MODELO PREDICTIVO PARA DETERMINAR EL FRACASO DE MATEMÁTICAS EN GRADO 11 USANDO MACHINE LEARNING

OMAR ALVARADO CASTILLO

SANTOS MIGUEL ZAMBRANO SAAVEDRA

UNIVERSIDAD DISTRITAL FRANCISCO JOSÉ DE CALDAS

FACULTAD TECNOLÓGICA

INGENIERIA EN TELEMATICA

BOGOTÁ

2020

MODELO PREDICTIVO PARA DETERMINAR EL FRACASO DE MATEMÁTICAS EN GRADO 11 USANDO MACHINE LEARNING

OMAR ALVARADO CASTILLO

20172678007

SANTOS MIGUEL ZAMBRANO SAAVEDRA

20171678029

Proyecto de grado presentado como requisito para optar por el título de

Ingeniero Telemático

PROYECTO DE DESARROLLO DE INGENIERÍA

TUTOR: JAIRO HERNÁNDEZ GUTIÉRREZ

Ingeniero de Sistemas

UNIVERSIDAD DISTRITAL FRANCISCO JOSÉ DE CALDAS

FACULTAD TECNOLÓGICA

INGENIERIA EN TELEMATICA

BOGOTÁ

2020

Nota de Aceptación

_____

_____

_____

_____

_____

_____

_____

Firma del Presidente del Jurado

_____

Firma del Jurado

Bogotá D.C., junio de 2020

AGRADECIMIENTO

Queremos agradecer a Dios por habernos bendecido con nuestras familias que nos han brindado su apoyo incondicional y acompañamiento en esta importante etapa de mi vida.

Agradecemos a nuestro tutor el Ingeniero Jairo Hernández Gutiérrez, por la colaboración prestada en el desarrollo de este proyecto.

Agradecemos al colegio Almirante Padilla de Usme, por facilitar el acceso a las notas de los estudiantes con el fin de desarrollar el respectivo análisis.

Dedicado a: Dios, a nuestras familias, compañeros y amigos por su paciencia, comprensión y ayuda; gracias a ellos se ha logrado cumplir un objetivo más en nuestras vidas.

TABLA DE CONTENIDO

LISTA DE FIGURAS

LISTA DE ECUACIONES

LISTA DE TABLAS

RESUMEN

El modelo predictivo para determinar el fracaso de matemáticas en grado once usando machine learning es el resultado de un análisis de datos en el Colegio Público Almirante Padilla de Usme

Este modelo se compone de un proceso de integración de datos, descripción de variables, limpieza de datos, eliminación de variables, análisis de correlaciones, transformación de datos y balanceo de datos. Estos módulos fueron implementados bajo el lenguaje de programación Python para construir varias aplicaciones inteligentes que utilizan el aprendizaje automático, este es un lenguaje de secuencias de comandos simple que facilita la interacción con los datos; además, tiene una amplia gama de paquetes que facilitan el inicio y la creación de aplicaciones, desde los más simples hasta los más complejos. Python se usa ampliamente en la industria y se está convirtiendo en el lenguaje de facto para la ciencia de datos en la industria., para el análisis de datos con machine learning. En relación a los datos de los estudiantes, estos fueron trabajados en formato .csv.

Para el proceso de análisis se optó por la metodología CRISP-DM (Cross Industry Standard Process for Data Mining) cubre las fases de un proyecto, sus tareas respectivas, y las relaciones entre estas tareas contempla el proceso de análisis de datos como un proyecto profesional, estableciendo así un contexto mucho más rico que influye en la elaboración de los modelos.

En primer capítulo se abordan los temas relacionados a la problemática de la deserción estudiantil, buscando así un mejor entendimiento del modelo del negocio, en el segundo capítulo se realiza el análisis y entendimiento de los datos, en el tercer capítulo se realiza la preparación de los datos, en el cuarto capítulo se realiza el modelado aplicando las técnicas de regresión y clasificación, en el quinto capítulo se revisan los resultados obtenidos con el fin de determinar cuál es mejor modelo y finalmente en sexto capítulo se evaluaran los resultados obtenidos.

ABSTRACT

The predictive model to determine the failure of mathematics in grade eleven using machine learning is the result of a data analysis at the Almirante Padilla Public School in Usme

This model is made up of a process of data integration, description of variables, cleaning of data, elimination of variables, analysis of correlations, transformation of data and balancing of data. These modules were implemented under the Python programming language to build various intelligent applications that use machine learning, this is a simple scripting language that facilitates interaction with data; In addition, it has a wide range of packages that make it easy to start and create applications, from the simplest to the most complex. Python is widely used in the industry and is becoming the de facto language for data science in the industry, for data analysis with machine learning. In relation to the student data, these will be worked in .csv format.

For the analysis process, the CRISP-DM (Cross Industry Standard Process for Data Mining) methodology was used to cover the phases of a project, their respective tasks, and the relationship between these tasks considers the data analysis process as a professional project. , thus establishing a much richer context that influences the elaboration of the models.

In the first chapter the topics related to the problem of student desertion are addressed, thus seeking a better understanding of the business model, in the second chapter the analysis and understanding of the data is carried out, in the third chapter the preparation of the data, in the fourth chapter the modeling is performed applying the regression and classification techniques, in the fifth chapter the results obtained are reviewed in order to determine which is the best model and finally in the sixth chapter the results obtained are evaluated.

INTRODUCCIÓN

El presente documento fue elaborado con el propósito de reconocer la importancia de crear un modelo de análisis de datos con machine learning con su respectiva documentación teniendo en cuenta las fases de desarrollo bajo la metodología CRISP-DM en el cual se buscó crear un modelo capaz de predecir los estudiantes con altas probabilidades de perder la asignatura de matemáticas

El aprendizaje automático o aprendizaje de máquinas (machine learning) han logrado generar grandes aportes a diferentes áreas tales como la estadística, la probabilidad, investigaciones profundas de datos y recuperación de información para reconocer patrones dentro de procesos matemáticos, de ingeniería y física (entre otros) necesarios para su buen desempeño.

Machine learning se basa en la recolección y análisis sistemático de datos para desarrollar un modelo predictivo, se define como una subdisciplina de la Inteligencia Artificial que tiene como principal objetivo descubrir comportamientos y establecer patrones o correlaciones para un fenómeno determinado.

A lo largo de este proyecto y de la realización del documento se vio la evolución de un modelo predictivo llevado a cabo mediante el lenguaje de programación Python, el conjunto de datos en extensión .csv y guiado por la metodología Cross Industry Standard Process for Data Mining (CRISP-DM), que a través del adelanto de cada uno de sus flujos de trabajo y de las fases que componen, permite de forma concreta y segura, garantizar el análisis; adicionalmente, se relacionan los tipos de aprendizaje supervisado de los cuales se usaron regresión y clasificación buscando identificar el más adecuado para la identificación de estudiantes con altas probabilidades de fracaso en matemáticas del grado once, esto con el fin de que se les brinden los soportes, refuerzos y recursos necesarios para mejorar.

1. FASE DE DEFINICIÓN, PLANEACIÓN Y ORGANIZACIÓN

1.1 TÍTULO DEL TRABAJO

MODELO PREDICTIVO PARA DETERMINAR EL FRACASO DE MATEMÁTICAS EN GRADO 11 USANDO MACHINE LEARNING

1.2 TEMA

En el desarrollo del proyecto son abordados los temas de metodología de desarrollo de proyectos (CRISP-DM), machine learning en Python, deserción estudiantil y metodologías implementadas que buscan mejorar la educación utilizando mecanismos de inteligencia artificial.

1.3 PLANTEAMIENTO DEL PROBLEMA

1.3.1 Descripción del Problema

Actualmente, la necesidad de mejorar el sistema educativo escolar, es cada vez más alto; esto conlleva a usar análisis de datos para facilitar la toma de decisiones correctamente en base a la información de los estudiantes, dado que estos ofrecen la facilidad de identificar puntos de mejora. Circunstancia que ha llevado al análisis de las causas o factores que influyen en el fracaso escolar en la materia de matemáticas del grado 11, ya que las matemáticas no son fáciles de aprender, su aprendizaje requiere la creación de significados abstractos, la codificación y descodificación de símbolos y la capacidad de hacer relaciones en el plano de lo posible.

Las dificultades en el aprendizaje de las matemáticas, no son debidas a una única causa, o un único tipo de dificultad. Existen diferentes factores que pueden dar lugar a diferentes dificultades en el aprendizaje de las matemáticas, tales como situaciones sociales que se estén presentado en su entorno, la pobreza, mala salud, nutrición del estudiante, distancia para llegar al centro educativo o exigencia

de útiles escolares entre otros, los cuales son factores que inciden en el fracaso de los estudiantes.

La deserción universitaria es una problemática en todas la Instituciones de Educación Superior (IES) a nivel nacional estudiado desde la década de los 60. Muchas de ellas han estudiado los factores que influyen en la decisión de abandonar los estudios por parte de una persona. En particular las matemáticas y su paso de la educación media a la educación superior marcan un fuerte cambio en los modelos de abstracción en los ejes temáticos, donde se evidencian los vacíos dejados en la educación media.

Aunque las universidades están haciéndose cargo de deficiencias por parte de los estudiantes cuando provienen del colegio, en áreas como lenguaje y matemáticas, algunos alumnos llegan al punto en el que se sobrecargan de trabajo, por lo que dejan sus estudios.

En Colombia existen estudiantes de colegios propensos a perder matemáticas o a finalizar su bachillerato sin las competencias adecuadas, por lo cual se evidencia la necesidad de contar con herramientas que permitan detectar oportunamente estos estudiantes, por medio, por ejemplo, de la identificación o reconocimiento de jóvenes con alto riesgo de fracaso.

1.3.2 Formulación del Problema

¿De qué manera es posible realizar un análisis predictivo que permita identificar a los estudiantes que van a tener un bajo rendimiento en matemáticas en el grado 11?

1.3.3 Justificación del Problema

En la gestión estudiantil, disponer de información actualizada y de calidad puede aportar grandes ventajas a la hora de tomar decisiones y detectar puntos de mejora y oportunidades de ahorro. Para ello, las notas pueden ser una fuente de información muy importante y valiosa, pero más allá de su creación y supervisión se abre otro estudio muy útil para la compañía: el análisis de los datos.

Las notas puede ser una fuente de información muy importante y valiosa para la gestión empresarial y la toma de decisiones, siempre que se sepa aprovechar bien. Los indicadores clave de rendimiento, tienen que aportar información fundamental en el proceso de toma de decisiones, así que la entidad educativa tiene que empezar por identificar los indicadores clave que le interesa analizar.

Utilizando la tecnología, la capacidad intelectual y la riqueza de la información se pueden descubrir patrones ocultos dentro de enormes paquetes de datos al igual que como se consiguen patrones en los diferentes trazos de información.

En la industria tecnológica se escucha cada vez más el término Machine Learning, y son muchas las empresas interesadas en saber utilizar esta técnica para obtener información.
El Machine Learning o Aprendizaje Automático es una rama dentro del campo de la Inteligencia Artificial cuyo objetivo es dotar a los ordenadores de la capacidad de aprender sin necesidad de ser programados. Entre los problemas evidenciados en la deserción universitaria se encuentran fallas en la formación de los estudiantes en la interdisciplinariedad de la matemática, tanto en carreras afines como no afines con otras áreas del conocimiento, el uso planeación basada en libros definidos limitan la capacidad de investigación y de inventiva no solo del estudiante, también del docente, ocasionando que los temas tratados en un periodo pocas veces tengan uso o influencia en el transcurso del año o incluso de los próximos ciclos de formación, esto es, algo que se estudia hoy tendrá poco uso en unos meses o años y pronto sea olvidado.

Es por esto, que se hace visible la necesidad de un modelo predictivo, desde la perspectiva de la ciencia de datos, para brindar apoyo en el fracaso escolar, brindar una mejor calidad en la educación, además de facilitar el aprendizaje de los estudiantes, debido a que un modelo utilizando machine learning puede brindar información de los estudiantes y en base a esta se pueden tomar planes de acción.

1.4 OBJETIVOS

1.4.1 Objetivo General

Realizar un modelo predictivo para determinar el fracaso de matemáticas en grado 11 usando machine learning

1.4.2 Objetivos Específicos

- Realizar el levantamiento, depuración y limpieza de información de los diferentes factores del bajo desempeño de los estudiantes en matemáticas del grado 11.

- Realizar el entrenamiento y evaluación del modelo de la data obtenida de la limpieza y depuración.

- Diseñar el modelo de los estudiantes que comienzan el grado 11, utilizando machine learning

- Evaluar el modelo con los datos obtenidos de un colegio de la localidad de Usme.

- De las técnicas supervisadas de minería de datos determinar cuál es más indicada para la solución Clasificación o Pronóstico/Predicción.

1.5 ALCANCES Y LIMITACIONES

1.5.1 Alcances

Se busca realizar el diseño de un modelo predictivo, que brinde la predicción de los estudiantes de matemáticas del grado 11 que tenga alta probabilidad de fracasar en matemáticas.

Este modelo permitirá a quien lo implemente la facilidad de identificar a qué estudiantes se le debe brindar una mejor atención o refuerzo en matemáticas para que este no pierda la asignatura de matemáticas.

1.5.2 Limitaciones

- **Técnica**. En el desarrollo de este modelo se utilizará herramientas de software libre como el lenguaje de programación Python y los datos serán tratados en formato de texto, que permitan el buen diseño y desempeño del proyecto.

  Lenguaje de programación Python: permite programación multiparadigma, ya que soporta orientación a objetos, programación imperativa y, en menor medida, programación funcional. Es un lenguaje interpretado, dinámico y multiplataforma.

  Documentos de Texto: son un tipo de documento en formato abierto sencillo para representar datos en forma de tabla, en las que las columnas se separan por comas

Tabla 1 Herramientas de Desarrollo

SOFTWARE	
APLICACIÓN	**DESCRIPCIÓN**
Windows 10	Sistema operativo
Ubuntu	Capa de compatibilidad desarrollada por Microsoft (WSL)
PyCharm	Entorno de desarrollo integrado utilizado en la programación de computadoras, específicamente para el lenguaje Python

- **Temática**. Para el diseño y desarrollo de este proyecto se abordarán temas relacionados con:
    - CRISP-DM (Cross Industry Standard Process for Data Mining): Utilizado para poner orden en los proyectos de Data Science.
    - Diseño de modelo de análisis de datos: Corresponde a la secuencia de pasos a seguir para el desarrollo del análisis.
    - Programación Python: es el lenguaje que se utilizara para realizar el análisis.
    - Deserción Estudiantil: Es una situación que se presenta en las instituciones educativas que se busca controlar.

- **Geográfica**. El desarrollo del modelo predictivo de para determinar el fracaso de matemáticas en grado 11 usando machine learning se llevará a cabo en la Universidad Distrital Francisco José de caldas.

- **Temporal**. El tiempo estimado para el desarrollo de sistema es de 6 meses a partir de la aprobación del proyecto.

1.6 ESTADO DEL ARTE

En la actualidad el avance de la tecnología en relación a la informática como una ciencia busca ayudar en las actividades del ser humano que requieran el procesamiento de datos, continuamente se buscan mejores formas en las que pueda ser utilizada, una de ellas está relacionada con la aplicación de una de las ramas de la inteligencia artificial el machine learning, que tiene como objetivo primordial abordar y resolver problemas prácticos mediante el desarrollo de técnicas que permiten que los sistemas o máquinas aprendan. [1]

En la búsqueda de esta forma de inteligencia se ha llegado a elaborar diferentes tipos de algoritmos de aprendizaje automático, especialmente enfocados en hallar información valiosa en grandes volúmenes de datos y/o predecir con altas probabilidades eventos que aún no ocurren, por tal motivo hace que sean empleados en los bancos permitiendo identificar clientes con perfiles de alto riesgo, en el gobierno para identificar patrones de conducta en los habitantes, en la salud para proporcionar diagnósticos mejorados, en marketing y ventas para personalizar la experiencia de comprar, en el transporte para que la regular la movilidad y es aquí donde radican los fundamentos teóricos que se deben tener en cuenta a la hora de plantear la solución en la educación.

Existen varios algoritmos que se pueden emplear para resolver el problema planteado, estos se clasifican como de aprendizaje supervisado y no supervisado, pero al compararlos se llega a la conclusión de que los algoritmos supervisados son más idóneos para la soluciona ya que se suelen usar en problemas de clasificación y de regresión.[2]

La educación es un tema de vital importancia para todo habitante de la ciudad ya que implica el desarrollo de capacidades intelectuales, morales y afectivas de las personas que conformarán la sociedad, el sistema de educación está regido por varios factores entre ellos está el estado de las instalaciones donde se imparten las clases, el número de estudiantes por curso y la supervisión adecuada de profesores.[3]

---

[1] Ian H. Witten and Eibe Frank (2011). Data Mining: Practical machine learning tools and techniques Morgan Kaufmann, 664 pág.

[2] Lise Getoor y Taskar Ben: Introducción a estadística de relación de aprendizaje, MIT Press, 2007.

[3] HERNANDEZ, Juan Miguel. Poca calidad y mucha deserción: ¿crisis en la educación media? [en línea]. (4 de agosto de 2019). [Consultado: 7 de mayo de 2020]. Disponible en: https://www.elespectador.com/arti-503

Los estudios realizados sobre la educación en Colombia muestran un gran aumento de la deserción estudiantil llegando a establecer cifras preocupantes donde uno de cada cinco estudiantes no continúa estudiando después de la primaria, el 12% queda por fuera en la básica secundaria y de cada 100 estudiantes que ingresan a la educación media, solo 44 se gradúan en las zonas rurales del país, el embarazo a temprana edad, el bullying, el desplazamiento forzado, la falta de motivación al proyectar su futuro y la baja calidad de la educación son algunas de las causas que incentivan a los estudiantes a abandonar su educación escolar y a verse en la necesidad de salir a trabajar o incluso pedir dinero para lograr el sostenimiento propio y el de sus familias,[4] adicionalmente se presenta que al graduarse no están preparados para enfrentar las demandas del mercado laboral, todo esto representa uno de los desafíos más grandes que enfrenta este país [5]

Para el prevenir la deserción escolar se han implementado diversas estrategias que buscan predecir la deserción y/o mejorar el rendimiento del estudiante utilizando inteligencia artificial, esto teniendo en cuenta variables de personalidad identificando las razones por las cuales el rendimiento académico (promedio) y la reprobación escolar pueden afectarse, desde el punto de vista psicológico observando que inicialmente aquellos alumnos con tendencias narcisistas mostraron peores evaluaciones a lo largo del tiempo, mientras aquellos con niveles promedio de autoestima mostraron una mejoría en sus evaluaciones.[6]

Otra de las estrategias empleadas es el análisis de los comportamientos disruptivos en el aula ya que según la literatura especializada, el rendimiento académico depende de factores como los estilos intelectuales, en este sentido se señala que una clase disciplinada es uno de los indicadores más importantes de la enseñanza exitosa, donde se determina que existen dos tipos bien diferenciados

---

[4] Prevenir la deserción escolar, una tarea de todos [en línea]. [Consultado: 7 de mayo de 2020]. Disponible en: http://technocio.com/prevenir-la-desercion-escolar-una-tarea-de-todos/

[5] La fórmula para combatir la deserción escolar en Colombia [en línea]. [Consultado: 7 de mayo de 2020]. Disponible en: https://www.eltiempo.com/vida/educacion/como-disminuir-la-desercion-escolar-en-colombia-459204

[6] RANGEL DE LA GARZA, Hugo Jaime. La predicción del rendimiento académico y la reprobación escolar a partir de variables de personalidad. Trabajo de investigación Doctor en Psicología. México DF. Universidad Iberoamericana. Facultad de Psicología, 2013. 7 p.

de conductas de indisciplina en el aula, denominadas instruccional y convencional demostrando así tener una incidencia distinta en los resultados de aprendizaje.[7]

El uso de métodos de inteligencia de negocios enfocados en el ámbito educativo también es de utilidad ya que se ha demostrado que el data warehouse como repositorio de datos permite la ejecución de consultas complejas y análisis estadísticos detallados permitiendo realizar seguimiento a los datos históricos de los estudiantes; inclusive a través de diversos lugares geográficos y realizar cruces de información entre diferentes bases de datos.[8]

La implementación de los algoritmos de aprendizaje automático en ciertas instituciones educación de una ciudad se ve reflejado en la disminución en la deserción estudiantil, su validación se puede realizar mediante la realización de un modelo predictivo para evitar la deserción escolar.

Muchas de las tecnologías empleadas en la actualidad para controlar la deserción y bajo rendimiento académico en las ciudades usan técnicas de predicción que permiten extraer conocimiento no implícito de los datos, la minería de Datos (Data Mining, DM) este es un campo multidisciplinario que resulta beneficioso al momento de descubrir patrones o diseñar modelos de predicción buscando así estudiar la situación actual de los estudiantes a fin de descubrir posibles patrones de deserción, no solo con motivos de evaluación sino principalmente para atacar y disminuir los índice de abandono escolar[9]

La alternativa de solución propuesta para la ciudad de sangolqui ubicada en el centro geográfico del estado de Guayas, Ecuador, por ejemplo, consta del uso de técnicas de clasificación usando árboles de decisión estos se enfocan en observaciones de un conjunto de variables que son estadísticamente analizadas para generar una base de predicción sobre un atributo cualitativo predefinido por valores previos; mediante su uso, se dio como resultado la identificación con éxito

---

[7] Gotzens Busquets, C., Cladellas Pros, R., Clariana Muntada, M., & Badia Martín, M. (2015). Indisciplina instruccional y convencional: su predicción en el rendimiento académico. Revista Colombiana de Psicología.

[8] Kimball, Ralph (2004). «1». The Data Warehouse ETL Toolkit (en inglés). Wiley. p. 23.

[9] CUJI CHACHA, Blanca Roció. Las Técnicas de Predicción y su Incidencia en la Detección de Patrones de Deserción Estudiantil en la Carrera de Docencia en Informática de la Facultad de Ciencias Humanas y de la Educación de la Universidad Técnica de Ambato. Trabajo de investigación. Magister en Gestión de Bases de Datos. México DF. Universidad Técnica de Ambato. Facultad de Ingeniería en Sistemas, Electrónica e Industrial 2016. 3 p.

de atributos relevantes de datos socio-demográficos, académicos e institucionales de estudiantes que permitieron reducir la deserción.[10]

Otro caso de solución propone el uso de árboles de decisión con parámetros optimizados, esta técnica no paramétrica clasifica una población en un modelo de segmentos de tipo de ramas que construyen un árbol invertido, y luego este modelo se utiliza para predecir una variable objetivo en este caso la deserción, su investigación concluyó en un árbol de decisión que clasifica correctamente un 81.3% de los casos. [11]

La diferencia principal entre la minería de datos y el aprendizaje automático es que, sin la participación humana la minería de datos, no puede funcionar, pero en el machine learning el esfuerzo humano está involucrado sólo en el momento en que se define el algoritmo.

1.7 JUSTIFICACIÓN

La incorporación de análisis avanzados como el modelado predictivo ayudará a los colegios a identificar estudiantes y cómo debe ser la comunicación con ellos, el tipo de educación, el momento indicado y las características de estudio que podrían interesarle.

El machine learning está ayudando a mejorar la calidad y la retención de los estudiantes. Además, mediante este proceso los colegios pueden analizar los comportamientos de los estudiantes tales como: transacciones entre colegios, actividades de estudio, prácticas en redes sociales y todo tipo de actividades medibles para identificar áreas de oportunidad, lo cual permitirá la optimización de estrategias y la personalización de la experiencia de estudio.

El análisis de datos y la creación de modelos predictivos se han convertido no sólo en una tendencia para las industrias en telecomunicaciones sino en una necesidad permanente para poder ofrecer mejora continua.

---

[10] GALLARDO CORRALES, Diego Eduardo. Análisis de Patrones de Deserción Estudiantil de la Unidad Educativa Lenin School Aplicando Minería de Datos. Magister en Gestión de Sistemas de Información e inteligencia de Negocios. Sangolqui. Departamento de Ciencias de la Computación. Facultad de Posgrados. 2005. 81 p.

[11] RAMIREZ, Patricio E.; GRANDON, Elizabeth. Predicción de la Deserción Académica en una Universidad Publica chilena a traces de la clasificación basada en Arboles de Decisión con Parámetros Optimizados. En: Revista Formación Universitaria. Santiago de Chile: Universidad Católica del Norte, diciembre-enero, 2018, nro. 16.

Es por esto que se hace visible la necesidad de implementar un modelo predictivo, desde la perspectiva del análisis, para brindar una adecuada identificación de estudiantes de grado con altas probabilidad de perder matemáticas del grado once, ya que se ha evidenciado que gran cantidad de estudiantes tiene falencia en esta materia en específico.

1.8 MARCO DE REFERENCIA

1.8.1 Marco Histórico

La analítica predictiva ha estado presente por varias décadas, es una tecnología cuyo momento ha llegado. Cada vez más organizaciones recurren a la analítica predictiva para mejorar su base de operación y lograr una ventaja competitiva en la actualidad se trabaja mucho debido a que se gestionan volúmenes y tipos de datos cada vez mayores, además de un mayor interés en el uso de datos para producir insights valiosos, se cuenta con computadoras más rápidas y económicas, el software es más fácil de usar y las condiciones económicas más difíciles junto con la necesidad de tener una diferenciación competitiva hacen que el software interactivo se vuelve más predominante, la analítica predictiva ya no es sólo del dominio de matemáticos y estadísticos. Analistas de negocios y expertos en línea de negocios utilizan también estas tecnologías.[12]

Los modelos predictivos son modelos de la relación entre el rendimiento específico de un sujeto en una muestra y uno o más atributos o características del mismo sujeto. El objetivo del modelo es evaluar la probabilidad de que un sujeto similar tenga el mismo rendimiento en una muestra diferente. Esta categoría engloba modelos en muchas áreas como el marketing, para este proyecto en particular se tendrá como énfasis la educación donde se buscan patrones de datos ocultos que respondan preguntas sobre el comportamiento del estudiante. Los modelos predictivos a menudo ejecutan cálculos durante las transacciones en curso, por ejemplo, para evaluar el riesgo o la oportunidad de un cliente o transacción en particular, de forma que aporte conocimiento a la hora de tomar una decisión. Gracias a los avances de ingeniería en el análisis de grandes volúmenes de datos estos modelos son capaces de simular el comportamiento humano frente a estímulos o situaciones específicas.

[12] Finlay, Steven (2014). Predictive Analytics, Data Mining and Big Data. Myths, Misconceptions and Methods (1st edición). Basingstoke: Palgrave Macmillan. p. 237.

A través del tiempo, el aprendizaje automático o machine learning ha demostrado ser de gran utilidad desde los años 60's con el algoritmo de "nearest neighbor" el cual permitió a las computadoras utilizar un reconocimiento de patrones básico, para los años 80's se instauró en la informática y la estadística que dio lugar a enfoques probabilísticos en la IA, esto generó un gran cambio en machine learning ya que se empezó a trabajar con más datos, pero fue para los años 90's donde se dio explosión en su uso y se acuñó el término "Deep Learning", con el que se explican nuevas arquitecturas y distinguir objetos y texto en imágenes y videos.[13]

1.8.2 Marco Teórico

Minería de Datos

Es un campo de la estadística y las ciencias de la computación referido al proceso que intenta descubrir patrones en grandes volúmenes de conjuntos de datos. Utiliza los métodos de la inteligencia artificial, aprendizaje automático, estadística y sistemas de bases de datos. La tarea de minería de datos real es el análisis automático o semi-automático de grandes cantidades de datos para extraer patrones interesantes hasta ahora desconocidos, como los grupos de registros de datos (análisis clúster), registros poco usuales (la detección de anomalías) y dependencias (minería por reglas de asociación).

El término es un concepto de moda, y es frecuentemente mal utilizado para referirse a cualquier forma de datos a gran escala o procesamiento de la información (recolección, extracción, almacenamiento, análisis y estadísticas), pero también se ha generalizado a cualquier tipo de sistema informático de apoyo a decisiones, incluyendo la inteligencia artificial, aprendizaje automático y la inteligencia empresarial. En el uso de la palabra, el término clave es el descubrimiento, comúnmente se define como "la detección de algo nuevo". Incluso el popular libro "La minería de datos: sistema de prácticas herramientas de aprendizaje y técnicas con Java" (que cubre todo el material de aprendizaje automático) originalmente iba a ser llamado simplemente "la máquina de aprendizaje práctico", y el término "minería de datos" se añadió por razones de marketing. A menudo, los términos más generales "(gran escala) el análisis de datos", o "análisis" -. o cuando se refiere a los métodos actuales, la inteligencia artificial y aprendizaje automático, son más apropiados.[14]

---

[13] Ian H. Witten and Eibe Frank (2011). Data Mining: Practical machine learning tools and techniques Morgan Kaufmann, 664 pág., ISBN 978-0-12-374856-0.

[14] Xingquan Zhu, Ian Davidson (2007). Knowledge Discovery and Data Mining: Challenges and Realities. (Descubrimiento del conocimiento y minería de datos: desafíos y realidades), Pág. 11

La tarea de minería de datos real es el análisis automático o semi-automático de grandes cantidades de datos para extraer patrones interesantes hasta ahora desconocidos, como los grupos de registros de datos (análisis clúster), registros poco usuales (la detección de anomalías) y dependencias (minería por reglas de asociación). Esto generalmente implica el uso de técnicas de bases de datos como los índices espaciales. Estos patrones pueden entonces ser vistos como una especie de resumen de los datos de entrada, y pueden ser utilizados en el análisis adicional o, por ejemplo, en el aprendizaje automático y análisis predictivo. Por ejemplo, el paso de minería de datos podría identificar varios grupos en los datos, que luego pueden ser utilizados para obtener resultados más precisos de predicción por un sistema de soporte de decisiones. Ni la recolección de datos, preparación de datos, ni la interpretación de los resultados y la información son parte de la etapa de minería de datos, pero que pertenecen a todo el proceso KDD como pasos adicionales.

El proceso de determinar por los siguientes pasos:

Selección del conjunto de datos: tanto en lo que se refiere a las variables objetivo (aquellas que se quiere predecir, calcular o inferir), como a las variables independientes (las que sirven para hacer el cálculo o proceso), como posiblemente al muestreo de los registros disponibles.

Análisis de las propiedades de los datos: en especial los histogramas, diagramas de dispersión, presencia de valores atípicos y ausencia de datos (valores nulos).

Transformación del conjunto de datos de entrada: se realizará de diversas formas en función del análisis previo, con el objetivo de prepararlo para aplicar la técnica de minería de datos que mejor se adapte a los datos y al problema, a este paso también se le conoce como pre procesamiento de los datos.

Selección y aplicación de la técnica de minería de datos: se construye el modelo predictivo, de clasificación o segmentación.

Extracción de conocimiento: mediante una técnica de minería de datos, se obtiene un modelo de conocimiento, que representa patrones de comportamiento observados en los valores de las variables del problema o relaciones de asociación entre dichas variables. También pueden usarse varias técnicas a la vez para generar distintos modelos, aunque generalmente cada técnica obliga a un pre procesado diferente de los datos.

Interpretación y evaluación de datos: una vez obtenido el modelo, se debe proceder a su validación comprobando que las conclusiones que arroja son válidas y suficientemente satisfactorias. En el caso de haber obtenido varios modelos mediante el uso de distintas técnicas, se deben comparar los modelos en busca de aquel que se ajuste mejor al problema. Si ninguno de los modelos

alcanza los resultados esperados, debe alterarse alguno de los pasos anteriores para generar nuevos modelos. [15]

Las técnicas de la minería de datos provienen de la inteligencia artificial y de la estadística, dichas técnicas, no son más que algoritmos, más o menos sofisticados que se aplican sobre un conjunto de datos para obtener unos resultados. Las técnicas más representativas son:

Redes neuronales: Son un paradigma de aprendizaje y procesamiento automático inspirado en la forma en que funciona el sistema nervioso de los animales. Se trata de un sistema de interconexión de neuronas en una red que colabora para producir un estímulo de salida.

Perceptrón: El modelo biológico más simple de un perceptrón es una neurona y viceversa. Es decir, el modelo matemático más simple de una neurona es un perceptrón.

Perceptrón multicapa: es una red neuronal artificial (RNA) formada por múltiples capas, de tal manera que tiene capacidad para resolver problemas que no son linealmente separables, lo cual es la principal limitación del perceptrón (también llamado perceptrón simple).

Mapa auto organizado: Es un tipo de red neuronal artificial (ANN por sus siglas en inglés), que es entrenada usando aprendizaje no supervisado para producir una representación discreta del espacio de las muestras de entrada, llamado mapa.

Regresión lineal: Es la más utilizada para formar relaciones entre datos. Rápida y eficaz pero insuficiente en espacios multidimensionales donde puedan relacionarse más de 2 variables.

Árboles de decisión: Un árbol de decisión es un modelo de predicción utilizado en el ámbito de la inteligencia artificial y el análisis predictivo, dada una base de datos se construyen estos diagramas de construcciones lógicas, muy similares a los sistemas de predicción basados en reglas, que sirven para representar y categorizar una serie de condiciones que suceden de forma sucesiva, para la resolución de un problema. Ejemplos:

Algoritmo ID3: es utilizado dentro del ámbito de la inteligencia artificial. Su uso se engloba en la búsqueda de hipótesis o reglas en él, dado un conjunto de ejemplos.

Algoritmo C4.5: es un algoritmo usado para generar un árbol de decisión desarrollado por Ross Quinlan.1 C4.5 es una extensión del algoritmo ID3 desarrollado anteriormente por Quinlan.

---

[15] Visual Data Mining: Allowing business users to mine and gain insight into the data. Springer, New York. NTC 1299. Pago. 10

Modelos estadísticos: Es una expresión simbólica en forma de igualdad o ecuación que se emplea en todos los diseños experimentales y en la regresión para indicar los diferentes factores que modifican la variable de respuesta

Agrupamiento o Clustering: Es un procedimiento de agrupación de una serie de vectores según criterios habitualmente de distancia; se tratará de disponer los vectores de entrada de forma que estén más cercanos aquellos que tengan características comunes. Ejemplos:

Algoritmo K-means: es un método de agrupamiento, que tiene como objetivo la partición de un conjunto de n observaciones en k grupos en el que cada observación pertenece al grupo cuyo valor medio es más cercano. Es un método utilizado en minería de datos.

Algoritmo K-medoids: es un algoritmo de agrupamiento (del inglés clustering) relacionado con los algoritmos k-means y medoidshift. Tanto el k-medoids como el k-means son algoritmos que trabajan con particiones (dividiendo el conjunto de datos en grupos) y ambos intentan minimizar la distancia entre puntos que se añadirían a un grupo y otro punto designado como el centro de ese grupo. En contraste con el algoritmo k-means, k-medoids escoge datapoints como centros y trabaja con una métrica arbitraria de distancias entre datapoints en vez de usar la norma L2 En 1987 se propuso este método para el trabajo con la norma L1 y otras distancias.

Reglas de asociación: Se utilizan para descubrir hechos que ocurren en común dentro de un determinado conjunto de datos.

Según el objetivo del análisis de los datos, los algoritmos utilizados se clasifican en supervisados y no supervisados (Weiss y Indurkhya, 1998):

Algoritmos supervisados (o predictivos): predicen un dato (o un conjunto de ellos) desconocido a priori, a partir de otros conocidos.

Algoritmos no supervisados (o del descubrimiento del conocimiento): se descubren patrones y tendencias en los datos.[16]

Machine Learning

Es el subcampo de las ciencias de la computación y una rama de la inteligencia artificial, cuyo objetivo es desarrollar técnicas que permitan que las computadoras aprendan. Se dice que un agente aprende cuando su desempeño mejora con la experiencia; es decir, cuando la habilidad no estaba presente en su genotipo o

[16] Sally Jo Cunningham, Geoffrey Holmes, The Driving Need for Analytics in a Big Data World. 2017. Pag.57.

rasgos de nacimiento. De forma más concreta, los investigadores del aprendizaje de máquinas buscan algoritmos y heurísticas para convertir muestras de datos en programas de computadora, sin tener que escribir los últimos explícitamente. Los modelos o programas resultantes deben ser capaces de generalizar comportamientos e inferencias para un conjunto más amplio (potencialmente infinito) de datos.

En muchas ocasiones el campo de actuación del aprendizaje automático se solapa con el de la estadística inferencial, ya que las dos disciplinas se basan en el análisis de datos. Sin embargo, el aprendizaje automático incorpora las preocupaciones de la complejidad computacional de los problemas. Muchos problemas son de clase NP-hard, por lo que gran parte de la investigación realizada en aprendizaje automático está enfocada al diseño de soluciones factibles a esos problemas. El aprendizaje automático también está estrechamente relacionado con el reconocimiento de patrones. El aprendizaje automático puede ser visto como un intento de automatizar algunas partes del método científico mediante métodos matemáticos. Por lo tanto, es un proceso de inducción del conocimiento.

El aprendizaje automático tiene una amplia gama de aplicaciones, incluyendo motores de búsqueda, diagnósticos médicos, detección de fraude en el uso de tarjetas de crédito, análisis del mercado de valores, clasificación de secuencias de ADN, reconocimiento del habla y del lenguaje escrito, juegos y robótica.

Algunos sistemas de aprendizaje automático intentan eliminar toda necesidad de intuición o conocimiento experto de los procesos de análisis de datos, mientras otros tratan de establecer un marco de colaboración entre el experto y la computadora.

Los modelos pueden también clasificarse como modelos de agrupamiento y modelos de gradiente. Los primeros tratan de dividir el espacio de instancias en grupos. Los segundos, como su nombre lo indica, representan un gradiente en el que se puede diferenciar entre cada instancia. Clasificadores geométricos como las máquinas de vectores de apoyo son modelos de gradientes.[17]

[17] Russell, Stuart; Norvig, Peter (2009), Inteligencia Artificial: Un Enfoque Moderno, 3ra edición. Redmond, Washington, p 24

1.8.3 Marco Conceptual

Modelo Predicativo: es un modelo de datos, basado en estadísticas inferenciales, que se utiliza para predecir la respuesta a una promoción de marketing o a una determinada inversión.

El modelo predictivo es creado habitualmente por los científicos de datos y utiliza estadísticas para predecir los resultados. La mayoría de las veces el evento que uno quiere predecir es en el futuro, pero el modelado predictivo puede aplicarse a cualquier tipo de evento desconocido, independientemente de cuándo ocurrió. En muchos casos el modelo se elige sobre la base de la teoría de la detección para tratar de adivinar la probabilidad de un resultado dada una cantidad establecida de datos de entrada, por ejemplo, dando un correo electrónico para determinar la probabilidad de que sea spam.[18]

Los modelos pueden utilizar uno o más clasificadores al intentar determinar la probabilidad de un conjunto de datos pertenecientes a otro conjunto. Cuando se despliega comercialmente, el modelado predictivo se refiere a menudo como análisis predictivo.

Python: Python es un lenguaje de scripting independiente de plataforma y orientado a objetos, preparado para realizar cualquier tipo de programa, desde aplicaciones Windows a servidores de red o incluso, páginas web. Es un lenguaje interpretado, lo que significa que no se necesita compilar el código fuente para poder ejecutarlo, lo que ofrece ventajas como la rapidez de desarrollo e inconvenientes como una menor velocidad. [19]

Ciencia de Datos: es un campo interdisciplinario que involucra métodos científicos, procesos y sistemas para extraer conocimiento o un mejor entendimiento de datos en sus diferentes formas, ya sea estructurados o no estructurados, lo cual es una continuación de algunos campos de análisis de datos como la estadística, la minería de datos, el aprendizaje automático, y la analítica predictiva.[20]

---

[18] https://www.arimetrics.com/glosario-digital/modelo-predictivo [Consulta: 7º. de junio de 2019].

[19] LUCA. ¿Qué es Python? (2020) Archivado desde el original el 24 de febrero de 2020. Consultado el 24 de febrero de 2020.

[20] Liu, Alex (17 de septiembre de 2015). «Data Science and Data Scientist» (en inglés). Consultado el 24 de septiembre de 2015.

También se define La ciencia de datos como "un concepto para unificar estadísticas, análisis de datos, aprendizaje automático, y sus métodos relacionados, a efectos de comprender y analizar los fenómenos reales", empleando técnicas y teorías extraídas de muchos campos dentro del contexto de las matemáticas, la estadística, la ciencia de la información, y la informática.[21]

1.8.4 Marco Metodológico

Metodología para poner orden en los proyectos de Data Science (CRISP-DM)

Las técnicas de Data Science o Data Analytics, que tanto interés despiertan hoy en día, en realidad surgieron en la década de los 90, cuando se usaba el término KDD (Knowledge Discovery in Databases) para referirse al (amplio) concepto de hallar conocimiento en los datos. En un intento de normalización de este proceso de descubrimiento de conocimiento, de forma similar a como se hace en ingeniería software para normalizar el proceso de desarrollo software, surgieron a finales de los 90 dos metodologías principales: CRISP-DM (Cross Industry Standard Process for Data Mining) y SEMMA (Sample, Explore, Modify, Model, and Assess). Ambas especifican las tareas a realizar en cada fase descrita por el proceso, asignando tareas concretas y definiendo lo que es deseable obtener tras cada fase. [22]

CRISP-DM (Cross Industry Standard Process for Data Mining) proporciona una descripción normalizada del ciclo de vida de un proyecto estándar de análisis de datos, de forma análoga a como se hace en la ingeniería del software con los modelos de ciclo de vida de desarrollo de software. El modelo CRISP-DM cubre las fases de un proyecto, sus tareas respectivas, y las relaciones entre estas tareas. En este nivel de descripción no es posible identificar todas las relaciones; las relaciones podrían existir entre cualquier tarea según los objetivos, el contexto, y el interés del usuario sobre los datos.

La metodología CRISP-DM contempla el proceso de análisis de datos como un proyecto profesional, estableciendo así un contexto mucho más rico que influye en

[21] Tukey, John W. (1962-03). «The Future of Data Analysis». The Annals of Mathematical Statistics (en inglés) 33 (1): 1-67. ISSN 0003-4851.
[22] SANDERS, William. CRISP-DM: La metodología para poner orden en los proyectos de Data Science, 2 ed. Portland, Oregon. MCGRAW-HILL Osborne Media. Pag. 54

la elaboración de los modelos. Este contexto tiene en cuenta la existencia de un cliente que no es parte del equipo de desarrollo, así como el hecho de que el proyecto no sólo no acaba una vez se halla el modelo idóneo (ya que después se requiere un despliegue y un mantenimiento), sino que está relacionado con otros proyectos, y es preciso documentarlo de forma exhaustiva para que otros equipos de desarrollo utilicen el conocimiento adquirido y trabajen a partir de él.

El ciclo de vida del proyecto de minería de datos consiste en seis fases mostradas en la figura 1 Metodología CRISP-DM. [23]

La secuencia de las fases no es rígida: se permite movimiento hacia adelante y hacia atrás entre diferentes fases. El resultado de cada fase determina qué fase, o qué tarea particular de una fase, hay que hacer después. Las flechas indican las dependencias más importantes y frecuentes.[24]

Figura 1 Metodología CRISP-DM

[23] Sutton, Richard S., Barto, Andrew G. Reinforcement Learning: An Introduction. The MIT Press. 2008. Pág. 83

[24] Shearer C., el modelo CRISP-DM: el nuevo plan para la minería de datos, almacenamiento de los datos J (2000); 5:13-22p

Fuente: Shearer, Carl. CRISP-DM 1.0. Fases del modelo de referencia CRISP-DM

El círculo externo en la figura simboliza la naturaleza cíclica de los proyectos de análisis de datos. El proyecto no se termina una vez que la solución se despliega. La información descubierta durante el proceso y la solución desplegada pueden producir nuevas iteraciones del modelo. Los procesos de análisis subsecuentes se beneficiarán de las experiencias previas.

1.9 SELECCIÓN DE LENGUAJE Y HERRAMIENTA

Las aplicaciones inteligentes usando Machine Learning son cada vez más usadas en los diferentes tipos de industria, algunos ejemplos:

Netflix: Su aplicación recolecta información de las visualizaciones que el cliente tiene para así clasificarlo y generar un seguimiento por único por cliente para dar recomendaciones de nuevas series y películas.

Amazon: La aplicación almacena los comportamientos de compra de sus clientes para generar ofertas dependiendo de lo que se haya buscado.

Creación de vacunas: Los científicos almacenan el comportamiento de los diferentes reactivos aplicado a un virus, para determinar el comportamiento de este y poder realizar simulaciones y encontrar la cura.

Aparte de las aplicaciones mencionadas hay un sin número de aplicaciones usando machine learning con diferentes herramientas que facilitan el uso de esta técnica.

Para la elección de la herramienta para la predicción de los estudiantes de 11 en matemáticas se hizo un paneo de las diferentes herramientas del mercado de las cuales se escogió Python debido a las ventajas que tiene sobre otros productos. A continuación, se describen otras herramientas:

1.9.1 Lenguaje de Programación R y R Studio

R es un lenguaje de programación con un enfoque al análisis estadístico, es uno de los lenguajes de programación más utilizados en investigación científica, siendo muy popular en los campos de aprendizaje automático (machine learning), minería de datos, investigación biométrica, bioinformación y matemáticas financieras.

Ya que R es un lenguaje de programación permite que los usuarios lo extiendan definiendo sus propias funciones, gran parte de las funciones de R están escritas en el mismo R, aunque para algoritmos computacionales exigentes es posible desarrollar biblioteca en C, C++ o Fortran que se cargan dinámicamente. Los usuarios más avanzados pueden manipular los objetos de R dinámicamente desde código desarrollado en R. Además, R puede integrarse con distintas bases de datos y existen bibliotecas que facilitan su utilización desde lenguajes de programación interpretados como Perl y Python.

R también puede usarse como herramienta de cálculo numérico, campo en el que puede ser tan eficaz como otras herramientas específicas como GNU Octave y su equivalente MATLAB, Se ha desarrollado una interfaz RWeka para interactuar con Weka que permite leer y escribir ficheros en el formato arff y enriquecer R con los algoritmos de minería de datos de dicha plataforma.

R forma parte de un proyecto colaborativo y abierto, Los usuarios pueden publicar paquetes que extienden su configuración básica. Dado el enorme número de nuevos paquetes, estos se han organizado en vistas (o temas), que permiten agruparlos según su naturaleza y función. Por ejemplo, hay un grupo de paquetes relacionados con estadística bayesiana, econometría, series temporales entre muchos otros.

Para facilitar el desarrollo de nuevos paquetes, se ha puesto a servicio de la comunidad una forja de desarrollo que facilita las tareas relativas a dicho proceso.

RStudio es un entorno de desarrollo integrado (IDE) para el lenguaje de programación R, dedicados a la computación estadística y gráficos. Incluye una consola, editor de sintaxis que apoya la ejecución de código, así como herramientas para el trazado, la diputación y la gestión del espacio de trabajo.

1.9.1.1 Ventajas

- Excelente gama de paquetes de código abierto y de alta calidad. R tiene un paquete para casi todas las aplicaciones cuantitativas y estadísticas imaginables. Esto incluye redes neuronales, regresión no lineal, filogenia, cartografía, mapas y muchos, muchos otros.
- La instalación básica viene con funciones y métodos estadísticos integrales muy complejos. R también maneja el álgebra y matrices particularmente bien.
- La visualización de datos es una fortaleza clave con el uso de bibliotecas como ggplot2.

1.9.1.2 Contras

- Rendimiento. R no es un lenguaje rápido. Esto no es accidente. R fue diseñado a propósito para facilitar el análisis de datos y las estadísticas. No fue diseñado para hacer la vida más fácil para la computadora.
- Especificidad de dominio. R es muy bueno para fines estadísticos y científicos de datos, pero no es tan bueno para programación de propósito general.

- Raro o extraño. R tiene algunas características poco frecuentes que pueden complicar un poco a los programadores con experiencia en otros idiomas.
- Por ejemplo, indexación desde 1, utilizando operadores de asignación múltiple, estructuras de datos no convencionales.

1.9.2 Lenguaje de Python

El lenguaje de programación Python permite construir diferentes tipos de aplicaciones inteligentes que usen Machine Learning, debido a que facilita la manipulación masiva de información y en el momento existen muchas librerías ya construidas que aceleran el proceso de implementar un algoritmo de machine learning además a nivel industrial es el lenguaje de programación líder que usan los científicos de datos.

1.9.2.1 Ventajas

- Conjunto de librerías para facilitar la implementación de los algoritmos de machine learning como:

  Matplotlib: Para la graficación de la información.

  Numpy: Para el cálculo matricial.

  Pandas: Para el cálculo matricial.

  TuriCreate: ya tiene implementado los algoritmos de machine learning como regresión y clasificación.

- Python y la mayoría de sus librerías tienen licencia GNU.
- El tiempo de procesamiento de la información es rápido.
- Puede trabajar con millones de registros.
- Buen salario para aquellos que sepan python

1.9.2.2 Contras

- No tiene una buena documentación.

1.9.3 Plataforma WEKA

Es una plataforma de software para el aprendizaje automático y la minería de datos escrito en java y desarrollado en la universidad waikato, weka es software libre distribuido bajo la licencia GNU-GPL, contiene una colección de herramientas de visualización y algoritmos para análisis de datos y modelos predictivos, unidos a una interfaz gráfica de usuario para acceder fácilmente a sus funcionalidades.

1.9.3.1 Ventajas

- Software libre y de sencilla interfaz gráfica.
- Numerosas técnicas de modelado y procesamiento de datos.
- Escrito en Java, lo que permite la compatibilidad con numerosas plataformas.

1.9.3.2 Contras

- Los resultados no dan la total confiabilidad de los mismos.
- Carece de algoritmos para el modelado de secuencias.

1.9.4 MATLAB

Es una plataforma de programación equipada con las más sofisticadas herramientas de análisis de datos, visualización y comparación de modelos que hay en el mercado lo que acelera el proceso para construir aplicaciones que usen

machine learning su principal uso es para análisis de señales, motores, creación de medicamentos entre otros la desventaja que tiene contra Python es que es licenciado y dependiendo de las herramientas que se vayan a usar así es el costo de la licencia.

1.9.4.1 Ventajas

- Conjunto de librerías para facilitar la implementación de los algoritmos de machine learning.
- El tiempo de procesamiento de la información y uso de recursos de la máquina en comparación con otras herramientas es muy superior.
- Puede trabajar con millones de registros.
- Buen salario para aquellos que sepan MATLAB

1.9.4.2 Contras

- Es licenciado y dependiendo de las funcionalidades requeridas es el valor de la licencia.

1.9.5 Azure Machine Learning

Es un servicio de analista predictivo en la nube que permite crear e implementar rápidamente tanto modelos predictivos como soluciones de análisis. Se trata de una herramienta que busca predecir intenciones de compra, realizar mantenimientos preventivos en los equipos e incluso predecir riesgos en transacciones económicas entren muchas otras predicciones.

Incorpora todas las herramientas necesarias para crear completas soluciones de análisis predictivo en la nube, desde una gran biblioteca de algoritmos a un estudio para la creación de modelos, pudiendo crear rápidamente modelos predictivas arrastrando, quitando y conectando los módulos que se necesiten.

1.9.5.1 Ventajas

- Cuenta con una interfaz fácil de entender e intuitiva que guía en el proceso de modelado de predicciones.

- Tiene una curva de aprendizaje muy corta un usuario con experiencia limitada en proyectos ML puede desarrollar soluciones en muy poco tiempo utilizando esta herramienta.

1.9.5.2 Contras

- Las últimas herramientas tienen la ventaja de una solución realmente abierta y flexible, que permite diseñar soluciones muy complejas, pero azure machine learning studio no cuenta con estas características y por lo tanto se tiende a usar para desarrollar proyectos con una dificultad media.

1.10 FACTIBILIDAD

1.10.1 Factibilidad Técnica

Para el desarrollo del proyecto se contó con dos equipos de cómputo con las siguientes características:

ESTACIÓN DE TRABAJO

- 5GB de espacio mínimo
- 4GB de memoria RAM
- Windows 10 Professional 64 bits
- PyCharm

1.10.2 Factibilidad Operativa

La utilización del modelo predictivo es viable, porque es desarrollado con software libre

Requisitos mínimos

- Procesador 500 MHz
- Memoria 256 Mb
- Disco duro 10 GB
- Sistema Operativo: Windows 98, 2000, NT, XP, VISTA, 7, 8,10.

El recurso humano es básico para el desarrollo del modelo predictivo que permite determinar el fracaso de matemáticas en estudiantes de grado 11 usando machine learning, para este fin se cuenta con:

El estudiante Omar Alvarado Castillo y Santos Miguel Zambrano Saavedra

Tutor Jairo Hernández Gutiérrez (U. Distrital).

1.10.3 Factibilidad Económica

A continuación, se presenta los presupuestos y fuentes de financiación necesarios para la implementación del proyecto:

Tabla 2 Factibilidad Económica

RECURSOS	ESTUDIANTE	UNIVERSIDAD	EMPRESA / ENTIDAD	OTRO
Descargas de herramientas	$ 30.000	0	0	0
Equipo de computo	$ 1.300.000	0	0	0
Papelería, Documentación y Transporte	$ 300.000	0	0	0
Bibliografía	$ 300.000	0	0	0
Ingeniero tutor horas 25.000/hora por 24	0	$ 600.000	0	0
Desarrollador: Estudiante tecnología. Valor 12.000 hora por 500 por 1 integrante	$ 6.000.000	0	0	0
Energía eléctrica	$ 100.000	0	0	0
Imprevistos	$ 400.000	0	0	0
Herramientas de Cómputo y Licencias	0	0	0	0
Total por recurso	$ 8.430.000	$ 600.000	0	0
Total general				$ 9.030.000

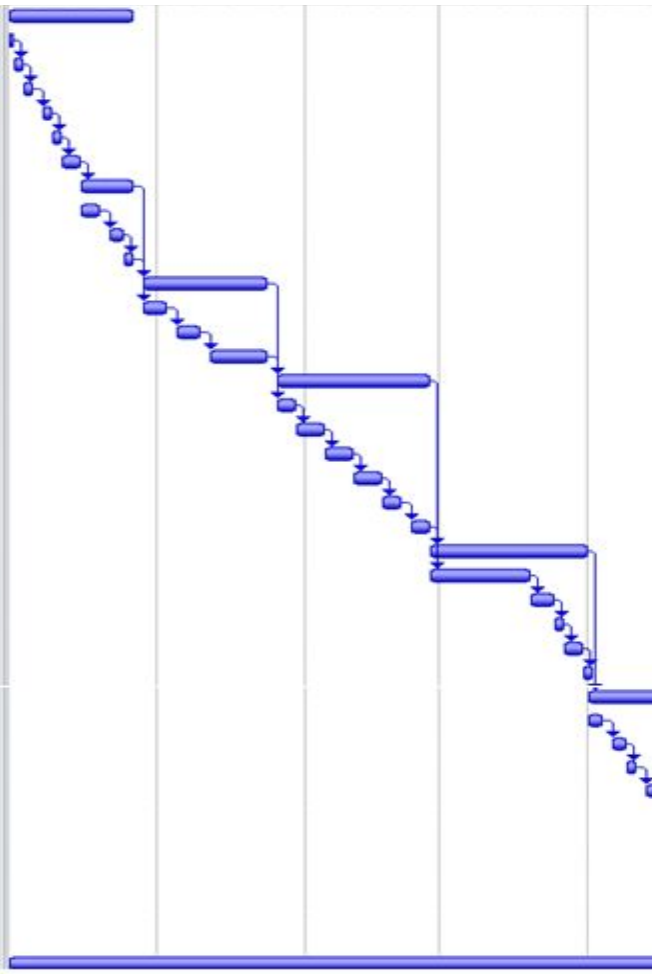
Fuente: Autores

De lo anterior se puede deducir que el Modelo predictivo para determinar el fracaso de matemáticas en grado 11 usando machine learning es viable, y por consiguiente se recomienda su desarrollo.

1.11 CRONOGRAMA

Figura 2 Cronograma de Actividades

FASE DE PLANEACIÓN Y ORGANIZACIÓN	**20 días**	**lun 01/12/14**	**vie 26/12/14**
Definición de Tema	1 día	lun 01/12/14	lun 01/12/14
Definición de la problemática	2 días	mar 02/12/14	mié 03/12/14
Determinar Alcances	2 días	jue 04/12/14	vie 05/12/14
Definicion de objetvos	2 días	lun 08/12/14	mar 09/12/14
Establecer marco de referencia	2 días	mié 10/12/14	jue 11/12/14
Establecer factblidad	2 días	vie 12/12/14	lun 15/12/14
Modelo organzacional del negocio	**9 días**	**mar 16/12/14**	**vie 26/12/14**
modelo de procesos	4 días	mar 16/12/14	vie 19/12/14
modelo del dominio	3 días	lun 22/12/14	mié 24/12/14
documentación modelo del dominio	2 días	jue 25/12/14	vie 26/12/14
REQUERIMIENTOS	**20 días**	**lun 29/12/14**	**vie 23/01/15**
Definción de actores	5 días	lun 29/12/14	vie 02/01/15
Lista preliminar de casos de uso	5 días	lun 05/01/15	vie 09/01/15
Depuración de casos de uso	10 días	lun 12/01/15	vie 23/01/15
ANALISIS	**24 días**	**lun 26/01/15**	**jue 26/02/15**
Diagrama de secuencia	4 días	lun 26/01/15	jue 29/01/15
Diagrama de colaboración	4 días	vie 30/01/15	mié 04/02/15
Diagrama de actividad	4 días	jue 05/02/15	mar 10/02/15
Diagrama de estado	4 días	mié 11/02/15	lun 16/02/15
Dagrama Lógico	4 días	mar 17/02/15	vie 20/02/15
Dagrama fisico	4 días	lun 23/02/15	jue 26/02/15
DISEÑO	**23 días**	**vie 27/02/15**	**mar 31/03/15**
Modelo de Base de datos	15 días	vie 27/02/15	jue 19/03/15
Lista de clases	3 días	vie 20/03/15	mar 24/03/15
Responsabilidad entre clases	2 días	mié 25/03/15	jue 26/03/15
Modelo Interfaz	2 días	vie 27/03/15	lun 30/03/15
Modelo lógico	2 días	mar 31/03/15	mié 01/04/15
IMPLEMENTACIÓN	**30 días**	**mié 01/04/15**	**mar 12/05/15**
Diagrama de desplegue	3 días	mié 01/04/15	vie 03/04/15
Diagrama de componentes	3 días	lun 06/04/15	mié 08/04/15
Diagrama de paquetes	2 días	jue 09/04/15	vie 10/04/15
Crear componentes de la aplicación	25 días	lun 13/04/15	vie 15/05/15
Proporcionar información requerda por la aplicación	2 días	lun 18/05/15	mar 19/05/15
PRUEBAS	**20 días**	**mié 20/05/15**	**mar 16/06/15**
Ensamblar coponentes en el servidor	4 días	mié 20/05/15	lun 25/05/15
Correciones del sistema	12 días	mar 26/05/15	mié 10/06/15
Puesta en marcha de pruebas indivduales	4 días	jue 11/06/15	mar 16/06/15
DOCUMENTACIÓN	**142 días?**	**lun 01/12/14**	**mar 16/06/15**



Fuente: Autores

2. DESARROLLO DEL ANÁLISIS DE DATOS

En el desarrollo del análisis de datos, según la metodología CRISP-DM está se compone por comprensión del negocio, comprensión de los datos, preparación de los datos, modelado y evaluación. Todos estos pasos son abordados en el presente documento, sin embargo, se agrupan de tal forma que se logren identificar fácilmente la solución de los objetivos propuestos y por lo tanto en el análisis de datos se unen la comprensión del negocio, la comprensión de los datos y la preparación de datos.

2.1 COMPRENSIÓN DEL NEGOCIO

En esta fase se busca entender la problemática de la deserción estudiantil y darle una perspectiva de negocio, a fin de convertir este conocimiento en la definición de un problema de minería de datos y plantear cómo se obtuvieron los datos.

La educación es la herramienta más eficaz que tiene un país para mejorar el estilo de vida de sus habitantes, unos ciudadanos íntegros en conocimiento se abren las puertas hacia el progreso, equidad y constante mejora a las falencias que se tienen en los procesos actuales que se ejecutan; por esta razón, las matemáticas juegan un papel importante en el desarrollo íntegro de una persona ya que sirven como lenguaje de comunicación con el universo, el desarrollo del pensamiento abstracto para la resolución de problemas y el cálculo de operaciones tanto cotidianas como en proyectos de gran escala; debido a esto, es necesario contar con herramientas que permitan la identificación automática de estudiantes con falencias en el aprendizaje de la materia para poder intervenir y ayudar a superar los inconvenientes.

Se utilizó la sábana de datos, proporcionada por el coordinador del Colegio Almirante Padilla de los años 2016, 2017 y 2018. El colegio inició sus labores en el año 1983, lleva el nombre Almirante Padilla pues el barrio donde se encuentra ubicado tiene ese nombre y por las relaciones cercanas que en sus inicios mantuvo con la Armada Nacional, el colegio ha contado con la fortuna de ser impulsado por la comunidad de la localidad 5, esfuerzo que llevó a la administración del año 1985 a incluir a Usme como beneficiario directo del plan Cidad Bolívar, y dentro del programa de educación a reconocer al Almirante Padilla como uno de los CEDID Centro de Enseñanza Diversificada Distrital. Está ubicado en la localidad quinta de Usme, Barrio Almirante Padilla (Calle 76 A Sur N.

1 D 59 Este), este es un buen lugar ya que cumple adecuadamente con las características de un colegio presencial de Bogotá.

Figura 3 Colegio Almirante Padilla

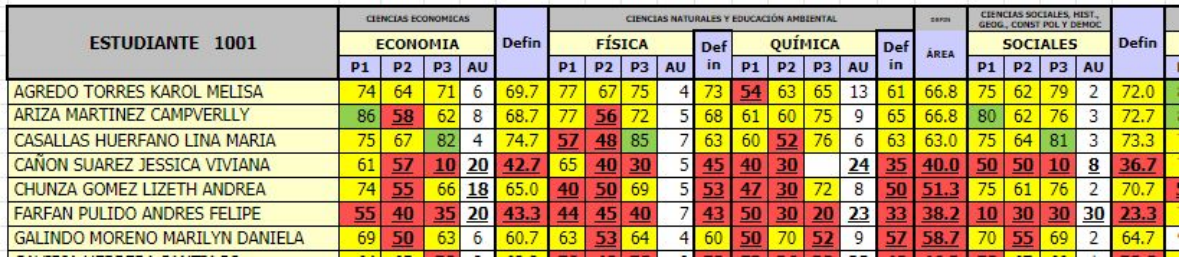
Fuente: Autores

2.2 COMPRENSIÓN DE LOS DATOS

En esta fase el objetivo principal fue hacer un entendimiento inicial de los datos a analizar para familiarizarse con ellos, identificar problemas de calidad en los mismos, detectar subconjuntos de datos para formular hipótesis específicas que se revisaron posteriormente con el análisis, adicionalmente se buscó identificar las primeras claves del conocimiento que se puede extraer de los datos.

La sábana de datos se compone por tres archivos, un archivo por cada año. Cada archivo contiene información de los cursos desde quinto de primaria hasta once con las notas de cada estudiante en las diferentes asignaturas, separado por cursos como se muestra en la figura 4 Notas Curso 1001 y en la figura 5 Notas Curso 1003:

Figura 4 Notas Curso 1001



ESTUDIANTE 1001	CIENCIAS ECONOMICAS				Defin	CIENCIAS NATURALES Y EDUCACIÓN AMBIENTAL										DEFIN ÁREA	CIENCIAS SOCIALES, HIST., GEOG., CONST POL Y DEMOC				Defin
	ECONOMIA					FÍSICA				Def in	QUÍMICA				Def in		SOCIALES				
	P1	P2	P3	AU		P1	P2	P3	AU		P1	P2	P3	AU			P1	P2	P3	AU	
AGREDO TORRES KAROL MELISA	74	64	71	6	69.7	77	67	75	4	73	54	63	65	13	61	66.8	75	62	79	2	72.0
ARIZA MARTINEZ CAMPVERLLY	86	58	62	8	68.7	77	56	72	5	68	61	60	75	9	65	66.8	80	62	76	3	72.7
CASALLAS HUERFANO LINA MARIA	75	67	82	4	74.7	57	48	85	7	63	60	52	76	6	63	63.0	75	64	81	3	73.3
CAÑON SUAREZ JESSICA VIVIANA	61	57	10	20	42.7	65	40	30	5	45	40	30		24	35	40.0	50	50	10	8	36.7
CHUNZA GOMEZ LIZETH ANDREA	74	55	66	18	65.0	40	50	69	5	53	47	30	72	8	50	51.3	75	61	76	2	70.7
FARFAN PULIDO ANDRES FELIPE	55	40	35	20	43.3	44	45	40	7	43	50	30	20	23	33	38.2	10	30	30	30	23.3
GALINDO MORENO MARILYN DANIELA	69	50	63	6	60.7	63	53	64	4	60	50	70	52	9	57	58.7	70	55	69	2	64.7

Fuente: Autores

Figura 5 Notas Curso 1003

ESTUDIANTE 1003	CIENCIAS ECONOMICAS				Defin	CIENCIAS NATURALES Y EDUCACIÓN AMBIENTAL										DEFIN ÁREA	CIENCIAS SOCIALES, HIST., GEOG., CONST POL Y DEMOC				Defin
	ECONOMIA					FÍSICA				Def in	QUÍMICA				Def in		SOCIALES				
	P1	P2	P3	AU		P1	P2	P3	AU		P1	P2	P3	AU			P1	P2	P3	AU	
ACOSTA CRUZ PAULA ALEJANDRA	40	50	40	2	43.3	30	55	40	3	42		40	15	10	28	34.6	30	50	55	6	45.0
AGUILERA GUERRERO JOSTEIN STEVEN	82	79	60	6	73.7	83	71	72	1	75	73	61	76	7	70	72.7	70	56	81	0	69.0
ARIAS ROPERO KEVIN ALEJANDRO	65	49	60	4	58.0	67	67	65	1	66	75	45	65	6	62	64.0	50	40	71	0	53.7
ARTUNDUAGA VARGAS LEIDER	69	65	73	4	69.0	73	69	60	2	67	66	60	68	3	65	66.0	50	60	76	2	62.0
BARBOSA CESPEDES BRAYAN ESTEBAN	76	59	55	4	63.3	77	63	72	2	71	65	50	70	6	62	66.2	75	68	80	1	74.3
BARBOSA SANCHEZ YULIETH STEFANNY	64	76	66	8	68.7	78	71	69	2	73	64	60	71	12	65	68.8	65	64	81	1	70.0
BAREÑO YALANDA HASVEIDY KATERINE	62	58	52	2	57.3	68	60	60	0	63	51	60	70	6	60	61.5	60	54	67	0	60.3

Fuente: Autores

Para el modelo se extrajo la información de los cursos décimo y once como se describe en la Tabla 3 Información de cursos extraída:

Tabla 3 Información de cursos extraída

Año	Grado 10	Grado 11
2016	x	
2017	x	x
2018		x

La información de los estudiantes de grado décimo se toma como entrada para predecir la nota en cálculo del grado 11, como solo se tiene la información de esos tres años y se necesita tanto la información de décimo como de once se hace necesario prescindir de los demás datos.

De los datos recolectados se determinó la siguiente información:

1. De los 215 registros que se obtuvieron, 32 no sirvieron; debido a que, el estudiante no continuo en el colegio para terminar el bachillerato.
2. Hay 8 estudiantes que repitieron décimo, de los cuales 6 son mujeres y 2 son hombres.
3. De los 182 registros utilizados para los modelos, 105 de los registros son mujeres y 77 son hombres como se muestra en la tabla 4 Cantidad hombres y mujeres.

Tabla 4 Cantidad hombres y mujeres

SEXO	CANTIDAD
MUJERES	105
HOMBRES	77

4. La nota más alta de la sábana de datos es de 93 y la obtuvo una mujer.
5. La nota más baja de la sábana de datos es de 40 y la obtuvo un hombre.
6. La media de la sábana de datos es de 68.10567766.
7. La moda de la sábana de datos es de 54.3 y la sacaron 40 estudiantes de los cuales 11 son hombres y 29 son mujeres.
8. De la sábana de notas la nota más alta de los hombres fue de 87.66666667.
9. De la sábana de notas, la nota más baja de las mujeres fue de 54.3.
10. La tabla 5 cantidad de hombres y mujeres con estrato 1 y 2 muestra la cantidad de hombres y mujeres con estrato 1 y 2 de la sábana de datos.

Tabla 5 Cantidad de hombres y mujeres con estrato 1 y 2

\ Estrato **Sexo **	**1**	**2**
MUJERES	12	93
HOMBRES	11	66

11. La tabla 6 Convenciones ministerio según rango de notas, muestra el valor literal que tiene una nota según la secretaría de educación.

Tabla 6 Convenciones ministerio según rango de notas

CONVENCIONES MINISTERIO	RANGO	# ESTUDIANTES
SUPERIOR	[90-100]	2
ALTO	[80-90)	21
BÁSICO	[60-80)	113
BAJO	[0-60)	46

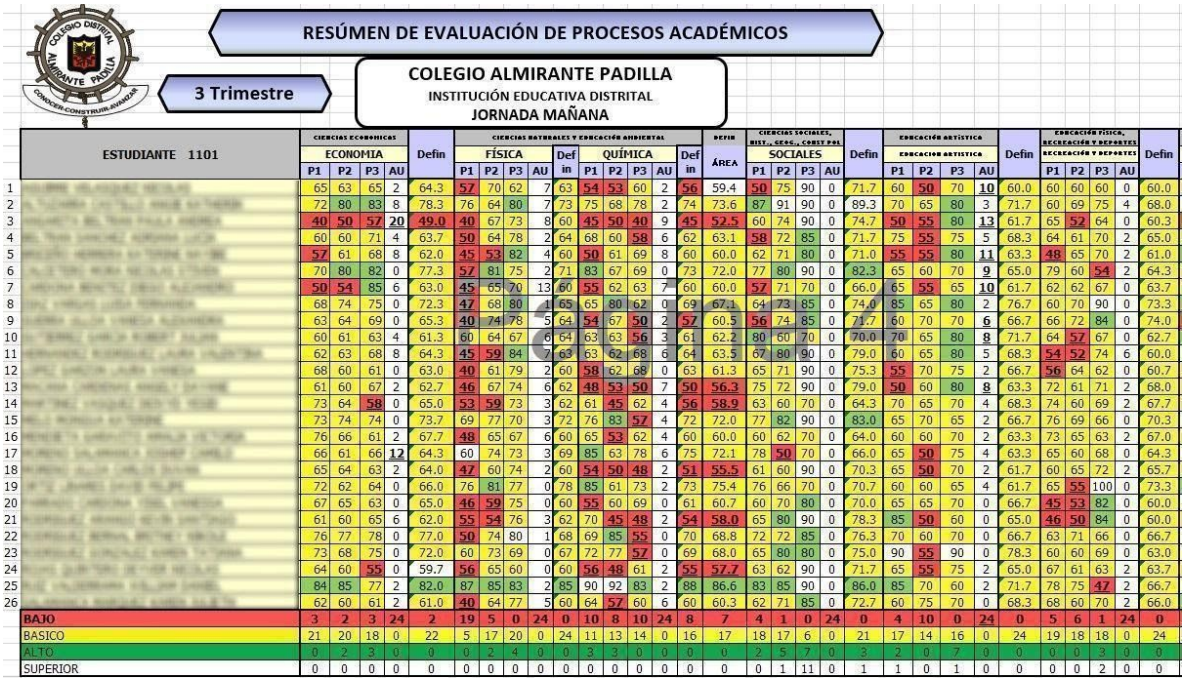
Fuente: Autores

2.3 PREPARACIÓN DE DATOS

Antes de que los datos en bruto suministrados por el Colegio Almirante Padilla de Usme se sometan a un análisis de machine learning, deben convertirse a una forma apropiada para tal análisis.

Se tienen tres archivos de Excel con formato .xlsx, en estos se encuentran los resúmenes de evaluación estudiantil de los años 2016, 2017 y 2018. Dentro de cada archivo hay varias hojas y en cada hoja se encuentran agrupadas las notas trimestrales de los estudiantes diferenciados por curso desde el grado quinto hasta el grado once; adicionalmente, por materia se tienen tres notas y de estas se obtiene una nota definitiva.

Figura 6 Resumen de Evaluación de Procesos Académicos



RESÚMEN DE EVALUACIÓN DE PROCESOS ACADÉMICOS

3 Trimestre

COLEGIO ALMIRANTE PADILLA
INSTITUCIÓN EDUCATIVA DISTRITAL
JORNADA MAÑANA

	ESTUDIANTE 1101	CIENCIAS ECONOMICAS				Definitiva	CIENCIAS NATURALES Y EDUCACIÓN AMBIENTAL										DEFINITIVA ÁREA	CIENCIAS SOCIALES, HIST., GEOG., CONST POL				Definitiva	EDUCACIÓN ARTISTICA				Definitiva	EDUCACIÓN FISICA, RECREACIÓN Y DEPORTES				Definitiva
		ECONOMIA					FÍSICA				Definitiva	QUÍMICA				Definitiva		SOCIALES					EDUCACION ARTISTICA					RECREACIÓN Y DEPORTES				
		P1	P2	P3	AU		P1	P2	P3	AU		P1	P2	P3	AU			P1	P2	P3	AU		P1	P2	P3	AU		P1	P2	P3	AU	
1		65	63	65	2	64.3	57	70	62	7	63	54	53	60	2	56	59.4	50	75	90	0	71.7	60	50	70	10	60.0	60	60	60	0	60.0
2		72	80	83	8	78.3	76	64	80	7	73	75	68	78	2	74	73.6	87	91	90	0	89.3	70	65	80	3	71.7	60	69	75	4	68.0
3		40	50	57	20	49.0	40	67	73	8	60	45	50	40	9	45	52.5	60	74	90	0	74.7	50	55	80	13	61.7	65	52	64	0	60.3
4		60	60	71	4	63.7	50	64	78	2	64	68	60	58	6	62	63.1	58	72	85	0	71.7	75	55	75	5	68.3	64	61	70	2	65.0
5		57	61	68	8	62.0	45	53	82	4	60	50	61	69	8	60	60.0	62	71	80	0	71.0	55	55	80	11	63.3	48	65	70	2	61.0
6		70	80	82	0	77.3	57	81	75	2	71	83	67	69	0	73	72.0	77	80	90	0	82.3	65	60	70	9	65.0	79	60	54	2	64.3
7		50	54	85	6	63.0	45	65	70	13	60	55	62	63	7	60	60.0	57	71	70	0	66.0	65	55	65	10	61.7	62	62	67	0	63.7
8		68	74	75	0	72.3	47	68	80	1	65	65	80	62	0	69	67.1	64	73	85	0	74.0	85	65	80	2	76.7	60	70	90	0	73.3
9		63	64	69	0	65.3	40	74	78	5	64	54	67	50	2	57	60.5	56	74	85	0	71.7	60	70	70	6	66.7	66	72	84	0	74.0
10		60	61	63	4	61.3	60	64	67	6	64	63	63	56	3	61	62.2	80	60	70	0	70.0	70	65	80	8	71.7	64	57	67	0	62.7
11		62	63	68	8	64.3	45	59	84	7	63	63	62	68	6	64	63.5	67	80	90	0	79.0	60	65	80	5	68.3	54	52	74	6	60.0
12		68	60	61	0	63.0	40	61	79	2	60	58	62	68	0	63	61.3	65	71	90	0	75.3	55	70	75	2	66.7	56	64	62	0	60.7
13		61	60	67	2	62.7	46	67	74	6	62	48	53	50	7	50	56.3	75	72	90	0	79.0	50	60	80	8	63.3	72	61	71	2	68.0
14		73	64	58	0	65.0	53	59	73	3	62	61	45	62	4	56	58.9	63	60	70	0	64.3	70	65	70	4	68.3	74	60	69	2	67.7
15		73	74	74	0	73.7	69	77	70	3	72	76	83	57	4	72	72.0	77	82	90	0	83.0	65	70	65	2	66.7	76	69	66	0	70.3
16		76	66	61	2	67.7	48	65	67	6	60	65	53	62	4	60	60.0	60	62	70	0	64.0	60	60	70	2	63.3	73	65	63	2	67.0
17		66	61	66	12	64.3	60	74	73	3	69	85	63	78	6	75	72.1	78	50	70	0	66.0	65	50	75	4	63.3	65	60	68	0	64.3
18		65	64	63	2	64.0	47	60	74	2	60	54	50	48	2	51	55.5	61	60	90	0	70.3	65	50	70	2	61.7	60	65	72	2	65.7
19		72	62	64	0	66.0	76	81	77	0	78	85	61	73	2	73	75.4	76	66	70	0	70.7	60	60	65	4	61.7	65	55	100	0	73.3
20		67	65	63	0	65.0	46	59	75	0	60	55	60	69	0	61	60.7	60	70	80	0	70.0	65	65	70	0	66.7	45	53	82	0	60.0
21		61	60	65	6	62.0	55	54	76	3	62	70	45	48	2	54	58.0	65	80	90	0	78.3	85	50	60	0	65.0	46	50	84	0	60.0
22		76	77	78	0	77.0	50	74	80	1	68	69	85	55	0	70	68.8	72	72	85	0	76.3	70	60	70	0	66.7	63	71	66	0	66.7
23		73	68	75	0	72.0	60	73	69	0	67	72	77	57	0	69	68.0	65	80	80	0	75.0	90	55	90	0	78.3	60	60	69	0	63.0
24		64	60	55	0	59.7	56	65	60	0	60	56	48	61	2	55	57.7	63	62	90	0	71.7	65	55	75	2	65.0	67	61	63	2	63.7
25		84	85	77	2	82.0	87	85	83	2	85	90	92	83	2	88	86.6	83	85	90	0	86.0	85	70	60	2	71.7	78	75	47	2	66.7
26		62	60	61	2	61.0	40	64	77	5	60	64	57	60	6	60	60.3	62	71	85	0	72.7	60	75	70	0	68.3	68	60	70	2	66.0
	BAJO	3	2	3	24	2	19	5	0	24	0	10	8	10	24	8	7	4	1	0	24	0	4	10	0	24	0	5	6	1	24	0
	BASICO	21	20	18	0	22	5	17	20	0	24	11	13	14	0	16	17	18	17	6	0	21	17	14	16	0	24	19	18	18	0	24
	ALTO	0	2	3	0	0	0	2	4	0	0	3	3	0	0	0	0	2	5	7	0	3	2	0	7	0	0	0	0	3	0	0
	SUPERIOR	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	11	0	1	1	0	1	0	0	0	0	2	0	0

Fuente: Autores.

Para facilidad del análisis de los datos se crea un nuevo archivo llamado DataEstudiantes en formato.csv, está teniendo en cuenta que el enfoque del modelo predictivo está dado solo para los estudiantes del grado once, se eliminan los demás cursos y se borran las notas diferentes a la nota definitiva de las materias; adicionalmente, no se cuenta con el sexo del estudiante, por tanto, este se determina con base en su nombre.

El archivo queda con la estructura que se representa en la tabla 7 estructura archivo entrada:

Tabla 7 Estructura archivo entrada

CARACTERÍSTICA	DESCRIPCIÓN
ESTUDIANTE	Nombre del estudiante
R	Y indica que repitió décimo N que es la primera vez.
S	Sexo biológico del estudiante
CURSO	El curso al que pertenecía en décimo
J10	Jornada en la que curso décimo
ESTRATO	Estrato del estudiante
AÑO	Año en que curso décimo

ECONOMÍA	Nota definitiva para la asignatura en décimo
FÍSICA	Nota definitiva para la asignatura en décimo
QUÍMICA	Nota definitiva para la asignatura en décimo
SOCIALES	Nota definitiva para la asignatura en décimo
EDUCACIÓN ARTÍSTICA	Nota definitiva para la asignatura en décimo
RECREACIÓN Y DEPORTES	Nota definitiva para la asignatura en décimo
RELIGIÓN	Nota definitiva para la asignatura en décimo
ÉTICA	Nota definitiva para la asignatura en décimo
FILOSOFÍA	Nota definitiva para la asignatura en décimo
INGLÉS	Nota definitiva para la asignatura en décimo
LENGUA CASTELLANA	Nota definitiva para la asignatura en décimo
TRIGONOMETRÍA	Nota definitiva para la asignatura en décimo
ADMINISTRACIÓN	Nota definitiva para la asignatura en décimo
MANEJO ORDENADORES	Nota definitiva para la asignatura en décimo
PREPARACIÓN TECNOLÓGICA	Nota definitiva para la asignatura en décimo
CÁLCULO	Nota definitiva para la asignatura en once
CONVENCIONES MINISTERIO	Escala en la que se encuentra la nota del estudiante.
C11	El curso al que pertenecía en once
A11	Año en que curso once

3. MODELADO DE DATOS

En esta fase se seleccionan y aplican las técnicas (algoritmos) de modelado de regresión y clasificación, midiendo los parámetros para conseguir valores óptimos. Para un mismo problema de minería de datos se tienen diferentes técnicas susceptibles de ser usadas y, dado que cada una de ellas puede tener requisitos diferentes en la forma en que deben presentarse los datos de entrada, es probable que sea necesario realizar ciclos adicionales o independientes de “preparación de los datos”.

3.1 APRENDIZAJE AUTOMÁTICO: REGRESIÓN

Para el desarrollo del modelo, se depuró de manera manual la tabla 7 dejando las características más relevantes para la aplicación del método. La tabla resultante es la Tabla 8 Características usadas en el método de regresión.

Tabla 8 Características usadas en el método de regresión

H	CARACTERÍSTICA	DESCRIPCIÓN
h_1	ECONOMÍA	Nota definitiva para la asignatura en décimo
h_2	FÍSICA	Nota definitiva para la asignatura en décimo
h_3	QUÍMICA	Nota definitiva para la asignatura en décimo
h_4	SOCIALES	Nota definitiva para la asignatura en décimo
h_5	EDUCACIÓN ARTÍSTICA	Nota definitiva para la asignatura en décimo
h_6	RECREACIÓN Y DEPORTES	Nota definitiva para la asignatura en décimo
h_7	RELIGIÓN	Nota definitiva para la asignatura en décimo
h_8	ÉTICA	Nota definitiva para la asignatura en décimo
h_9	FILOSOFÍA	Nota definitiva para la asignatura en décimo
h_{10}	INGLÉS	Nota definitiva para la asignatura en décimo
h_{11}	LENGUA CASTELLANA	Nota definitiva para la asignatura en décimo
h_{12}	TRIGONOMETRÍA	Nota definitiva para la asignatura en décimo
h_{13}	ADMINISTRACIÓN	Nota definitiva para la asignatura en décimo
h_{14}	MANEJO ORDENADORES	Nota definitiva para la asignatura en décimo
h_{15}	PREPARACIÓN TECNOLÓGICA	Nota definitiva para la asignatura en décimo

Hay 16 características, lo que permite tener un total de 65,536 posibles modelos, este valor se obtiene de la ecuación 1 número de posibles modelos.

Ecuación 1 Número de posibles modelos

$$M = 2^{D+1}$$

Donde D es el número de características. Computacionalmente no es viable probar todos los modelos para obtener el mejor de todos. Así que se utilizó el

método de LASSO (Least Absolute Shrinkage and Selection Operator) para descartar aquellas características que no le dan valor al modelo o reducir el valor de sus coeficientes.

Para esto se inicia la búsqueda del modelo utilizando todas las características. La ecuación 2 muestra cómo sería un modelo utilizando las 15 características:

Ecuación 2 Expansión para modelo con 15 características

$$y_i = w_0 h_0(x_i) + w_1 h_1(x_i) + w_2 h_2(x_i) + w_3 h_3(x_i) + w_4 h_4(x_i) + ... + w_{15} h_{15}(x_i) + \varepsilon_i$$

w_j es el coeficiente o peso de la característica. A cada característica se le debe encontrar su coeficiente.

$h_j(x_i)$ es el valor de cada característica.

y_i es el valor que da al tener todas las características.

ε_i es el error faltante para que la predicción sea igual a y_i.

Para entender el concepto la tabla 9 utiliza tres características con la nota de cálculo que es la variable objetivo(y_i) las 3 características sirven para predecir la nota de cálculo:

Tabla 9 Tres características regresión

h_1	h_2	h_3	y_i
64	57	72	72
71	97	84	80
81	78	90	67
60	74	86	85
63	75	78	67
60	72.7	75.7	64

Teniendo como entrada la tabla 9 el modelo con las tres características del primer registro sería la ecuación 3

Ecuación 3 Modelo con tres características del registro 1

$$72 = w_0 + w_1 64 + w_2 57 + w_3 72$$

La ecuación 4 representa de manera simplificada la ecuación 2.

Ecuación 4 Modelo simplificado de regresión

$$\sum_{j=0}^{D} w_j h_j(x_i) + \varepsilon_i$$

Para encontrar el valor de los coeficientes, para el posterior uso de LASSO. Se utilizó la técnica coordenada descendente. Iniciando con la normalización de las características primero se encuentra la magnitud de cada columna. como se muestra en la ecuación 5.

Ecuación 5 Magnitud características

$$\left|h_j\right| = \sqrt{\sum_{i=1}^{15} h_j(x_i)^2}$$

Siguiendo con los valores de la tabla 9, la magnitud para cada columna se muestra en la tabla 10.

Tabla 10 Magnitud tres características

h_1	h_2	h_3
163.9115615	187.4254252	198.87305

Luego se divide cada valor por la magnitud correspondiente para dejar las características normalizadas como se muestra en la ecuación 6.

Ecuación 6 Características normalizadas de regresión

$$\sum_{i=1}^{15} \frac{h_j(x_i)}{|h_j|}$$

En la tabla 11 muestra las características normalizadas.

Tabla 11 Características normalizadas

h_1	h_2	h_3	y_i
0.3904544585	0.3041209588	0.362040005	72
0.4331604149	0.5175391754	0.4223800058	80
0.4941689241	0.4161655225	0.4525500062	67
0.3660510549	0.3948237009	0.4324366726	85
0.3843536076	0.4001591563	0.3922100054	67
0.3660510549	0.3878876088	0.3806448386	64

Después de normalizada la información se procedió con los siguientes pasos:

1. Se asignó por cada coeficiente el valor de 0 para este caso sería $[0, 0, 0]$.
2. Se realizó la predicción de la nota de cálculo operando los coeficientes con las características.
3. Se calculó la coordenada aplicando la ecuación 7.

Ecuación 7 Coordenada descendente

$$ro = h_j(x_i) * (SR + w_i h_j(x_i))$$

ro es el valor de la coordenada utilizado para determinar el nuevo valor del coeficiente, se calcula realizando el producto punto entre los valores de cada característica con la suma entre la diferencia y la multiplicación del coeficiente con los valores de la característica.

SR es la diferencia de la nota de cálculo y_i con la nota que se predijo $\hat{y}_i$ La tabla 10 Valores suma residual muestra el resultado.

Tabla 10 Valores suma residual

h_1	h_2	h_3	y_i	$\hat{y}_i$	SR
0.3904544585	0.3041209588	0.362040005	72	0	72
0.4331604149	0.5175391754	0.4223800058	80	0	80
0.4941689241	0.4161655225	0.4525500062	67	0	67
0.3660510549	0.3948237009	0.4324366726	85	0	85
0.3843536076	0.4001591563	0.3922100054	67	0	67
0.3660510549	0.3878876088	0.3806448386	64	0	64

4. Después de calculada la ro se procede a actualizar el nuevo valor de los coeficientes siguiendo la lógica de la ecuación 8.

Ecuación 8 Actualización de coeficientes

$$w[i] = \begin{cases} (ro[i] + \lambda/2) & si\ ro[i] < -\lambda/2 \\ 0 & si\ -\lambda/2 <= ro[i] <= \lambda/2 \\ (ro[i] - \lambda/2) & si\ ro[i] > \lambda/2 \end{cases}$$

λ es el parámetro para balancear los coeficientes sacando de la ecuación aquellos que no aportan, a una buena predicción del modelo. para esta explicación se toma el valor de λ=10, en la próxima unidad se explica cómo se obtuvo el valor.

La tabla 11 muestra el nuevo valor de los coeficientes para la primera iteración:

Tabla 11 Nuevos valores coeficientes iteración 1

ro	w	w+1	r
176.168171	0	171.168171	171.168171
176.3784181	0	176.3784181	176.3784181
177.5745884	0	177.5745884	177.5745884

La columna r es la diferencia entre los valores nuevos con los antiguos del coeficiente. La razón de esta operación es para determinar la finalización del proceso, el cual termina hasta que el máximo de los valores de la diferencia sea menor a 1 la representación se visualiza en la ecuación 9.

Ecuación 9 Máximo valor de la diferencia

$$Max(r) = 177.574$$

Mientras Max(r)>1 se repetirán los pasos del 2 hasta el 4. Cuando Max(r)<1 obtiene los coeficientes finales de los cuales solo se utilizan aquellos que cumplan con el umbral.

3.1.1 Entrenamiento del Modelo

Para el entrenamiento del modelo se utilizó la división de los datos de manera aleatoria como se representa en la tabla 12.

Tabla 12 Porcentaje de división de la información

	%	Cantidad
Datos Entrenamiento	70	127
Datos Prueba	30	55

Con los datos de entrenamiento se obtienen los coeficientes, y se usa λ para reducir el valor de aquellos coeficientes que no aportan a la predicción. Cuando se tienen los coeficientes se calcula el RSS (Residual Sum Square) con los datos de prueba. para este proyecto se utilizaron 4000 valores de λ y se escogió el valor con el RSS menor. La tabla 13 muestra los 20 menores valores para RSS:

Tabla 13 Primeros λ con menor RSS en los datos de validación

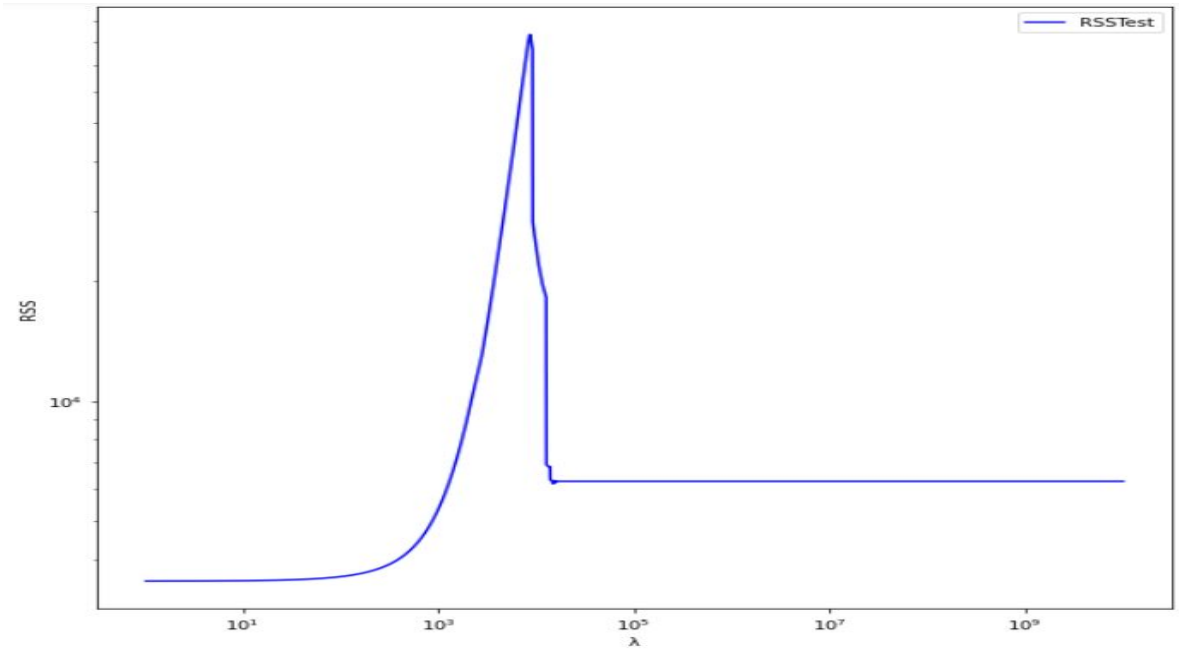
RSSTest	λ
3528,27	1,02329
3528,27	1,0292
3528,28	1,03513
3528,28	1,0411
3528,29	1,05315
3528,29	1,04711
3528,3	1,05923
3528,3	1,06534
3528,31	1,07148
3528,31	1,07766
3528,32	1,08388
3528,32	1,09013
3528,33	1,09642
3528,33	1,10275
3528,34	1,10911
3528,34	1,11551

3528,35	1,12194
3528,35	1,12841
3528,36	1,13492
3528,37	1,14147

De los 4000 registros se escoge el valor de λ=1,2329 ya que el RSS fue el menor con los datos de prueba.

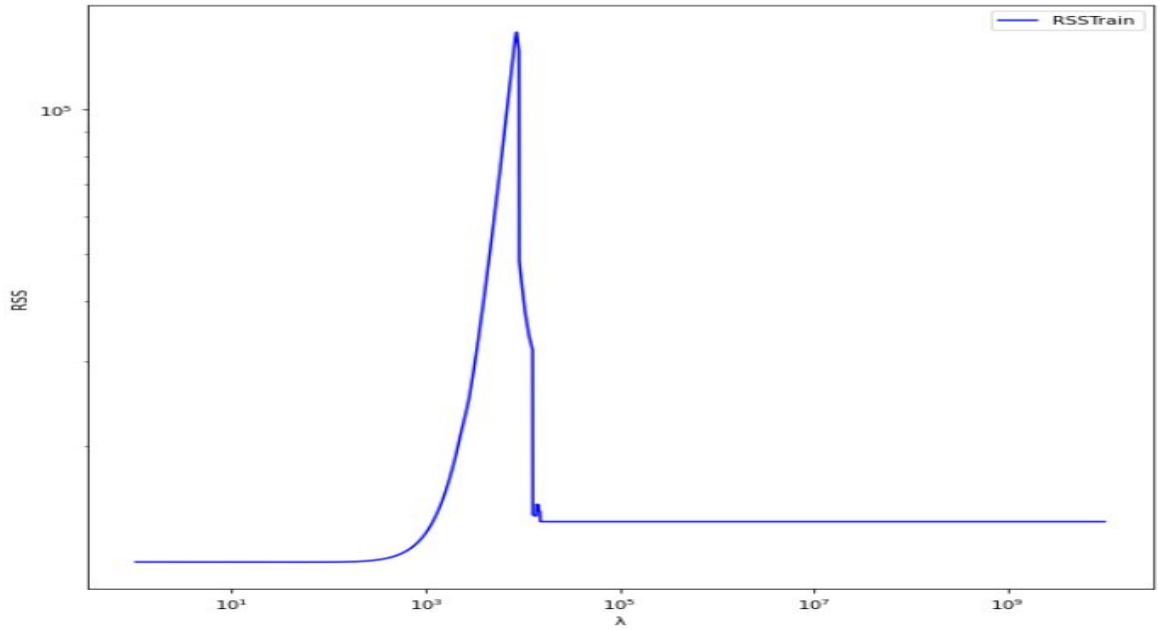
La figura 7 muestra el RSS en función de λ usando los datos de evaluación.

Figura 7 RSS Datos Evaluación.



Se puede observar que el error más alto se obtiene entre 1000<λ<100000. Para verificar que el valor de λ elegido es el correcto la figura 8 muestra el comportamiento del RSS utilizando los datos de entrenamiento.

Figura 8 RSS Datos Entrenamiento



La gráfica tiene el mismo comportamiento con esto se asegura que el valor de λ escogido es el correcto. La siguiente ecuación representa el modelo obtenido

Ecuación 10 Modelo Regresión

$$4{,}459582334843983 + 0{,}06578458944494692\,H1(x) + 0{,}06560394971957408\,H2(x) + 0{,}06786995939266334\,H3(x) + 0{,}06556924722504778\,H4(x) + 0{,}06217762903795568\,H5(x) + 0{,}06547172612555534\,H6(x) + 0{,}05918163761574663\,H7(x) + 0{,}05872600328758504\,H8(x) + 0{,}0602466477207836\,H9(x) + 0{,}06631271290705533\,H10(x) + 0{,}06155374139713069\,H11(x) + 0{,}06621075804539084\,H12(x) + 0{,}04156184661128062\,H13(x) + 0{,}0559417652602988\,H14(x) + 0{,}06125040776814061\,H15(x)$$

3.1.2 Evaluación del Modelo

Para evaluar el modelo se utiliza la ecuación 10. con las notas de los estudiantes de décimo del año 2018, el resultado de la predicción se muestra en la Tabla 13.

Tabla 13 Predicción estudiantes grado décimo

CALCULO	CONVENCIONES MINISTERIO	PREDICCION	CON MINISTERIO PREDICCION	ASERTO
60,8003	BASICO	60,5663	BASICO	1
64,1944	BASICO	63,9514	BASICO	1
57,13	BAJO	56,912	BAJO	1
63,8657	BASICO	63,6451	BASICO	1
60,4465	BASICO	60286	BASICO	1
72,2251	BASICO	71,9994	BASICO	1
69,7935	BASICO	69,544	BASICO	1
67,7787	BASICO	67,5147	BASICO	1
60,2474	BASICO	60,0135	BASICO	1
59,9482	BAJO	59,7228	BAJO	1
63,4423	BASICO	63,2016	BASICO	1
60,7093	BASICO	60,4737	BASICO	1
72,0214	BASICO	71,7664	BASICO	1
56,5364	BAJO	56,2972	BAJO	1
64,8629	BASICO	64,6037	BASICO	1
59,8876	BAJO	59,7129	BAJO	1
61,2936	BASICO	61,0855	BASICO	1
54,3635	BAJO	54,1383	BAJO	1
67,0403	BASICO	66,7923	BASICO	1
74,0158	BASICO	73,7656	BASICO	1
65,989	BASICO	65,7758	BASICO	1
61,7026	BASICO	61,4841	BASICO	1
58,8481	BAJO	58,5998	BAJO	1
70,7333	BASICO	70,4654	BASICO	1
61,6229	BASICO	61,4007	BASICO	1
69,1201	BASICO	68,8849	BASICO	1
73,8115	BASICO	73,5813	BASICO	1
50,9092	BAJO	50,6815	BAJO	1
66,3807	BASICO	66,1482	BASICO	1
69,4482	BASICO	69,3774	BASICO	1
63305	BASICO	63,2928	BASICO	1
62627	BASICO	62,6441	BASICO	1
64888	BASICO	64,8482	BASICO	1
57,0634	BAJO	57,0135	BAJO	1

66,4584	BASICO	66,4228	BASICO	1
74019	BASICO	73,9669	BASICO	1
66222	BASICO	66,1779	BASICO	1
70,3463	BASICO	70,3295	BASICO	1
71,0983	BASICO	71,0439	BASICO	1
66,1992	BASICO	66,1614	BASICO	1
67,3784	BASICO	67,3444	BASICO	1
62,9677	BASICO	62,9549	BASICO	1
61,8979	BASICO	61,8948	BASICO	1
60,0747	BASICO	60,0636	BASICO	1
66,4957	BASICO	66,4751	BASICO	1
65,3857	BASICO	65,3565	BASICO	1
65,0651	BASICO	65,0649	BASICO	1
56,0474	BAJO	56,0463	BAJO	1
57,7138	BAJO	57,6565	BAJO	1
59,9928	BAJO	59,9883	BAJO	1
53,3018	BAJO	53,2873	BAJO	1
75,1906	BASICO	75,1522	BASICO	1
76,5446	BASICO	76,4645	BASICO	1
66,5244	BASICO	66,4821	BASICO	1
66,2359	BASICO	66,2223	BASICO	1
47,4659	BAJO	47,4913	BAJO	1
84063	ALTO	83988	ALTO	1
65,8469	BASICO	65,8576	BASICO	1
66,7047	BASICO	66,6764	BASICO	1
63,3294	BASICO	63,3272	BASICO	1
59,8404	BAJO	59,8091	BAJO	1
70,1331	BASICO	70,1122	BASICO	1
66,4735	BASICO	66,4634	BASICO	1
69,4948	BASICO	69,4818	BASICO	1
67,3688	BASICO	67,3694	BASICO	1
66,6379	BASICO	66,6364	BASICO	1
69,6209	BASICO	69,5956	BASICO	1
77,2077	BASICO	77,1501	BASICO	1
64,1586	BASICO	64,1578	BASICO	1
67,1611	BASICO	67,1318	BASICO	1
65,5652	BASICO	65,5308	BASICO	1

64,9551	BASICO	64,9353	BASICO	1
63,8978	BASICO	63,8601	BASICO	1
65,5764	BASICO	65,5395	BASICO	1
70,0419	BASICO	70,0235	BASICO	1
83,9926	ALTO	83,9389	ALTO	1
64,0278	BASICO	64017	BASICO	1
67,9235	BASICO	67921	BASICO	1
66,8814	BASICO	66,8801	BASICO	1
66,8938	BASICO	66,9066	BASICO	1
63,8184	BASICO	63,8151	BASICO	1
63,7075	BASICO	63,701	BASICO	1
67,3589	BASICO	67,3382	BASICO	1
69,349	BASICO	69,302	BASICO	1

La columna ‘ASERTÓ’ indica 1 si la predicción fue correcta y 0 si no. Lo cual se concluye que de los 84 registros evaluados en 84 registros se obtuvo la predicción.

De los registros en los que se falló el modelo los estudiantes tenían una nota muy baja lo que indica que estaría bien que se les hiciera seguimiento a estos estudiantes.

3.2 APRENDIZAJE AUTOMÁTICO: CLASIFICACIÓN

3.2.1 Regresión Logística

Para el desarrollo del modelo, se depuro de manera manual la tabla 7 dejando las características categóricas más relevantes para la aplicación del método. La tabla resultante es la tabla 16.

Tabla 16 Características Usadas en el método de Clasificación

H	CARACTERÍSTICA	DESCRIPCIÓN
h_1	R	Y indica que repitió décimo N que es la primera vez
h_2	S	Sexo biológico del estudiante
h_3	J10	Jornada en la que curso décimo
h_4	ECONOMÍA	Nota definitiva para la asignatura en décimo
h_5	FÍSICA	Nota definitiva para la asignatura en décimo
h_6	QUÍMICA	Nota definitiva para la asignatura en décimo
h_7	SOCIALES	Nota definitiva para la asignatura en décimo
h_8	EDUCACIÓN ARTÍSTICA	Nota definitiva para la asignatura en décimo
h_9	RECREACIÓN Y DEPORTES	Nota definitiva para la asignatura en décimo
h_{10}	RELIGIÓN	Nota definitiva para la asignatura en décimo
h_{11}	ÉTICA	Nota definitiva para la asignatura en décimo
h_{12}	FILOSOFÍA	Nota definitiva para la asignatura en décimo
h_{13}	INGLÉS	Nota definitiva para la asignatura en décimo
h_{14}	LENGUA CASTELLANA	Nota definitiva para la asignatura en décimo
h_{15}	TRIGONOMETRÍA	Nota definitiva para la asignatura en décimo
h_{16}	ADMINISTRACIÓN	Nota definitiva para la asignatura en décimo
h_{17}	MANEJO ORDENADORES	Nota definitiva para la asignatura en décimo
h_{18}	PREPARACIÓN TECNOLÓGICA	Nota definitiva para la asignatura en décimo

Al igual que la regresión, la clasificación inicia la búsqueda del modelo utilizando todas las características. La ecuación 10 muestra cómo sería un modelo utilizando n características

Ecuación 10 Expansión para modelo con n características

$$y_i = w_0 h_0(x_i) + w_1 h_1(x_i) + w_2 h_2\,(x_i) + ... + w_n h_n(x_i)$$

w_j es el coeficiente o peso de la característica. A cada característica se le debe encontrar su coeficiente.

$h_j(x_i)$ es el valor de cada característica.

y_i es el valor que da al tener todas las características.

Para y_i se segmentan en dos grupos los valores de los datos de la siguiente forma:

Si ($Nota_{(x)}$) > 60 :

Entonces: y_i = +1

Sino:

Entonces: y_i = -1

$Nota_{(x)}$ es la nota con la cual se determina si el estudiante pasa la materia o no.

Para entender el concepto la tabla 14 contiene 6 registros con la nota de cálculo que es la variable objetivo(y_i) y 3 características con las cuales se va a predecir la nota de cálculo:

Tabla 14 Tres características clasificación

h_1	h_2	h_3	y_i
64	57	72	1
71	97	84	1
81	78	90	1
60	40	45	-1
63	75	78	1
60	63	65	1

Teniendo como entrada la Tabla 14 el modelo con las tres características del primer registro sería el que se visualiza en la ecuación 11.

Ecuación 11 Modelo con tres características del registro 1

$$+1 = w_0 h_0(x_i) + w_1 64 + w_2 57 + w_3 72$$

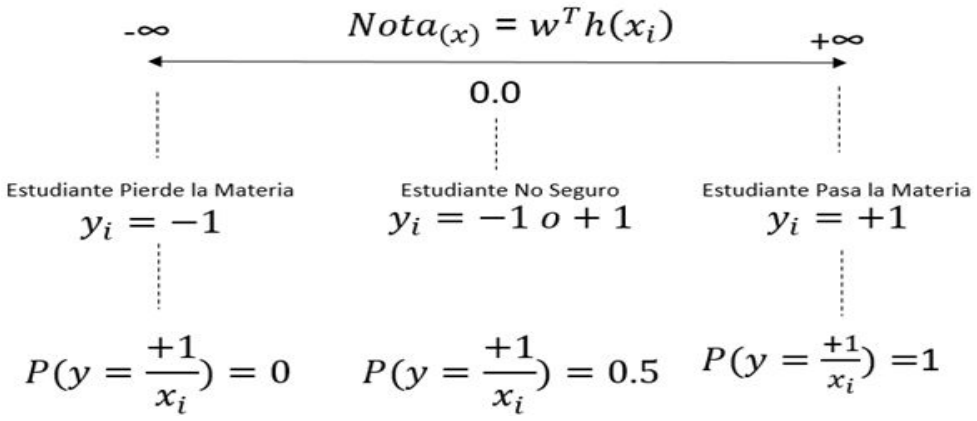
La ecuación 12 representa de manera simplificada la Ecuación 11.

Ecuación 12 Modelo simplificado de clasificación

$$\sum_{j=0}^{D} w_j h_j(x_i) = w^T h(x_i)$$

Para encontrar el valor de los coeficientes se usan probabilidades en clasificación y para eso se le da la interpretación que se visualiza en la figura 9.

Figura 9 Interpretación de Probabilidad de Notas



Fuente: Autores

P es la probabilidad medida de 0 a 1.

Teniendo en cuenta lo anterior, con base en la probabilidad obtenida se podrá determinar la nota del estudiante como se visualiza en la tabla 14.

Tabla 15 Interpretación de Notas

$Notas$	$Nota_{(x)}$
-6	0
-4	20
-2	40
0	60
2	75
4	85
6	100

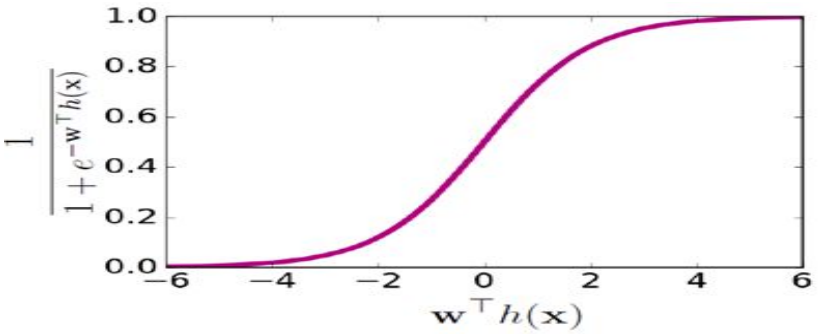
Cuando se entienden las probabilidades estas se igualan a la función de sigmoide, esta se usa en muchos procesos naturales y curvas de aprendizaje de sistemas complejos muestran una progresión temporal desde unos niveles bajos al inicio, hasta acercarse a un clímax transcurrido un cierto tiempo; la transición se produce en una región caracterizada por una fuerte aceleración intermedia. La función sigmoide permite describir esta evolución y debido a esto se usa la ecuación 13.

Ecuación 13 Modelo de regresión logística

$$P\left(y = \frac{+1}{x,w}\right) = \frac{+1}{1+e^{-w^T h(x)}}$$

Para entender el modelo de regresión logística se puede visualizar de forma gráfica con la figura 10.

Figura 10 Función Sigmoide



Fuente: FUNSION SIGMOIDE. Wikipedia [sitio web]. Fundación Wikimedia; [Consultado: 15 de mayo de 2020]. Disponible en: https://es.wikipedia.org/wiki/Funci%C3%B3n_sigmoide

Aplicando la función sigmoide se obtiene $\hat{y}_i$ que es la probabilidad de que estudiante pierda la materia esto se puede apreciar en la tabla 15.

Tabla 16 Aplicación de Característica y ˆ_i

h_1	h_2	h_3	y_i	$\hat{y}_i$
64	57	72	+1	0. 5206286360250189
71	97	84	+1	0. 9446912694798453
81	78	90	+1	0.9999999999997958
60	40	45	-1	-0. 9446912694798453
63	75	78	+1	0. 6772714555592905
60	62.7	65.7	+1	0. 0004862670271344

3.2.1.1 Entrenamiento del Modelo

Para el entrenamiento del modelo se pretendía utilizar el principio de Pareto, el cual indica que las entradas y salidas tiene una relación desigual, este explica que el 20% del refuerzo es responsable por 80% de los resultados; es decir que; el 80% de las consecuencias están dadas del 20% de las causas; sin embargo, tras

realizar una prueba con el 70% en datos de entrenamiento y 30% con datos de prueba se evidenció una mayor precisión dejando una distribución como se representa en la tabla 17.

Tabla 17 Distribución de datos para entrenamiento

	%	Cantidad
Datos de Entrenamiento	70	118
Datos de Prueba	30	64

Para el entrenamiento del modelo se usa el método logistic_classifier () de la librería turi, este es un paquete de Python que permite a los programadores realizar análisis de datos de gran escala de extremo a extremo, este método se utiliza para predecir la clase de una variable objetivo discreta (binaria o multiclase) basada en un modelo de probabilidad de clase como función logística de una combinación lineal de las características.

Una vez se tiene el modelo se pueden obtener los coeficientes utilizando el método coefficients el cual crea la tabla 18 donde se representan los pesos de cada coeficiente y se determina a qué clase tiene influencia.

Tabla 18 Peso de coeficientes del modelo de clasificación

Nombre de Coeficiente	# Valor
ECONOMIA	-0.03368391711600948
FÍSICA	0.022528114297520118
QUÍMICA	-0.015623545105073373
SOCIALES	0.03545297476478828
EDUCACION ARTISTICA	-0.01036769606464626
RECREACIÓN Y DEPORTES	0.007549084195047999
RELIGION	0.006570400129667652
ETICA	-0.11245415750266886
FILOSOFIA	0.06619867245258083
INGLES	-0.0009616540187199522
LENGUA CASTELLANA	0.03462140625584672
TRIGONOMETRIA	0.026553704483145903
ADMINISTRACION	-0.04582845137917302
MANEJO ORDENADORES	0.007335060116618236
PREPARACIÓN TECNOLÓGICA	0.01091936894226757
S	0.14733403326632408
R	-5.224278499682006

Los coeficientes positivos mayores que cero corresponden a pesos que causan notas positivas, mientras que los pesos negativos corresponden a notas negativas.

El coeficiente de correlación de Pearson, que se representa en la ecuación 14 la cual da como resultados valores entre -1 y 1, una buena correlación entre dos variables sugiere que existe algún tipo de dependencia entre ellas, si cambia una el cambio se verá reflejado en la otra.

La correlación es una relación matemática que se establece entre dos variables debido a su naturaleza de comportamiento, cuando el coeficiente es más pequeño que 0.6 o 0.64 es una correlación que es mera causalidad y no es relevante a partir de 0.68 y tendiendo a 1 es cuando la correlación tiene sentido, por otro lado, una correlación negativa será cuyo valor de coeficiente sea de –1 a 0 y significara que tiene una relación inversa entre las variables, por tanto cuando una incremente la otra decrece y si una decrece la otra incrementara para valores menores a - 0.6, esta se correlación se representa en la tabla 19, en la cual se puede identificar que las notas de cálculo no tienen mucha correlación con las otras materias

Ecuación 14 Coeficiente de Correlación de Pearson

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \ \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Tabla 19 Correlación de variables

	ECONOMIA	FÍSICA	QUÍMICA	SOCIALES	EDUCACION ARTISTICA	RECREACIÓN Y DEPORTES	RELIGION	ETICA	FILOSOFIA	INGLES	LENGUA CASTELLA	TRIGONOMETRIA	ADMINISTRACION	MANEJO ORDENADORE	PREPARACIÓN TECNOLÓGICA	CALCULO
ECONOMIA	1.000.000	0.65456	0.66266	0.568937	0.512738	0.419377	0.464416	0.492083	0.553461	0.682503	0.740452	0.539239	0.402085	0.630284	0.587749	0.41342
FÍSICA	0.65456	1.000.000	0.646565	0.51918	0.536394	0.525484	0.468681	0.53202	0.488587	0.697686	0.680097	0.639865	0.220494	0.615061	0.656335	0.438967
QUÍMICA	0.66266	0.646565	1.000.000	0.515767	0.522393	0.446326	0.447482	0.402374	0.541596	0.69092	0.669528	0.524839	0.30461	0.548704	0.59466	0.437705
SOCIALES	0.568937	0.51918	0.515767	1.000.000	0.538828	0.37575	0.556477	0.470238	0.620119	0.525302	0.689563	0.593761	-0.147766	0.520013	0.452037	0.438208
EDUCACION ARTISTICA	0.512738	0.536394	0.522393	0.538828	1.000.000	0.527466	0.497146	0.424762	0.556738	0.507173	0.60611	0.549845	-0.005689	0.486597	0.536888	0.410474
RECREACIÓN Y DEPORTES	0.419377	0.525484	0.446326	0.37575	0.527466	1.000.000	0.382537	0.279597	0.449259	0.486549	0.512944	0.538684	0.087713	0.486393	0.420001	0.237051
RELIGION	0.464416	0.468681	0.447482	0.556477	0.497146	0.382537	1.000.000	0.463187	0.478877	0.442039	0.589606	0.477554	0.049104	0.354483	0.429625	0.287764
ETICA	0.492083	0.53202	0.402374	0.470238	0.424762	0.279597	0.463187	1.000.000	0.497641	0.550391	0.533262	0.470689	0.183785	0.560148	0.494082	0.302786
FILOSOFIA	0.553461	0.488587	0.541596	0.620119	0.556738	0.449259	0.478877	0.497641	1.000.000	0.494767	0.675481	0.618068	-0.104232	0.520034	0.422126	0.447931
INGLES	0.682503	0.697686	0.69092	0.525302	0.507173	0.486549	0.442039	0.550391	0.494767	1.000.000	0.675294	0.575051	0.365003	0.639461	0.64183	0.457809
LENGUA CASTELLANA	0.740452	0.680097	0.669528	0.689563	0.60611	0.512944	0.589606	0.533262	0.675481	0.675294	1.000.000	0.688497	0.117295	0.659398	0.615378	0.4561
TRIGONOMETRIA	0.539239	0.639865	0.524839	0.593761	0.549845	0.538684	0.477554	0.470689	0.618068	0.575051	0.688497	1.000.000	-0.119174	0.660365	0.522556	0.473654
ADMINISTRACION	0.402085	0.220494	0.30461	-0.147766	-0.005689	0.087713	0.049104	0.183785	-0.104232	0.365003	0.117295	-0.119174	1.000.000	0.244827	0.247979	0.069108
MANEJO ORDENADORES	0.630284	0.615061	0.548704	0.520013	0.486597	0.486393	0.354483	0.560148	0.520034	0.639461	0.659398	0.660365	0.244827	1.000.000	0.632746	0.467335
PREPARACIÓN TECNOLÓGICA	0.587749	0.656335	0.59466	0.452037	0.536888	0.420001	0.429625	0.494082	0.422126	0.64183	0.615378	0.522556	0.247979	0.632746	1.000.000	0.416741
CALCULO	0.41342	0.438967	0.437705	0.438208	0.410474	0.237051	0.287764	0.302786	0.447931	0.457809	0.4561	0.473654	0.069108	0.467335	0.416741	1.000.000

3.2.1.2 Evaluación del Modelo

Para evaluar el modelo es importante observar los tipos de errores puede cometer un clasificador y una forma de hacerlo es a través de lo que se llama la matriz de confusión, que en el campo de la inteligencia artificial es una herramienta que permite la visualización del desempeño de un algoritmo que se emplea en aprendizaje supervisado.[25] Cada columna de la matriz representa el número de predicciones de cada clase, en el caso actual hay 2 clases, el estudiante pierde la materia o el estudiante pasa la materia, mientras que cada fila representa a las instancias en la clase real.

Se busca identificar la relación entre el estudiante que pasa la materia y cualquiera que sea el clasificador que prediga, entonces, si el estudiante que pasa la materia, pasa la materia y se predice que el estudiante pasa la materia, se llama un verdadero positivo porque se predijo bien; del mismo modo, si el estudiante que pierde la materia y se predice que la pierde se llama un verdadero negativo.

Hay dos tipos de errores que se pueden cometer, si el estudiante que pasa la materia, pasa la materia, pero se predijo que perdería a este se le llama falso negativo; igualmente, si el estudiante que pasa la materia, pierde, cuando se predijo que pasaría a esto se le llama falso positivo.

De lo anteriormente mencionado en la tabla 20 se identifican cuatro casos distintos para los datos de entrenamiento

Tabla 20 Matriz de confusión con los datos de entrenamiento

82	3	Aprueba
5	28	Reprueba

[25] APRENDIZAJE AUTOMATICO: Matriz de confusión [en línea]. Wikipedia. [Consultado: 7 de mayo de 2020]. Disponible en: https://es.wikipedia.org/wiki/Matriz_de_confusi%C3%B3n

Como resultado del método utilizando los datos de entrenamiento y los datos de prueba se obtienen los siguientes resultados:

La exactitud es de 0.9322033898305084

La precisión del modelo es de 0.9425287356321839

La sensibilidad o recall es de 0.9647058823529412

La tabla de confusión con los datos de pruebas se representa en la tabla 21.

Tabla 21 Matriz de confusión con los datos de pruebas

46	5	Aprueba
1	12	Reprueba

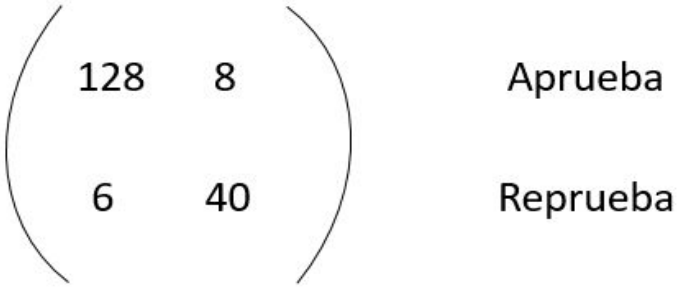
La exactitud es de 0.90625

La precisión del modelo es de 0.9787234042553191

La sensibilidad o recall es de 0.9019607843137255

Consolidando los valores de las tablas 20 y 21, se obtiene la tabla 22.

Tabla 22 Matriz de confusión consolidada



Teniendo en cuenta los datos anteriormente mencionados se puede visualizar en la tabla 23 otras métricas de evaluación que incrementan la capacidad para discriminar los casos positivos, de los casos negativos; estas son, la sensibilidad que es la proporción de casos positivos que fueron correctamente identificados por el algoritmo y la especificad que son los casos negativos que el algoritmo ha clasificado correctamente. Las filas de la matriz

Para la evaluación del modelo se usa el método evaluate de la librería turi, el cual internamente realiza una matriz de confusión, este método recibe dos vectores, uno donde se encuentran los datos de la clase de verdad básica y en otro se encuentra la predicción que corresponde al valor objetivo, como resultado un SFrame el cual es marco de datos escalable, que permite trabajar con conjuntos de datos que son más grandes que la cantidad de RAM en el sistema, este contiene recuentos de ‘variable objetivo’, ‘predicción’ y ‘conteo’ correspondiente a cada par de etiquetas verdaderas y predichas.

Como resultado del método utilizando los datos de entrenamiento y los datos de prueba se obtienen los siguientes resultados:

La exactitud dada es 0.9230769230769231 la cual es el número de predicciones correctas realizadas por el modelo total de registros.

La precisión del modelo es de 0.9552238805970149 en esta se evalúan los datos por el desempeño de predicciones positivas.

La sensibilidad o recall es de 0.9411764705882353, se calcula como el número de predicciones positivas correctas dividido por el número de positivo, es una medida que dice que la proporción de estudiantes que pasaron fue predicho por el algoritmo como si hubiera pasado.

La especificidad es exactamente lo contrario a la sensibilidad es decir que da un resultado de 0.10227231 es la tasa negativa verdadera se calcula como el número de predicciones negativas correctas dividido por el número total de negativos, esta medida permite obtener la proporción de estudiantes que no pasaron fueron predichos por el modelo como que no pasarían.

Cuando la precisión está por arriba del 0.9 se puede determinar que es fiable el modelo de predicción.

Los errores en la clasificación no se pueden medir en términos de suma cuadra como en la regresión ya que las entradas están dadas en valores de -1 o +1. El algoritmo clasificador lo que hace es aprender los pesos de los estudiantes, entonces, realizando un proceso de interacción aprenderá que un estudiante con una nota de 47 es malo y uno con una nota de 85 es bueno, luego, estos pesos se utilizan para calificar cada elemento en el conjunto de datos de prueba y así entrenar identificando qué tan bien se está haciendo en términos de clasificación.

Para realizar el análisis del error se utiliza la ecuación 14 esta se mide tomando la cantidad total de errores y dividiéndolo por el número total de estudiantes de prueba. Entonces, por ejemplo, si se tienen 100 estudiantes de prueba y se cometen 10 errores el error sería 0.1 o 10%.

Ecuación 15 Ecuación de error

$$\varepsilon = \frac{\#\ de\ Errores}{\#\ de\ Estudiantes}$$

En la evaluación del modelo, de los 84 registros se encontraron 37 errores; por lo tanto, el error es de 0.440, esto se debe a que en la conversión de los datos a valores categóricos se pierde precisión.

Tabla 23 Evaluación modelo clasificación

CÁLCULO	CONVENCIONES MINISTERIO	PREDICCIÓN	ASERTÓ
60.8003	BÁSICO	Perdió	0
64.1944	BÁSICO	Perdió	0
57.13	BAJO	Perdió	0
63.8657	BÁSICO	Perdió	0
60.4465	BÁSICO	Perdió	0
72.2251	BÁSICO	Perdió	0
69.7935	BÁSICO	Pasó	1
67.7787	BÁSICO	Pasó	1
60.2474	BÁSICO	Pasó	1
59.9482	BAJO	Pasó	0
63.4423	BÁSICO	Pasó	1
60.7093	BÁSICO	Perdió	0
72.0214	BÁSICO	Pasó	1
56.5364	BAJO	Perdió	1
64.8629	BÁSICO	Pasó	1
59.8876	BAJO	Pasó	0
61.2936	BÁSICO	Perdió	0
54.3635	BAJO	Perdió	1
67.0403	BÁSICO	Perdió	0
74.0158	BÁSICO	Pasó	1
65989	BÁSICO	Pasó	1
61.7026	BÁSICO	Perdió	0
58.8481	BAJO	Pasó	1
70.7333	BÁSICO	Pasó	1
61.6229	BÁSICO	Perdió	0
69.1201	BÁSICO	Pasó	1
73.8115	BÁSICO	Pasó	1
50.9092	BAJO	Pasó	0
66.3807	BÁSICO	Pasó	1
69.4482	BÁSICO	Perdió	1
63305	BÁSICO	Perdió	0
62627	BÁSICO	Perdió	0
64888	BÁSICO	Pasó	1
57.0634	BAJO	Pasó	0
66.4584	BÁSICO	Pasó	1

74019	BÁSICO	Pasó	1
66222	BÁSICO	Pasó	1
70.3463	BÁSICO	Pasó	1
71.0983	BÁSICO	Perdió	0
66.1992	BÁSICO	Perdió	0
67.3784	BÁSICO	Pasó	1
62.9677	BÁSICO	Perdió	0
61.8979	BÁSICO	Perdió	0
60.0747	BÁSICO	Perdió	0
66.4957	BÁSICO	Pasó	1
65.3857	BÁSICO	Perdió	0
65.0651	BÁSICO	Pasó	1
56.0474	BAJO	Perdió	1
57.7138	BAJO	Pasó	0
59.9928	BAJO	Pasó	0
53.3018	BAJO	Pasó	0
75.1906	BÁSICO	Pasó	1
76.5446	BÁSICO	Pasó	1
66.5244	BÁSICO	Perdió	0
66.2359	BÁSICO	Perdió	0
47.4659	BAJO	Perdió	1
84063	ALTO	Pasó	1
65.8469	BÁSICO	Pasó	1
66.7047	BÁSICO	Pasó	1
63.3294	BÁSICO	Pasó	1
59.8404	BAJO	Pasó	0
70.1331	BÁSICO	Pasó	1
66.4735	BÁSICO	Pasó	1
69.4948	BÁSICO	Pasó	1
67.3688	BÁSICO	Pasó	1
66.6379	BÁSICO	Perdió	0
69.6209	BÁSICO	Pasó	1
77.2077	BÁSICO	Pasó	1
64.1586	BÁSICO	Perdió	0
67.1611	BÁSICO	Pasó	1
65.5652	BÁSICO	Pasó	1
64.9551	BÁSICO	Pasó	1
63.8978	BÁSICO	Perdió	0
65.5764	BÁSICO	Pasó	1

70.0419	BÁSICO	Pasó	1
83.9926	ALTO	Perdió	0
64.0278	BÁSICO	Pasó	1
67.9235	BÁSICO	Pasó	1
66.8814	BÁSICO	Perdió	0
66.8938	BÁSICO	Perdió	0
63.8184	BÁSICO	Perdió	0
63.7075	BÁSICO	Perdió	0
67.3589	BÁSICO	Pasó	1
69349	BÁSICO	Pasó	1

La columna ‘ASERTÓ’ indica 1 si la predicción fue correcta y 0 si no. Lo cual se concluye que de los 84 registros evaluados en 47 registros se obtuvo la predicción correcta y en 37 se falló.

3.2.2 Redes Neuronales Perceptrón Multicapa

Para el desarrollo del modelo, se utilizaron los mismos datos que han sido depurados y revisados en el capítulo de regresión logística, adicionalmente se trabaja con una separación de 70% para datos de entrenamiento y 30% para datos de prueba.

Se usa la herramienta de análisis Weka en su versión 3.8.

Figura 11 Interfaz de Weka

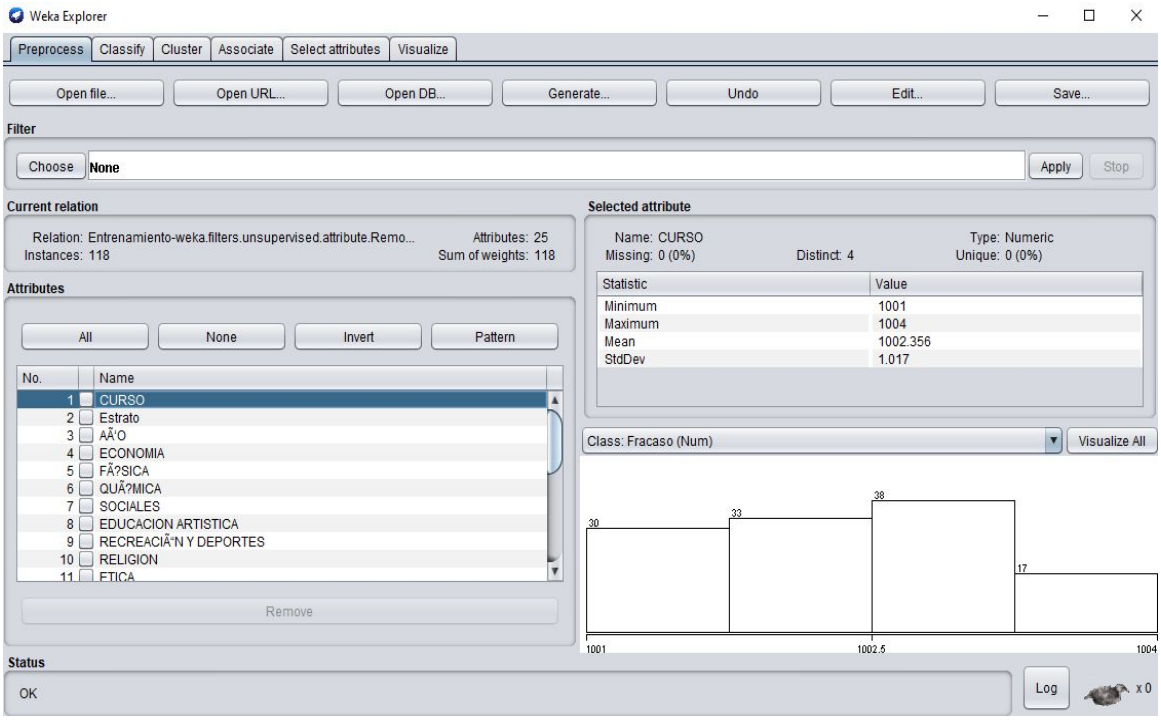
Se crean dos documentos en formato .csv uno para los datos de entrenamiento y el otro para los datos de prueba, ya que Weka trabaja con formato .arff, se realiza la conversión de los datos utilizando el ArttViewer una de las herramientas de Weka.

En la interfaz principal de Weka que se visualiza en la figura 11 se selecciona la opción Explorer.

3.2.2.1 Entrenamiento y Evaluación del Modelo

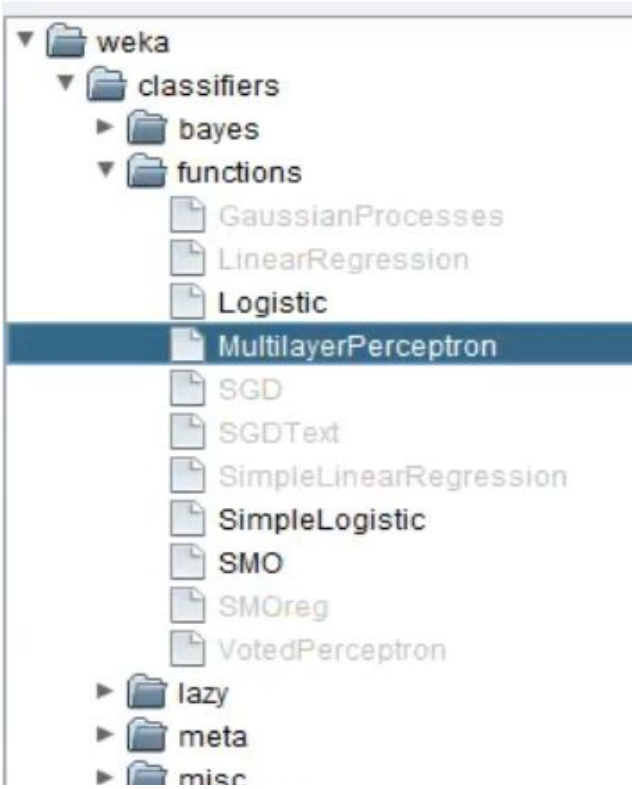
En el weka explorer que se visualiza en la figura 12 se realiza el cargue de los datos seleccionado la opción Open File y buscando el archivo Entrenamiento.arff

Figura 12 Weka Explorer



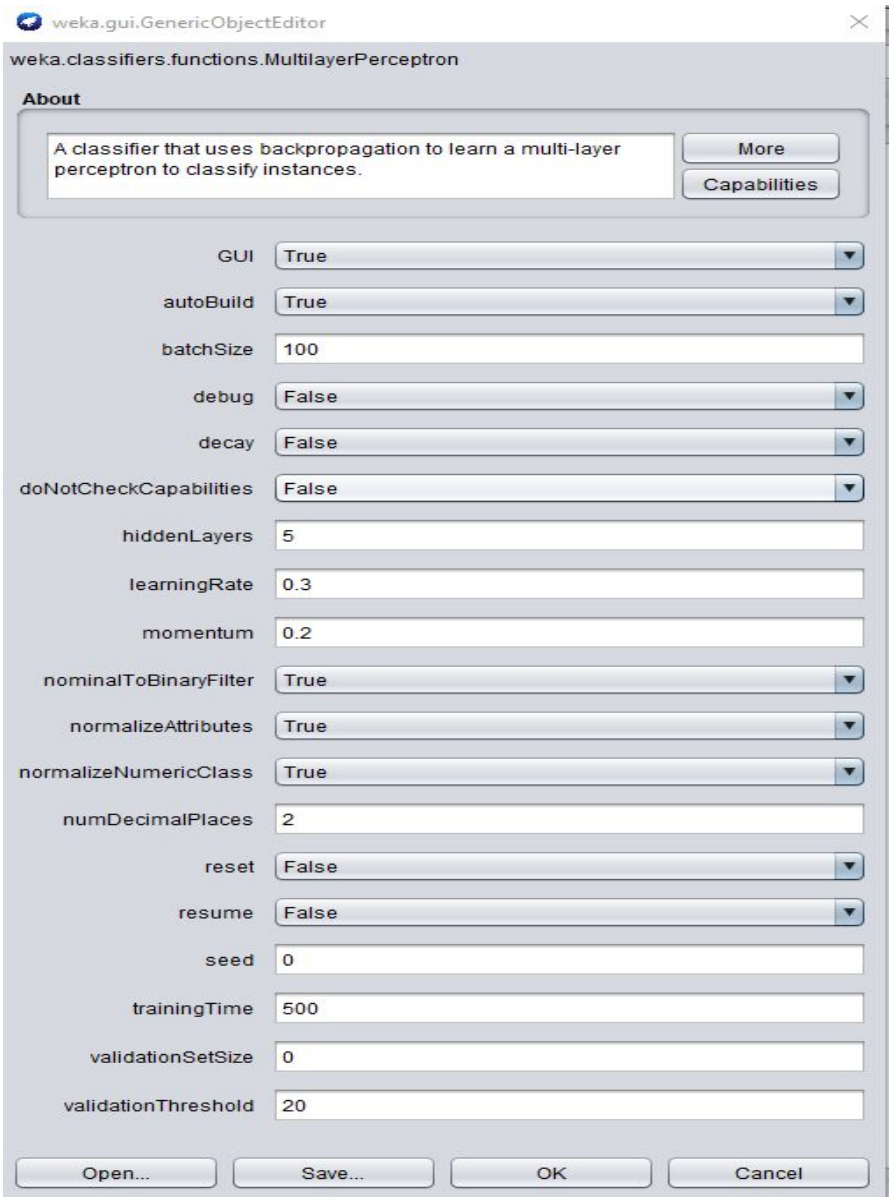
En la pestaña de clasificación se selecciona el método de redes neuronales con perceptrón multicapa como se representa en la figura 13.

Figura 13 MultilayerPerceptron



A continuación, se establece la configuración de la red neuronal como se visualiza en la figura 14 dejando como true el valor de GUI para que se represente una interfaz gráfica, se establecen como capas ocultas el valor de 5.

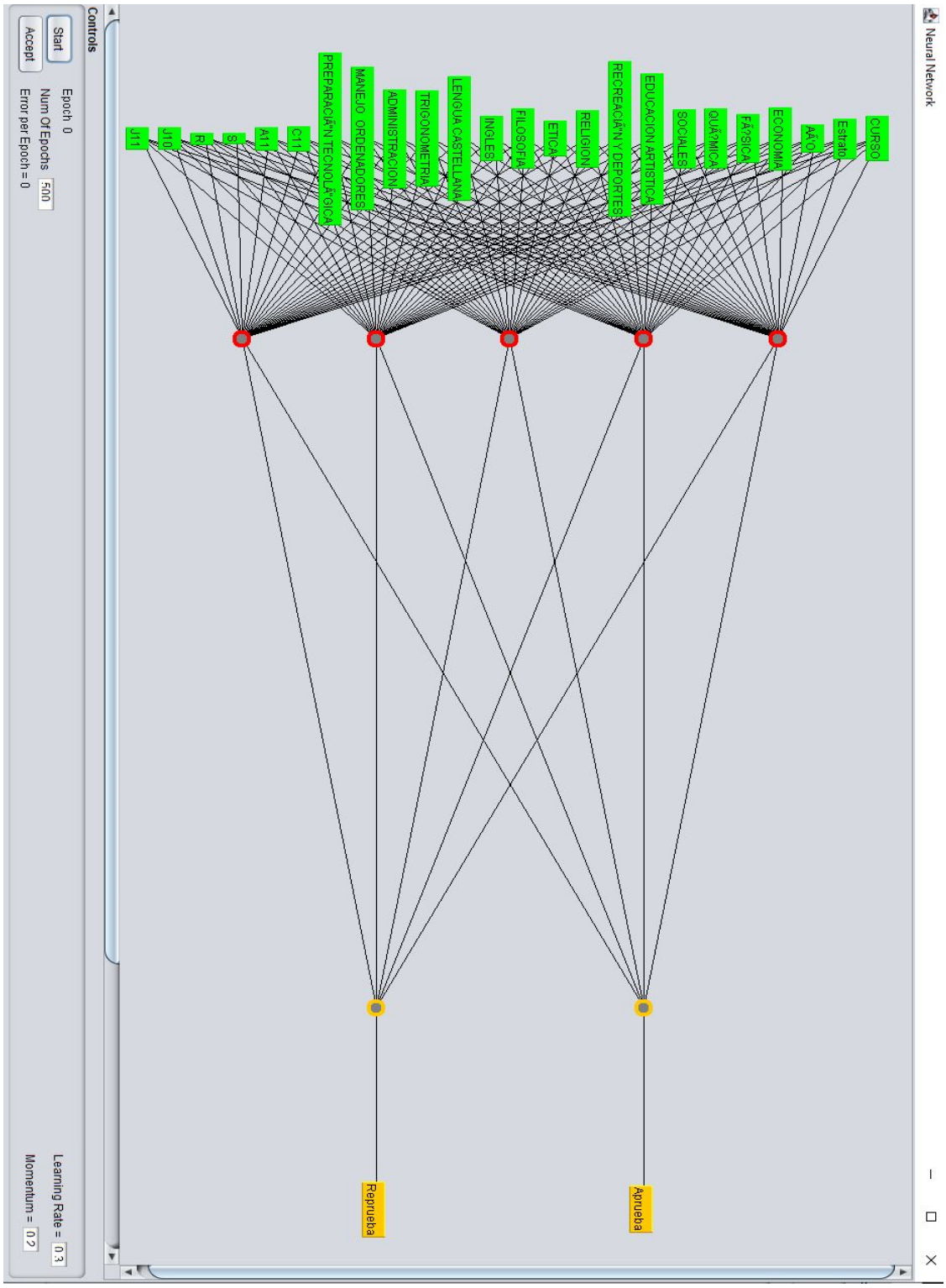
Figura 14 Funciones de MultilayerPerceptron



Luego se selecciona la opción para usar el entrenamiento de los datos una vez seleccionado se da clic en start

Las capas ocultas se determinan como 5 y se deja true el valor para visualizar la representación gráfica de la red neuronal que se evidencia en la figura 15 donde se tiene dos salidas y cinco capas ocultas, luego se selecciona en el botón start donde se hace una iteración de 500 épocas y se obtiene el resultado del error.

Figura 15 Representación gráfica



Como resultado se obtiene la figura 16 donde se clasificó correctamente el 97% de las 118 instancias disponibles.

Figura 16 Resultados con datos de entrenamiento

```
Time taken to build model: 85.24 seconds

=== Evaluation on training set ===

Time taken to test model on training data: 0 seconds

=== Summary ===

Correctly Classified Instances         115               97.4576 %
Incorrectly Classified Instances         3                2.5424 %
Kappa statistic                          0.9351
Mean absolute error                      0.038
Root mean squared error                  0.1598
Relative absolute error                  9.3986 %
Root relative squared error             35.598  %
Total Number of Instances              118

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1,000    0,091    0,966      1,000    0,983      0,937    0,947     0,969     Aprueba
                 0,909    0,000    1,000      0,909    0,952      0,937    0,947     0,945     Reprueba
Weighted Avg.    0,975    0,065    0,975      0,975    0,974      0,937    0,947     0,963

=== Confusion Matrix ===

  a  b   <-- classified as
 85  0 |  a = Aprueba
  3 30 |  b = Reprueba
```

Se puede concluir que este modelo presenta un buen desempeño ya que el margen de error es de solo 2.5 % ya que solo en 3 casos se clasificó mal el pronóstico.

Para utilizar los datos de prueba que se encuentran en el archivo Prueba.arff se selecciona la opción Supplied test set donde se visualizará nuevamente la representación gráfica y realizando nuevamente 500 épocas se obtiene un nuevo resultado como se puede apreciar en la figura 17.

Figura 17 Resultados con datos de pruebas

```
Time taken to build model: 60.71 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0 seconds

=== Summary ===

Correctly Classified Instances          55               85.9375 %
Incorrectly Classified Instances         9               14.0625 %
Kappa statistic                          0.5528
Mean absolute error                      0.1428
Root mean squared error                  0.3558
Relative absolute error                 38.457  %
Root relative squared error             86.7274 %
Total Number of Instances               64

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,922    0,385    0,904      0,922    0,913      0,553    0,807     0,935     Aprueba
                 0,615    0,078    0,667      0,615    0,640      0,553    0,807     0,693     Reprueba
Weighted Avg.    0,859    0,322    0,856      0,859    0,857      0,553    0,807     0,886

=== Confusion Matrix ===

  a  b   <-- classified as
 47  4 |  a = Aprueba
  5  8 |  b = Reprueba
```

A comparación del resultado anterior en este modelo con los datos de prueba se tiene una 86 % de validación fue correcta y con un error de 14 %

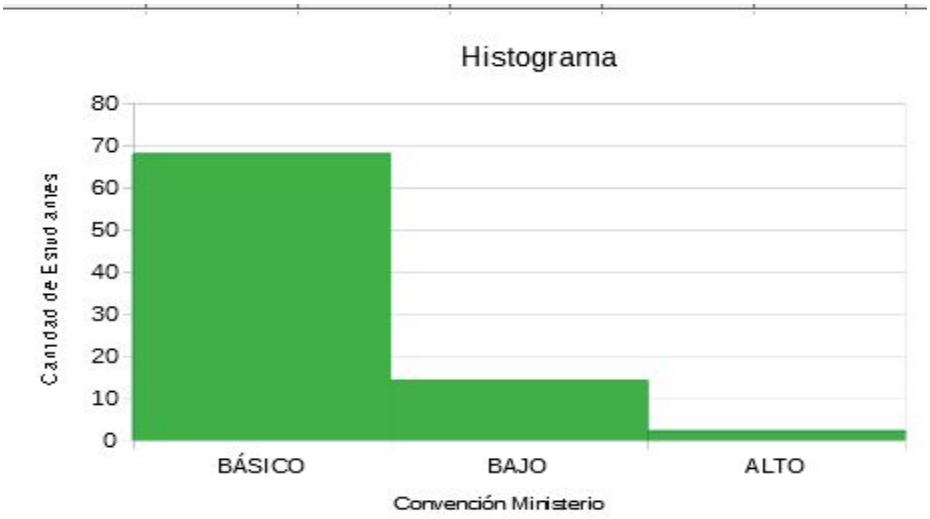
4. COMPARACIÓN DE LOS MODELOS

En esta fase se busca comparar los resultados de los modelos de regresión y clasificación, con el objetivo de determinar cual tiene un mejor desempeño, esto se realiza de la siguiente forma:

Una vez que se tienen construidos los modelos se busca elegir el que presenta una mayor calidad desde la perspectiva del análisis de datos, para evaluar cada uno de los modelos se utilizó la sábana de datos de las notas de décimo del año 2018 para predecir la nota en cálculo, estos datos no intervinieron en ninguna parte del proceso de entrenamiento y evaluación ya que se busca encontrar el comportamiento real de la predicción.

Antes de comparar los modelos es necesario entender el comportamiento de los datos con los que se va a realizar la comparación. Para ello la figura 18 muestra la cantidad de estudiantes por cada rango definido por el ministerio. En el histograma se puede observar que 68 estudiantes pasaron cálculo en la escala de BÁSICO, 14 la perdieron y solo 2 la pasaron con ALTO.

Figura 18 Histograma notas cálculo 2019



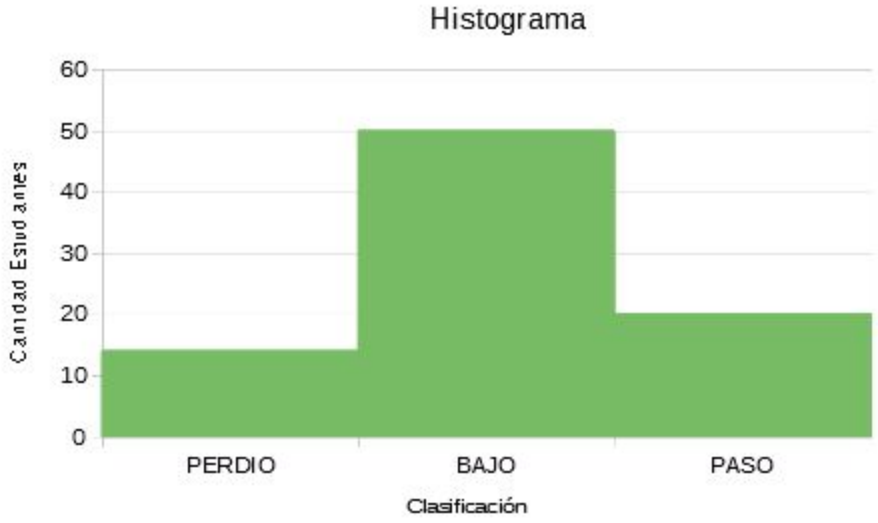
De los datos de cálculo de los estudiantes de 2019 los rangos que interesan para el modelo son de [0-70) ya que son estudiantes que perdieron la materia o la pasaron con mucha dificultad. La Tabla 20 muestra la nueva escala con la cual se va a calificar a los estudiantes para la evaluación de los modelos.

Tabla 24 Clasificación rango notas

Clasificación	Rango Notas
PERDIÓ	[0-60)
BAJO	[60-70)
PASÓ	[70-100]

Con la información de la Tabla 20 se construye un nuevo histograma donde se evidencia una gran cantidad de estudiantes que pasaron con una nota muy baja lo que impactaría a la hora de entrar en una universidad ya que sus habilidades en matemáticas son muy bajas.

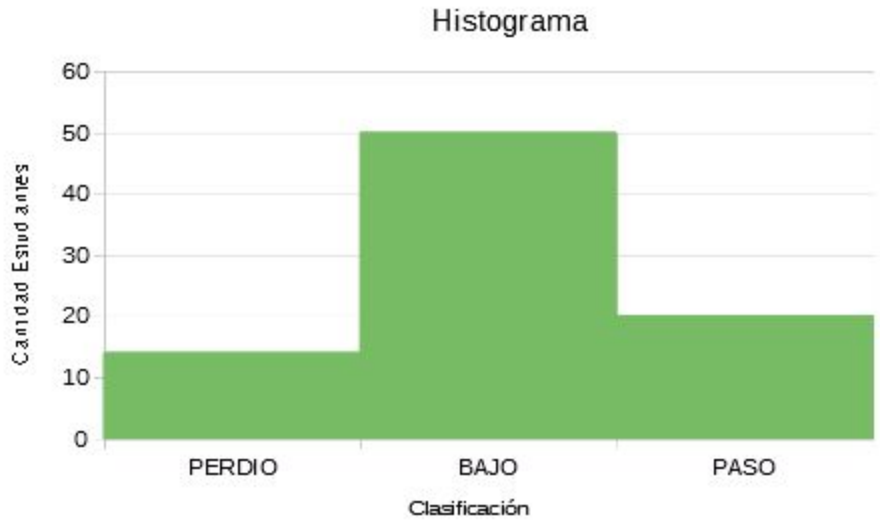
Figura 19 Histograma para evaluación modelos



De la figura 19 se puede concluir que 50 estudiantes, aunque pasaron tienen una nota por debajo de 70 lo cual indica que se les debe dar un seguimiento, y no descartarlos en los modelos construidos.

Teniendo la consideración de los nuevos rangos y los resultados de la tabla 13 el histograma para la predicción del modelo utilizando Regresión se muestra en la figura 20.

Figura 20 Histograma modelo regresión



Con el modelo de regresión se puede concluir que la predicción es exacta.

Tabla 25 Notas de cálculo con su predicción regresión

CÁLCULO	PREDICCIÓN	CLASIFICACIÓN
60,8003	60,5663	BAJO
64,1944	63,9514	BAJO
57,13	56,912	PERDIÓ
63,8657	63,6451	BAJO
60,4465	60,286	BAJO
72,2251	71,9994	PASO
69,7935	69,544	BAJO
67,7787	67,5147	BAJO
60,2474	60,0135	BAJO
59,9482	59,7228	PERDIÓ

63,4423	63,2016	BAJO
60,7093	60,4737	BAJO
72,0214	71,7664	PASO
56,5364	56,2972	PERDIÓ
64,8629	64,6037	BAJO
59,8876	59,7129	PERDIÓ
61,2936	61,0855	BAJO
54,3635	54,1383	PERDIÓ
67,0403	66,7923	BAJO
74,0158	73,7656	PASO
65,989	65,7758	BAJO
61,7026	61,4841	BAJO
58,8481	58,5998	PERDIÓ
70,7333	70,4654	PASO
61,6229	61,4007	BAJO
69,1201	68,8849	BAJO
73,8115	73,5813	PASO
50,9092	50,6815	PERDIÓ
66,3807	66,1482	BAJO
69,4482	69,3774	BAJO
63305	63,2928	BAJO
62627	62,6441	BAJO
64888	64,8482	BAJO
57,0634	57,0135	PERDIÓ
66,4584	66,4228	BAJO
74019	73,9669	PASO
66222	66,1779	BAJO
70,3463	70,3295	PASO
71,0983	71,0439	PASO
66,1992	66,1614	BAJO
67,3784	67,3444	BAJO
62,9677	62,9549	BAJO
61,8979	61,8948	BAJO
60,0747	60,0636	BAJO
66,4957	66,4751	BAJO
65,3857	65,3565	BAJO

65,0651	65,0649	BAJO
56,0474	56,0463	PERDIÓ
57,7138	57,6565	PERDIÓ
59,9928	59,9883	PERDIÓ
53,3018	53,2873	PERDIÓ
75,1906	75,1522	PASO
76,5446	76,4645	PASO
66,5244	66,4821	BAJO
66,2359	66,2223	BAJO
47,4659	47,4913	PERDIÓ
84063	83988	PASO
65,8469	65,8576	BAJO
66,7047	66,6764	BAJO
63,3294	63,3272	BAJO
59,8404	59,8091	PERDIÓ
70,1331	70,1122	PASO
66,4735	66,4634	BAJO
69,4948	69,4818	BAJO
67,3688	67,3694	BAJO
66,6379	66,6364	BAJO
69,6209	69,5956	BAJO
77,2077	77,1501	PASO
64,1586	64,1578	BAJO
67,1611	67,1318	BAJO
65,5652	65,5308	BAJO
64,9551	64,9353	BAJO
63,8978	63,8601	BAJO
65,5764	65,5395	BAJO
70,0419	70,0235	PASO
83,9926	83,9389	PASO
64,0278	64017	PASO
67,9235	67921	PASO
66,8814	66,8801	BAJO
66,8938	66,9066	BAJO
63,8184	63,8151	BAJO
63,7075	63,701	BAJO

67,3589	67,3382	BAJO
69,349	69,302	BAJO

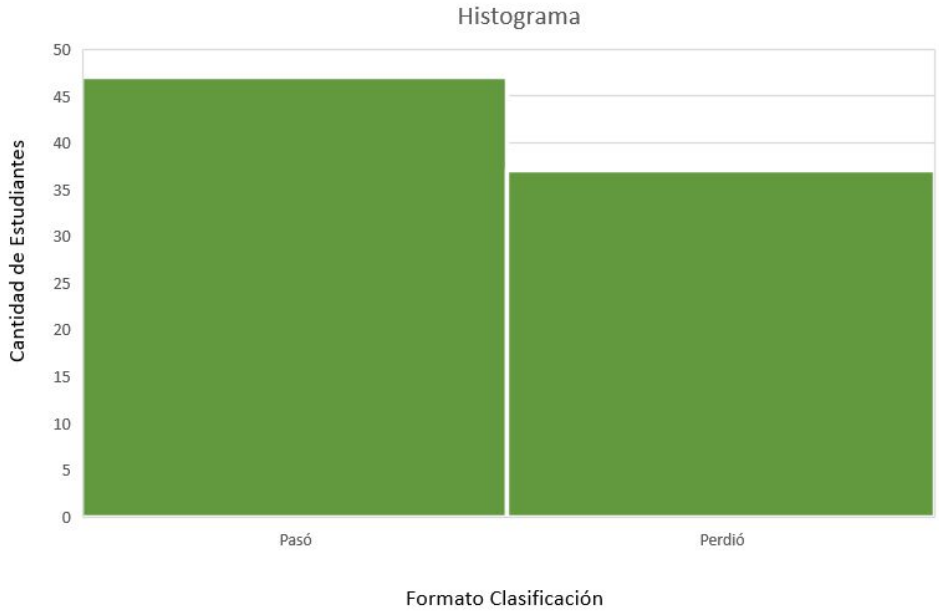
En la tabla 21 se encuentra la predicción realizada aplicando el modelo de regresión logística donde se clasifican los estudiantes.

Tabla 26 Clasificación de estudiantes

Formato Clasificación	Frecuencia
PERDIÓ	47
PASÓ	37

Teniendo en cuenta la información de la Tabla 22 se construye el histograma de la figura 21 donde se evidencia que un poco más del 60% de los estudiantes pasaron la asignatura y el porcentaje restante la perdieron o no tienen conocimientos sólidos sobre la asignatura.

Figura 21 Histograma modelo clasificación



Con el modelo de regresión logística se puede concluir que la predicción agregó 20 nuevos estudiantes a la clasificación de “PERDIÓ” lo cual prácticamente resulta real ya que la nota está entre [60-70], se presenta valor atípico en la nota 47.4659 el cual no se tiene en cuenta dentro del rango.

Tabla 27 Notas de cálculo con su predicción clasificación

CÁLCULO	PREDICCIÓN
60.8003	PERDIÓ
64.1944	PERDIÓ
57.13	PERDIÓ
53.8657	PERDIÓ
60.4465	PERDIÓ
60.7093	PERDIÓ
56.5364	PERDIÓ
61.2936	PERDIÓ
54.3635	PERDIÓ
61.7026	PERDIÓ
61.6229	PERDIÓ
63.305	PERDIÓ
62.627	PERDIÓ
62.9677	PERDIÓ
61.8979	PERDIÓ
60.0747	PERDIÓ
55.3857	PERDIÓ
56.0474	PERDIÓ
56.5244	PERDIÓ
47.4659	PERDIÓ
66.6379	PERDIÓ
64.1586	PERDIÓ
53.8978	PERDIÓ
53.9926	PERDIÓ
53.8184	PERDIÓ
53.7075	PERDIÓ

El resultado de la comparación de los 85 registros, da como resultado la tabla 28 donde se evidencia un mejor resultado en el algoritmo de regresión al tener 56 predicciones reales respecto a 47 del algoritmo de clasificación.

Tabla 28 Resultado de comparación de modelos

Total Regresión	Total Clasificación
84	80

5. CONCLUSIONES Y RECOMENDACIONES

- En este proyecto se desarrolló un modelo sobre un colegio público de Bogotá con metodología presencial al cual se le hizo el análisis de notas, logrando observar que los estudiantes tienen bajas calificaciones en matemáticas en comparación con otras asignaturas.

- El desarrollo de un modelo predictivo se facilita y agiliza utilizando las librerías de Python, lo que permite que los colegios interesados que cuentan con pocos recursos económicos viabilicen soluciones tecnológicas que respondan a las necesidades particulares de cada institución.

- Al finalizar la fase de comparación de modelos se puede concluir que se logró alcanzar los objetivos y satisfacer las necesidades impuestas en el proyecto permitiendo obtener un modelo predictivo para determinar el fracaso de matemáticas en estudiantes de grado once.

- El método LASSO deja en 0 los coeficientes que aportan poco a la predicción, para este caso todos los coeficientes quedaron con un valor pequeño ya que en los modelos donde había coeficientes en 0 aumentaba el error.

- Con el método de regresión se logró una predicción correcta del 67%, el otro 33% aunque no fue exacta sirve para tener en cuenta a aquellos estudiantes que pasaron con una baja nota, y requieren refuerzo.

- El algoritmo de clasificación presenta un mayor rendimiento cuando la mayoría de sus variables son categóricas y no cuantitativas; por tal razón, el algoritmo de regresión presenta un mejor modelo predictivo.

- En trabajos futuros se puede enfocar en realizar el modelo predictivo usando otros algoritmos de aprendizaje supervisado.

- Una de las recomendaciones a futuro para mejorar la experiencia del usuario final de la predicción, es la creación de una aplicación web o móvil, en la que se presenten los datos de forma sencilla y fácil de interpretar.

- Otra de las recomendaciones a considerar para hacer la predicción más precisa, sería necesario de disponer de información relativa a otras causas de deserción como, por ejemplo, ingreso mensual familiar, cantidad de hermanos, embarazo, baja motivación, discapacidad, bullying, religión y muchos otros.

- Hacer un modelo utilizando datos de la Universidad Distrital para evaluar el desempeño de los profesores de cálculo, el comportamiento de eficiencia de los estudiantes en la materia en los diferentes horarios del día y la frecuencia de asistencia de tutorías.

6. LISTA DE ANEXOS

Anexo A Sábana de Datos de Entrenamiento.

DataEstudiantes.txt (/Anexo/DataEstudiantes.txt)

Anexo B Sábana de Datos de Evaluación.

DatosEvaluacionModelo.xls (/Anexo/DatosEvaluacionModelo.xls)

Anexo C Implementación Método Regresión.

ModeloRegresionDataEstudiantes.ipynb(/Anexo/ModeloRegresionDataEstudiantes.ipynb)

Anexo D Implementación Modelo Clasificación.

ModeloClasificacionDataEstudiantes.ipynb(/Anexo/ModeloClasificacionDataEstudiantes.ipynb)

Anexo E Resultado de las 2000 Iteraciones Usando el Método de Regresión.

ResultadoModeloL1.xls (/Anexo/ResultadoModeloL1.xls)

INFOGRAFÍA

UNIVERSIDAD DE CORUÑA. [en línea] Madrid. Cómo evitar el fracaso escolar, [Consultado: 5 de mayo de 2020]. Disponible en: https://www.serpadres.es/familia/tiempo-libre/articulo/evitar-fracaso-escolar

UNIVERSIDAD DEL TOLIMA. [en línea] Ibagué. Diseño de estrategias de retención para disminuir la deserción escolar de estudiantes del grado sexto del instituto politécnico de Bucaramanga [en línea] [Consultado: 7 de febrero de 2020] Disponible en: http://45.71.7.21/handle/001/1154

UNIVERSIDAD NACIONAL DE QUILMES. [en línea] Argentina. Relaciones entre composición estudiantil, proceso escolar y el logro en matemáticas en la educación secundaria en Argentina [Consultado: 13 de marzo de 2020] Disponible en:
http://www.scielo.org.mx/scielo.php?pid=S160704120030001000004&script=sci_abstract&tlng=en

PLATAFORMA VIRTUAL COURSERA. [en línea] Universidad de Stanford. Lasso Regression: Regularization for feature selection [Consultado: 9 de enero de 2020], Disponible
en:https://d3c33hcgiwev3.cloudfront.net/_ad29bd587ef6b352cec39e3b35db83c1_week5_lassoregressionAnnotated.pdf?
Expires=1589328000&Signature=KX6MkXfsmbBL6FSJJyVa6N4pK71xBuWcGwhpV17v8jM2g0d7WBa1~SfmmtLnOMOxrHeddT81BlsVcWNq04UzNZoOFG5O4bIEyhG5YEAWknXWvnpZ1Qg3FQgTXFDutsI8_&KeyPairId=APKAJLTNE6QMUY6HBC5A

PLATAFORMA VIRTUAL COURSERA. [en línea] Universidad de Stanford. Multiple Regresion: Linear regression with multiple features [Consultado: 23 de enero de 2020], Disponible en: https://d3c33hcgiwev3.cloudfront.net/_16e63b9e55b963ed3eede1759991b812_week2_multipleregressionannotated.pdf?Expires=1589414400&Signature=BzfkCBiqCEBBKPsEY~bfQyI6EIsEcKO2VtZSHdOtFfZvGE6VLrQJFpz2yiyy9liGzwJ7B0nzQliwDh6uxpMrV0hBNYYKlGR3n3M~b7sORCJIguVWGdRc9w5poo6iYfj-MnmBBOguU_&Key-Pair-Id=APKAJLTNE6QMUY6HBC5A

BIBLIOGRAFÍA

- SHEARER C. (2015), el modelo CRISP-DM: el nuevo plan para la minería de datos, almacenamiento de los datos J; 5:13-22.
- INSTITUTO COLOMBIANO DE NORMAS TÉCNICAS. Normas Colombianas para la presentación de tesis, trabajos de grado y otros trabajos de Investigación. Quinta actualización. Bogotá: ICONTEC, 2002.
- MARTELLI, Alex (2007). Python. Guía de referencia. Gorjón Salvador, Bruno (1 edición). Anaya Multimedia-Anaya Interactiva.
- EIBE Frank (2011). Data Mining: Practical machine learning tools and techniques Morgan Kaufmann, 664 p.
- HAN, Richard (2019). Matemáticas del Aprendizaje Automático: Introducción a la analítica de datos e inteligencia artificial, 215 p. ISBN: 1731265387
- JONES, Herbert (2019). Ciencia de Los Datos: Lo Que Saben Los Mejores Científicos de Datos Sobre el análisis de Datos, Minería de Datos, Estadísticas, Aprendizaje Automático y Big Data - Que Usted Desconoce, 134 p. ISBN: 1797989243

- IAN Davidson (2007). Knowledge Discovery and Data Mining: Challenges and Realities. Hershey, New Your. 108 p.
- JAUME ARNAU, Constantino y ATO GARCIA, Manuel E. Métodos y Técnicas Avanzadas de Análisis de Datos en Ciencias del Comportamiento. 2 ed. España: Publicacions I Edicions De La Universitat De Barcelona. 2000, 366 p. ISBN: 849220043.
- WILKINSON, Christopher. Ciencia de Datos Python: Una guía definitiva para que los principiantes aprendan los fundamentos de la ciencia de datos con Python. 1 ed. Washington: Independiente. 2020, 207 p. ISBN: 9781654192389.

Anexo F Formato trabajo de grado culminado cumplimiento de objetivos.

ResultadoModeloL1.xls (/Anexo/ FORMATO TRABAJO DE GRADO CULMINADO CUMPLIMIENTO DE OBJETIVOS.pdf)

Bogotá D. C., 25 de mayo de 2020

Señores:

CONSEJO CURRICULAR

TECNOLOGÍA EN SISTEMATIZACIÓN DE DATOS E INGENIERÍA EN TELÉMATICA

Universidad Distrital F.J.C

Bogotá

Respetados Señores:

En cumplimiento del reglamento de trabajo de grado, atentamente me permito presentar el trabajo de grado titulado modelo predictivo para determinar el fracaso de matemáticas en grado 11 usando machine LEARNING, en la modalidad de monografía, desarrollado por los estudiantes Omar Alvarado Castillo, 20172678007, 1010200410 y Santos Miguel Zambrano Saavedra, 20171678029, 1022986833, para optar al título de ingeniero en telemática, el cual cumple con los objetivos de la propuesta inicial.

Cordialmente,

Jairo Hernández Gutiérrez
Docente Universidad Distrital
Facultad Tecnológica