
# Universidad de Bogotá Jorge Tadeo Lozano

Facultad de Ciencias Naturales e Ingeniería
Maestría en Modelado y Simulación Computacional
MM&S

## Modelos de clasificación supervisados de zonas metropolitanas y no metropolitanas en estudiantes de educación virtual por medio de la plataforma Moodle

**Sergio Morales Dussan**
MSc(c) Maestría en Modelado y Simulación Computacional

Universidad de Bpgptá Jorge Tadeo Lozano
Facultad de Ciencias Naturales e Ingeniería
Bogotá, Colombia
June 2, 2021

# Universidad de Bogotá Jorge Tadeo Lozano

Facultad de Ciencias Naturales e Ingeniería
Maestría en Modelado y Simulación Computacional
MM&S

## Modelos de clasificación supervisados de zonas metropolitanas y no metropolitanas en estudiantes de educación virtual por medio de la plataforma Moodle

Trabajo de grado presentado como requisito para optar al título en :
Magíster en Modelado y Simulación Computacional

Director:
PhD. Olmer Garcia Bedoya [1]

Universidad de Bogotá Jorge Tadeo Lozano
Facultad de Ciencias Naturales e Ingeniería
Bogotá, Colombia June 2, 2021

[1]Docente de tiempo completo E-mail olmer.garciab@utadeo.edu.co

***Dedicatoria...***

*A mis Padres:*
*Hugo Morales y Maudy Dussan*

Porque sin ustedes nada de esto sería posible los amo.

*A mi esposa:*

*Jenny Alejandra Beltran*

Por su Apoyo, Comprensión y Paciencia...

Y a ti Martina mi hija que aunque no has nacido ya eres el motor de mi vida.

# Agradecimientos

Gracias a la vida por permitirme esta oportunidad, a mi familia que siempre a creído en mí y me ha apoyado en cada uno de los pasos que he dado.

Gracias a mis compañeros de clase, a la universidad, al profesor Ixent Galpin y a Mauricio Leon que contribuyeron arduamente en el artículo basado en esta tesis y especialmente al profesor Olmer el cual guió esta última parte del proceso de aprendizaje donde veo expresado el esfuerzo y dedicación de mucho tiempo.

# Resumen

En este trabajo, se evalúa la capacidad de varios modelos de aprendizaje automático para clasificar a los estudiantes según vivan o no en un área metropolitana, utilizando datos de comportamiento de los estudiantes obtenidos de los registros de Moodle de una Universidad de educación a distancia en Antioquia, Colombia.
Usando F1-score, la métrica que se consideró como la más adecuada dada la naturaleza desequilibrada de los datos, se encontró que el algoritmo del árbol de decisión XGBoost produce el mejor rendimiento de clasificación.
El alto desempeño de los algoritmos de clasificación evaluados confirma la brecha digital urbano-rural prevalente en Colombia, así como también permite inferir las relaciones que existen en la variables que hace que los algoritmos tomen una decisión de clasificar en área metropolitana o no metropolitana a un estudiante.
Los resultados de este trabajo pueden permitir que un entorno de aprendizaje inteligente se ajuste de forma adaptativa a las competencias y necesidades específicas de los alumnos en función de sus características socio-económicas.

**Abstract**

In this paper, we evaluate the ability of various machine learning models to classify students according to whether they live in a metropolitan area or not, using student behavior data obtained from Moodle logs obtained from a distance learning University in Antioquia, Colombia.
UsingF1-score, the metric that we deem to be most suitable given the unbalanced nature of the data, we find that the XGBoost decision tree algorithm yields the best classification performance.
The high performanceof the classification algorithms evaluated confirms the urban-rural digi-tal divide prevalent in Colombia.
The findings of this work can enable a smart learning environment to adaptively adjust to specific learners'competencies and needs based on their socioeconomic characteristics.



**Palabras clave**
Educación virtual, analítica de datos, modelos de clasificación, aprendizaje supervisado, Moodle, logs, Métricas de evaluación, Variables geográficas.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introducción

El presente trabajo de grado fue financiado por el proyecto *Prototipo de herramienta basada en técnicas de Big Data que contribuya a la permanencia de los Estudiantes en procesos de Educación Virtual para el departamento de Antioquia* de minciencias , realizado mediante la alianza de los grupos de investigación de la UNAD (UBUNTU, SIGCIENTY, CIDLIS), la Universidad de Bogotá Jorge Tadeo Lozano (ID&SI) y con la alianza estratégica con la empresa Queos S.A.S. y la financiación de Minciencias Convocatoria 804 de 2018, "Convocatoria Regional proyectos de I+D que contribuyan al fortalecimiento de la formación virtual en el Departamento de Antioquia, Occidente".

En la actualidad la educación virtual es una realidad que trae consigo una serie expectativas y problemas propios de los cuales la minería de datos y el Big data son herramientas que permiten abordar dichas problemáticas en pro de poder describir, cuantificar, clasificar y pronosticar diversas variables que sirvan como insumo para la toma de decisiones frente a estos desafíos.

De esta manera, las plataformas virtuales con que cuentan las universidades, permiten recolectar información del comportamiento de los estudiantes frente al desarrollo de las sesiones de clase, midiendo variables como tiempos en la plataforma, consulta a los materiales del curso entre otros que permiten tener una noción global del actuar académico del estudiante, estas variables por separado no necesariamente brindan una información que permita tomar decisiones frente a problemáticas como el bajo desempeño académico el cual pueda llevar a la deserción estudiantil, ya que como menciona Tinto (1989) el análisis de variables socio-económicas y geográficas contribuyen a un mejor entendimiento de la situación académica del estudiante.

De esta manera, el objetivo principal es determinar si los estudiantes pertenecen o viven en una zona metropolitana o no metropolitana y así poder identificar la incidencia de las variables en la toma de decisiones por parte del algoritmo, de esta forma analizar que diferencias se encuentran entre los dos grupos que puedan incidir en el rendimiento educativo.

Por otro lado, esta metodología brinda información asociada a aspectos sociales de los estudiantes que pueden influir en la toma de decisiones por parte de las universidades y contribuir en la solución de problemas asociados a la deserción estudiantil.

Dado lo anterior, se evaluarán tres algoritmos de clasificación tomando como entrada varias métricas obtenidas de la plataforma Moodle, que tienen como objetivo determinar si los estudiantes viven en un área metropolitana o no metropolitana. Los algoritmos usados son Regresión logística, Máquinas de vectores de soporte (SVM) y el algoritmo del árbol de decisiones de XGBoost (Chen et al., 2015); donde se proporcionará una comparación analizando las características matemáticas

y estadísticas relevantes, así como la significancia y eficiencia de la solución para este tipo de problemas, describiendo el uso de cada algoritmo en función de su aplicación y validación.

Es importante resaltar que para el modelado se analizó el problema del desbalanceo de datos y se desarrollaron diferentes tipos de modelos que solucionaran dicho problema, también se identificó las diferentes métricas que permitiera validar de mejor manera el modelo solución.

Este documento está estructurado de la siguiente manera:la Sección 1 y la introducción la Sección 2 presenta el problema a desarrollar, la sección 3 nos muestra el marco teórico donde se las definiciones técnicas de los algoritmos y su desarrollo matemático, en la sección 4 encontraremos el estado del arte donde se encontraran trabajos relacionados con el análisis y la predicción de datos de estudiantes. En la Sección 5 se presenta la justificación del trabajo, en la sección 6 encontraremos los objetivos, mientras en la sección 7 se presentara una caracterización de la población estudiantil utilizada en el estudio, donde seguimos ampliamente las fases enmarcadas en la metodología CRISP-DM (Wirth y Hipp, 2000) para el desarrollo de proyectos de data mining. En la Sección 8 se describe la fase de modelado de cada algoritmo. La Sección 9 presenta la evaluación de los modelos y el análisis del modelo usado y finalmente en la sección 10 se presentaran los resultados y conclusiones.

**NOTA**: la diferentes tablas y gráficas encontradas en este documento están en ingles dado al articulo "Exploring the Colombian Digital Divide using Moodle Logs through Supervised Learning" del cual este trabajo de grado fue base para su construcción.

# Chapter 2

# Problema de Desarrollo

Una problemática que se ha estudiado en la educación superior a nivel global es la deserción estudiantil, este problema se presenta en diferentes tipos de centros universitarios de toda índole, siendo una preocupación que permite la reflexión y el estudio de las causas asociadas a dicho problema. A nivel internacional según cifras del banco mundial la deserción ronda entre el 10% y 20%, en América Latina es del 37% y específicamente en Colombia esa cifra alcanza el 52% siendo el segundo país en la región con el porcentaje más alto de deserción estudiantil universitaria. De esta manera se han desarrollado diversos estudios con el fin de determinar los factores que contribuyen a que los estudiantes no continúen sus carreras universitarias.

Entre los resultados encontrados podemos destacar que las variables asociadas a la deserción se encuentran en dos grandes grupos, como nos muestra el ministerio de educación, uno son los factores de tipo social que tenga un estudiante como el entorno familiar, la edad, el tiempo, si tiene hijos, si vive en una zona rural o urbana, etc. Por otro lado, tenemos factores de tipo académico como lo son metodología de los cursos, didáctica de los docentes, estrategias pedagógicas acordes y rendimiento académico entre otros.

Como menciona Tinto (1989) donde plantea que la deserción es causada por la interacción de elementos individuales, sociales e institucionales. En relación con los aspectos sociales, el autor destaca la importancia de analizar dichos aspectos en forma conjunta y no como variables individuales en los procesos académicos.

Una variable que particularmente evidencia diferencias significativas entre los estudiantes es la zona de residencia, es decir, estudiantes de zonas alejadas de los cascos urbanos presentan necesidades diferentes a estudiantes de zonas urbanas y esto lleva a que sus procesos académicos sean diferentes, principalmente se ha distinguido que personas de un Área metropolitana tiene algunas facilidades en la conexión de internet, en el uso de aparatos tecnológicos y hasta en apropiación de habilidades digitales, de esta forma el análisis de este tipo de variables puede contribuir a la hora de entender el desempeño de un estudiante.

En países como Colombia, donde aún existe una división digital considerable, el acceso a Internet en áreas rurales puede sufrir problemas de latencia o intermitencia, en contraste con áreas urbanas donde tecnologías como la fibra óptica y 5G garantizan una banda ancha confiable y rápida. Esto es especialmente preocupante dado que, según Tinto (1989), los orígenes sociales constituyen uno de los factores más importantes para determinar el éxito académico. Además, Berrio-Zapata and Rojas-Hernández (2014) señala que las diferencias marcadas entre grupos de estudiantes pueden llevar a procesos académicos que necesiten especial atención de un grupo respecto a otro, por ende es necesario comprender qué características o variables influyen en estudiantes de una zona de

residencia respecto a otra.

Para el caso de la educación virtual, la cual es una modalidad que viene en un aumento progresivo y con unas características especiales y diferentes frente a los procesos de formación presencial que han desarrollado de igual manera procesos de identificación de factores asociados a la deserción, teniendo en cuenta que la interacción con el estudiante se realiza a partir de plataformas virtuales de aprendizaje (E-learning) la forma de poder identificar estas variables vienen condicionada a la información que se pueda extraer de dichas plataformas.

En general, la información que brindan estas plataformas comúnmente son el tiempo de conexión, el ingreso a las diversas actividades, la entrega de trabajos y foros, etc; de esta manera estos insumos hacen parte del análisis de la información ligada al desarrollo del desempeño de los estudiantes, pero además pueden verse de manera desligada entre ellos, no permitiendo que el docente o directivo puedan identificar de manera temprana y oportuna alerta de bajos rendimientos en los estudiantes para tomar decisiones que lleven a identificar posibles factores que están impidiendo un buen desarrollo académico.

Las plataformas virtuales en sus informes de estadísticas basan sus resúmenes en información de tiempo de los estudiantes y en entregables de los mismos, dejando de lado el proceso académico de las competencias que debería estar adquiriendo los estudiantes, por otro lado, la información al estar desligada pues muestra datos de variables individuales, presenta una dificultad para la interpretación de los resultados en términos de tomar correctivos a nivel pedagógico, metodológico y didáctico frente al proceso de aprendizaje que podría contribuir al bajo desempeño de un estudiante y posteriormente una posible alerta de deserción.

Para este caso la plataforma Moodle ofrece una serie de plugin que permite el análisis de información para identificar el comportamiento del estudiante en el ámbito académico, pero como se mencionaba anteriormente esta información queda desligada y no produce una retroalimentación que posibilite validar cual podría ser una potencial solución a un problema que este teniendo un estudiante en el ámbito educativo.

De esta forma el analizar qué tipos de variables inciden en que un estudiante pueda ser clasificado en una zona metropolitana o una zona no metropolitana a partir de su proceso académico, permitiría tomar medidas por parte de la comunidad educativa en pro de identificar alertas tempranas de bajo rendimiento que contribuyan a detectar casos de deserción estudiantil en formación universitaria de modalidad virtual, con base en esto podemos formular las siguientes preguntas de investigación.

¿Es posible construir modelos basados en técnicas de aprendizaje automático supervisado que permitan clasificar estudiantes por zona metropolitana y no metropolitana en la plataforma Moodle ?.

¿Es posible identificar qué incidencia tienen las diferentes variables en los modelos realizados para clasificar un estudiante en zona metropolitana y no metropolitana?

# Chapter 3

# Marco teórico

## 3.1 Deserción estudiantil

La deserción estudiantil es un problema que se ha querido caracterizar a lo largo de los años siendo una problemática presente en todos los niveles educativos de aquí parte la pregunta principal ¿Qué es deserción? Y características tiene, para este trabajo se toma la definición de Tinto (1989), quien la menciona de la siguiente manera: La deserción universitaria entendida como " ... el fracaso para completar un determinado curso de acción o alcanzar una meta deseada, en pos de la cual el sujeto ingresó a una particular Institución de educación superior ..." , esta definición como tal es muy amplia y simplemente plantea que la deserción como la no permanencia de una persona en la universidad, objetivo de este trabajo; de allí se desprende ciertas características que determinan las causas por la cuales se da este hecho.

En diversos trabajos los autores han tratado de identificar cuáles son las causas para que un estudiante no pueda seguir su proceso académico y se han creado múltiples categorías para ello como lo define Canales and De los Ríos (2007) "por un lado, factores personales, culturales, sociales y económicos de los alumnos y sus familias, y, por otro, factores académicos e institucionales", de esta forma se puede precisar que hay dos grandes grupos en los cuales recaen dichos factores los de carácter social y los de carácter académico. En ese orden de ideas al referirse al ámbito académico implicará identificar que se entiende por este término.

## 3.2 Deserción estudiantil en ambientes virtuales de aprendizaje

La diferencia entre las características que tiene un formato presencial de aprendizaje a uno virtual, justifica que se deba identificar y se aclare qué características tienen una y otra en términos de la deserción la educación virtual como lo menciona Tinto (1989) pues afirma que "un sistema de educación a distancia se destaca por tener alumnos adultos, estudiantes de tiempo completo, trabajadores de tiempo completo, con responsabilidades familiares y que viven en zonas rurales o alejadas, en este sentido los factores o barreras más importantes para culminar con éxito un programa de pregrado a distancia se refieren a las características personales, el tipo de programa y el soporte que da la institución a esta clase de estudiantes", de esta forma los factores antes mencionados no serían los únicos que se pueden mencionar, entrarían otro tipo de factores asociados a esta modalidad ya que su funcionamiento es diferente, por ende traería nuevos desafíos para la comunidad estudiantil. Factores como la carencia de tiempo, escasa tutoría, poca información sobre el proceso de enseñanza-aprendizaje, falta de soporte y dificultad de comunicación con las instituciones, son de los más comunes que se puede definir como una nueva categoría que no estaba

presente en el formato presencial como las dificultades de tipo tecnológico.

En general las plataformas virtuales permiten darle un mayor porcentaje a las valoraciones que hace el docente frente al trabajo realizado por el estudiante, ya que, al no haber una relación presencial, los valores de apreciación que tienen los docentes como forma de identificar ciertos aspectos como su motivación, su gusto y dedicación por el trabajo, de esta manera poder cuantificar el trabajo académico e indagar si el estudiante esta cumpliendo con los propósitos de los cursos que está tomando.

## 3.3 Frente a la plataforma Moodle

Moodle es una plataforma que sirve como soporte para implementar educación virtual propiamente en el sentido pedagógico se llama un entorno virtual de aprendizaje (EVA) donde tiene una serie de herramientas y espacios que permiten la interacción de los estudiantes, el docente y el saber de esta manera Moodle es un entorno que cumple con las siguientes características.

"Moodle es destacable por facilitar una estructura modular donde el profesor puede optar entre tres formatos de curso: semanal, por temas o por foro social, dependiendo de las necesidades e intereses del mismo. Se caracteriza por ser flexible y modular en cuanto que permite al docente incorporar, escoger o eliminar los materiales y recursos que considere sean necesarios para su curso." Llorente-Cejudo (2007).

Así como tiene grandes ventajes en el uso de herramientas, Moodle proporciona cierta información para que el docente o directivo de los cursos pueda tomar algún tipo de información de sus estudiantes, en este caso llamaremos a esa información como los LOGS , los cuales con registros que usualmente vienen dados en formatos de tiempo, estos LOGS tiene la función de mostrar la información de los estudiantes en la plataforma, como la cantidad de ingresos, el tiempo en cada actividad el porcentaje de completado una actividad o el curso, etc.

### 3.3.1 Plugin Analíticos Moodle

"Los análisis del aprendizaje (Learning Analytics ) son piezas de información que pueden ayudarle a un usuario de un Sistema de Gestión del Aprendizaje ( LMS= Learning Management System ) a mejorar los resultantes de su aprendizaje. Los usuarios incluyen estudiantes, profesores, administradores y las personas que toman las decisiones". Moodle (2018).

Los plugin que brinda Moodle permiten el análisis de información del estudiante frente a diferentes procesos como son frecuencia de entrada, frecuencia de tiempos, cantidad de entregas, monitoreo de procesos, graficas de interacciones de foros y entregas, exportación de datos a Google Analytics y Pywick, entre otros; esta serie de plugin permite identificar estados de los estudiantes frente a su interacción con el entorno de aprendizaje, muchos de ellos tienen un componente de encontrar patrones de reconocimiento de si un estudiante esta ejecutando las tareas asignadas y ha ingresado a los diferentes componentes de la plataforma, y sus mediciones permiten ver el avance de los diferentes contenidos temáticos de un estudiante, de allí la importancia en como a partir de diversos tipo de información se pueden reconocer características importantes para posteriormente mostrar información para toma de decisiones.

Los plugin de análisis de Moodle manejan unas dimensiones que permiten identificar ciertas características de la población dado el requerimiento de las personas a cargo expresadas de la siguiente forma.

"Los Análisis del Aprendizaje son un concepto que ha estado emergiendo bajo diferentes nombres en las décadas recientes. Sus orígenes descansan en la investigación en minería de datos y sistemas tutoriales inteligentes. Las herramientas para Análisis de Aprendizaje pueden categorizarse en varias formas": Moodle (2018)

- Descriptiva
- Predictiva
- Diagnóstica

- Prescriptiva

Los últimos plugin de Moodle tienen como foco usar el Big Data y la analítica de datos como soporte a diferentes informes, para que los diferentes usuarios tengan una información de la cual puedan soportar diversas decisiones, la condiciones de ellos tienen un sentido descriptivo que se va volviendo predictivo para atender a problemáticas propias de los entornos de aprendizaje.

## 3.4 Algortimos de clasificacion supervisada

### 3.4.1 Logistic Regression

La regresión logística es un modelo de regresión y clasificación que se basa en predecir la probabilidad de un suceso de forma binaria, en donde las variables predictoras toman pesos sobre la regresión y por medio de una transformación logarítmica la probabilidad toma dos resultados 1 o 0. Este modelo se dio a conocer en el año de 1958 por David Cox, el cual sugiere un avance respecto a los modelos lineales y multivariados, usando el criterio de máxima verosimilitud para evaluar el ajuste del modelo.Cox (1958)

### 3.4.2 SVM (Support vector machine)

La definición de las máquinas de vectores de soporte se basan en la formación de hiperplanos donde agrupa una cantidad de puntos que representa datos, dicha agrupación se realiza dado la optimización de distancia respecto a una función (kernel) la cual permite construir los diferentes hiperplanos, dicho algoritmo creado en 1995 por Cortes y Vapnick Cortes and Vapnik (1995) inicialmente ingresaba vectores de forma lineal en espacios de dimensión alta y por medio de una función de optimización, en este caso lineal se clasificaba por los diferentes grupos, la función lineal separaba los grupos dado una distancia óptima que diferenciara los grupos a clasificar, esta idea permitía una generalización de dicho proceso simplificando errores a comparación de algoritmos clásicos.

De igual manera, las maquinas de soporte han ido evolucionando frente al uso de hiperparametros que permiten variar las distancias, los intervalos de confianza entre otros, además de usar diferentes tipo de funciones de optimización (lineales, polinomiales y RBF) los cuales brindan mejores opciones de clasificación.

### 3.4.3 XGboost

El modelo Xgboost es un modelo de ensamble, llamado así porque es la unión de varios modelos de clasificación "débiles" en los cuales se busca optimizar resultados probando con diversos parámetros de los diferentes tipos de modelos con el fin de crear el modelo más estable y con mejor poder de predicción o clasificación.

El modelo Xgboost es relativamente actual ya que fue publicado en 2016 por Chen, Tianqi, and Carlos Guestrin Chen and Guestrin (2016) su nombre se debe al uso del algortimo de solución de la función de optimización de perdida el cual se llama gradiente descendente.

Este modelo se basa en la construcción de árboles de decisión con refuerzo en donde se entrena el algoritmo y se busca minimizar el error en las hojas, este proceso se realiza de forma iterativa añadiendo variables de entrada y optimizando los resultados de las clasificaciones de esta manera los árboles nuevos tiene memoria de anteriores para poder discriminar todas las posibles combinaciones de las variables en el árbol.

## 3.5 Selección de parámetros

Por otro lado, es importante en el desarrollo de cualquier modelo de clasificación los parámetros que acompañan el modelo, en definitiva, los modelos matemáticos de los cuales se soportan dichos algoritmos tienen diferentes formulaciones que depende de la forma de las variables de entrada, la cantidad de datos y los clasificadores que queremos tener de esta forma se describirán algunos de ellos.

### 3.5.1 SVM (Support vector machine)

Para el caso de este modelo los hiperpámetros usados usualmente en la construcción de los clasificadores son:

- Kernel: Es el núcleo de transformación de los datos, la cual dará forma a la manera en cómo se clasificará los elementos aquí se encontrarán funciones (lineales, polinomiales, Rbf entre otras, siendo esta las más utilizadas)
- Valor C (valor de regularización): Es el parámetro que permite la cantidad errores en la clasificación, esto permite la optimización en el modelo, valores altos en este parámetro permitirá mejor la clasificación de los datos, aunque se puede incurrir en sobreajuste.
- Gamma: Este parámetro considera la distancia de los elementos clasificados a la línea de separación de los grupos, es decir entre más bajo sea el valor de gamma los puntos más lejanos podrían clasificarse en dicho grupo y viceversa.

### 3.5.2 XGboost

Los hiperparámetros que se encuentran en este modelo son variados dependiendo el problema y sus variables en este caso se han tomado los siguientes.

- N_stimators: El número de estimadores
- Learning_rate: La tasa de aprendizaje suele ir enfocado al rendimiento del modelo relacionado con los árboles de clasificación.
- Max_depht: Representa la profundidad de cada árbol, es decir cuantas características diferentes estarán presente en cada uno de los árboles.
- Colsample_bytree: El número de columnas usadas en cada árbol, es importante determinar un ajuste apropiado a la cantidad de variables de entrada para no incurrir en sobreestimación.

### 3.5.3 Regresión Logística

Solo se tendrá en cuenta un hiperparámetros

- C: Es el parámetro que permite la cantidad errores en la clasificación, esto permite la optimización en el modelo, valores altos en este parámetro permitirá mejor la clasificación de los datos, aunque se puede incurrir en sobre-ajuste.

## 3.6 Selección de evaluación de métricas

Dentro de todos los modelos anteriormente visto existen una serie de métricas que permiten identificar y cuantificar los diferentes modelos, en estos casos no cualquier métrica es un buen indicador en la toma de decisiones frente a la comparación de un modelo u otro , de esta manera se describirán las principales:

| Métrica | Definición | Fortalezas | Debilidades |
|---|---|---|---|
| Matriz de confusión | No es una métrica como tal, pero permite identificar errores en la variable de salida cuando se realiza clasificaciones. Se basa en mostrar si tanto la clasificación (pronóstico) como el valor original es correcta y los cuantifica, así como los valores que no se clasificaron correctamente. | Muestra el contraste de los resultados positivos y negativos de acuerdo con el tipo de predicción si es positiva o negativa según su clase original. | Puede presentar errores cuando las probabilidades no se encuentran equilibradas es decir que el conjunto de datos es desbalanceado, en ese caso estas métricas se muestran fuertemente sesgadas en favor de la clase mayoritaria García Jiménez (2010) |
| Accuracy | Es el porcentaje de elementos clasificados correctamente, se encuentran la diagonal principal de la matriz de confusión. | Es el indicador más usado en términos generales siempre y cuando los datos se encuentren balanceados, ya que permite identificar globalmente el comportamiento del modelo. | Presenta debilidades cuando las muestras de datos son desbalanceadas, ya que tendera a predecir la que presenta mayor frecuencia. |
| Recall | Es el valor de elementos positivos que son correctamente clasificados. | Nos brinda información sobre los falsos negativos es decir cuántos fallaron respecto a una clase o variable de salida | Depende de los que se quisiera maximizar o minimizar en este caso, si queremos minimizar falsos negativos este score debe ser lo más alto posible. |
| Precisión | Es el valor de los elementos clasificados positivos de los casos positivos predichos. | Nos brinda información sobre los falsos positivos es decir cuántos pronósticos correctos respecto a una clase o variable de salida | Depende de los que se quisiera maximizar o minimizar en este caso, si queremos minimizar falsos positivos este score debe ser lo más alto posible. |
| F1 -score | Esta medida representa la media armónica entre Precisión y Recall. | Es empleada en la evaluación de entornos desbalanceados, ya que un valor de alto de F1 indica que tanto Precisión como Recall también serían altos García Jiménez (2010) | Deja de lado índices sobre las clases negativas, es decir no tiene en cuenta partes de la matriz de confusión. |

Table 3.1: Métricas de evaluación de modelos

# Chapter 4

# Estado del Arte

Como se ha mencionado anteriormente en los últimos 20 años las ciencias computacionales han tenido gran impacto a todos los niveles científicos siendo este campo de acción punto de partido para diversos tipos de investigaciones si bien la analítica de datos y el Big data son términos un poco recientes, los algoritmos como serie de pasos para la solución de un evento es algo que ha estado en la historia de la ciencia.

Se podría nombrar a Alan Turing como uno de los precursores en este campo quien brindó ciertas pautas para la creación y exploración de algoritmos que se usarán en la programación de máquinas como forma de analizar patrones de datos y realizar con rapidez procesos que probablemente si en estos equipos y algoritmos todavía no podríamos solucionar.

Los usos que se le ha dado a estas prácticas han sido muchos y la inteligencia artificial al servicio del hombre ha facilitado múltiples tareas, obviamente esto trae consigo la inclusión de varias áreas del conocimiento como lo son las ciencias matemáticas, estadísticas y computacionales en servicio de otras tantas.

En particular en el caso de la predicción se ha usado en áreas como la predicción de eventos climatológicos, de enfermedades, de aspectos económicos, y hasta condiciones sociales en poblaciones, para el entorno de la educación quizás no se encuentre una relación directa de esta práctica, pero ya se ha avanzado en varios estudios de esta índole, se puede encontrar en el caso de estudios de la deserción estudiantil diferentes propuestas desde el punto de vista estadística donde muestran una serie de porcentajes de personas que han desertado y han cuantificado algunas de sus causas, por otro en los procesos de educación virtual aparte de tener dichos datos estadísticos se encuentran diversos trabajos que mencionan algunos acercamientos al Big data y la Analítica de datos en esta problemática como los son:

Dicovskiy Riobóo and Pedroza (2018). Minería de datos, una innovación de los métodos cuantitativos de investigación, en la medición del rendimiento académico universitario. En este trabajo de investigación se hace un análisis de algunos factores que pueden llevar a la deserción estudiantil haciendo usos de Analítica de datos resaltando que se poseía información histórica y que promedio de estas técnicas pudieron determinar descriptivamente el comportamiento de la población en el tema de la deserción.
Formia et al. (2013) caracterización de la deserción universitaria en la UNRN utilizando Minería de Datos. En esta investigación se crea un modelo basado en analítica de datos que permita identificar qué factores pueden llevar a la deserción estudiantil de una universidad en Argentina.

Si bien aquí se recopilan dos ejemplos de trabajos realizados en el contexto educativo haciendo uso de algoritmos de aprendizaje automatizado, se debe realizar una búsqueda exhaustiva que muestre los avances recientes en temas como, ¿cuáles son los modelos que más se desarrollan?, ¿Qué problemas surgen regularmente?, ¿cómo se validan dichos modelos?, etc.

## 4.1 Uso de algoritmos de clasificación

La analítica de datos es un campo que se ha venido trabajando con gran auge en los últimos años dado a la facilidad que hemos encontrado con el uso de softwares computacionales y equipos que permiten hacer simulaciones de modelos en tiempos cortos y ahorrando en costos para cualquier investigador, dado a que la computación y la realización de algoritmos de programación ha servido como puente para abordar diferentes tipos de problemática que inicialmente eran analizados desde otra perspectiva metodológica podemos incluir la analítica de datos como una opción muy importante para los investigadores en diversos campos.

La justificación de este se puede ampliar dado lo que menciona Dicovskiy Riobóo and Pedroza (2018). la Minería de Datos. La cual se basa en el uso de algoritmos computacionales que permiten extraer nuevos conocimientos, de grandes bases de datos que surgen de la acumulación de información que se generan de las actividades cotidianas de las organizaciones. Este conocimiento permite entre otros, conocer anomalías no esperadas y tomar decisiones sobre nuevas situaciones generadas". Otra definición que nos muestra la importancia de la minería de datos es la propuesta por Perez Lopez y Santín González, (2007) "La minería o exploración de datos, conocida también como "Data Mining", se puede definir como: un proceso de descubrimiento de nuevas y significativas relaciones, patrones y tendencias al analizar grandes volúmenes de datos. Este proceso se utiliza hoy en diferentes campos de la ciencia, incluidos aplicaciones financieras, análisis de mercados y comercio, seguros, educación, etc.".

De esta manera los algoritmos de clasificación son usados a partir de identificar patrones que muestren el comportamiento de un usuario o un grupo de usuarios específicos, investigaciones como Cárdenas Liebethal et al. (2019). Donde clasificaron clientes potenciales a partir de su historial crediticio usando algoritmos de scoring y clasificación, o propuestas como las de González (2015). Donde a partir de datos históricos clasifica pacientes con problemas reumatológicos dan cuenta de las diversas aplicaciones de este tipo de algoritmos.

Si bien dichos trabajos de investigación muestran algunas técnicas para la clasificación de variables haciendo uso de aprendizaje automático, las técnicas pueden variar según el tipo de problema que se quiere abordar, el tipo de variables entrada que se tengan y hasta la cantidad de las mismas, así como cuales son lo clases de salidas que se quieren clasificar, de esta forma se analizaron los modelos de clasificación que trate el problema de la retención y educación de manera global, para poder identificar principalmente cuales son los más usado para este tipo de problemas.

## 4.2 Aprendizaje automatizado en la Educación

Una gran cantidad de artículos en la literatura describen el trabajo que involucra la clasificación usando datos de los estudiantes. Como se describe en esta sección, existe una diversidad significativa en (a) las fuentes de datos utilizadas (b) los atributos de clasificación objetivo (c) los modelos de aprendizaje automático evaluados y (d) las métricas utilizadas para evaluar el rendimiento del modelo.

Un enfoque principal de la investigación existente implica predecir el rendimiento de los estudiantes. Trabajo temprano (Al-Radaideh et al., 2006; Baradwaj and Pal, 2012) utiliza los datos disponibles de los registros de los estudiantes para predecir la calificación final de un estudiante. Por ejemplo, Al-Radaideh et al. (2006) emplea atributos que incluyen el grado de la escuela secundaria, el tipo de financiación y el género para comparar el rendimiento del árbol de decisiones ID3 y C4.5, así como los algoritmos Naive-Bayes, utilizando la precisión de clasificación como la métrica para medir el desempeño.

Un trabajo más reciente ha predicho el rendimiento de los estudiantes utilizando datos recopilados automáticamente de Learning Management Systems utilizando modelos de aprendizaje automático más sofisticados. Por ejemplo, Xu and Yang (2016) usa datos que involucran actividades de los estudiantes, como la actividad de reproducción de videos y la actividad del foro del curso de los MOOC en Coursera, para predecir la motivación de los estudiantes y si obtendrán una certificación, utilizando SVM. Tsai et al. (2011) evalúa la capacidad de varias técnicas de agrupamiento para derivar árboles de decisión para evaluar la competencia informática de los estudiantes. Lopez et al. (2012) analiza la participación de los estudiantes en los foros de Moodle, incluyendo variables como número de mensajes enviados, número de mensajes leídos en el foro, tiempo de permanencia en el foro y centralidad de grado del alumno, para estimar la nota final. En este

trabajo, se evalúan los algoritmos de clustering FarthestFirst, HierarchicalClusterer, sIB, SimpleKMeans y XMeans.

Horvat et al. (2015) analiza la información del registro de Moodle, incluida la fecha de apertura de la tarea (OpenD), la fecha en que los estudiantes ven la tarea por primera vez (FirstviewD), la fecha de envío de la tarea (SubmissionD), la fecha de vencimiento de la tarea (Fecha límite), para explorar el papel de la procrastinación del estudiante tiene en el grado obtenido. Rizvi et al. (2019) analiza el papel de la información demográfica, incluida la región geográfica de residencia, la banda de privación y la discapacidad declarada en la predicción del grado utilizando varios algoritmos de árbol de decisión. Detoni et al. (2016) usa las primeras tres semanas de la actividad de Moodle del estudiante para predecir si un estudiante aprobará o reprobará un curso. Los modelos evaluados son SVM, Naive Bayes y Adaboost Decision Trees.

Las tasas de retención y deserción de estudiantes también son objeto de preocupación para la investigación en el área. Yadav et al. (2012) usa información de los registros de los estudiantes, incluido el promedio y las calificaciones anteriores, para predecir si un estudiante abandonará la escuela. Se comparan los algoritmos del árbol de decisión ID3, C4.5 y ADT, y C4.5 presenta la mayor precisión.

Otro trabajo relacionado con la clasificación de estudiantes, tiene como objetivo predecir si un estudiante tiene una discapacidad de aprendizaje. Wu et al. (2008) obtiene resultados de varias pruebas diseñadas para identificar discapacidades de aprendizaje y utiliza un algoritmo de selección de características basado en algoritmos genéticos para preprocesar los datos. Posteriormente, se lleva a cabo una comparación de la red neuronal artificial (ANN) y los clasificadores SVM, encontrándose que las ANN tienen una tasa de identificación correcta (CIR) más alta.

# Chapter 5

# Justificación

Las diferentes investigaciones tratan el tema de la deserción como la categorización de razones y a partir de ellos aplican correctivos a dichas problemáticas, de esta forma el grupo estudiado no tendrá una solución pronta y quedará como un dato más, por otro lado cuando se hace uso de herramientas de analítica de datos la mayoría de investigaciones hace una análisis descriptivo de la población en general pero no se centra en alguna problemática particular, en este orden de idea los plugin son lo más cercano a la contribución de esta problemática en los entornos virtuales.

Como podemos ver los plugin que utiliza Moodle y se utilizan técnicas algorítmicas de analítica de datos para la predicción de comportamiento atípicos para un buen proceso académico, pero ninguno nos brinda una información relacionada de cada uno de estos tópicos, ni tampoco se realiza un estudio frente al desarrollo del aprendizaje del estudiante.

Las métricas usadas en estos plugin no son pensadas para determinar que tanto ha aprendido un estudiante o que tanto no aprendió y así poder determinar una alerta temprana, de hecho, un plugin que tuviera un reconocimiento de estas diferentes variables podría tener como resultado un posible ayuda oportuna y eficaz que permitiera ayudar al estudiante en función de su bajo rendimiento y posiblemente ayudar a solucionar un posible caso de deserción a futuro.

Este trabajo busca que los modelos realizados no solamente cumplan la labor de clasificar y predecir acertadamente variables socio-económicas a partir del comportamiento del estudiante en la plataforma y permitan contribuir con información con la cual se pueda analizar si el estudiante pueda no permanecer en su proceso académico, pero por otro lado aporta a la discusión de las técnicas de aprendizaje automáticos ya se busca comparar tres modelos y argumentar su funcionalidad y pertinencia a partir del uso de hiperpárametros, métricas de evaluación y variables de entrada.

# Chapter 6

# Objetivo General.

Construir modelos basados en técnicas de aprendizaje automático supervisado que permitan clasificar estudiantes por zona metropolitana y no metropolitana en la plataforma Moodle.

## 6.1 Objetivos Específicos

- Identificar y describir las variables en la plataforma Moodle que permitan la construcción de modelos de clasificación.
- Construir y validar modelos de clasificación teniendo en cuenta parámetros acordes a la problemática.
- Identificar qué incidencia tienen las diferentes variables en los modelos realizados para clasificar un estudiante en zona metropolitana y no metropolitana.

# Chapter 7

# Metodología

Este trabajo contempla una serie de fases enmarcadas en la metodología Crisp-Dm Wirth and Hipp (2000) para el desarrollo de proyectos de análisis de datos el cual contempla seis fases: Comprensión del problema, comprensión de los datos, preparación de los datos, modelado, evaluación e implementación. Una interesante propiedad de esta metodología es que contempla la minería de datos como un proceso iterativo.

La comprensión del problema descrita en las sesiones anteriores, consistió en el entendimiento de como funcionan los LMS y que el objetivo es poder pronosticar la zona donde vive el estudiante a partir de modelos de clasificación supervisados que permitan contribuir con información con la cual se pueda analizar el desempeño académico, de esta manera se busca comparar tres modelos y argumentar su funcionalidad y pertinencia dado una serie de datos que proporciona el aula virtual. A continuación son presentadas las otras fases del desarrollo de la metodología.

## 7.1 Comprensión y preparación de los datos

La fuente de información para la construcción de los modelos de clasificación parten de dos fuentes de datos, la primera fuente se llaman LOGS, la cual viene directamente de la base de datos del aula virtual donde se consigna todas las acciones que realiza el estudiante en el proceso académico, dicha base de datos cuenta con dieciocho variables:

| Variable | Descripción |
|---|---|
| VISAE06 | Escuela o facultad a la que está asociado |
| id | Número de identificación de cada logs |
| eventname | lista de todos los tipos de eventos |
| component | Filtración de componentes específicos |
| action | Niveles de los componentes |
| target | El objetivo en que se realiza la acción se calcula a partir del nombre de la acción |
| objecttable | Nombre de la base de datos que representa el objeto del evento |
| objectid | Id del registro del objeto |
| edulevel | Nivel educativo del evento. |
| contextid | Id del resgistro del contexto (actividad) |
| contextlevel | Nivel de contexto, este indica si fue un curso, actividad, categoría del curso,etc. |
| contextinstanceid | Dependiendo del nivel de contexto, es la identificación del curso, la identificación del módulo del curso,etc. |
| userid | id del usuario. |
| courseid | Codigo del curso. |
| relateduserid | No se encontró información. |
| anonymous | Eventos en anónimo establecido en 1. |
| timecreated | Tiempo en que se creó el evento. |

Table 7.1: Variables

A estos datos se les hizo un análisis descriptivo con el fin de entender, su naturaleza, observar relaciones e identificar que tanta información brinda cada variable, esto es necesario para determinar si todo el conjunto de datos es necesario para la construcción de los modelos o si se pueden dejar de lado variables que no contribuyan esto permite mejorar la eficiencia ya que se podrá dejar de lado información que causen ruido y por ende distorsionen los resultados esperados.

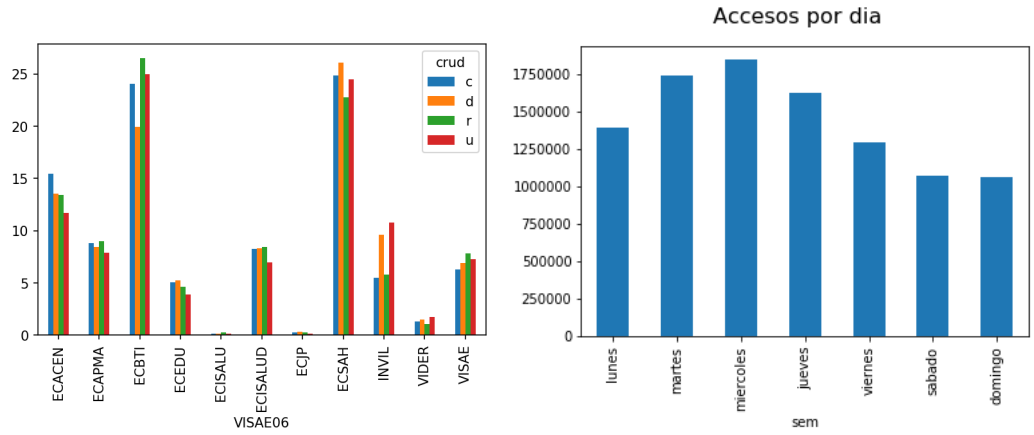



Figure 7.1: Cantidad de LOGS por escuelas y accesos por días de la semana

Para hacer este tipo de análisis es indispensable conocer muy bien los datos materia de estudio, para este caso se han tenido en cuenta 23 variables entre las que se incluyen los promedios de visitas a la plataforma así como los conteos de acceso a la misma y los tiempos promedios de permanencia en esta.

Es importante anotar que para hacer esta clasificación se segmentaron los alumnos según su ubicación, esta medida en términos de, si el estudiante esta ubicado en zona metropolitana o no metropolitana; como es de esperarse esta clasificación tiene incidencia directa en el resultado del modelo.

| Indicador | Descripción |
|---|---|
| 2.2 | *Número de LOGS generados por el estudiante* |
| 2.6 | *Número de logs generados en los días hábiles de la semana x Estudiante* |
| 2.7 | *Número de logs generados en los fines de semana* |
| 2.8.1.1 | *Número de logs generados franja (0 - 6 )* |
| 2.8.1.2 | *Número de logs generados franja (6 - 12 )* |
| 2.8.1.3 | *Número de logs generados franja (12 - 18 )* |
| 2.8.1.4 | *Número de logs generados franja (18 - 24 )* |
| 2.10 | *Radio IP* |
| 2.18 | *Cantidad de cursos a los que el estudiante accede durante una semana* |
| 2.11 | *Frecuencia de accesos por semana* |
| 2.5 | *Número de visitas a los materiales del curso* |
| 2.12 | *Número de interacciones promedio semanal con otros miembros del curso* |
| 2.13 | *Número de sesiones promedio por semana* |
| 2.17 | *Número de eventos en que modifica el sistema* |
| 2.15.1 | *Lunes - Número de diferentes días que el estudiante se loguea en el campus* |
| 2.15.2 | *Martes - Número de diferentes días que el estudiante se loguea en el campus* |
| 2.15.3 | *Miercoles - Número de diferentes días que el estudiante se loguea en el campus* |
| 2.15.4 | *Jueves - Número de diferentes días que el estudiante se loguea en el campus* |
| 2.15.5 | *Viernes - Número de diferentes días que el estudiante se loguea en el campus* |
| 2.15.6 | *Sabado - Número de diferentes días que el estudiante se loguea en el campus* |
| 2.15.7 | *Domingo - Número de diferentes días que el estudiante se loguea en el campus* |
| 2.3 | *Tiempo Total gastado por estudiante en el CMS* |

Table 7.2: Indicadores de entrada

En esta fase se decidió obtener una serie de indicadores basados en el análisis descriptivo donde cada indicador evidenciara ciertas características del comportamiento del estudiante, siendo estos indicadores el insumo de entrada para los posteriores modelos.

Los indicadores que encontramos anteriormente es suministrada por aula virtual para una cantidad de 3317 estudiantes agrupadas en periodos de tiempo semanales en este caso se tiene la información de 20 semanas, cada indicador queda depositada en una base de datos que cuenta con el valor del indicador y el promedio acumulado por semana esto con la intención de poder cuantificar tanto los valores individuales de cada estudiante por periodo de tiempo así como su recorrido acumulado de semanas anteriores.

Por otro lado la segunda fuente de información se centra en la variable de salida , está variable es la zona donde vive el estudiante y se clasifica de dos formas:

- Zona metropolitana
- Zona no metropolitana

Estos datos son proporcionados por la universidad dado la información geográfica del lugar de residencia de cada estudiante, de esta manera se puede establecer si el estudiante pertenece a una zona metropolitana o no , estos datos no están dentro del aula virtual así que no hacen parte de los LOGS es información externa de las bases de datos de la universidad.

En la gráfica a continuación podemos observar que los estudiantes ubicados en zona metropolitana casi duplican a los estudiantes ubicados en zona metropolitana, por tal razón se presenta un des-balanceo en los datos y así mismo en el resultado del entrenamiento del modelo.

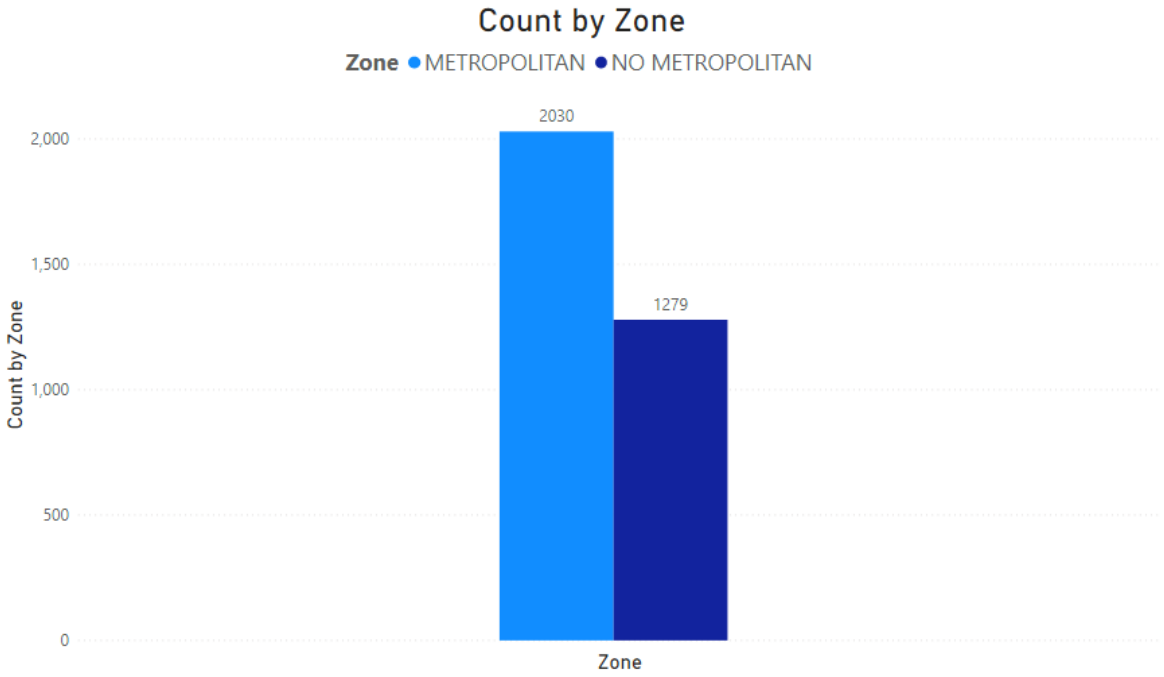



Figure 7.2: Histograma cantidad de personas por zona.

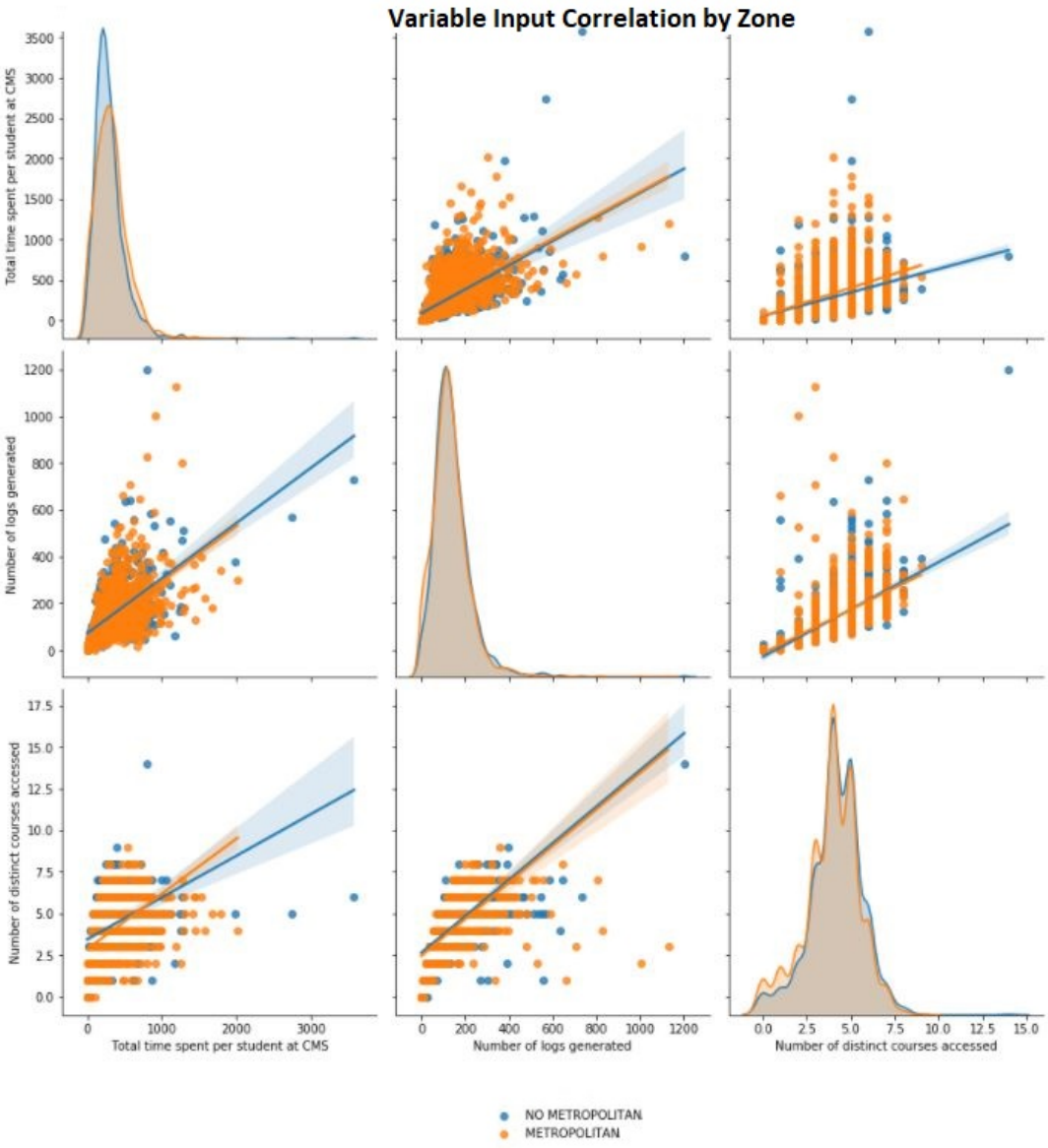
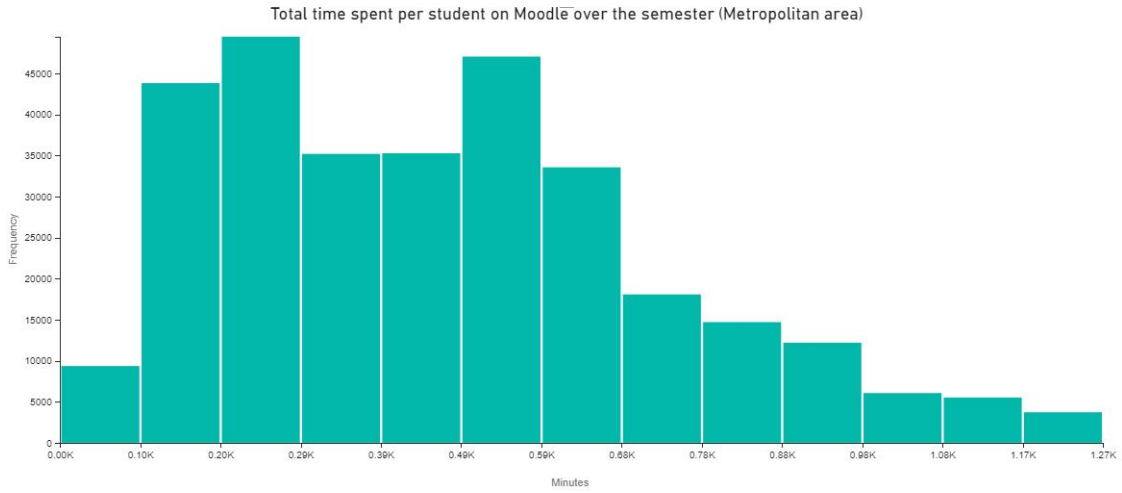
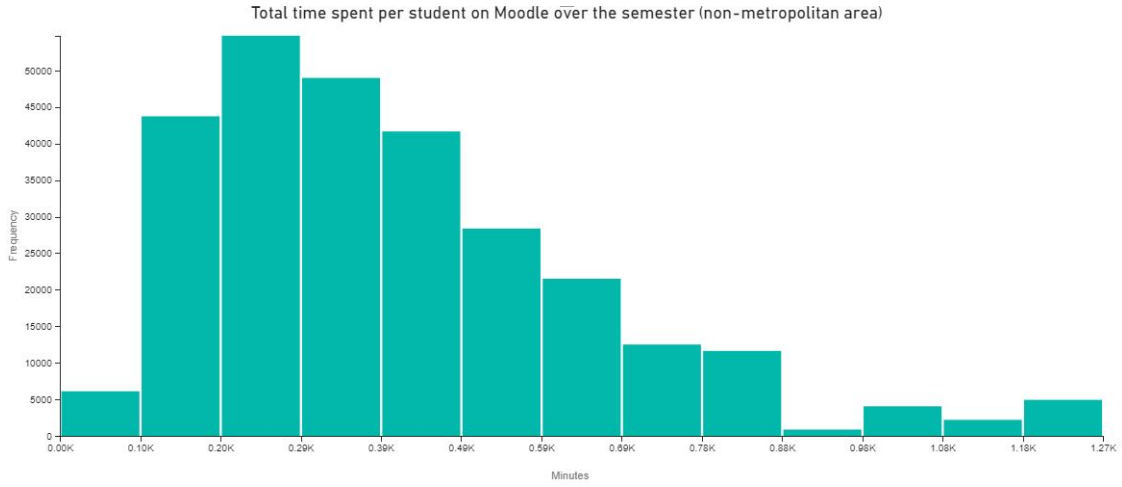



Figure 7.3: Correlación de variables por zona

La Figura 7.3 muestra la correlación de tres variables (número de Logs, Tiempo en la plataforma, número de cursos) concernientes al área. Existe una relación proporcional entre el número de cursos versus el tiempo en la plataforma y el número de registros. Esta relación es ligeramente mayor en el área metropolitana y con menor dispersión que en el área no metropolitana. Algunos casos atípicos son los casos en los que el número de cursos no es tan alto y el tiempo empleado, y el número de logs no es típico en este tipo de grupos.




(a) Metropolitan area




(b) Non metropolitan area

Figure 7.4: Histograma tiempo total por alumno que usa el entorno de aprendizaje durante este semestre por zona .

Por otro lado observando la figura 7.4 se realiza una comparación de la variable de cantidad de tiempo respecto a cada grupo, se puede observar una distribución cercana a la normal con sesgo hacia la izquierda para los dos grupos esto quiere decir que su promedio es menor a la mediana, esto se debe a que hay valores muy altos en algunos tiempos causado por lecturas dentro de la plataforma donde los estudiantes pueden dejarla abierta pero no necesariamente estén haciendo uso de sus herramientas de aprendizaje, de esta forma los tiempos se concentran en un intervalo de tiempo para los dos grupos, pero cuando observamos los picos de la cresta en la distribución podemos identificar que los estudiantes de la zona metropolitana tienen una mayor frecuencia de acceso en promedio respecto a los de la zona no metropolitana.

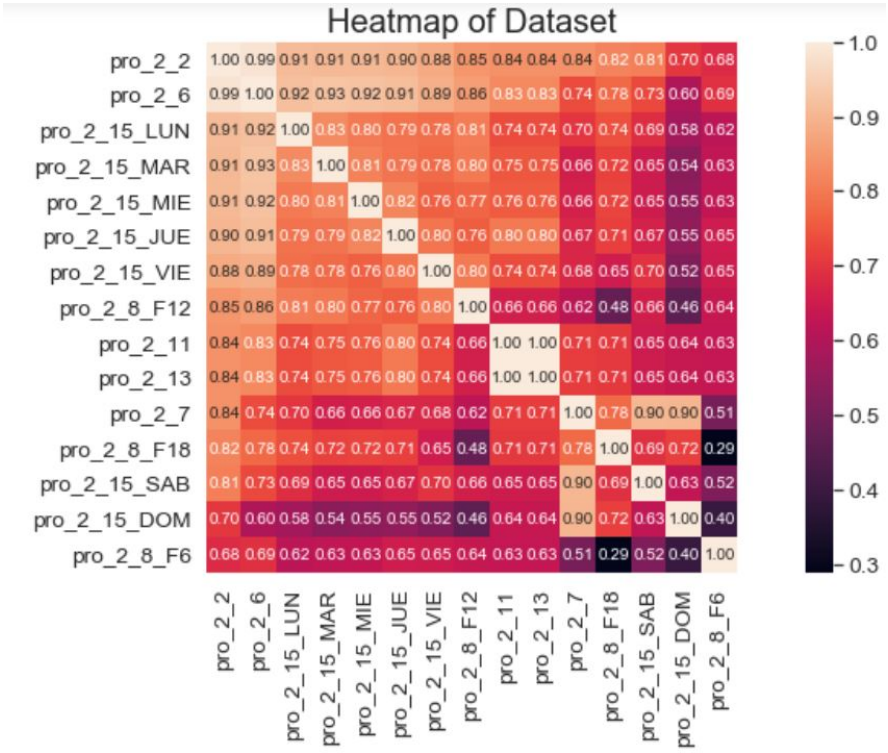



Figure 7.5: Mapa de calor que muestra las 15 variables con mayor correlación entre ellas.

En general la mayoría de variables tienen un valor de correlación por encima de 0.5 respecto a la variable cantidad de Logs, en la figura 7.5 se muestra que la variable cantidad de Logs esta fuertemente correlacionada con la variable cantidad de Logs generados en los días hábiles y a su vez esta correlación se da con las variables cantidad de Logs para los días lunes, martes, miércoles, jueves y viernes, algo que se esperaría dado a como se realizo el preprocesamiento de los Logs para la identificación de frecuencia en diversos aspectos tales como los días de la semana, de esta forma para la construcción de los modelos se podría prescindir de la información del indicador de días hábiles, ya que esta información estaría descrita en los Logs de cada día, a su vez, también podemos identificar correlaciones significativas entre el número de Logs y el número de sesiones promedio por semana así como para algunas franjas horarias como las de 12 a 6 pm y la de 6 pm a 12 am, variables como la pro2_2_F6 que se ven dentro de la figura 7.3 tiene una correlación de 0.68 la cual no es alta respecto a las demás pero aparece dentro de las mejores 15 seleccionadas para este gráfico. Dichas variables estarían midiendo características comunes de las cuales se podrían prescindir para simplificar el modelo, por otro lado variables como el tiempo en la plataforma, la cantidad de cursos, la cantidad de IP diferentes usadas semanalmente y algunas franjas horarias no evidencia valores altos de correlación entre ellas, por tanto estas variables permitirían mostrar información relevante, en el los clasificadores del modelo supervisado.

# Chapter 8

# Modelado

Para la fase de modelado se partió de la idea de trabajar con tres algoritmos (Regresion Logistica , SMVs y Xgboost), en cada uno de estos algoritmos se realizaron diferentes modelos donde se trabajarían variables de entrada, ajuste de hiperparametros etc, y se validaba a partir de la métrica seleccionada F1- score.

Inicialmente para todos los modelos se realizó una partición de la base de datos en dos grupos, un set de entrenamiento y uno de prueba con un proporción del 80 % y 20 % respectivamente, en total se tuvieron en cuenta nueve modelos, tres para cada tipo de algoritmo, donde se varían algunos indicadores de entrada dado a lo encontrado en el apartado de preparación de la data. Algunos modelos se les aplico ajuste de hiperparametros y balanceo de las clases, dichas posibilidades brinda diferentes enfoques de modelado las cuales serán validados mediante validación cruzada, a continuación se describirá cada modelo.

- **Logistic Regression**

  Modelo 1. Se usaron las 22 variables originales y se realizó una inspección de los hiperparametros en este caso para c con los valores (1, 10 , 100, 1000,10000), con clases no balanceadas y solver = liblinear.

  Model 2. Se eliminaron las variables Número de logs generados en los días laborables y promedio semanal de interacciones con otros miembros del curso. El valor de c itera sobre los valores (1, 10 , 100, 1000,10000).

  Model 3. Lo mismo del modelo anterior ajustando los pesos de los datos de entrada inversamente proporcionales a las frecuencias de clase, balanceo de datos.

- **Support vector machine SVM**

  Model 4.Las 22 variables originales se utilizaron con kernel rbf c en uno y escalando en las variables de entrada con datos no balanceados.

  Model 5. Lo mismo del modelo cuatro pero los hiperparámetros se probaron en cuadrícula optimizada c=(1, 10, 100, 1000,10000) and gamma = (0.1,1e-3, 1e-4) con datos desbalanceados.

  Model 6. lo mismo del modelo cinco con clases balanceadas ajustando los pesos de los datos de entrada inversamente proporcionales a las frecuencias de clase.

- **XGBoost**

  Model 7.Las 22 variables originales se utilizaron con los parámetros predeterminados de la biblioteca.

Model 8. En este modelo se probaron los hiperparámetros a través de una cuadrícula con learning-rate=[0.01,0.1,0.5,0.9], n-estimators=[200,100], subsample=[0.3,0.5,0.9], max-depth=[3,4,5],colsample-bytree=[0.3,0.5,0.8,1]

Model 9. Se eliminaron las variables Número de logs generados en los días laborables y Número medio semanal de interacciones con otros miembros del curso Se creó la misma cuadrícula del modelo octavo pero se balanceó la información ajustando los pesos de los datos de entrada inversamente proporcional a la clase frecuencias.

## 8.1 Evaluación de los Modelos

Para la evaluación, se realizó una comparación entre los diferentes modelos usando k-fold cross-validation (Fushiki, 2011) usando cinco divisiones del conjunto de datos de entrenamiento. La medida de evaluación para este problema de clasificación fue F1-score. Observamos que, dado que las clases de datos están desequilibradas, la precisión no es muy útil en este caso. Esto se debe a que la clase más representativa influye mucho en la precisión (García Jiménez, 2010). La tabla 8.1 complementa la comparación mediante métricas de precisión, recuperación y exactitud.

| **Models / Metrics** | **F1 - Score** | | **Accuracy** | | **Precision** | | **Recall** | |
|---|---|---|---|---|---|---|---|---|
| | **Mean** | **Std** | **Mean** | **Std** | **Mean** | **Std** | **Mean** | **Std** |
| **Model 1** | 0.64 | 0.02 | 0.62 | 0.02 | 0.66 | 0.02 | 0.67 | 0.02 |
| **Model 2** | 0.64 | 0.02 | 0.62 | 0.02 | 0.66 | 0.02 | 0.67 | 0.02 |
| **Model 3** | 0.64 | 0.02 | 0.62 | 0.02 | 0.66 | 0.02 | 0.67 | 0.02 |
| **Model 4** | 0.48 | 0.01 | 0.51 | 0.00 | 0.62 | 0.10 | 0.51 | 0.00 |
| **Model 5** | 0.57 | 0.03 | 0.56 | 0.03 | 0.60 | 0.02 | 0.56 | 0.03 |
| **Model 6** | 0.65 | 0.02 | 0.64 | 0.02 | 0.65 | 0.02 | 0.64 | 0.02 |
| **Model 7** | 0.63 | 0.02 | 0.60 | 0.02 | 0.62 | 0.02 | 0.63 | 0.02 |
| **Model 8** | 0.67 | 0.01 | 0.64 | 0.01 | 0.67 | 0.01 | 0.68 | 0.01 |
| **Model 9** | 0.66 | 0.02 | 0.64 | 0.02 | 0.67 | 0.02 | 0.68 | 0.02 |

Table 8.1: Desviación estándar y media por métrica y modelo

La tabla 8.1 presenta las medias y las desviaciones estándar de los nueve modelos para cada una de las métricas mencionadas anteriormente. En general, podemos observar que todos los modelos tienen comportamientos similares en cuanto al valor medio de las métricas. Específicamente, la puntuación F1 tiene una pequeña desviación estándar, lo que significa que el entrenamiento de los cinco subconjuntos está dando valores similares y que los datos son homogéneos.

Los modelos 1, 2 y 3 son algoritmos de regresión logística. Los valores para cada métrica son los mismos, lo que indica que los ajustes del hiperparámetro, la supresión de algunas variables correlacionadas o el equilibrio de los datos no están cambiando el resultado. Los otros modelos basados en Suport Vector Machine o XGBoost tienen diferencias significativas. Por ejemplo, el Modelo 4 que tiene todas las variables y valores predeterminados, tiene un promedio muy por debajo de los modelos 5 y 6 a los que se ajustaron sus hiperparámetros por cuadrícula. En el caso de los modelos 7, 8 y 9 asociados a XGBoost, su comportamiento es similar entre sí, aunque presenta una mejora en la puntuación cuando se ajustan los hiperparámetros y se equilibran las clases. De esta forma, los modelos 8 y 9 tienen las mejores métricas promedio.

Para tomar una decisión frente a los dos modelos que tienen las mejores métricas, se realizó un prueba de hipótesis para determinar si había diferencias significativas en cada modelo y elegir el de mejor desempeño, en este caso se utilizo la prueba t de Student pareada para cuantificar la diferencia entre la media de dos muestras de datos dependientes.

Dado a que su comportamiento es dependiente a causa de que se toman las mismas filas y columnas para su entrenamiento se realiza una algoritmo llamado $5x2ProcedureWithMLxtend$, el cual permite realizar una validación cruzada independiente y formular la prueba t-Student. El resultado de esta prueba arrojo como resultado que lo dos algoritmos no presentaban diferencias significativas así que cualquiera de los dos modelos funcionarían de igual manera y no se vería reflejado en la métrica de evaluación.

De esta manera se escogió el modelo 8 el cual tiene un rendimiento ligeramente mejor que el modelo 9 esto representado en la media F1-score y con menor desviación estándar.

## 8.2 Análisis del Modelo Seleccionado

En esta sección, describimos la evaluación del Modelo 8, nuestro modelo seleccionado, con más detalle.

Los algoritmos Xgboost pertenecen a un tipo de modelos aditivos o de impulso para desarrollar una clasificación o predicción, en este caso el modelo realizado se basa en optimizar una función de pérdida, esta función es la que determina la cantidad de error en la clasificación, para a partir de ella ajustarla y obtener un modelo nuevo que sea mejor que el anterior, en este procesos el algoritmo Xgboost trabaja en base a un modelo débil como el de los arboles de decisión y va adicionando arboles para mejorar la clasificación.

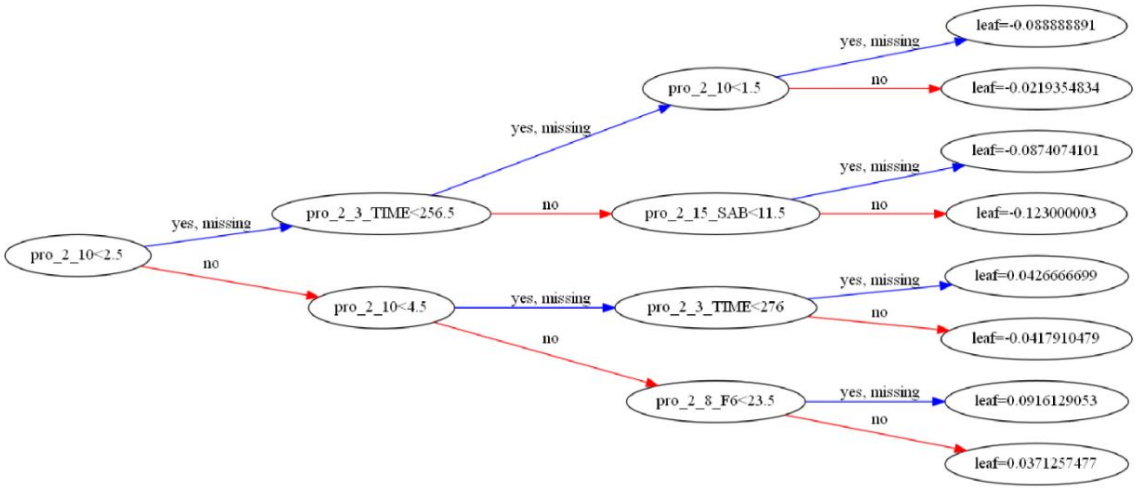



Figure 8.1: Árbol inicial para la construcción del modelo.

Como se mencionaba anteriormente el modelo realizado inicia con un árbol de decisión, en la figura 8.1 podemos observar el proceso de división de las variables para el primer caso o momento 0, donde el modelo toma una de las variables y empieza a preguntarse dado un score como es el comportamiento de dicha variable respecto al clasificador, seguido a ello se va combinando con otra variable que escoge el algoritmo dado su importancia y sigue el proceso hasta asignar finalmente valores a las hojas las cuales tienen un peso especifico en el modelo, la suma de estos valores sera el valor del clasificador, esta conformación del primer árbol simplemente brinda una información con una serie de errores de clasificación que busca que el siguiente árbol trate de corregirlos, de esta manera este procesos sigue secuencialmente creando cierta cantidad de arboles, hasta que la función de pérdida optimiza el valor de error, dicho valor es uno de los hiperparámetros ajustables pero no necesariamente se debe tomar un valor muy pequeño de error, si no depende de las condiciones del modelo, ya que se podría hacer tan pequeño que las condiciones de las variables no permitan una clasificación correcta.

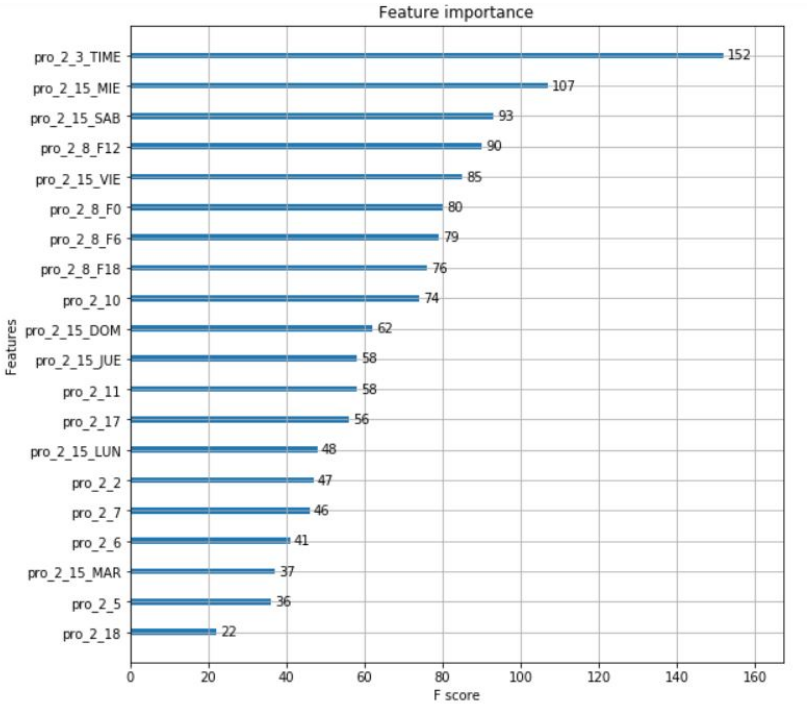



Figure 8.2: Mayor número de frecuencias de uso de las variables en el modelo

En la construcción del modelo las diferentes variables cobran una importancia que se ve reflejada en los arboles de desición, es decir no todas las variables están presentes en todos los arboles ni se evalúan la misma cantidad de veces esto depende de la selección de variables que toma el algoritmo y la discriminación para evaluar la clasificación, de esta manera se puede observar en la figura 8.2 que las variables con mayor cantidad de frecuencias para el modelo es el tiempo, la cantidad de Logs del día miércoles, la cantidad de Logs del día sábado y la cantidad de Logs para la franja horaria de las 12 a 6 pm , esto significa que estas variables son las más usadas por el modelo en cada árbol construido, también se puede evidenciar que todas las variables en algún momento han sido parte de algún árbol y han aportado en la construcción final del modelo.

Por otro lado en la figura 8.3 se puede identificar las variables con mayor contribución relativa al modelo dicha contribución, se calcula para cada variable por árbol usado en el modelo, es decir que la métricas con los valores más altos permiten inferir una mayor contribución, en ese orden de ideas no necesariamente la frecuencia de aparición en los árboles de decisión es directamente proporcional a la contribución de cada variable, ya que si se observa la figura 8.3 la variable que contribuye en mejor medida para la clasificación es la cantidad de IP usadas semanalmente, seguido el tiempo usado las cuales no son las que presentan mayor frecuencias de uso, en general todas la variables tienen una contribución al modelo y no se ve una diferencia grande entre ellas excepto la variable de cantidad de IP que para este modelo si denota una gran diferencia en la contribución.

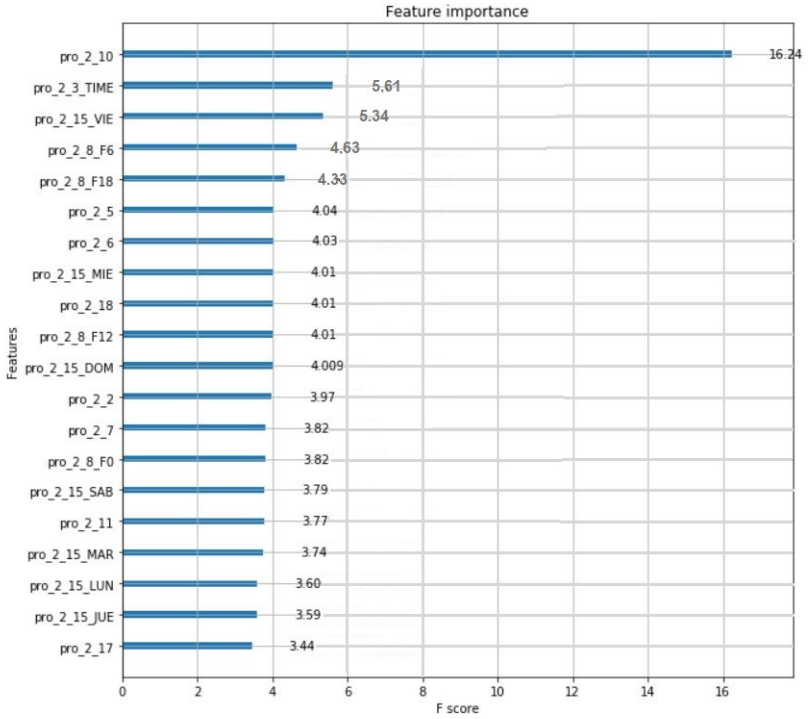



Figure 8.3: Cantidad de contribución de cada variable al modelo

Ya descrito la construcción del modelo se evalúa la efectividad de clasificación de las variables zona metropolitana y no metropolitana, de esta manera el modelo 8 después de ser entrenado se realizó la clasificación (predicción) de entradas del set de test que equivalía al 20% de la data total y se encontraron un total de 662 predicciones, con el fin de contrastarlos con los datos reales almacenados en las salidas de test; a continuación se mostrara 20 estudiantes y su comparación, se resaltan en color rojo los valores donde las predicciones no concuerdan con el valor real para el estudiante indicado. Comparando la predicciones del modelo con la data de valores reales se evidencia como es su comportamiento general, para esto, si se observa la matriz de confusión correspondiente a la tabla 4 se evidenciaron 334 verdaderos positivos, es decir, clasificaciones por parte del modelo con resultado de la variable metropolitana de estudiantes que realmente pertenece a esta zona, 97 verdaderos negativos que corresponden a clasificaciones por parte del modelo con resultado no metropolitana en estudiantes que corresponden a la zona no metropolitana, 133 falsos negativos que corresponden a estudiantes que son de la zona metropolitana y el modelos clasifico como zona no metropolitana y por ultimo 98 falsos positivos los cuales son estudiantes de la zona no metropolitana y que el modelo clasifico como pertenecientes a la zona metropolitana.

| | | Predicted Class | |
|---|---|---|---|
| | | **Positive** | **Negative** |
| Actual Class | **Positive** | 334 | 98 |
| | **Negative** | 133 | 97 |

Table 8.2: Matriz de confusión del modelo

De los 662 estudiantes el modelo clasifico correctamente cerca del 66% el cual era un valor esperado si contrastamos con el F1-score de la validación cruzada el cual nos mostraba un puntaje del 0.67 o 67% , ahora bien si se analizan los puntajes por cada clasificador (Zona metropolitana, zona no metropolitana) se identifica que los puntajes son más altos tanto en la precisión, recall y F1- score de la zona metropolitana en este caso se debe al desequilibrio de clases; si se observa la tabla 5 que muestra 432 valores reales de la zona metropolitana, mientras que solo hay 230 de la zona no metropolitana, este desequilibrio hace que el sesgo a clasificar la zona metropolitana sea más alto y por eso la cantidad de falsos negativos es mayor a la de falsos positivos.

| | **precision** | **recall** | **f1** | **support** |
|---|---|---|---|---|
| METROPOLITAN | 0.72 | 0.77 | 0.74 | 432 |
| NON-METROPOLITAN | 0.50 | 0.42 | 0.46 | 230 |
| | | | | |
| **accuracy** | | | 0.65 | 662 |
| macro avg | 0.61 | 0.60 | 0.60 | 662 |
| weighted avg | 0.64 | 0.65 | 0.64 | 662 |

Table 8.3: Informe de clasificación

# Chapter 9

# Discusión

El modelo que mostró el mejor rendimiento con respecto a la producción de clasificaciones correctas fue el algoritmo XGBoost que utiliza la optimización de la cuadrícula de hiperparámetros. Con este modelo se clasificó correctamente el 78% de los estudiantes metropolitanos y el 59% de los no metropolitanos. El éxito de la clasificación por este modelo se debe a la comprensión de las variables por el algoritmo del modelo. Esto permite identificar diferencias en el conjunto de indicadores, lo que permite clasificar correctamente las dos poblaciones. El modelo XGBoost tomó en cuenta las siguientes variables como principales variables para la toma de decisiones: tiempo promedio, frecuencia de accesos, número de IPs diferenciadas, número de eventos modificados por el sistema, franjas horarias y días de la semana específicos de actividad del usuario. De esta forma, estas variables nos permiten identificar diferencias que pueden ayudar a comprender por qué existe una brecha entre las personas que se encuentran en el área metropolitana de Antioquia y las que no lo están.

Willems (2019) señala que la brecha digital existe en todas las regiones del mundo y que impacta la oportunidad que tiene la sociedad de ingresar al sistema educativo o tener acceso a Internet de manera no intermitente. El acceso a Internet brinda la oportunidad de obtener un mayor conocimiento de aspectos relacionados con el mundo digital. Además, se observa que las personas del área metropolitana ingresan, en promedio, un 3.8% más a menudo, y se conectan a Moodle en promedio 4.8% más que las personas del área no metropolitana.

Esto podría deberse a que las personas del área metropolitana tienen mayor facilidad de acceso a equipos informáticos e Internet. Además, como señalan Gil et al. (2017), las áreas alejadas de los centros urbanos pueden tener un suministro de electricidad menos robusto que las de los centros urbanos, así como tener menos computadoras y oportunidades de acceso a Internet. Por otro lado, si tenemos en cuenta el número de IP distintas que se utilizan, en promedio, la cifra es superior para los usuarios del área no metropolitana en un 17,3%. Esto puede indicar que menos estudiantes tienen conexión a Internet en casa y, por lo tanto, tienen que salir de casa para obtener acceso a Internet (por ejemplo, cibercafés o puntos de acceso wifi municipales) para realizar sus actividades académicas.

De acuerdo con la información brindada por el Departamento de Planeación de Antioquia (Planeación, 2019), la cobertura de Internet residencial es de 66.4% para el área metropolitana frente al 25.0% para las áreas no metropolitanas. Sin embargo, la variación de los datos es notable: algunas ciudades dentro de las áreas metropolitanas alcanzan un porcentaje de conectividad de casi el 89%, mientras que algunos municipios no metropolitanos tienen un mínimo del 7%. En otras palabras, la desigualdad en términos de conexiones a Internet revela una enorme brecha digital entre las dos poblaciones.

Esta brecha digital se puede entender en el contexto de la definición de área metropolitana que da (de Estudios Urbanos, 2016), que define un área metropolitana como una serie de municipios integrados en un núcleo (en este caso, una ciudad). A su vez, implica que se fortalecen el desarrollo y la prestación de los servicios públicos, así como los aspectos de programación y coordinación del desarrollo sostenible, humano y tecnológico. Como resultado, la infraestructura, el acceso a dispositivos electrónicos y los programas educativos se brindan a los niños en edad escolar como parte de políticas que favorecen a los de las áreas metropolitanas, ampliando la brecha con los estudiantes de las áreas no metropolitanas.

Como señala San Nicolás et al. (2012), es necesario promover el desarrollo de las competencias informáticas con el fin de mejorar la adaptación y el aprendizaje de los estudiantes para afrontar sus actividades académicas y profesionales, así como dar respuesta a las demandas sociales derivadas del cambio y la evolución tecnológica.

# Chapter 10

# Conclusión

Se puede concluir que la información que aporta un aula virtual como Moodle consignada en sus Logs es en principio bastante amplia para analizar el comportamiento académico de un estudiante, de esta manera es importante adecuar un preprocesa-miento de datos dependido al análisis que se quiera hacer, en este caso fue importante procesar la información en indicadores que describieran el comportamiento de los estudiantes en función de tiempo, cantidad de accesos, cantidad de asignaturas, etc.

Por otro lado los modelos de clasificación pueden variar depende de la cantidad y forma de la variables de entrada así como el balance que tenga las variables de salida, en este caso el modelo Xgboost se adecua a las condiciones de las variables de la zona metropolitana y no metropolitana arrojando clasificaciones positivas en la mayoría de casos; dicho modelos al justar sus hiperpárametros mejoran la eficiencia de clasificación en contraste con otros modelos usados ejemplo SVM, en donde le es difícil al algoritmo encontrar una función que separa los grupos óptimamente.

En la aplicación del modelo, al realizar el proceso entrenamiento se observo que los clasificadores estaban teniendo un comportamiento sesgado en la variable metropolitana, esto causado a que la data de entrada era desbalanceada ya que había una cantidad mayor en la variable metropolitana 2030 salidas, mientras que en la no metropolitana eran solo 1279, de esta forma la optimización por hiperpárametros permitió suplir en algunos casos este problema, por ejemplo en los modelos de regresión logística y Xgboost los resultados después de la optimización mejoraron respecto al entrenamiento con parámetros por defecto, por otro lado la métrica de puntuación usada fue F1-score ya que medía de mejor manera las puntuaciones de cada variable independientemente de que estuviera desbalanceada, mientras que el modelos SVM tenia dificultades al encontrar una función de separación es tal que en el modelo 1 con parámetros por defecto el algoritmo al detectar el sesgo del clasificador tuvo un resultado muy alto para la variable metropolitana, pero bajo para la no metropolitana.

Como se mencionaba anteriormente el uso de técnicas de (hypertuning) y la selección correcta de métricas de puntación permite tener mejores clasificaciones y posteriormente un mejor modelo en casos donde los clasificadores estén desblanceados, en este caso particular las clasificaciones sesgadas a la variable metropolitana fueron suplidas por la optimización de hiperparametros, mientras que la evaluación del modelo dado a la medida F1- Score representaba de mejor manera los resultados de las clasificaciones ya que no se basaba en una medida global, teniendo en cuenta la presición y el recall, quitando los valores extremos que potencialmente se podían dar por el desbalance en los clasificadores.

En trabajos futuros se podrían abordar otro tipo de variables socio-económicas sin tener que acceder a información personal del estudiante, si no simplemente ver su comportamiento en el aula virtual, de igual manera se podría generar alertas al clasificar a los estudiantes por diversas variables socio-económicas con el fin de tomar acciones frente a un perfil que presente diversas dificultades. Así mismo, el modelo aquí empleado podría ser una guía frente a desarrollos similares en donde se pueda mejorar su precisión de clasificación aumentando la cantidad y calidad de datos , o variando algunas de las técnicas acá descritas, esto con el fin de poder abarcar escenarios mucho más complejos y diversos afines a la educación virtual y a la plataforma Moodle.

# Bibliography


Al-Radaideh, Q. A., Al-Shawakfa, E. M., and Al-Najjar, M. I. (2006). Mining student data using decision trees. In *International Arab Conference on Information Technology (ACIT'2006), Yarmouk University, Jordan*, pages 1–5.

Baradwaj, B. K. and Pal, S. (2012). Mining educational data to analyze students' performance. *arXiv preprint arXiv:1201.3417*.

Berrio-Zapata, C. and Rojas-Hernández, H. (2014). The digital divide in the university: The appropriation of ict in higher education students from bogota, colombia. *Comunicar*, 22(43):133–142.

Canales, A. and De los Ríos, D. (2007). Factores explicativos de la deserción universitaria. *Calidad en la Educación*.

Cárdenas Liebethal, J. A. et al. (2019). *Predicción de potenciales clientes de inmuebles para Aitué, basado en datos históricos de sus clientes*. PhD thesis, Universidad de Concepción. Facultad de Ingeniería. Departamento de ....

Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794.

Chen, T., He, T., Benesty, M., Khotilovich, V., and Tang, Y. (2015). Xgboost: extreme gradient boosting. *R package version 0.4-2*, pages 1–4.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3):273–297.

Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society: Series B (Methodological)*, 20(2):215–232.

de Estudios Urbanos, I. (2016). *Dinamicas de las Areas Metropolitanas de Colombia*. Universidad Nacional de Colombia, Sede Bogotá.

Detoni, D., Cechinel, C., Matsumura, R. A., and Brauner, D. F. (2016). Learning to identify at-risk students in distance education using interaction counts. *Revista de Informática Teórica e Aplicada*, 23(2):124–140.

Dicovskiy Riobóo, L. M. and Pedroza, M. E. (2018). Minería de datos, una innovación de los métodos cuantitativos de investigación, en la medición del rendimiento académico universitario. *Revista Científica de FAREM-Estelí*, (24):143–152.

Formia, S., Lanzarini, L. C., and Hasperué, W. (2013). Caracterización de la deserción universitaria en la UNRN utilizando Mineríade Datos. *Revista Iberoamericana de Educación en Tecnología y Tecnología en Educación*, 11:92–98.

Fushiki, T. (2011). Estimation of prediction error by using k-fold cross-validation. *Statistics and Computing*, 21(2):137–146.

García Jiménez, V. (2010). Distribuciones de Clases No Balanceadas: Métricas, Análisis de Complejidad y Algoritmos de Aprendizaje. *TDX (Tesis Doctorals en Xarxa)*.

Gil, H. A. P., Castro, K. A. C., Bermúdez, G. M. T., et al. (2017). La brecha digital en colombia: Un análisis de las políticas gubernamentales para su disminución. *Redes De Ingeniería*, .:59–71.

González, F. A. (2015). Modelos de aprendizaje computacional en reumatologia. *Revista Colombiana de Reumatologia*, 22(2):77–78.

Horvat, A., Dobrota, M., Krsmanovic, M., and Cudanov, M. (2015). Student perception of Moodle learning management system: a satisfaction and significance analysis. *Interactive Learning Environments*, 23(4):515–527.

Llorente-Cejudo, M. d. C. (2007). Moodle como entorno virtual de formación al alcance de todos. *Comunicar: Revista Científica de Comunicación y Educación*, 14(28):197–202.

Lopez, M. I., Luna, J. M., Romero, C., and Ventura, S. (2012). Classification via clustering for predicting final marks based on student participation in forums. *International Educational Data Mining Society*.

Planeación, D. A. (2019). Anuario Estadístico de Antioquia 2018.

Rizvi, S., Rienties, B., and Khoja, S. A. (2019). The role of demographics in online learning: A decision tree based approach. *Computers & Education*, 137:32–47.

San Nicolás, M. B., Vargas, E. F., and Moreira, M. A. (2012). Competencias digitales del profesorado y alumnado en el desarrollo de la docencia virtual: El caso de la Universidad de la Laguna. *Revista Historia de la Educación Latinoamericana*, 14(19).

Tinto, V. (1989). Definir la desercion: Una cuestion de perspectiva. *Revista de Educación Superior*, 18(71):160.

Tsai, C.-F., Tsai, C.-T., Hung, C.-S., and Hwang, P.-S. (2011). Data mining techniques for identifying students at risk of failing a computer proficiency test required for graduation. *Australasian Journal of Educational Technology*, 27(3).

Willems, J. (2019). Digital equity: Considering the needs of staff as a social justice issue. *Australasian Journal of Educational Technology*, 35(6):150–160.

Wirth, R. and Hipp, J. (2000). Crisp-dm: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, pages 29–39. Springer-Verlag London, UK.

Wu, T.-K., Huang, S.-C., and Meng, Y.-R. (2008). Evaluation of ANN and SVM classifiers as predictors to the diagnosis of students with learning disabilities. *Expert Systems with Applications*, 34(3):1846–1856.

Xu, B. and Yang, D. (2016). Motivation classification and grade prediction for MOOCs learners. . *Computational Intelligence and Neuroscience*, 44.

Yadav, S. K., Bharadwaj, B., and Pal, S. (2012). Mining Education data to predict student's retention: A comparative study. *arXiv preprint arXiv:1203.2987*.