atributos, conocidos como atributos redundantes. En este método los subconjuntos preferidos son aquellos altamente correlacionados con el atributo que define las clases y con poca correlación entre ellos. Un método de búsqueda que determina la forma de realizar la búsqueda de conjuntos. La evaluación exhaustiva de todos los posibles subconjuntos se convierte en un problema combinatorio inabordable cuando el número de atributos es elevado. Por lo tanto, se necesita una estrategia de búsqueda que sea eficiente, una de las estrategias de búsqueda más efectiva, por su rapidez, es el ForwardSelection, que basa en elegir primero el mejor atributo, y realizar un proceso interactivo de ir añadiendo atributos que aporten más información hasta llegar a la situación en la que añadir un nuevo atributo empeora los resultados [7].

- Balanceo de clases. Se puede definir como conjunto de datos desbalanceados, a aquellos que presentan una desproporción notable en el número de instancias pertenecientes a cada clase. Esto provoca un sesgo en el desempeño de los clasificadores hacia el reconocimiento de las clases mayoritarias, en detrimento de las clases minoritarias [8]. Dos de los métodos más usados para resolver el problema de balanceo de clases es: sobre muestreo (oversampling) y submuestreo (undersampling). El sobre muestreo consiste en igualar las clases minoritarias a las clases mayoritarias,

duplicando las muestras hasta igualar las clases, esto podría representar un problema de sobre ajuste del modelo generado. El submuestreo es un método en el cual la clase mayoritaria se igualan a la clase minoritaria, la dificulta con este método es la perdida de excesiva de registros.

- **Evaluación de modelos**

  Es preciso revisar el proceso de modelado, teniendo en cuenta los resultados obtenidos y representados en la matriz de confusión. La matriz de confusión es una tabla que describe el rendimiento de un modelo supervisado de minería de datos, se llama “matriz de confusión” porque hace que sea fácil detectar dónde el sistema está confundiendo dos clases. A continuación, se presentan la matriz de confusión y los términos que la componen.

Actual	Predicción	
	TP	**FP**
	FN	**TN**

True Positives (TP): cuando la clase real de datos era 1 (Verdadero) y la predicha es también 1 (Verdadero).

True Negatives (TN): cuando la clase real de datos fue 0 (Falso) y el pronosticado también es 0 (Falso).

False Positives (FP): cuando la clase real de datos era 0 (False) y el pronosticado es 1 (True).

False Negatives (FN): Cuando la clase real de datos era 1 (Verdadero) y el valor predicho es 0 (Falso).

A partir de la matriz de confusión es posible obtener las siguientes métricas:

Exactitud (Accuracy). La exactitud es la medida de desempeño que es la proporción de la observación predicha correctamente al total de observaciones. Se puede pensar que, si el modelo tiene una alta exactitud, el modelo es el mejor. La exactitud es más significativa cuando el conjunto de datos es simétricos donde los valores de falso positivo y falso negativo son casi iguales.

$$\frac{TP + TN}{TP + TN + FP + FN}$$

Sensibilidad (Recall). Es la relación entre las observaciones positivas correctamente pronosticadas y todas las observaciones en la clase real.

$$\frac{TP}{TP + FN}$$

La precisión es la relación de las observaciones positivas predichas correctamente al total de observaciones positivas predichas.

$$\frac{TP}{TP + FP}$$

- **Despliegue**

  Una vez que el modelo ha sido construido y validado, se transforma el conocimiento obtenido en acciones dentro del proceso de negocio, ya sea que el analista recomiende acciones basadas en la observación

del modelo y sus resultados, ya sea aplicando el modelo a diferentes conjuntos de datos o como parte del proceso. Generalmente un proyecto de minería de datos no concluye en la implementación del modelo, se deben documentar y presentar los resultados de manera comprensible para el usuario, con el objetivo de lograr un incremento del conocimiento.

4.2 TRABAJOS RELACIONADOS

En el ambiente de la educación son muchas las tareas que se han resuelto mediante el uso de minería de datos, gran parte de ellas han sido orientadas a analizar el proceso de enseñanza aprendizaje. Por ejemplo: R. Baker [11] sugiere cuatro áreas de conocimiento para hacer minería de datos en educación las cuales son: mejorar el modelo educativo, mejorar el domino de la educación, mejorar el soporte pedagógico de estudio en el aprendizaje a través de software, mejorar la investigación científica en el aprender a aprender. Usando cinco enfoques o métodos de DM: predicción, agrupamiento, relación entre datos, destilación de datos para el juicio humano y el descubrimiento de modelos, F. Castro [12] sugiere dos enfoques: evaluar el rendimiento y el aprendizaje, para proveer aplicaciones encaminadas a la adaptación de estudiantes de acuerdo con su aprendizaje.

Muchos de los trabajos relacionados con la minería de datos en la educación han estado encaminados a prevenir el riesgo académico, entendiendo como riesgo académico la posibilidad de que un estudiante fracase académicamente o que deserte del sistema educativo. Razón por la cual los objetivos de dichos trabajos se han enfocado en la predicción del desempeño académico y la predicción de deserción escolar. A continuación, se mencionan algunos trabajos relacionados.

En un estudio realizado en Universidad de Minho en Portugal [9], sobre el uso de la minería de datos para predecir el rendimiento académico de los estudiantes de secundaria en las materias de matemáticas y lenguaje de colegio oficiales, donde se analizaron los datos socioculturales, socioeconómicos y académicos de los estudiantes de secundaria graduados en los dos años anteriores, empleando cuatro técnicas de minería de datos (Decision Trees, Random Forest, Neural Networks y Support Vector Machines) [9]. Se logró predecir el rendimiento académico de los estudiantes con rendimientos superiores al 86%, evidenciando como las características particulares de cada estudiante tiene una influencia sobre rendimiento académico.

En un estudio realizado en la universidad de Bosque en Bogotá Colombia [1], se analizaron tres algoritmos (J48, PAR, RIDOR) con el objetivo de predecir el rendimiento académico, se demostró que el algoritmo (J48) logró clasificar correctamente alrededor de 78% de las nuevas instancias. Las reglas descubiertas por el árbol de decisión J48 pueden ser resumidas como sigue:

- Los estudiantes cuyo estrato social es 2 o no definido son clasificados dentro de riesgo académico
- Los estudiantes cuyo estrato social es mayor que 3 son clasificados dependiendo de otras variables como estado civil.
- En relación con lo anterior, cuando los estudiantes son solteros, la clasificación depende de otras variables como educación de la madre y género del estudiante
- Los estudiantes que se encuentran en unión libre pero no casados son clasificados en riesgo académico por defecto.

Adicionalmente se establecieron otros factores que influyen sobre el rendimiento académicos como lo son: el puntaje individual del ICFES, educación de los padres, número de Hermanos y si el estudiante registra tener madre o no. Más interesante

fue encontrar que si un estudiante no es casado y su estrato social es 4 es inmediatamente es clasificado en riesgo académico. También para los estratos sociales 5 y 6 (más altos) la mayoría de reglas dieron por defecto una clasificación en desempeño sobresaliente.

En otro estudio realizado en Thaksin University, Phattalung, Thailand [10], el cual buscaba predecir el rendimiento académico en estudiantes universitarios de ingeniería sistemas de los semestres 1 y 2, con algoritmos (ADTree, J48, ID3), con 280 datos de validación de estudiantes anónimos, se obtuvo una precisión de los algoritmos (J48 con 90.625%) de precisión y (ID3 y ADTree con 96.875%) de precisión en la clasificación.

El sistema educativo colombiano reconoce la teoría de las inteligencias múltiples de Howard Gardner, al igual que la mayoría de modelos pedagógicos educativos a nivel mundial, dicha teoría reconoce que los estudiantes poseen habilidades e inteligencias particulares, las cuales pueden ser: Inteligencia lógico-matemática, Inteligencia lingüística-verbal, Inteligencia visual-espacial, Inteligencia corporal-cinestésica , Inteligencia musical, Inteligencia interpersonal, Inteligencia intrapersonal e Inteligencia naturalista, por estas particularidades inherentes al individuo, se tomó la decisión de realizar los experimento con las notas de matemáticas, porque si a un estudiantes la predicción eventualmente lo clasifica en rendimiento bajo o básico, no significa una incapacidad del estudiante para aprender la matemática, lo que más bien se sugiere es que el estudiante posee otras capacidades intelectuales, y le permite a la institución educativa buscar otras estrategias pedagógicas para fortalecer el proceso de aprendizaje de la matemáticas.

5. DESCRIPCIÓN DE LA SOLUCIÓN

En el proceso de minera de datos se aplicó el modelo de referencia CRISP-DM (Cross-Industry Standard Process for Data Mining) [11], el cual consta de seis fases. Los resultados de cada fase se describen a continuación:

5.1 Comprensión del negocio

Crear un modelo de minería de datos, para predecir el rendimiento académico de los estudiantes de Institución Educativa Libardo Madrid Valderrama, que permita a docentes y directivos docentes de la Institución establecer estrategias para mejorar la condición de los estudiantes a los cuales la predicción del modelo los clasifique con rendimiento bajo y básico, de esta forma contribuir en la prevención de la deserción académica. En proceso de creación de modelo entrenado se implementó con el uso de técnicas de minerías de datos, en el proceso se evaluaron tres modelos.

5.2 Comprensión de los datos

Los datos para realización del trabajo fueron recopilados en tres etapas:

a) Se realizaron tres reuniones informativas con el propósito de presentar el proyecto y los objetivos de este a la comunidad educativa. La primera reunión se realizó con la participación de directivos y docentes para explicar los alcances del proyecto y conocer las políticas que tiene la institución sobre el tratamiento y uso de datos ver anexo 1 (política de protección de datos, Libardo Madrid Valderrama).

   La segunda reunión se realizó con los padres de familia o acudientes de los estudiantes, a los cuales se entregó un consentimiento informado el

cual podrían diligenciar y firmar en caso de probar la participación del estudiante en el proyecto, ver anexo 2 (consentimiento informado). La tercera reunión se realizó con los estudiantes para explicarles los alcances del proyecto y la importancia de contestar las preguntas de la forma más honesta posible.

b) La información socioeconómica y sociocultural de los estudiantes fue recogida por medio de la aplicación de una encuesta web, implementada en un formulario de Google.com ver anexo 3 (formulario de Google.com), cada estudiante debía responder el formulario de 17 preguntas, las repuestas de los estudiantes fueron exportadas a un archivo de Excel. En total se recopilaron 290 registros de los estudiantes de los grados 9°,10° y 11° del año lectivo 2017, la descripción de los atributos se detalla en la Tabla 1.

c) Las notas académicas de los estudiantes fueron suministradas por el rector del colegio, estas notas incluyen los resultados por período académico de la materia de matemáticas de los grados 8°,9° y 10° del año lectivo 2016 y la nota final de matemáticas de los mismos estudiantes en año lectivo 2017.

d) La información obtenida de la encuesta y las notas académicas de matemáticas de los cuatro periodos de 2016 y la final del año lectivo 2017, fueron consolidadas en una base de datos MySQL, para posteriormente cruzar la información con los datos de la encuesta. En este proceso se presentaron algunos inconvenientes debido a que en el sistema de servicios educativos “ZETI”, no se registran los estudiantes con el documento de identificación nacional “tarjeta de identidad”, se hacen por nombres y apellidos. Adicionalmente los reportes de las notas fueron entregados en formato “PDF”, este formato fue convertido a excel, se agregó una columna con el número de identificación de cada estudiante,

una vez realizado este proceso se exportó la información a la base de datos. Luego de cruzar la información académica y de la encuesta en la base de datos, se generó un archivo consolidado de con el nombre “F4N17.csv”, el archivo contiene 290 registros de los cuales 68 registros no poseen las notas de matemáticas de los cuatro periodos académicos del año lectivo anterior 2016, debido a que eran estudiantes nuevos en la institución. Estos registros se descartaron, el archivo consolidado quedo con 222 registros con 22 atributos, cada uno, distribuidos en 17 categóricos y 5 numéricos.

Los atributos se describen en la Tabla 1, los atributos 1 al 17, fueron recopilados por medio de la encuesta, del 18 al 22 fueron obtenidos del sistema educativo “ZETI”, el atributo 22 es la nota final de matemáticas para el siguiente año lectivo 2017, la cual es clase objetivo. Dicha nota se convirtió en un atributo categórico, las categorías se definieron con los siguiente rangos: Bajo” B”(>=0 & <=2,94) ; Básico “BS” (>2,94 & <=3,44) y Alto “A” (>3,44 &<=5)

.

Item	Pregunta	Variable	Valor
1	Tipo de vivienda en la que vive el estudiante	TVE	Arrendada; Familiar; Propia; Otro
2	El estudiante convive con la madre	COVM	Si ; No
3	El estudiante convive con el padre	COVP	Si ; No
4	Educación de la madre del estudiante	EDUM	primaria; secundaria; profesional; posgrado; no sabe
5	Educación del padre del estudiante	EDUP	primaria; secundaria; profesional; posgrado; no sabe
6	La madre del estudiante trabaja	MTRAB	Si ; No
7	El padre del estudiante trabaja	PTRAB	Si ; No
8	Personas que conviven con el estudiante	#PCONV	numérico 0 a 99
9	Posición entre los hermanos que ocupa	PHERMANO	numérico 0 a 99
10	Población Tipo	TPOBLACION	Afrodescendiente; Desplazado; Desplazado por la violencia; indígena; Jóvenes vulnerables; Ninguna; Raizal
11	Tiempo que demora en llegar al Colegio	TDC	Menos de 15 Minutos; Entre 15 a 30 Minutos; Entre 30 minutos a 1 hora.
12	Horas de estudio extra clases semanal	#HEEC	Entre 1 a 2 horas; Entre 2 a 3 horas; Entre 3 a 5 horas; Entre 5 a 10 horas; Mas de 10 horas
13	El estudiante Trabaja	ET	Si ; No
14	Tiene computador en su casa	LAPTOP	Si ; No
15	Tiene internet en su casa	INTERNET	Si ; No
16	Practica algún deporte	DEPOR	Si ; No
17	Tiene una relación sentimental	RSENTI	Si ; No
18	Nota de primer periodo 2016	P1-16	numérico 0 a 5
19	Nota de segundo periodo 2016	P2-16	numérico 0 a 5
20	Nota de tercer periodo 2016	P3-16	numérico 0 a 5
21	Nota de cuarto periodo 2016	P4-16	numérico 0 a 5
23	Nota promedio Final 2017	NF	B; BS; A

Tabla 1 Descripción de Atributos del archivo F4N17.csv

5.3 Preparación de datos

Con la información consolidada en el archivo “F4N17.csv”, se procedió a realizar las tareas de preprocesamiento implementadas en Python usando la librería para el análisis de minería de datos Scikit-learn. Las tareas se realizaron en cuatro etapas que se describen a continuación:

- Carga y codificación de atributos: Los datos fueron cargados en Python como se muestra en la Figura 2. En este proceso se eliminaron 68 registros de estudiantes que no poseían notas del periodo académico anterior, año lectivo 2016, en total se cargaron 222 registros.

COVP	EDUM	EDUP	MTRAB	PTRAB	#PCONV	PHERMANO	TPOBLACION	...	ET	LAPTOP	INTERNET	DEPOR	RSENTI	P1-16	P2-16	P3-16	P4-16	NF
Si	Primaria	Primaria	No	Si	9	0	Ninguna	...	No	Si	No	No	No	2.9	3.4	3.4	3.6	A
Si	No sabe	No sabe	Si	Si	3	3	Ninguna	...	No	Si	Si	Si	No	2.6	2.4	2.8	3.9	A
Si	Secundaria	Primaria	No	Si	5	3	Ninguna	...	No	No	Si	No	Si	3.1	3.4	3.5	3.9	A
Si	Secundaria	Secundaria	No	Si	4	1	Ninguna	...	No	No	Si	No	Si	3.0	3.1	3.5	3.7	A
No	No sabe	No sabe	Si	Si	2	4	Ninguna	...	No	Si	No	No	Si	2.7	2.9	2.1	3.1	BS

Figura 2, Carga y codificación de atributos en Python.

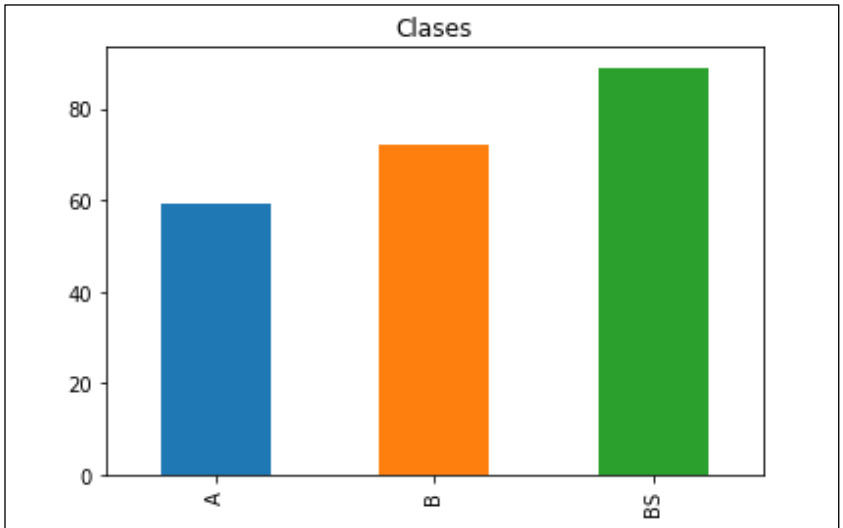
Se convirtieron los valores categóricos a numéricos. El atributo nota final promedio 2017 “NF”, se renombro atributo “C”, como se muestra en la Figura 3.

COVM	COVP	EDUM	EDUP	MTRAB	PTRAB	#PCONV	PHERMANO	TPOBLACION	...	ET	LAPTOP	INTERNET	DEPOR	RSENTI	P1-16	P2-16	P3-16	P4-16	C
0	0	0	0	0	0	9	0	0	...	0	0	0	0	0	2.9	3.4	3.4	3.6	A
1	0	1	1	1	0	3	3	0	...	0	0	1	1	0	2.6	2.4	2.8	3.9	A
1	0	2	0	0	0	5	3	0	...	0	1	1	0	1	3.1	3.4	3.5	3.9	A
1	0	2	2	0	0	4	1	0	...	0	1	1	0	1	3.0	3.1	3.5	3.7	A
1	1	1	1	1	0	2	4	0	...	0	0	0	0	1	2.7	2.9	2.1	3.1	BS

Figura 3, Codificación de atributos en Python.

La distribución de las clases del atributo “C”, y total de registros por clase fue: Alto,(A) = 59, Bajo,(B) = 72 y Básico,(BS) = 89, Figura 4

.



.

Figura 4 Distribución por clases del atributo “C”

- Representación gráfica de atributos por medio de histogramas, Figura 5,6 y 7. En este proceso no identificaron valores atípicos.

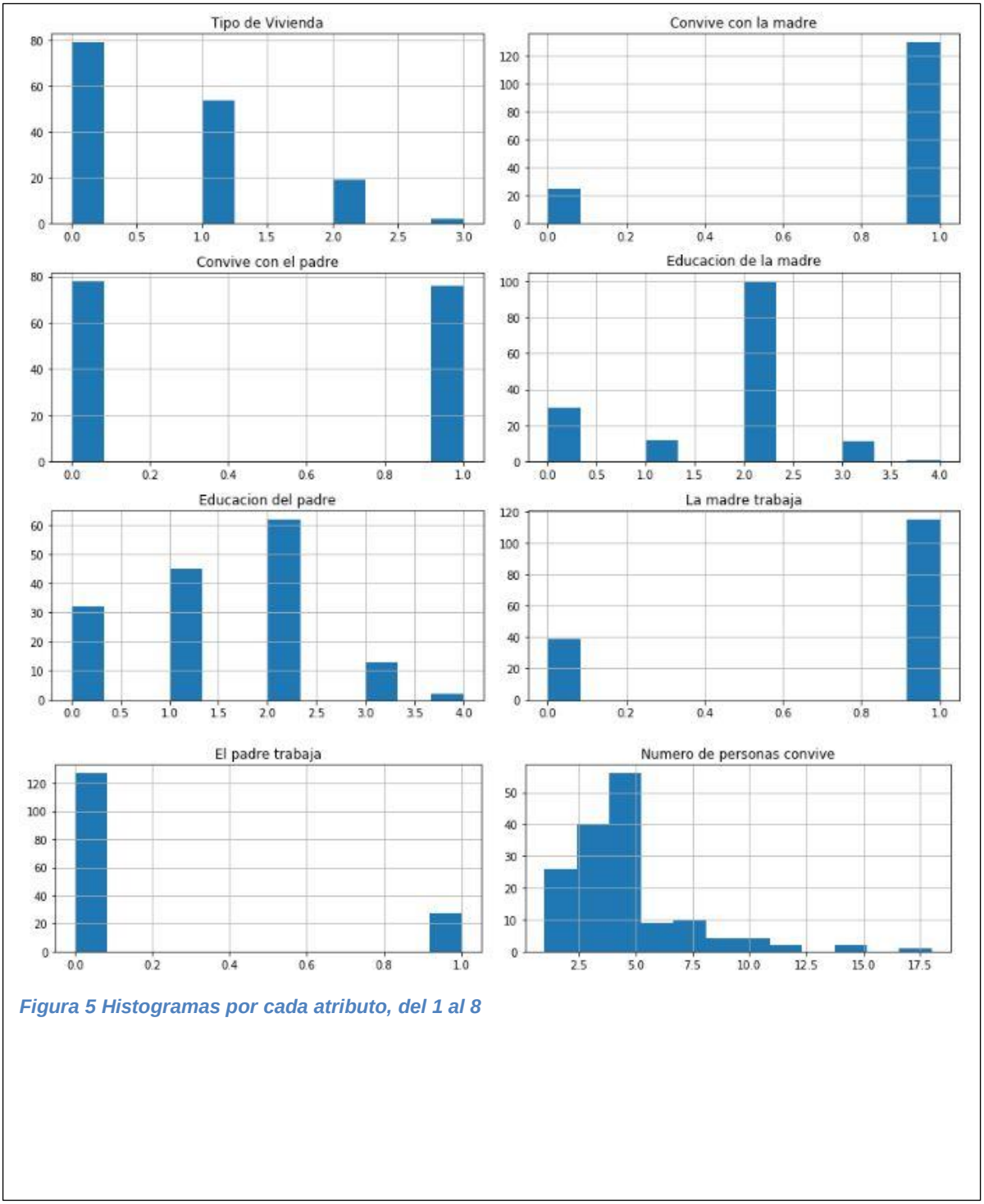


Figura 5 Histogramas por cada atributo, del 1 al 8

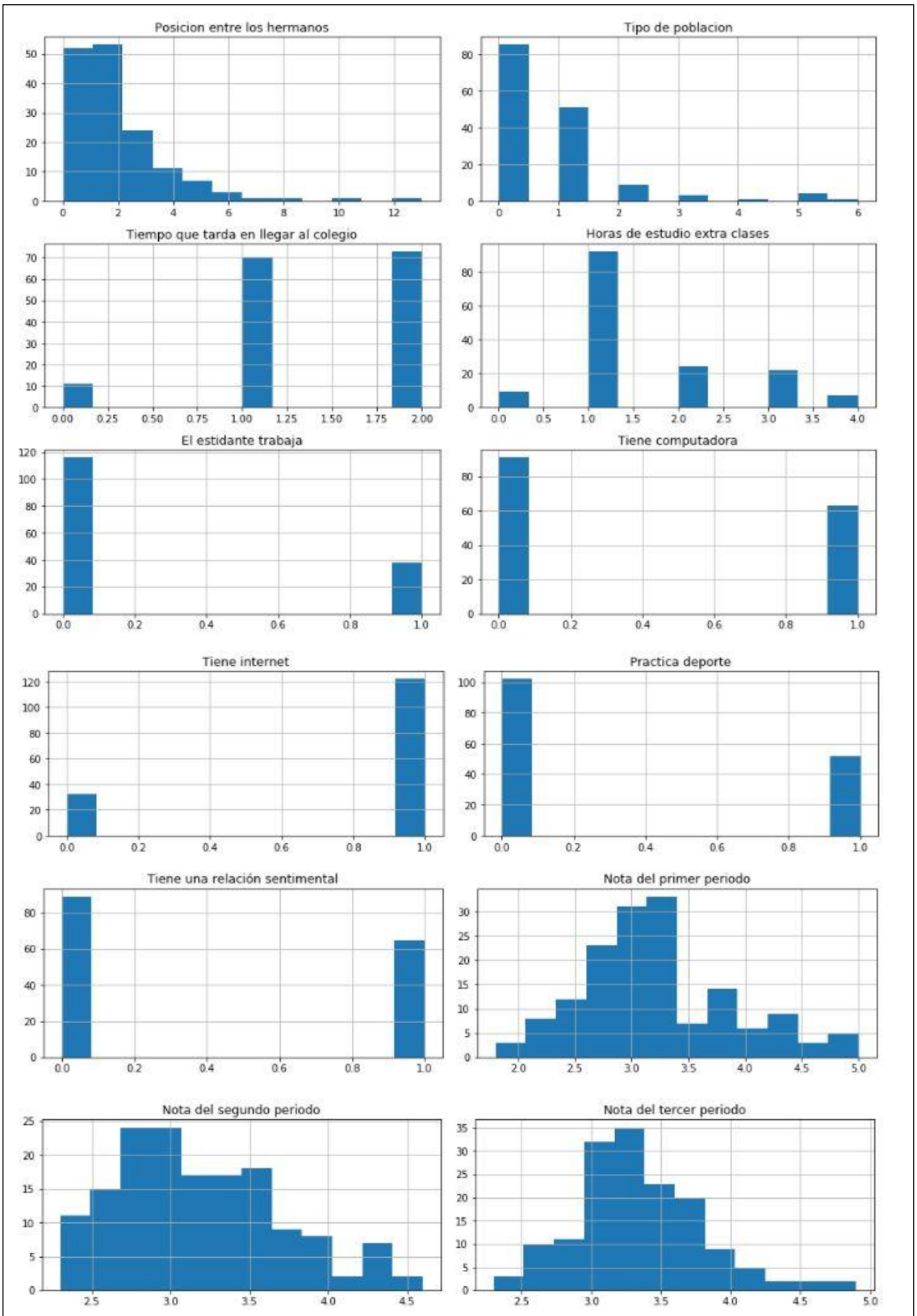


Figura 6 Histogramas por cada atributo, del 9 al 20

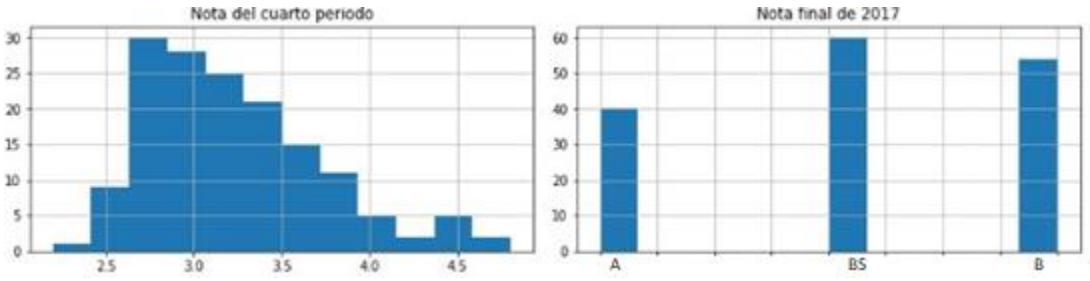


Figura 7 histogramas por cada atributo, del 21 al 22

- Se dividió el conjunto de datos en dos subconjuntos usando la técnica de muestreo por clases. Uno para entrenamiento con 70% de los datos y otro para test con el 30% de los datos, en la Figura 8, se muestran los resultados de este proceso.

```
Conteo de clases conjunto train Counter({'BS': 60, 'B': 54, 'A': 40})
Conteo de Clases conjunto test  Counter({'BS': 29, 'A': 19, 'B': 19})
```

Figura 8 Distribución por clases en los conjuntos de entrenamiento y test

5.4 Modelado

De acuerdo con las características del trabajo, se planteó una tarea de clasificación supervisada multiclases. Se evaluaron tres técnicas de clasificación: Máquinas de vectores de soporte (SVM), K-vecinos más cercanos (KNN), Redes neuronales artificiales perceptrón multicapa (MLP). Estas técnicas fueron seleccionadas por la capacidad que poseen para resolver problemas que no son linealmente separable, además de lo reducido del número de ejemplos con que se cuentan para entrenar los modelos. Cada técnica fue probada con diferentes configuraciones con el objetivo de establecer con cuál configuración se obtendría el mejor rendimiento del modelo. En el ajuste de parámetros se implementó la validación cruzada, con el

objetivo estimar la precisión de los modelos, esta técnica consiste en dividir el conjunto de entrenamiento en K iteraciones o (K-fold), los datos se dividen en K subconjuntos. Uno de los subconjuntos se utiliza como datos de prueba y el resto como datos de entrenamiento. El proceso de validación cruzada es repetido durante k iteraciones, con cada uno de los posibles subconjuntos de datos de entrenamiento. Finalmente se realiza la media aritmética de los resultados de cada iteración para obtener la exactitud del entrenamiento. La elección del número de iteraciones depende del tamaño de conjunto de datos. En el proceso se utilizó la validación cruzada de 10 iteraciones [16].

- Máquinas de vectores de soporte (SVM).

  Cuando se clasifican datos con una SVM es necesario fijar un margen de separación entre observaciones. El coste o parámetro “C” se debe ajustar en una SVM al igual que el tipo de kernel a emplear. “C” es un parámetro de regularización que controla la compensación entre maximizar el margen y minimizar el término de error de entrenamiento. Si “C” es demasiado pequeño, se colocará un esfuerzo insuficiente para ajustar los datos de entrenamiento. Si “C” es demasiado grande, entonces el algoritmo se ajustará a los datos de entrenamiento lo que es conocido como overfitting [12].
  El parámetro gamma puede verse como el inverso del radio de influencia de las muestras seleccionadas por el modelo como vectores de soporte. Si el gamma es demasiado grande, el radio del área de influencia de los vectores de soporte solo incluye el propio vector de soporte y ninguna regularización con “C” podrá prevenir overfitting [12].

  Se realizaron pruebas con diferentes variaciones de los parámetros “C” y gamma los resultados se muestran en la Figura 9. El parámetro gamma

fue variado entre 1,10,100, y 1000 obteniendo mejor resultado con gamma = 1.

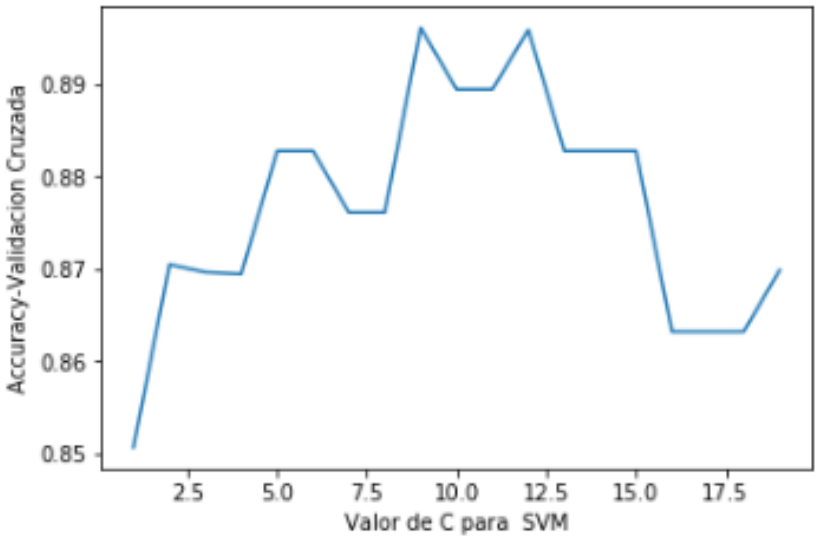


Figura 9 Estimación parámetro C SVM, kernel lineal, gamma = 1

Como se puede observar en la Figura 9, el mejor desempeño de la SVM se obtuvo con el kernel lineal, y los parámetros C=8 y gamma =1. Estos dos parámetros serán implementados en los experimentos de la sección 6.

- k- vecinos más cercanos (KNN).

Es una técnica de clasificación supervisada, en la cual la optimización del modelo requiere la selección de mejor “k”, el parámetro “k” encierra los vecinos más cercanos. La elección del parámetro “k” depende fundamentalmente de los datos; generalmente los valores grandes de “k” reducen el efecto de ruido en la clasificación, pero pueden crear límites falsos entre clases parecidas. Los resultados del proceso se presentan en la Figura 10.

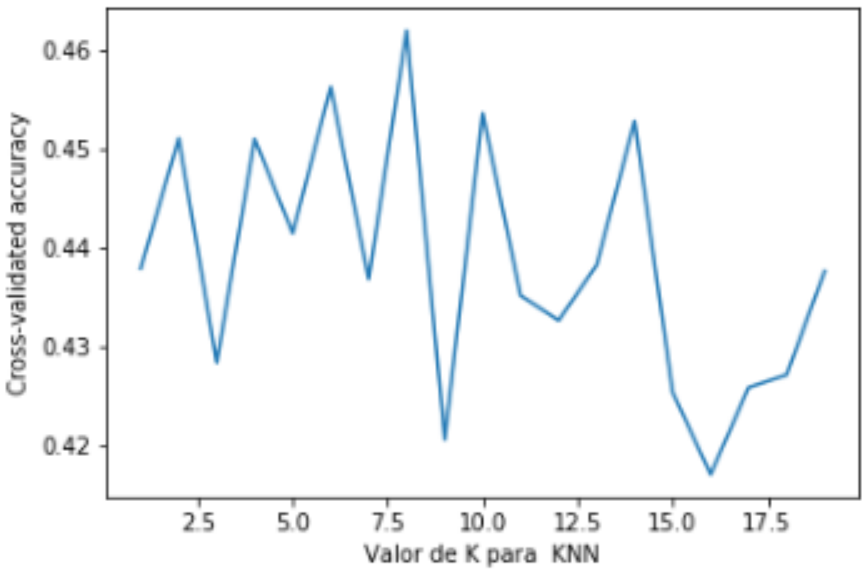


Figura 10 KNN estimación parámetro K

Como se muestra en la figura 10, el mejor resultado del modelo se obtiene con k =8, por tal motivo esta configuración se utilizará en los experimentos de sección 6.

- Perceptrón multicapa (MLP)

El Perceptrón multicapa (MLP), es un algoritmo de aprendizaje supervisado que aprende una función al entrenar un conjunto de datos. El diseño implica: la determinación de la función de activación a emplear, el número de neuronas y el número de capas ocultas de la red. El número de capas ocultas y el número de neuronas en estas capas deben ser elegidos por el diseñador. No existe un método o regla que determine el número óptimo de neuronas ocultas para resolver un problema dado [13]. En la mayor parte de las aplicaciones, estos parámetros se determinan por ensayo y error. Partiendo de una arquitectura ya entrenada, se

realizan cambios aumentando o disminuyendo el número de neuronas ocultas y el número de capas hasta conseguir una arquitectura adecuada. Se realizaron diferentes diseños para establecer cuál obtenía el mejor comportamiento, los resultados se muestran en la Figura 11.

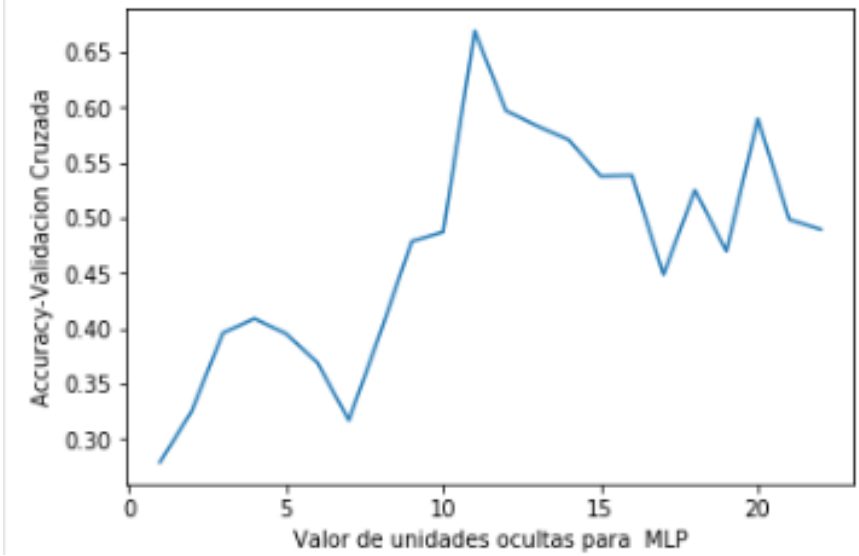


Figura 11 MLP estimación de parámetros

En las pruebas realizadas no se incluyen las variaciones en las unidades ocultas, solo las variaciones realizadas en las capas ocultas. El mejor rendimiento de la (MLP) ocurre con 10 capas ocultas y 13 unidades ocultas, por esta razón se utilizará este diseño en los experimentos sección 6.

5.5 Evaluación

Para la evaluación de la capacidad de generalización de los diferentes modelos generados, se incluyeron métricas como Exactitud (Accuracy), precisión y sensibilidad (Recall), destacando la importancia de la sensibilidad, dado que el

interés principal trabajo es conocer cuales estudiantes se clasifican correctamente en Bajo, (B) y cuales, en Básico, (BS). Las métricas serán evaluadas en cada uno de los experimentos a realizar. Al finalizar la etapa de experimentos se mostrará una tabla comparativa de los resultados obtenidos destacando el o los modelos con mejor desempeño.

5.6 Despliegue

Se creó un prototipo WEB, que incluye un modelo entrenado de minería de datos para predecir el desempeño académico en matemáticas de los estudiantes para el siguiente año lectivo. El prototipo permite cargar las cuatro notas de periodos académicos del año lectivo anteriormente cursado, más los diez y siete datos socioeconómicos y socioculturales del estudiante. Como resultado el prototipo web retorna la predicción de rendimiento académico por estudiante, el cual puede ser bajo, básico o alto, para el siguiente año lectivo en materia de matemáticas.

Para facilitar la documentación del proyecto y como estrategia de trabajo, los experimentos fueron realizados en jupyter-notebook, bajo el lenguaje de programación Python, y la librería Scikit-learn. Scikit-learn es una biblioteca de aprendizaje automático de uso gratuito implementado en Python.

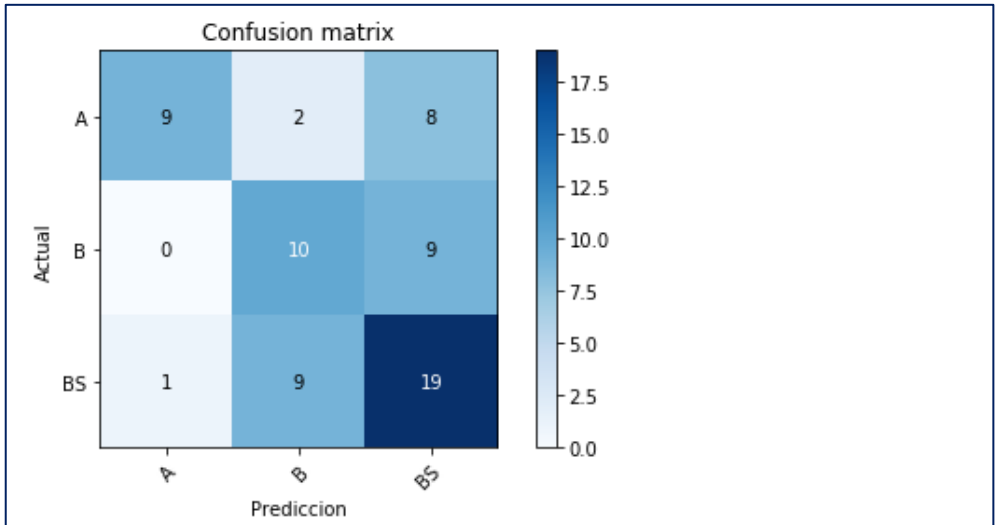
6 PRUEBAS

Los experimentos buscan predecir la clasificación de los estudiantes para el siguiente año lectivo 2017, con base a los 21 atributos, distribuidos de la siguiente forma del 1 a los 17 datos socioculturales y socioeconómicos tomados de la encuesta, del 18 al 21 las notas académicas de matemáticas del anterior año lectivo

2016. El experimento (1) toma todo el conjunto de datos, en el experimento (2) se aplica selección de atributos al conjunto de datos, en el experimento (3) se hace el balanceo de carga con la técnica sobre muestreo (oversampling) al conjunto de datos, en el experimento (4) se hace el balanceo de carga con submuestreo (undersampling) al conjunto de datos. Todos los experimentos se hacen con las técnicas SVM, KNN, y MLP, se busca comparar los resultados, para escoger la mejor técnica.

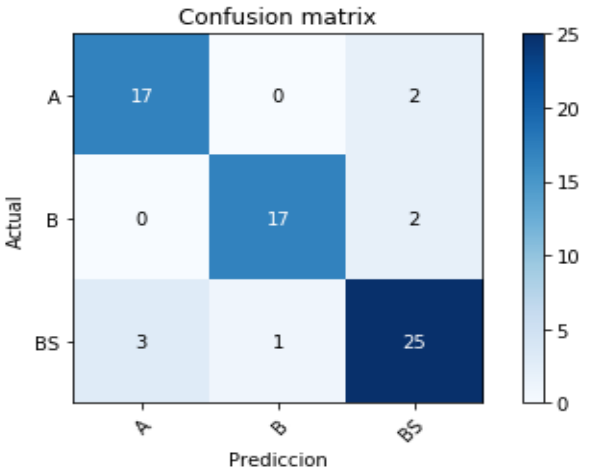
6.1 Experimento 1

El entrenamiento de los modelos se realizó con un conjunto de 21 atributos y 154 registros, los modelos se validaron con el conjunto de prueba de 21 atributos y 67 registros, los resultados se presentan en las Figuras 12 y 13,

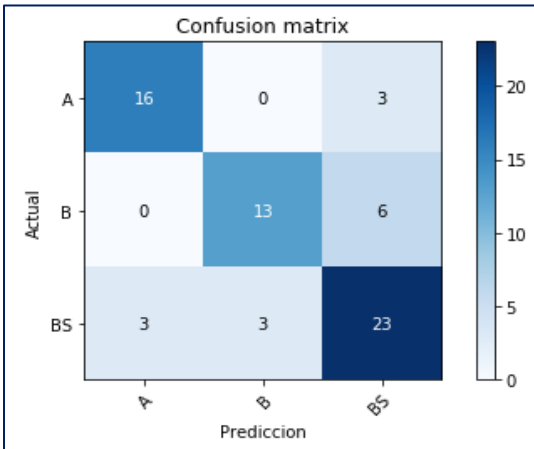



```
############## Accuracy KNN  K=8 ####################################
0.5671641791044776

############## Metricas KNN  K=8 ####################################
              precision    recall  f1-score   support

           A       0.90      0.47      0.62        19
           B       0.48      0.53      0.50        19
          BS       0.53      0.66      0.58        29
```

Figura 12 Resultados del experimento 1 modelo KNN, Clases Alto (A), Bajo (B), Básico (BS)




```
############### Accuracy SVM Lineal C=8, gamma=1 ################
0.8805970149253731

############### Metricas SVM Lineal C=8, gamma=1 ################
              precision    recall  f1-score   support

           A       0.85      0.89      0.87        19
           B       0.94      0.89      0.92        19
          BS       0.86      0.86      0.86        29
```




```
### Accuracy MLP capas ocultas 10 ,unidades ocultas 13 ###################
0.7761194029850746

### Metricas MLP capas ocultas 10 ,unidades ocultas 13 ###################
              precision    recall  f1-score   support

           A       0.84      0.84      0.84        19
           B       0.81      0.68      0.74        19
          BS       0.72      0.79      0.75        29
```

Figura 13 Resultado del experimento 1 modelos SVM y MLP, Clases Alto (A), Bajo (B), Básico (BS)

Cómo se puede observar en las Figura 12 y 13, el modelo que mejor rendimiento obtuvo en el experimento 1, fue el modelo de las SVM, además de obtener una mayor exactitud (accuracy) que los otros dos modelos KNN y MLP, la exactitud no es la métrica más significativa para el objetivo del trabajo, debido a que esta representa la proporción entre el número de predicciones correctas (tanto positivas como negativas) sobre el total de predicciones , por esta razón se dio mayor peso a la métrica sensibilidad (recall) de las clases básico y bajo, porque esta métrica evalúa la tasa de verdaderos positivos clasificados por el modelo. Los experimentos que se diseñaron para el trabajo están orientados a obtener el modelo de clasificación que mejor sensibilidad obtenga, especialmente en las clasificaciones estudiantes en las clases de Bajo, (B) y Básico, (BS). Estas dos clasificaciones resultan de mayor importancia para el objetivo del trabajo, puesto que a partir de esta clasificación se pueden desplegar las estrategias pedagógicas pertinentes de acuerdo a la clasificación obtenida por el estudiante.

6.2 Experimento 2

El objetivo del experimento es mejorar las predicciones del experimento 1, aplicando una técnica de selección de atributos.

El entrenamiento de los modelos se realizó con un conjunto de 18 atributos y 154 registros, los modelos se validaron con el conjunto de prueba de 18 atributos y 67 registros. Existen dos técnicas para la evaluación de atributos, en la primera se utiliza directamente un clasificador específico para medir la calidad del subconjunto de atributos a través de la tasa de error del clasificador. Estas técnicas necesitan un proceso completo de entrenamiento y evaluación en cada caso de búsqueda, lo que resulta con elevado coste computacional.

La segunda técnica no utiliza un clasificador específico, se basa en calcular la correlación de la clase con cada atributo, y eliminar atributos que tienen una correlación redundante entre atributos. En esta técnica los subconjuntos preferidos

son aquellos altamente correlacionados con el atributo que define las clases y con poca correlación entre ellos, en la Figura 14 se presenta un matriz de correlación.

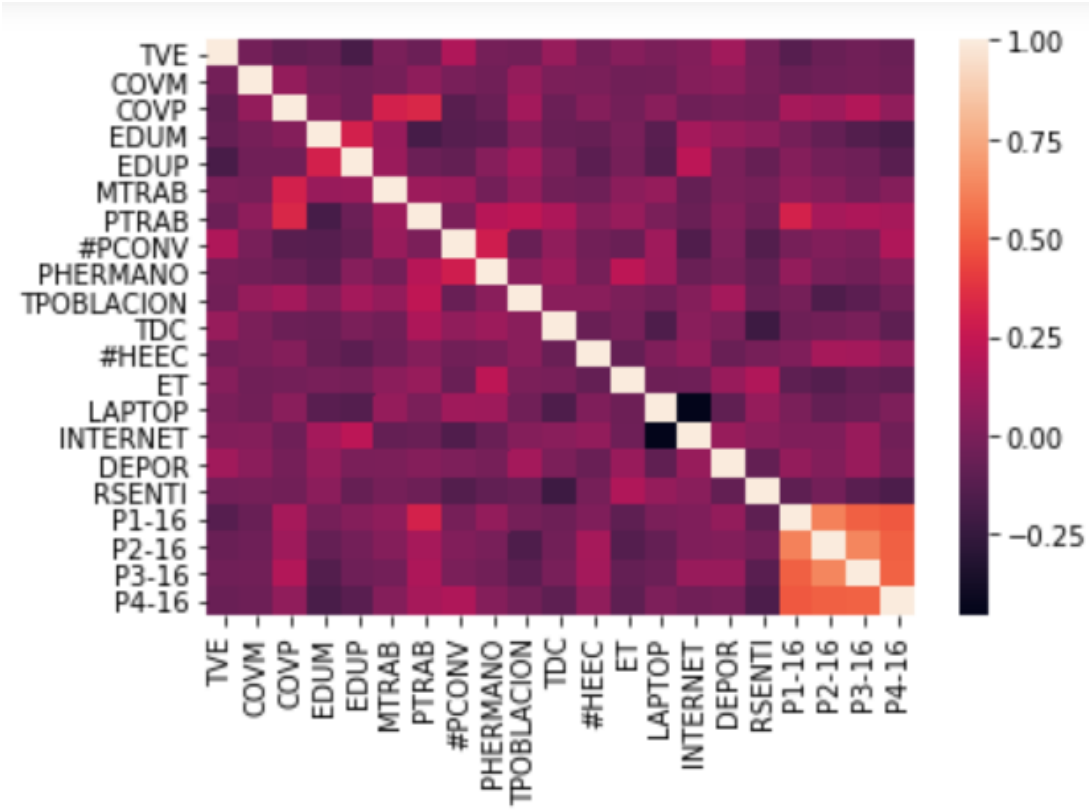
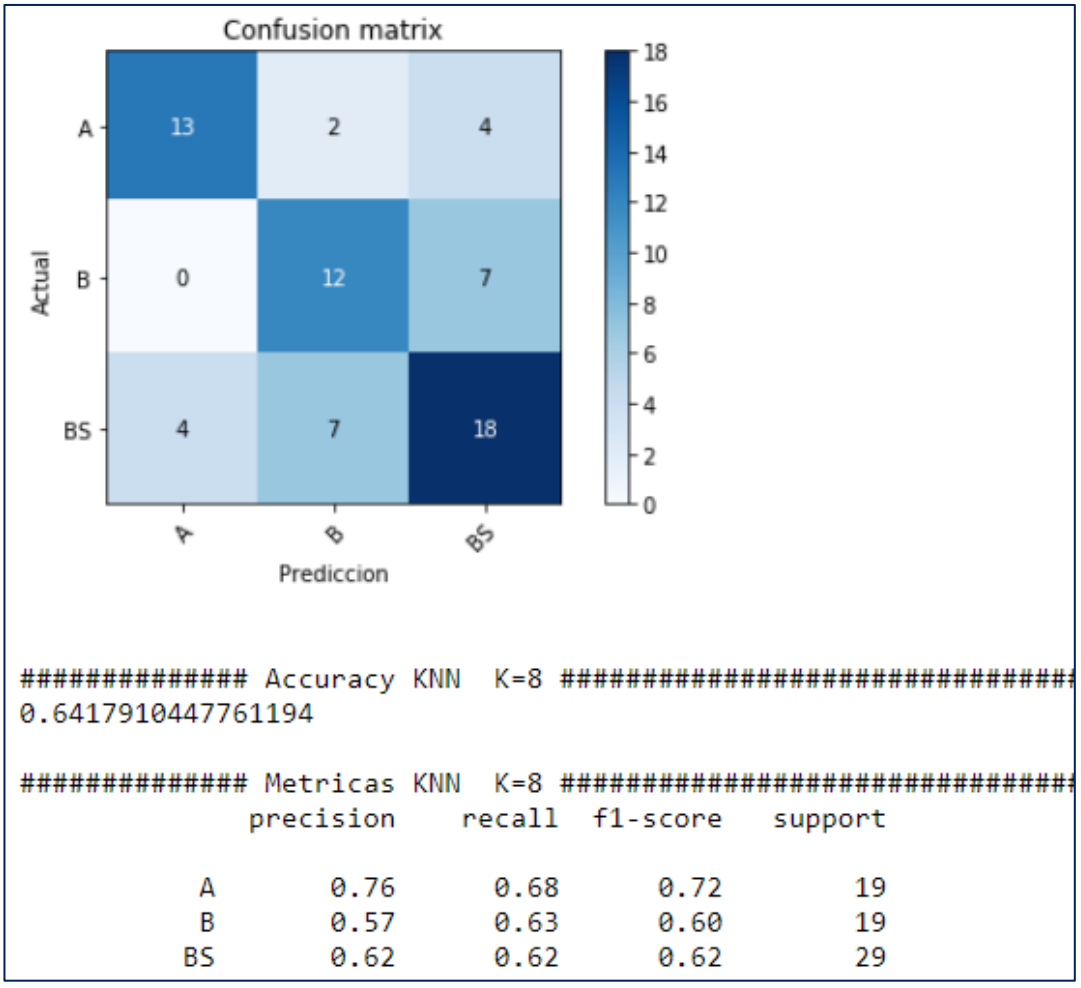


Figura 14 matriz de correlación

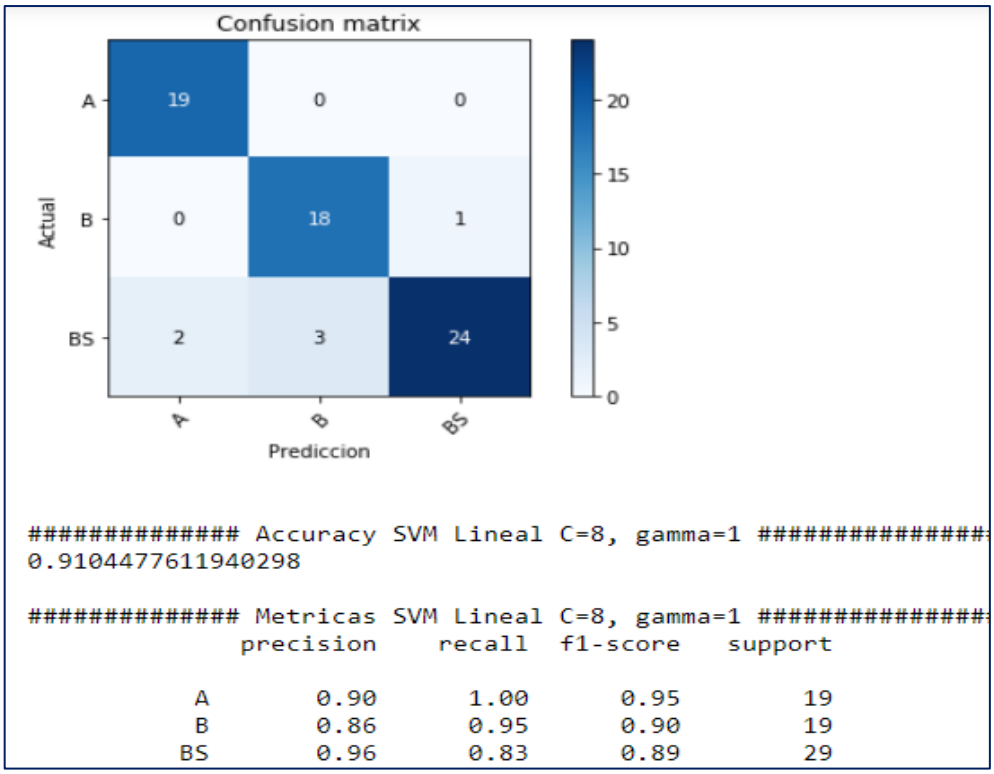
Como se puede observar en la matriz de correlación, hay varios atributos que presentan una alta correlación entre sí, como los atributos de las notas, sin embargo, por ser el principal indicador del desempeño del estudiante, las notas aunque presentan una alta correlación no son atributos candidatos a ser eliminadas, otros atributos que presentan una alta correlación con otros atributos y que son redundantes son: si el estudiantes convive con padre, (COVP) con el atributo: el padre del estudiante trabaja (PTRAB), la educación de la madre (EDUM) con el atributo: la educación del padre (EDUP), los estudiantes posee laptop (LAPTOP)

con el atributo: los que tienen internet (INTERNET) en la casa. Por lo cual se eliminaron del conjunto de datos los atributos (COVP, EDUM y LAPTOP), los resultados obtenidos después de la eliminación de estos atributos se presentan en la Figura 15 y 16.

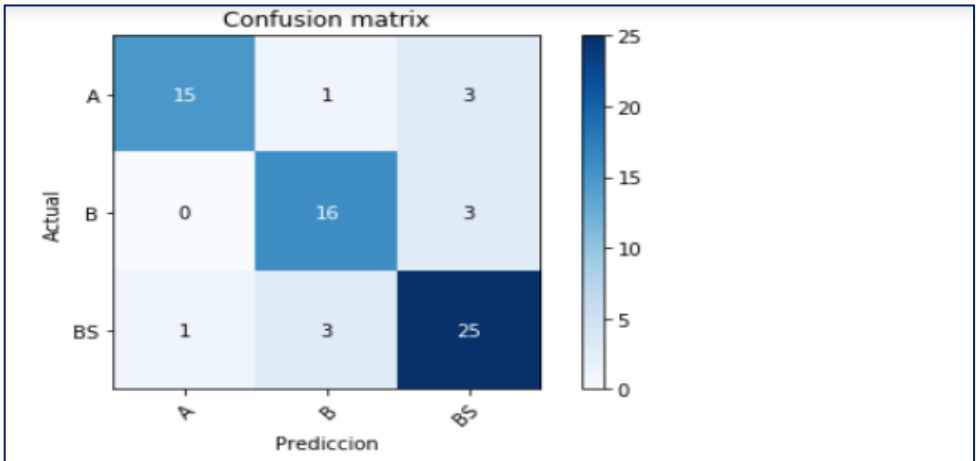



```
############### Accuracy KNN  K=8 ###################################
0.6417910447761194

############### Metricas KNN  K=8 ###################################
              precision    recall  f1-score   support

           A       0.76      0.68      0.72        19
           B       0.57      0.63      0.60        19
          BS       0.62      0.62      0.62        29
```

Figura 15 Resultados del experimento 2 modelo KNN, Clases Alto (A), Bajo (B), Básico (BS)




```
############### Accuracy SVM Lineal C=8, gamma=1 ################
0.9104477611940298

############### Metricas SVM Lineal C=8, gamma=1 ################
              precision    recall  f1-score   support

           A       0.90      1.00      0.95        19
           B       0.86      0.95      0.90        19
          BS       0.96      0.83      0.89        29
```




```
### Accuracy MLP capas ocultas 10 ,unidades ocultas 13 ######
0.835820895522388

### Metricas MLP capas ocultas 10 ,unidades ocultas 13 ######
              precision    recall  f1-score   support

           A       0.94      0.79      0.86        19
           B       0.80      0.84      0.82        19
          BS       0.81      0.86      0.83        29
```

Figura 16 Resultado del experimento 2 modelos SVM y MLP, Clases Alto (A), Bajo (B), Básico (BS)

Como se puede ver en la Figuras 15 y 16, los resultados de los de los tres modelos mejoran con respecto a los resultados obtenidos en el experimento 1, sobre sale el rendimiento obtenido por el modelo de la SVM, el cual consigue una mayor exactitud (accuracy) y mejora en la sensibilidad (recall), mejorando en todas las clases.

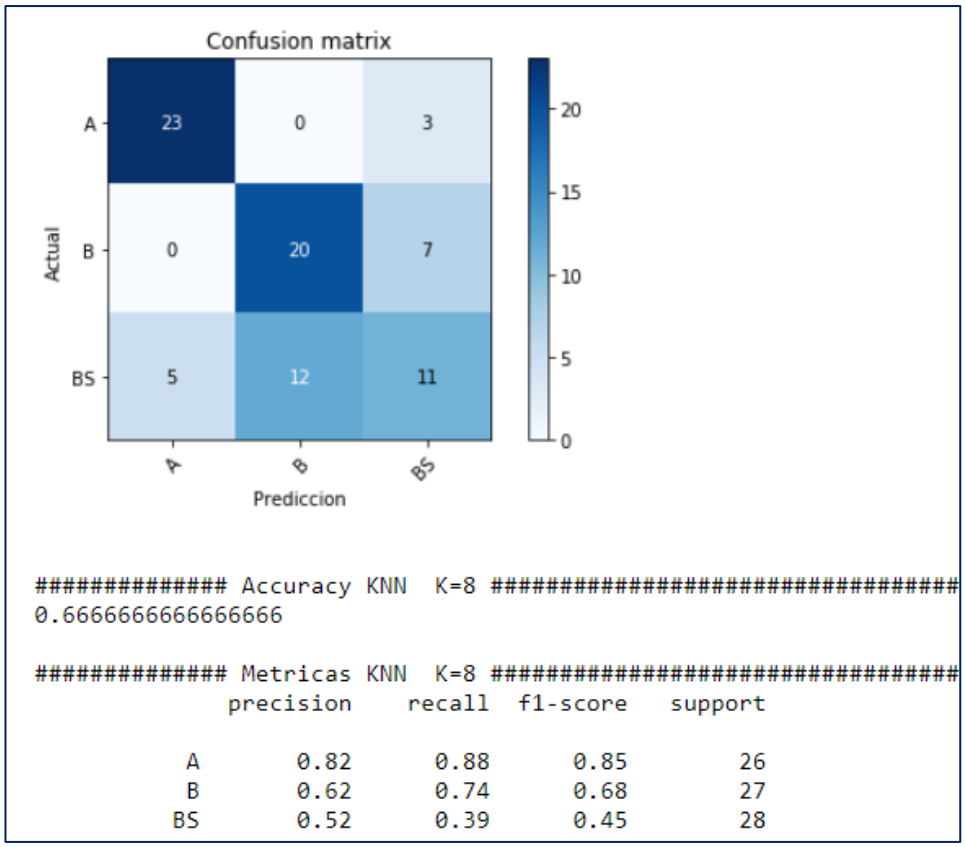
6.3 Experimento 3

El objetivo de este experimento fue tratar de mejorar los resultados obtenidos en el experimento 2, utilizando la técnica de sobre muestreó. Se tomaron los mismos conjuntos de datos de entrenamiento y prueba del experimento 2. Para aplicar la técnica de sobre muestreo.

Existen dos técnicas para trabajar con conjuntos de datos con clases desbalanceadas, la primera es conocida como sobre muestreo “oversampling” y consiste en crear muestras sintéticas de las clases minoritarias para igualarla a la clase mayoritaria, la segunda es conocida con el nombre submuestreo “undersampling”, esta técnica consiste en igualar todas las clases al número de muestras de la clase minoritaria.

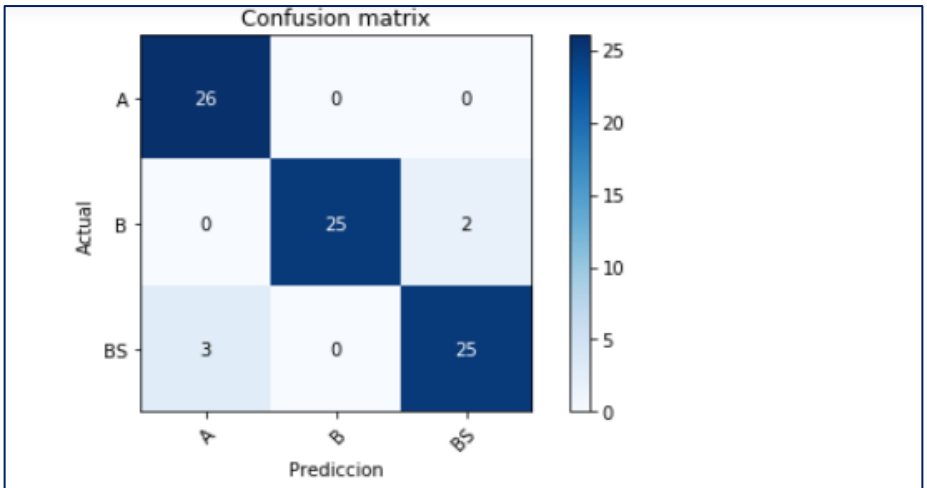
El sobre muestreo “oversampling”, es el proceso de duplicar aleatoriamente las observaciones de la clase minoritaria para reforzar su señal. Existen varias formas de hacerlo, pero la forma más simplemente es volver a muestrear con reemplazo. El proceso se realizó con el módulo de sobre muestreo resample de Scikit-Learn en Python. A continuación, se describen los pasos realizados en el proceso, primero se separan las observaciones de cada una de las clases en diferentes conjuntos de datos, segundo muestrear las clases minoritaria con reemplazo, estableciendo la cantidad de muestras para que coincida con la de la clase mayoritaria, tercer combinar los conjuntos de datos obtenidos en uno solo conjunto de datos.

La distribución de las clases del conjunto de datos antes de aplicar la técnica sobre muestreo es la siguiente: Alto, (A) = 59; Bajo (B) =73; Básico, (BS) = 89, y después de aplicar técnica, las clases quedaron (A) = 89; (B)= 89 ; (BS)=89. Los resultados obtenidos se presentan a continuación en las Figuras 17 y 18.

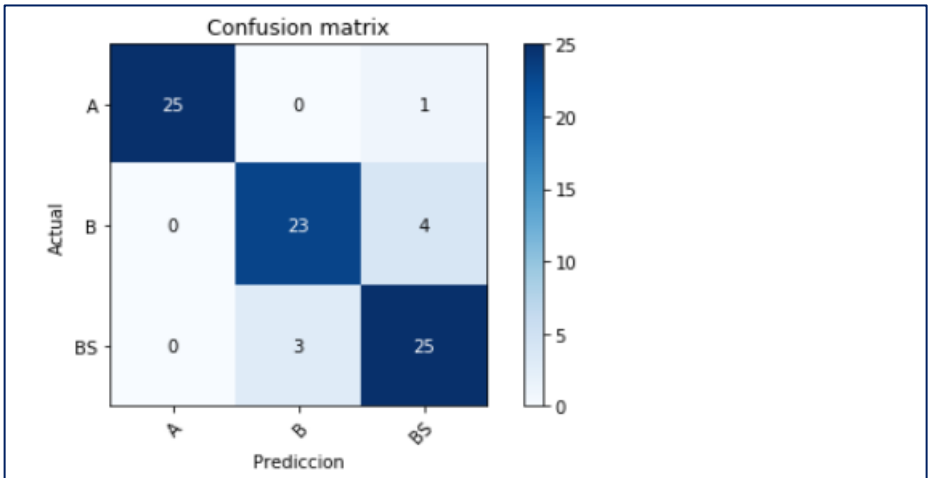



```
############## Accuracy KNN  K=8 ###################################
0.6666666666666666

############## Metricas KNN  K=8 ###################################
              precision    recall  f1-score   support

           A       0.82      0.88      0.85        26
           B       0.62      0.74      0.68        27
          BS       0.52      0.39      0.45        28
```

Figura 17 Resultados del experimento 3 KNN, Clases Alto (A), Bajo (B), Básico (BS)




```
############### Accuracy SVM Lineal C=8, gamma=1 #############
0.9382716049382716

############### Metricas SVM Lineal C=8, gamma=1 #############
              precision    recall  f1-score   support

           A       0.90      1.00      0.95        26
           B       1.00      0.93      0.96        27
          BS       0.93      0.89      0.91        28
```



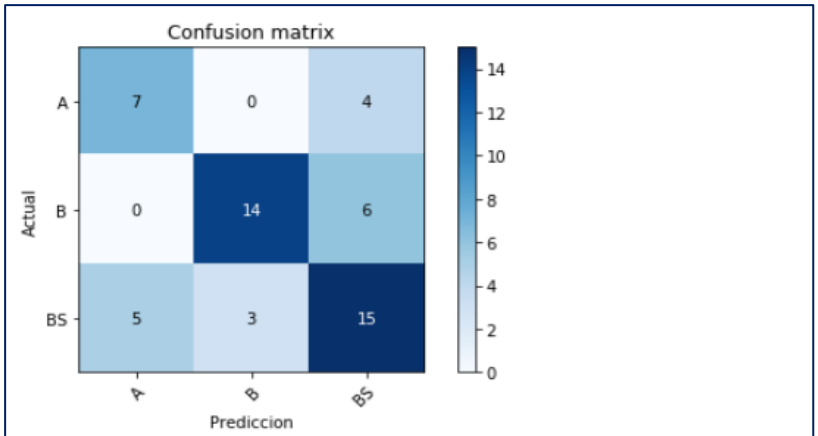

```
### Accuracy MLP capas ocultas 10 ,unidades ocultas 13 ##########
0.9012345679012346

### Metricas MLP capas ocultas 10 ,unidades ocultas 13 ##########
              precision    recall  f1-score   support

           A       1.00      0.96      0.98        26
           B       0.88      0.85      0.87        27
          BS       0.83      0.89      0.86        28
```

Figura 18 Resultados del experimento 3 modelos SVM y MLP, Clases Alto (A), Bajo (B), Básico (BS)

De acuerdo a los resultados mostrados en las Figuras 17 y 18, y comparando la métrica de la sensibilidad (recall) de los tres modelos, es importante destacar que el modelo de KNN se ve importantemente afectado por el uso de técnica de sobre muestreo, basado en que la clase mayoritaria Básico (BS), obtiene el resultado más bajo en comparación con los experimentos 1 y 2, mientras que las técnicas de SVM y MLP, parece mantener estable su funcionamiento. Es decir que se observa pequeña mejora en el funcionamiento de estos dos modelos.

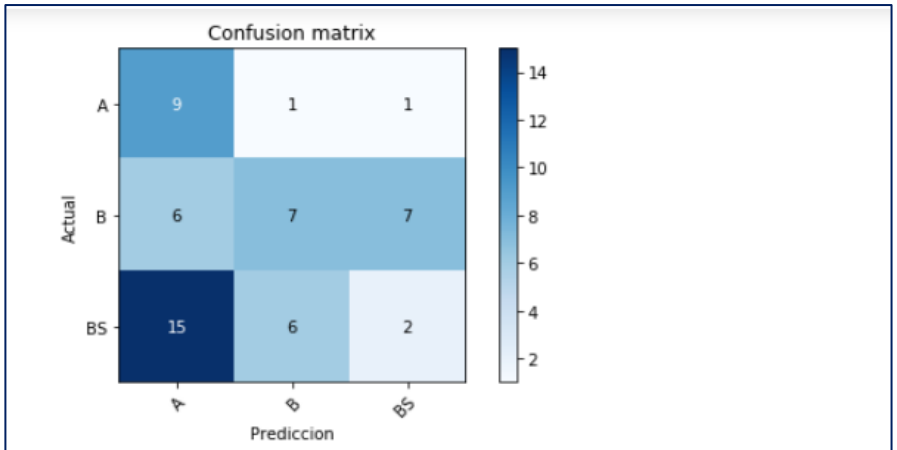
6.4 Experimento 4

El objetivo de este experimento es mejorar los resultados obtenidos en los experimentos anteriores aplicando la técnica de submuestreo (undersampling).
Este experimento fue realizado con el mismo conjunto de datos y atributos de entrenamiento y prueba utilizado en el experimento 2, la técnica de submuestreo consiste en igualar las clases del conjunto de datos, al número muestras de la clase minoritaria que en el conjunto de datos es la clase Alto (A). los resultados se presentan en las Figuras 19 y 20.

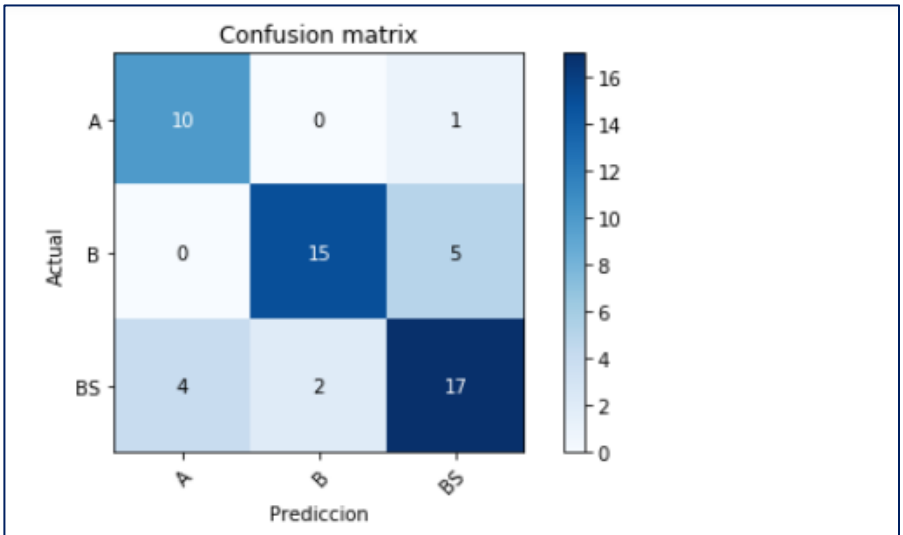



```
### Accuracy MLP capas ocultas 10 ,unidades ocultas 13 ##########
0.6666666666666666

### Metricas MLP capas ocultas 10 ,unidades ocultas 13 ##########
              precision    recall  f1-score   support

           A       0.58      0.64      0.61        11
           B       0.82      0.70      0.76        20
          BS       0.60      0.65      0.63        23
```

Figura 19 Resultados del experimento 4 modelo MLP, Clases Alto (A), Bajo (B), Básico (BS)




```
############## Accuracy KNN  K=8 ###################################
0.3333333333333333

############## Metricas KNN  K=8 ###################################
              precision    recall  f1-score   support

           A       0.30      0.82      0.44        11
           B       0.50      0.35      0.41        20
          BS       0.20      0.09      0.12        23
```




```
############## Accuracy SVM Lineal C=8, gamma=1 ###########
0.7777777777777778

############## Metricas SVM Lineal C=8, gamma=1 ###########
              precision    recall  f1-score   support

           A       0.71      0.91      0.80        11
           B       0.88      0.75      0.81        20
          BS       0.74      0.74      0.74        23
```

Figura 20 Resultados del experimento 4 modelos KNN y SVM, Clases Alto (A), Bajo (B), Básico (BS)

De acuerdo con los resultados presentados en las Figuras 19 y 20, todas las métricas se ven afectadas en su desempeño, debido a la reducción en el tamaño de conjunto de datos, el cual después de aplicar la técnica de submuestreo queda con 177 registros, de los cuales 123 quedan para entrenar los modelos y 46 para validación del modelo. El modelo KNN es el más afectado por el uso de la técnica, la clase mayoritaria (BS), en este modelo obtiene el valor más bajo de sensibilidad obtenido en todos los experimentos.

6.5 Evaluación

Para la evaluación de modelos en los diferentes experimentos se realizaron dos tablas comparativas en las se incluyen las métricas de exactitud (accuracy), la sensibilidad (recall) en la Tabla 2 y la precisión Tabla 3.

	Sensibilidad (Recall) y Exactitud (accuracy)			
	KNN			
Clase	Experimento 1	Experimento 2	Experimento 3	Experimento 4
A	0,47	0,68	0,88	0,82
B	0,53	0,63	0,74	0,35
BS	0,66	0,62	0,39	0,09
Accuracy	0,56	0,64	0,66	0,33
	SVM			
A	0,89	1,00	1,00	0,91
B	0,89	0,95	0,93	0,75
BS	0,86	0,83	0,89	0,74
Accuracy	0,88	0,91	0,93	0,77
	MLP			
A	0,84	0,79	0,96	0,64
B	0,68	0,84	0,85	0,70
BS	0,79	0,86	0,89	0,65
Accuracy	0,77	0,83	0,90	0,66

Tabla 2 Resultados de los 4 experimentos métricas recall y accuracy, Clases Alto (A), Bajo (B), Básico (BS)

	Precision			
	KNN			
Clase	Experimento 1	Experimento 2	Experimento 3	Experimento 4
A	0,90	0,76	0,82	0,30
B	0,48	0,57	0,62	0,50
BS	0,53	0,62	0,52	0,20
	SVM			
A	0,85	0,90	0,90	0,71
B	0,94	0,86	1,00	0,88
BS	0,86	0,96	0,93	0,74
	MLP			
A	0,84	0,94	1,00	0,58
B	0,81	0,80	0,88	0,82
BS	0,72	0,81	0,83	0,60

Tabla 3 Resultados de los 4 experimentos métrica precisión , Clases Alto (A), Bajo (B), Básico (BS)

La métrica que responde a la pregunta, ¿Qué proporción de positivos reales se identificaron correctamente? Se selecciono como métrica de decisión la sensibilidad (recall), para establecer el modelo con mejor desempeño en los diferentes experimentos.

De las tres clases posibles de clasificación, Alto (A), Básico (BS) y Bajo (B), en los diferentes modelos, las clases Básico (BS) y Bajo (B), adquieren una mayor importancia para el objetivo del proyecto, debido a que los estudiantes clasificados en esta clase será objeto de intervención por parte de directivos y docentes de la Institución Libardo Madrid Valderrama, creando las estrategias educativas pertinentes. Por tal motivo el experimento 3, se considera como el experimento en que los modelos obtuvieron el mejor desempeño.

Los modelos obtenidos con las SVM obtienen el mejor desempeño en todos los experimentos, se incluye una tabla para comparar la diferencia entre el recall, de los modelos SVM en los diferentes experimentos vs el recall de KNN y MLP, como se muestra en la Tabla 4.

	Diferencia SVM vs KNN			
Clase	Experimento 1	Experimento 2	Experimento 3	Experimento 4
A	0,42	0,32	0,12	0,09
B	0,36	0,32	0,19	0,40
BS	0,20	0,21	0,50	0,65
	Diferencia SVM vs MLP			
A	0,05	0,21	0,04	0,27
B	0,21	0,11	0,08	0,05
BS	0,07	-0,03	0,00	0,09

Tabla 4 Resultados de porcentaje diferencia 4 experimentos, metrica recall SVM vs KNN y SVM vs MLP

En los resultados mostrados en la Tabla 4, el experimento 3, donde se aplicó la técnica de sobre muestreó (oversampling), SVM vs MLP, la diferencia entre las dos técnicas fue mínima, incluso una de las clases Básico (BS), la diferencia es de (0), mientras para las clases Bajo (B) la diferencias (0,08) y la clase Alto (A) es de (0,04). El modelo KNN obtiene una mejora significativa en su desempeño que es mayor para este experimento que los demás experimentos.

El modelo SVM entrenado bajo las condiciones del experimento 3, fue elegido para implementar en el despliegue del prototipo de aplicación web, la información del prototipo se muestra en los anexos 4 y 5 .

7. CONCLUSIONES

Son varias las conclusiones que se pueden destacar del trabajo realizado. La primera es que, los modelos basados en distancia como KNN, estuvieron aproximadamente 32%, por debajo de rendimiento de los modelos generados con las SVM y MLP, en promedio en los cuatro experimentos. Los modelos generados con las SVM estuvieron aproximadamente 10%, por encima del rendimiento de los modelos generados con MLP, en promedio en los cuatro experimentos. En el experimento 3, se obtuvieron los mejores resultados incluso la diferencia entre el modelo generado con SVM solo estuvo un 4% por encima de los resultados obtenidos por el MLP.

Se logro desarrollar el prototipo de una aplicación Web que incluye un modelo entrenado de SVM, con una sensibilidad del 94% aproximadamente. Este puede ser usado por la Institución Educativa Libardo Madrid Valderrama, para hacer la predicción del rendimiento académico en matemáticas de los estudiantes para el siguiente año lectivo. Este trabajo requiere la articulación con la practicas diarias de la educación, el compromiso y la retro alimentación de directivos, docentes y estudiantes.

Por el momento no es posible aplicar el trabajo realizado para todas las materias que cursan los estudiantes en los diferentes grados. Esto debido a la falta de organización en sistema de registro de matrícula de los estudiantes (ZETI), no existe un estándar para matricular un estudiante, algunos son registrados con documento de identidad, otros con el nombre, otros con un código estudiante. Esto genera dificultad para hacer el cruce de información académica de otras materias y los datos de la encuesta.

8 TRABAJO FUTURO

Un trabajo futuro, sería crear modelos para otras áreas del conocimiento, y poder, predecir la clasificación de los estudiantes en diferentes materias simultáneamente.

9. REFERENCIAS BIBLIOGRÁFICAS

[1] S. M. Merchán and J. A. Duarte, “Analysis of Data Mining Techniques for Constructing a Predictive Model for Academic Performance,” IEEE Lat. Am. Trans., vol. 14, no. 6, pp. 2783–2788, 2016.

[2] P. Sectorial, “Encuentro Regional 2011.” [Online]. Available: http://www.mineducacion.gov.co/cvn/1665/articles-279754_archivo_pdf_ministra.pdf. [Accessed: 22-May-2017].

[3] Tan, M. & Shao, P. (2015). Prediction of Student Dropout in E-Learning Program Through the Use of Machine Learning Method. International Journal of Emerging Technologies in Learning (iJET), 10(1), 11-17. Kassel, Germany: International Association of Online Engineering

[4] “Zeti.” [Online]. Available: https://zeti.net.co/. [Accessed: 16-May-2017].

[5] J. S. Henao Parra, “Las redes neuronales y su desempeño bajo la estrategia de Neuroevolución,” web, 2013. [Online]. Available: https://repository.javeriana.edu.co/handle/10554/12100?show=full. [Accessed: 22-May-2017].

[6] D. Santín González, “Detección de alumnos de riesgo y medición de la eficiencia de centros escolares mediante redes neuronales / Daniel Santín González,” Madrid : Facultad de Ciencias Económicas y Empresariales, Universidad Complutense de Madrid, 1999, 2008. [Online].Available: http://cisne.sim.ucm.es/search~S6*spi/?searchtype=h&searcharg=W28+%289902%29&searchscope=6&sortdropdown=-&SORT=D&extended=0&SUBMIT=Buscar&searchlimits=&searchorigarg=hW28+%289902%29. [Accessed: 29-May-2017].

[7] P. Cortez and A. Silva, “Using Data Mining To Predict Secondary School Student Performance,” 5th Annual Future Business Technology. Conf., vol. 2003, no. 2000, pp. 5–12, 2008.

[8] N. Pukkhem, “A semantic-based approach for representing successful graduate predictive rules,” 16th Int. Conf. Advanced Communications

Technology., pp. 222–227, 2014.

[9] Castro, F., Vellido, A., Nebot, À., & Mugica, F. (2007). Applying data mining techniques to e-learning problems. In Evolution of teaching and learning paradigms in intelligent environment (pp. 183-221). Springer Berlin Heidelberg.

[10] “Manual CRISP-DM de IBM SPSS Modeler.” [Online]. Available: ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/15.0/es/CRISP-DM.pdf. [Accessed: 17-May-2017].

[11] R. Jindal and M. D. Borah, “A SURVEY ON EDUCATIONAL DATA MINING AND RESEARCH TRENDS" Rajni vol. 5, no. 3, pp. 53–73, 2013.

[12] Baker, R.S.J.D., & Yacef, K. (2009). The State of Educational Data Mining in 2009: A review and future visions. Journal of educational Data Mining, 1, 3-17.

[13] P. G. Espejo, S. Ventura and F. Herrera, "A Survey on the Application of Genetic Programming to Classification," in IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 40, no. 2, pp. 121-144, March 2010.

[14] “Análisis comparativo de algoritmos de aprendizaje para predecir la evolución de pacientes con Daño Cerebral Adquirido.” [Online]. Available: http://oa.

[15] “KDD Proceso de Extracción de conocimiento WebMining.” [Online]. Available:http://www.webmining.cl/2011/01/proceso-de-extraccion-de-conocimiento/. [Accessed: 31-May-2017].

[16] K. Tan, Steinbach, “Data Mining Classification: Basic Concepts, Decision Trees, and Model Evaluation Lecture Notes for Chapter 4 Introduction to DataMining.”[Online]Avaliable:http://www.users.cs.umn.edu/~kumar/dmbook/dmslides/chap4_basic_classification.pdf . [Accessed: 31-May-2017].

10. ANEXOS

Anexo 1 Política de protección de datos de Institución educativa Libardo Madrid Valderrama

POLÍTICA DE PROTECCIÓN DE DATOS PERSONALES DE LOS TITUARES DEL COLEGIO LIBARDO MADRID VALDERRAMA

En cumplimiento a lo dispuesto en la Ley estatutaria 1581 de 2012 y a su Decreto Reglamentario 1377 de 2013, La Institución Educativa Libardo Madrid Valderrama informa la política aplicable a la entidad para el tratamiento de protección de datos personales.

I.IDENTIFICACIÓN:

NOMBRE DE LA INSTITUCIÓN: La Institución Educativa Libardo Madrid Valderrama
DIRECCIÓN: Cl 40 41 F Esq. B/Unión de Vivienda Popular Cali
CORREO ELECTRÓNICO: admin@libardomadridcali.edu.com.co
TELEFÓNO DEL RESPONSABLE: 572-3283422

La Instrucción Educativa Libardo Madrid Valderrama de Carácter Oficial Ubicada en comuna 16 de la ciudad Cali, Ofrece un servicio educativo inclusivo en los niveles de preescolar, básica, media académica y educación formal para adultos que favorece el desarrollo humano integral desde la dimensión artística, deportiva y las competencias del siglo XXI. A través de la implementación de diversas estrategias que propicien el aprender a conocer, a ser, a hacer y a vivir juntos, para el beneficio de la ciudadanía

II.DEFINICIONES:

AUTORIZACIÓN: consentimiento previo, expreso e informado del titular para llevar a cabo el tratamiento de datos personales.
AVISO DE PRIVACIDAD: comunicación verbal o escrita generada por el responsable dirigida al titular para el tratamiento de sus datos personales, mediante la cual le informa acerca de la existencia de las políticas de tratamiento de información que le serán aplicables, la forma de acceder a las mismas y las finalidades del tratamiento que se pretende dar a los datos personales.
BASE DE DATOS: conjunto organizado de datos personales que sea objeto de tratamiento.
DATO PERSONAL: cualquier pieza de información vinculada a una o varias personas determinadas o determinables o que puedan asociarse a una persona natural o jurídica.
DATO PÚBLICO: es el dato que no sea semiprivado, privado o sensible. Son considerados datos públicos, entre otros, los datos relativos al estado civil de las personas, a su profesión u oficio y a su calidad de comerciante. Por su naturaleza, los datos públicos pueden estar contenidos, entre otros, en registros públicos, documentos públicos, gacetas y boletines oficiales y sentencias judiciales debidamente ejecutoriadas que no estén sometidas a reserva. Dicha información será manejada por la Direcciones de Talento Humano, de Admisiones y de Secretaría Académica.
DATOS SENSIBLES: se entiende por datos sensibles aquellos que afectan la intimidad del titular o cuyo uso indebido puede generar su discriminación, tales como aquellos que revelen el origen racial o étnico, la orientación política, las convicciones religiosas o filosóficas, la pertenencia a sindicatos, organizaciones sociales, de derechos humanos o que promueva intereses de cualquier partido político o que garanticen los derechos y garantías de partidos políticos de oposición, así como los datos relativos a la salud, a la vida sexual, y los datos biométricos. Dicha información será manejada por la Direcciones de Talento Humano, de Admisiones y de Secretaría Académica.

DATOS INDISPENSABLES: se entienden como aquellos datos personales de los titulares imprescindibles para llevar a cabo la actividad de educación en docencia, investigación y extensión. Los datos de naturaleza indispensable deberán ser proporcionados por los titulares de los mismos o los legitimados para el ejercicio de estos derechos. Dicha información será manejada por la Dirección de Talento Humano.

ENCARGADO DEL TRATAMIENTO: persona natural o jurídica, pública o privada que por sí misma o en asocio con otros, realice el Tratamiento de datos personales por cuenta del Responsable del Tratamiento.

LEY DE PROTECCIÓN DE DATOS: es la Ley 1581 de 2012 y sus Decretos reglamentarios o las normas que los modifiquen, complementen o sustituyan.

HABEAS DATA: derecho de cualquier persona a conocer, actualizar y rectificar las informaciones que se hayan recogido sobre ellas en el banco de datos y en archivos de entidades públicas y privadas.

RESPONSABLE DEL TRATAMIENTO: persona natural o jurídica, pública o privada que por sí misma o en asocio con otros, decida sobre la base de datos y/o Tratamiento de los datos.

TITULAR: persona natural cuyos datos personales sean objeto de Tratamiento.

TRATAMIENTO: cualquier operación o conjunto de operaciones sobre datos personales, tales como la recolección, almacenamiento, uso, circulación o supresión.

TRANSFERENCIA: la transferencia de datos tiene lugar cuando el responsable y/o encargado del tratamiento de datos personales, ubicado en Colombia, envía la información o los datos personales a un receptor, que a su vez es responsable del tratamiento y se encuentra dentro o fuera del país.

TRANSMISIÓN: tratamiento de datos personales que implica la comunicación de los mismos dentro o fuera del territorio de la República de Colombia cuando tenga por objeto la realización de un tratamiento por el encargado por cuenta del responsable.

III. PRINCIPIOS: En el desarrollo, interpretación y aplicación de la ley 1581 de 2012 por la cual se dictan disposiciones generales para la protección de datos personales y las normas que la complementan, modifican o adicionan, se aplicarán de manera armónica e integral los siguientes principios rectores:

a) **PRINCIPIO DE LA LEGALIDAD**: el Tratamiento de datos es una actividad regulada que debe sujetarse a lo establecido en la ley y las demás disposiciones que la desarrollen.
b) **PRINCIPIO DE FINALIDAD**: el tratamiento debe obedecer a una finalidad legítima de acuerdo con la Constitución y la Ley, la cual debe ser informada al titular. En lo correspondiente a la recolección de datos personales, La Institución Educativa Libardo Madrid Valderrama se limitará a aquellos datos que sean pertinentes y adecuados para la finalidad con la cual fueron recolectados o requeridos; la rectoría, secretarias, coordinación administrativa, admisiones, contabilidad etc.,
   Deberán informar al titular el motivo por el cual se solicita la información y el uso específico que se le dará a la misma.
c) **PRINCIPIO DE LIBERTAD**: el tratamiento solo puede ejercerse con el consentimiento previo, expreso, e informado del titular. Los datos personales no podrán ser obtenidos o divulgados sin previa autorización, o en ausencia de mandato legal o judicial que releve el consentimiento.
d) **PRINCIPIO DE VERACIDAD O CALIDAD**: la información sujeta a tratamiento debe ser veraz, completa, exacta, actualizada, comprobable y comprensible. Se prohíbe el tratamiento de datos parciales, incompletos, fraccionados o que induzcan a error.
e) **PRINCIPIO DE TRANSPARENCIA**: en el tratamiento debe garantizarse el derecho del titular a

b. **Publicitarios**: Para promocionar nuevos sistemas o desarrollos educativos que el colegio implementa.

c. **Educativos**: Para orientar a la institución en los aspectos necesarios para el cumplimiento de sus fines.

d. **Comerciales**: Para el cumplimiento de las obligaciones derivadas de las relaciones contractuales existentes con los grupos de interés.

e. **Formativos:** Para dar a conocer los eventos académicos, publicaciones.

f. **Investigativos**: Para informar los avances de los procesos de investigación que se desarrollan en la institución.

e) **ALMACENAMIENTO**: El almacenamiento de la información se realizará de manera electrónica, en formatos xls.

f) **CIRCULACIÓN**: El flujo de la información contenida en la base de datos, solo puede ser utilizada dentro de la Institución Educativa Libardo Madrid Valderrama para los fines que se recolectó, tal como está descrito en la Política de Tratamiento de la Información Personal, bien sea para temas Informativos, Publicitarios, Educativos, Comerciales, Formativos e Investigativos, ya que está expresamente prohibido transferir información contenida en las bases de datos, a terceros no autorizados.

g) **SUPRESIÓN**: La eliminación de la información suministrada por parte del titular que se encuentre en las bases de datos la Institución Educativa Libardo Madrid Valderrama, se hará, cuando éste expresamente lo solicite, indicando de manera expresa que no le interesa recibir ninguna clase de comunicación, y que desea que no continúe con el tratamiento de sus datos.

h) **REPORTES**. El reporte de novedades debe hacerse ante la Superintendencia de Industria y Comercio. Ello está a cargo del oficial de protección de datos, el cual tiene como finalidad reportar las novedades de los reclamos presentados por los titulares y los incidentes de seguridad que afectan la información personal.

V. DERECHOS QUE LE ASISTEN AL TITULAR DE LA INFORMACIÓN

El titular de los datos personales tendrá los siguientes derechos:

a) Conocer, actualizar y rectificar sus datos personales frente a la Institución Educativa Libardo Madrid Valderrama, en su condición de responsable del tratamiento. Este derecho se podrá ejercer, entre otros, frente a datos parciales, inexactos, incompletos, fraccionados, que induzcan a error, o aquellos cuyo tratamiento esté expresamente prohibido o no haya sido autorizado.
b) Solicitar prueba de la autorización otorgada a la Institución Educativa Libardo Madrid Valderrama, salvo cuando expresamente se exceptúe como requisito para el tratamiento (casos en los cuales no es necesaria la autorización).

c) Ser informado por la Institución Educativa Libardo Madrid Valderrama, previa solicitud, respecto del uso que le ha dado a sus datos personales.
d) Presentar ante la Superintendencia de Industria y Comercio quejas por infracciones a lo dispuesto en la ley 1581 de 2012 y las demás normas que la modifiquen, adicionen o complementen.
e) Revocar la autorización y/o solicitar la supresión del dato cuando en el Tratamiento no se respeten los principios, derechos y garantías constitucionales y legales.
f) Acceder en forma gratuita a sus datos personales que hayan sido objeto de tratamiento.

VI. DERECHOS DE LOS NIÑOS Y ADOLESCENTES

En el Tratamiento se asegurará el respeto a los derechos prevalentes de los niños, niñas y adolescentes. Queda proscrito el Tratamiento de datos personales de niños, niñas y adolescentes, salvo aquellos datos que sean de naturaleza pública.

Es tarea del Estado y las entidades educativas de todo tipo proveer información y capacitar a los representantes legales y tutores sobre los eventuales riesgos a los que se enfrentan los niños, niñas y adolescentes respecto del Tratamiento indebido de sus datos personales, y proveer de conocimiento acerca del uso responsable y seguro por parte de niños, niñas y adolescentes de sus datos personales, su derecho a la privacidad y protección de su información personal y la de los demás.

VII. DEBERES DE LA INSTITUCIÓN EDUCATIVA LIBARDO MADRID VALDERRAMA

En virtud de la presente política de tratamiento y protección de datos personales son deberes de la institución los siguientes, sin perjuicio de las disposiciones previstas en la ley.

a) Garantizar al titular, en todo tiempo, el pleno y efectivo ejercicio del derecho de hábeas data.
b) Solicitar y conservar, copia de la respectiva autorización otorgada por el titular.
c) Informar debidamente al titular sobre la finalidad de la recolección y los derechos que le asisten en virtud de la autorización otorgada.
d) Conservar la información bajo las condiciones de seguridad necesarias para impedir su adulteración, pérdida, consulta, uso o acceso no autorizado o fraudulento.
e) Garantizar que la información sea veraz, completa, exacta, actualizada, comprobable y comprensible.
f) Actualizar la información, atendiendo de esta forma todas las novedades respecto de los datos del titular. Adicionalmente, se deberán implementar todas las medidas necesarias para que la información se mantenga actualizada.
g) Rectificar la información cuando sea incorrecta y comunicar lo pertinente.
h) Respetar las condiciones de seguridad y privacidad de la información del titular.
i) Tramitar las consultas y reclamos formulados en los términos señalados por la ley.
j) Identificar cuando determinada información se encuentra en discusión por parte del titular.
k) Informar a solicitud del titular sobre el uso dado a sus datos.
l) Informar a la autoridad de protección de datos cuando se presenten violaciones a los códigos de seguridad y existan riesgos en la administración de la información de los titulares.
m) Cumplir los requerimientos e instrucciones que imparta la Superintendencia de Industria y

Comercio sobre el tema en particular.

n) Usar únicamente datos cuyo tratamiento esté previamente autorizado de conformidad con lo previsto en la ley 1581 de 2012.

o) La Institución Educativa Libardo Madrid Valderrama hará uso de los datos personales del titular solo para aquellas finalidades para las que se encuentre facultada debidamente y respetando en todo caso la normatividad vigente sobre de datos personales.

VIII. EL REGISTRO NACIONAL DE BASES DE DATOS

El Registro Nacional de Bases de Datos (RNBD), es el directorio público de las bases de datos sujetas a Tratamiento que operan en el país y será administrado por la Superintendencia de Industria y Comercio y será de libre consulta para los ciudadanos. Una vez el Gobierno Nacional reglamente la información mínima que debe contener el Registro, y los términos y condiciones bajo los cuales se deben inscribir en éste, El Colegio Libardo Madrid Valderrama aportará a la Superintendencia de Industria y Comercio las bases de datos sujetas a tratamiento en el tiempo indicado.

IX. AUTORIZACIONES Y CONSENTIMIENTO DEL TITULAR

Sin perjuicio de las excepciones previstas en la Ley, en el tratamiento de datos personales del titular se requiere la autorización previa e informada de éste, la cual deberá ser obtenida por cualquier medio que pueda ser objeto de consulta posterior.

X. MEDIO Y MANIFESTACIÓN PARA OTORGAR LA AUTORIZACIÓN DEL TITULAR

La Institución Educativa Libardo Madrid Valderrama en los términos dispuestos en la Ley generó un aviso en el cual se comunica a los titulares que pueden ejercer su derecho al tratamiento de los datos personales a través de la página http://ielibardomadridvalderrama.blogspot.com.co/ y a través del correo electrónico admin@libardomadridcali.edu.com.co

XI. EVENTOS EN LOS CUALES NO ES NECESARIA LA AUTORIZACIÓN DEL TÍTULAR DE LOS DATOS PERSONALES

La autorización del titular de la información no será necesaria en los siguientes casos:

a) Información requerida por una entidad pública o administrativa en ejercicio de sus funciones legales o por orden judicial.
b) Datos de naturaleza pública.
c) Casos de urgencia médica o sanitaria.
d) Tratamiento de información autorizado por la ley para fines históricos, estadísticos o científicos. Datos relacionados con el Registro Civil de las personas.

XII. LEGITIMACIÓN PARA EL EJERCICIO DEL DERECHO DEL TITULAR

Los derechos de los titulares establecidos en la Ley podrán ejercerse por las siguientes personas:

a) Por el titular, quien deberá acreditar su identidad en forma suficiente por los distintos medios que le ponga a disposición de La Institución Educativa Libardo Madrid Valderrama

b) Por los causahabientes del titular, quienes deberán acreditar tal calidad.
c) Por el representante y/o apoderado del titular, previa acreditación de la representación o apoderamiento
d) Por estipulación a favor de otro o para otro. Los derechos de los niños, niñas y adolescentes se ejercerán por las personas que estén facultadas para representarlos.

XIII. TRATAMIENTO AL CUAL SERÁN SOMETIDOS LOS DATOS Y FINALIDAD DEL MISMO

El tratamiento para los datos personales indispensables de estudiantes, padres de familia, docentes, trabajadores y/o contratistas, egresados estará enmarcado en el orden legal y en virtud de la condición del Institución de Educación, básica y media academica, y serán todos los necesarios para el cumplimiento de la misión institucional de docencia, investigación y extensión.

Para el caso de datos personales sensibles, se podrá hacer uso y tratamiento de ellos cuando:

a) El Titular haya dado su autorización explícita a dicho Tratamiento, salvo en los casos que por ley no sea requerido el otorgamiento de dicha autorización;
b) El Tratamiento sea necesario para salvaguardar el interés vital del Titular y este se encuentre física o jurídicamente incapacitado. En estos eventos, los representantes legales deberán otorgar su autorización;
c) El Tratamiento sea efectuado en el curso de las actividades legítimas y con las debidas garantías por parte de una fundación, ONG, asociación o cualquier otro organismo sin ánimo de lucro, cuya finalidad sea política, filosófica, religiosa o sindical, siempre que se refieran exclusivamente a sus miembros o a las personas que mantengan contactos regulares por razón de su finalidad. En estos eventos, los datos no se podrán suministrar a terceros sin la autorización del Titular;
d) El Tratamiento se refiera a datos que sean necesarios para el reconocimiento, ejercicio o defensa de un derecho en un proceso judicial;
e) El Tratamiento tenga una finalidad histórica, estadística o científica. En este evento deberán adoptarse las medidas conducentes a la supresión de identidad de los Titulares. El tratamiento de datos personales de niños, niñas y adolescentes está prohibido, excepto cuando se trate de datos de naturaleza pública, y cuando dicho tratamiento cumpla con los siguientes parámetros y/o requisitos:
    1. que respondan y respeten el interés superior de los niños, niñas y adolescentes.
    2. que se asegure el respeto de sus derechos fundamentales.

Cumplidos los anteriores requisitos, el representante legal de los niños, niñas o adolescentes otorgará la autorización, previo ejercicio del menor de su derecho a ser escuchado, opinión que será valorada teniendo en cuenta la madurez, autonomía y capacidad para entender el asunto la Institución

Educativa Libardo Madrid Valderrama velará por el uso adecuado del tratamiento de los datos personales de los niños, niñas o adolescentes

XIV. PERSONAS A QUIENES SE LES PUEDE SUMINISTRAR LA INFORMACIÓN

La información que reúna las condiciones establecidas en la ley podrá suministrarse a las siguientes personas:

a) A los titulares, sus causahabientes (cuando aquellos falten) o sus representantes legales.
b) A las entidades públicas o administrativas en ejercicio de sus funciones legales o por orden judicial.
c) A los terceros autorizados por el titular o por la ley.

XV. PERSONA O ÁREA RESPONSABLE DE LA ATENCIÓN DE PETICIONES, CONSULTAS Y RECLAMOS

La Institución Educativa Libardo Madrid Valderrama ha designado como área responsable de velar por el cumplimiento de esta política al interior de la institución en cabeza de La Coordinación Administrativa con el apoyo de la Secretaría Académica. Áreas Funcionales que manejan los Datos Personales de los Titulares y profesional es en Seguridad de la Información.

Esta dependencia estará atenta para resolver peticiones, consultas y reclamos por parte de los titulares y para realizar cualquier actualización, rectificación y supresión de datos personales, a través del correo electrónico admin@libardomadridcali.edu.com.co

XVI. PROCEDIMIENTO PARA LA ATENCIÓN DE CONSULTAS, RECLAMOS Y PETICIONES

a) **Consultas**: Los Titulares o sus causahabientes podrán consultar la información personal del Titular que repose en La Institución Educativa Libardo Madrid Valderrama quien suministrará toda la información contenida en el registro individual o que esté vinculada con la identificación del Titular. La consulta se formulará a través del correo admin@libardomadridcali.edu.com.co. La consulta será atendida en un término máximo de diez (10) días hábiles contados a partir de la fecha de recibo de la misma. Cuando no fuere posible atender la consulta dentro de dicho término, se informará al interesado, expresando los motivos de la demora y señalando la fecha en que se atenderá su consulta, la cual en ningún caso podrá superar los cinco (5) días hábiles siguientes al vencimiento del primer término.

b) **Reclamos**: El Titular o sus causahabientes que consideren que la información contenida en una base de datos debe ser objeto de corrección, actualización o supresión, o cuando adviertan el presunto incumplimiento de cualquiera de los deberes contenidos en la ley, podrán presentar un reclamo ante la Institución Educativa Libardo Madrid Valderrama el cual será tramitado bajo las siguientes reglas:

Anexo 2 Consentimiento Informado

Para la transferencia internacional de datos personales de los titulares, La Institución Educativa Libardo Madrid Valderrama tomará las medidas necesarias para que los terceros conozcan y se comprometan a observar esta Política, bajo el entendido que la información personal que reciban, únicamente podrán ser utilizada para asuntos directamente relacionados con la institución y solamente mientras ésta dure y no podrá ser usada o destinada para propósito o fin diferente.

Para la transferencia internacional de datos personales se observará lo previsto en el artículo 26 de la Ley 1581 de 2012. Las transmisiones internacionales de datos personales que efectúe la institución, no requerirán ser informadas al Titular ni contar con su consentimiento cuando medie un contrato de transmisión de datos personales de conformidad al artículo 25 del Decreto 1377 de 2013.

XVIII. VIGENCIA Y ACTUALIZACIÓN:

La presente Política entra en vigencia a partir de su aprobación, Se articularán las acciones conducentes a la protección de datos personales dentro de la Institución Educativa Libardo Madrid Valderrama el cual realizará revisiones periódicas de la correcta ejecución de la Política.

La versión aprobada de esta Política se publicará en la página oficial de la Institución Educativa Libardo Madrid Valderrama.

Es un deber de los empleados, padres de familia y colaboradores conocer esta Política y realizar todos los actos conducentes para su cumplimiento, implementación y mantenimiento.

La presente Política de Protección de Datos Personales fue aprobada el día 14 de agosto de 2017.

Orlando Quintero
Rector
Institución Educativa Libardo Madrid Valderrama

CONSENTIMIENTO INFORMADO PARA LA RECOLECCION DE DATOS.
PROYECTO: Uso De Minería de datos para predecir el rendimiento académico de estudiantes de la Institución Educativa Libardo Madrid Valderrama.
CIUDAD: Santiago de Cali.

Yo._____
identificado(a) con el número de cédula que aparece al pie de mi firma, actuando a mi nombre y en calidad de (Acudiente), acepto que el estudiante _____identificado(a) con tarjeta de identidad número _____ participe de manera voluntaria en el proceso de recolección de datos para el proyecto: "Uso De Minería de datos para predecir el rendimiento académico de estudiantes de la Institución Educativa Libardo Madrid Valderrama", La propuesta de investigación pretende predecir el rendimiento académico de los estudiantes de la Institución Educativa Libardo Madrid Valderrama con el uso de técnicas de minería de datos a partir de: (1) Las notas de otros periodos académicos cursados por el estudiante, (2) Los datos tomados de la aplicación de un cuestionario WEB que permita conocer las características socioeconómicas y socioculturales del estudiante. El proyecto es realizado por el investigador y Estudiante de Maestría en Ingeniería Andrés Racines Bolaños identificado con cedula de ciudadanía 94.507.629, de la Pontificia Universidad Javeriana Cali.

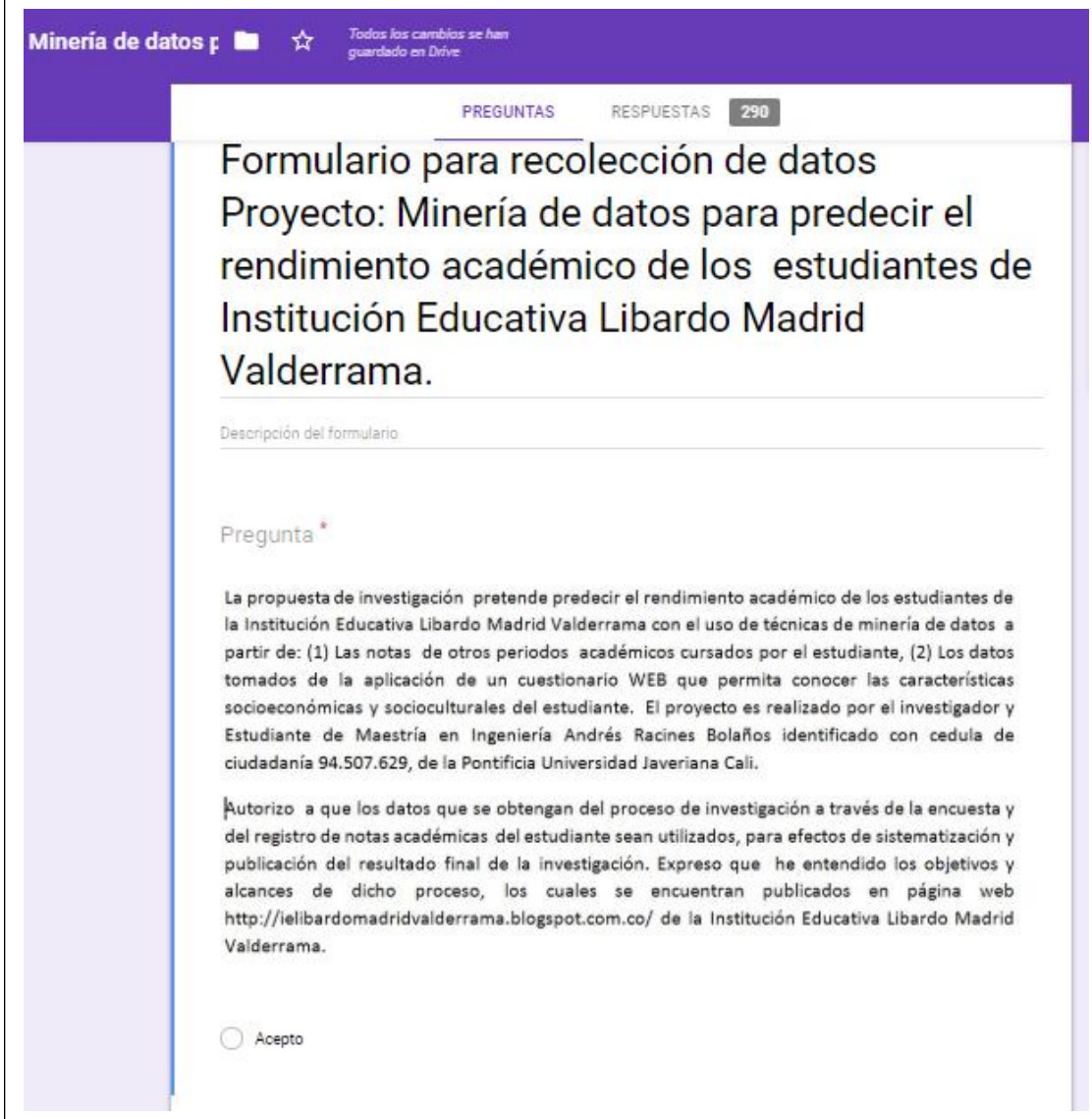
Autorizo a que los datos que se obtengan del proceso de investigación a través de la encuesta y del registro de notas académicas del estudiante sean utilizados, para efectos de sistematización y publicación del resultado final de la investigación. Expreso que he entendido los objetivos y alcances de dicho proceso, los cuales se encuentran publicados en página web http://ielibardomadridvalderrama.blogspot.com.co/ de la Institución Educativa Libardo Madrid Valderrama.

FIRMA DEL ACUDIENTE: _____

CEDULA:_____

FECHA:_____

Anexo 3 Formulario de Google.com



Documento Numero *

Ingresar el numero de Identidad sin puntos ni comas

Texto de respuesta corta

Nombres *

Ingrese su primer y segundo nombre

Texto de respuesta corta

Apellidos *

Ingrese su primer y segundo apellido tal como aparece en su documento de identidad

Texto de respuesta corta

Grupo *

Ingrese el grupo al cual pertenece separado por un guion ejemplo 6-1

Texto de respuesta corta

Tipo de vivienda en la que vive el estudiante *

Debe de seleccionar una de las siguientes opciones

1. Propia
2. Arrendada
3. Familiar
4. Otro

El estudiante convive con madre *

Debe de seleccionar una de las siguientes opciones

1. Si
2. No

El estudiante convive con padre *

Debe de seleccionar una de las siguientes opciones

1. Si
2. No

Educación de la madre del estudiante *

Debe de seleccionar una de las siguientes opciones

1. Primaria
2. Secundaria
3. Profesional
4. Posgrado
5. No sabe

Educación del padre del estudiante *

Debe de seleccionar una de las siguientes opciones

1. Primaria
2. Secundaria
3. Profesional
4. Posgrado
5. No sabe

La madre del estudiante trabaja *

Debe de seleccionar una de las siguientes opciones

1. Si
2. No

Número de personas que conviven con el estudiante *

Escriba el numero de personas con las conviven actualmente

Texto de respuesta corta

La posición entre los hermanos que ocupa *

Escriba el numero de personas con las conviven actualmente

Texto de respuesta corta

Tipo de población a la cual pertenece el estudiante *

Si cree que pertenece alguna de las siguientes poblaciones seleccione una en caso de no saber seleccione ninguna

1. Afrodescendiente
2. indígena
3. Raizal
4. Jóvenes vulnerables
5. Desplazado
6. Desplazado por la violencia
7. Desplazado por fenómeno natural
8. Ninguna

Tiempo que demora en llegar al Colegió *

Seleccione una de las opciones

1. Menos de 15 Minutos
2. Entre 15 a 30 Minutos
3. Entre 30 minutos a 1 hora
4. Mas de 1 hora

Número de horas de estudio extra clases semanal *

Seleccione una de las opciones

1. Entre 1 a 2 horas
2. Entre 2 a 3 horas
3. Entre 3 a 5 horas
4. Entre 5 a 10 horas
5. Mas de 10 horas

El estudiante Trabaja *

Seleccione una de las opciones

1. Si
2. No

El estudiante tiene computador en su casa *

Seleccione una de las opciones

1. Si
2. No

El estudiante tiene internet en su casa *

Seleccione una de las opciones

1. Si
2. No

El estudiante practica algún deporte *

Seleccione una de las opciones

1. Si
2. No

El estudiante esta en una relación sentimental *

Seleccione una de las opciones

1. Si
2. No

Anexo 4 Requerimiento de instalación de la aplicación web

Los requerimientos de la aplicación web de minería de datos se describen a continuación.

Requerimiento de Hardware

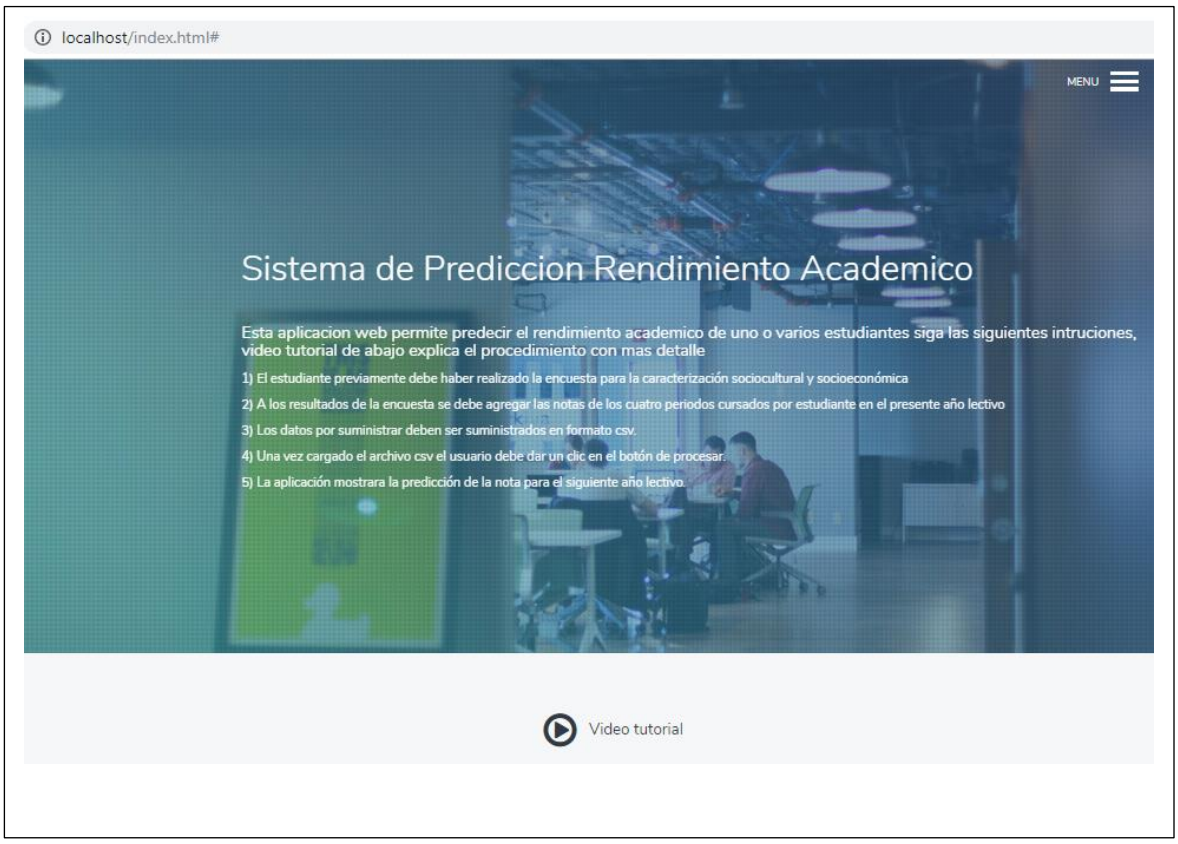
Procesador a 1 GHz con soporte para instrucciones PAE, NX y SSE2.
1 GB RAM (32 bits) o 2 GB RAM (64 bits)
16 GB (32 bits) o 20 GB (64 bits) de espacio en disco duro.
Tarjeta gráfica compatible con DX9 y controlador WDDM.

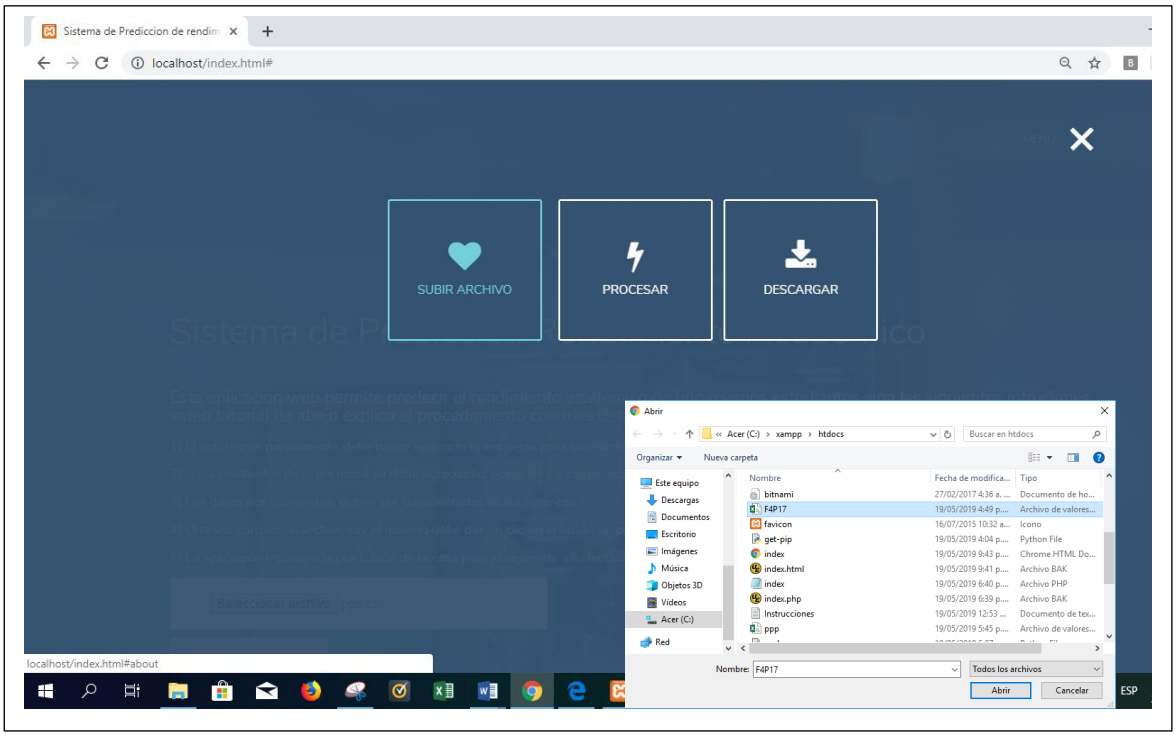
Requerimiento de Software

Windows 10
Python 3.7.3
NumPy.
SciPy
Pip
scikit-learn.

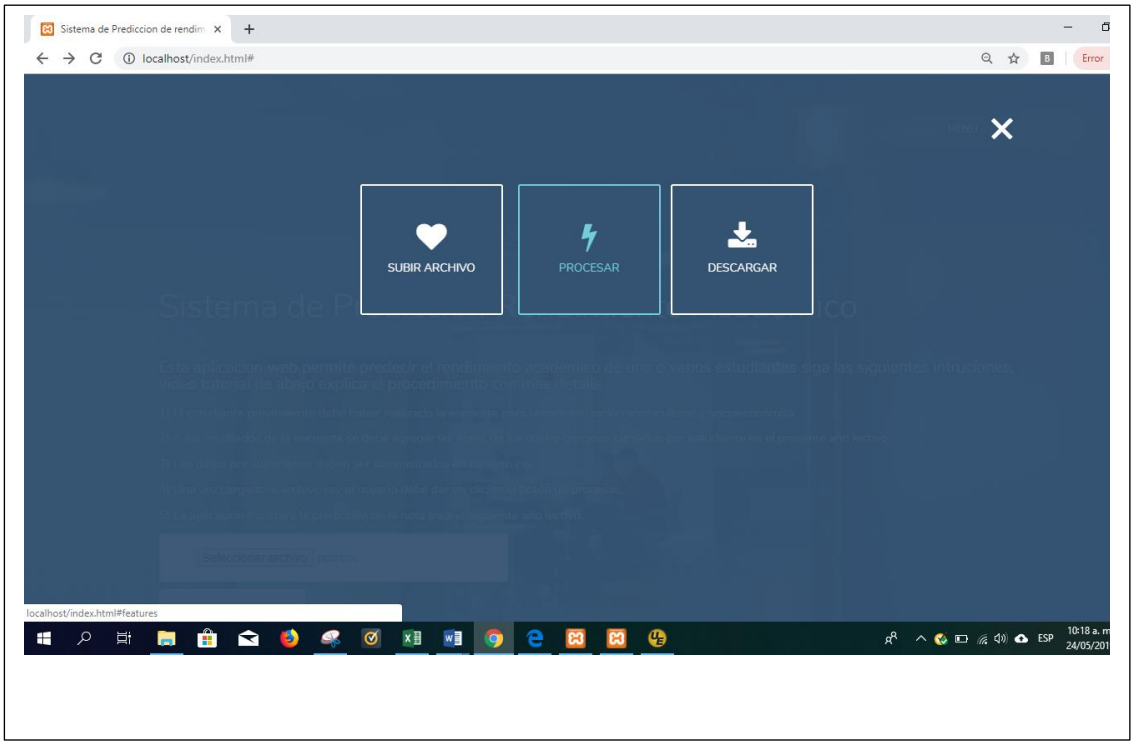
Anexo 5 Aplicación web para la predicción de rendimiento académico

El sistema de predicción de rendimiento académico posee un modelo de entrenado de minería de datos para la predicción de los estudiantes, recibe como dato de entrada un archivo csv con los 17 atributos socioculturales y socioeconómicos recopilados en la encuesta web, más la nota de los cuatro periodos académicos del año lectivo anterior. La aplicación web cuenta con tres opciones: subir archivo, procesar y descargar. Primero se debe subir el en formato csv separado por (,),

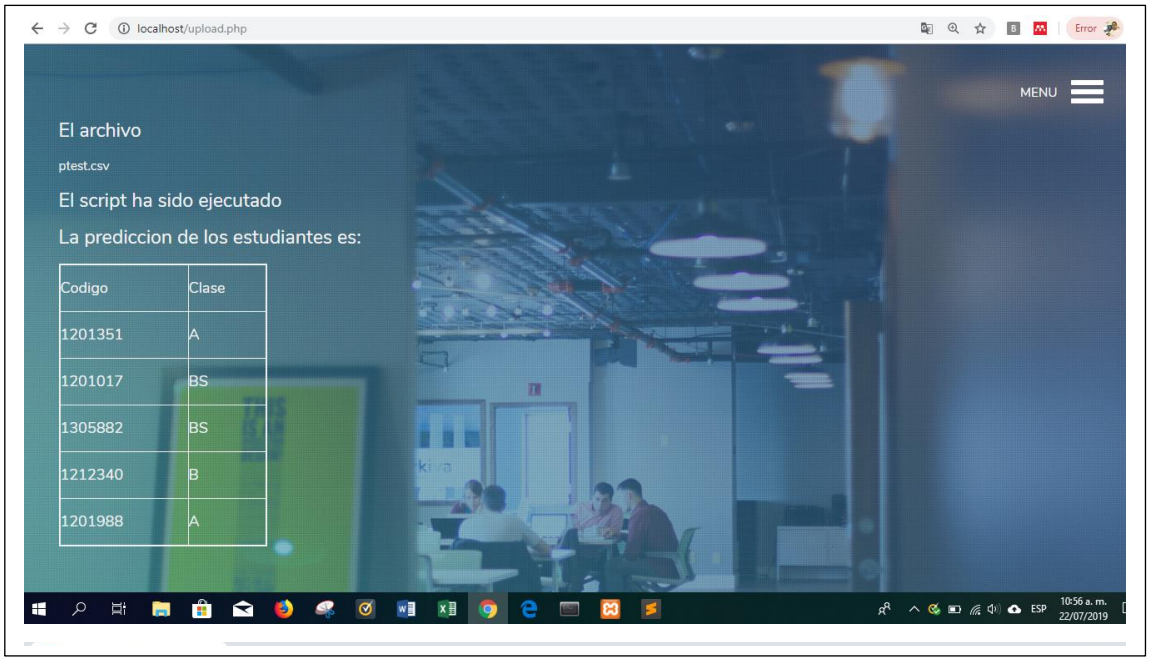




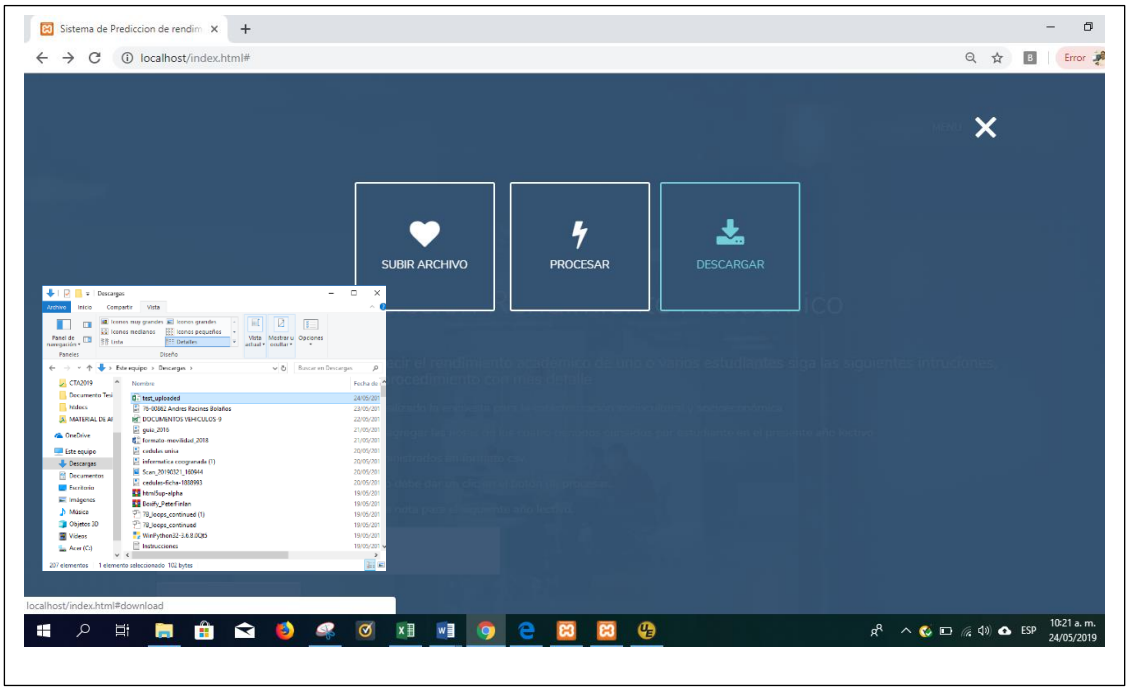
Luego de subir el archivo se debe dar clic en la opción de procesar para obtener la predicción.



A continuación, se muestra el resultado de predicción como se muestra en la siguiente imagen.



Adicionalmente la predicción se puede descargar como se muestra en la imagen.



El archivo con la predicción es generado como test_upload.csv en el cual se muestra el número de estudiante y la predicción generada como se muestra en a continuación.