Universidad del Valle de Guatemala

Facultad de Ingeniería

Generación de un sistema predictivo para estudiantes de Cálculo I de la Universidad del Valle de Guatemala

Trabajo de graduación presentado por Juan Pablo Pineda Melendez en modalidad de trabajo profesional para optar al grado académico de Licenciatura en Ingeniería de Ciencias de la Computación y Tecnologías de la Información.

Guatemala,
2024

Universidad del Valle de Guatemala

Facultad de Ingeniería

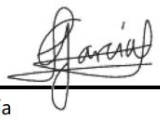
Generación de un sistema predictivo para estudiantes de Cálculo I de la Universidad del Valle de Guatemala

Trabajo de gracuación presentado por Juan Pablo Pineda Melendez en modalidad de trabajo profesional para optar al grado académico de Licenciatura en Ingeniería de Ciencias de la Computación y Tecnologías de la Información.

Guatemala,
2024

Hoja de firmas

Licda. Lynette García

X 

Lynette García
Ing

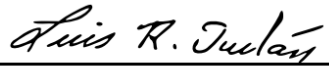
Lic. Douglas Barrios

Lic. Douglas Barrios

X 

Douglas Barrios
Ing.

Lic. Luis Furlán

X 

Luis Furlan
Ing.

09 de diciembre de 2024

ÍNDICE

RESUMEN

El presente trabajo aborda la alta tasa de reprobación en el curso de Cálculo 1 que imparte la universidad, una problemática que impacta tanto a los estudiantes como a la institución. Se desarrolló un sistema predictivo el cual, basado en modelos de aprendizaje automático, permite identificar de manera temprana a los estudiantes en riesgo de reprobación del curso, utilizando sus calificaciones acumuladas con el paso de los meses que dura el curso.

Se evaluaron y compararon múltiples algoritmos de aprendizaje automático, siendo el modelo de *Light Gradient Boosting Machine* (LGBM) el que mejor desempeño demostró, destacando en la minimización de falsos positivos y la optimización de la sensibilidad.

La implementación de esta herramienta, no solo beneficia a los estudiantes, permitiéndoles tomar decisiones informadas sobre su desempeño, sino que también ayuda a la institución a optimizar sus recursos académicos y mejorar el rendimiento académico general.

Este proyecto contribuye al desarrollo de soluciones innovadoras para mejorar la calidad educativa que provee la institución y a apoyar al éxito académico de los estudiantes en cursos clave en las fases tempranas de su formación profesional.

I. INTRODUCCIÓN

La Universidad del Valle, como muchas otras instituciones de educación superior, enfrenta el reto de la alta tasa de reprobación en cursos críticos como Cálculo I. Según estudios, esta materia presenta un índice de reprobación considerable, lo que genera un panorama desafiante tanto para los estudiantes como para la universidad en su conjunto.

En respuesta a esta problemática, este proyecto propone un enfoque innovador basado en el empleo de diversos modelos de aprendizaje automático para desarrollar un sistema predictivo del rendimiento académico en Cálculo I. El objetivo principal reside en la creación de una herramienta que permita predecir si un estudiante aprobará o reprobará la materia en cualquier momento del ciclo académico, basándose en sus calificaciones acumuladas hasta ese punto.

Para la consecución del proyecto planteado, se propone la implementación y evaluación de algoritmos aprendizaje automático de vanguardia para comparar sus rendimientos individuales y determinar cuál muestra un mejor desempeño de acuerdo con las métricas adecuadas.

La implementación de estos modelos de predicción brindará información valiosa a los estudiantes, permitiéndoles:

- Identificar su situación académica de manera oportuna: conocer su riesgo de reprobación en cualquier momento del ciclo académico, les facultará para tomar medidas proactivas para mejorar su desempeño.
- Implementar estrategias de estudio personalizadas: al comprender sus fortalezas y debilidades, los estudiantes podrían enfocar sus esfuerzos de estudio de manera más efectiva.
- Solicitar apoyo académico de manera preventiva: la identificación temprana de estudiantes en riesgo permitiría a la universidad brindarles el apoyo y la tutoría necesarios para optimizar su rendimiento.

A su vez, este proyecto también aportaría beneficios significativos a la universidad:

- Reducción de la tasa de reprobación en Cálculo I: al identificar a los estudiantes en riesgo con anticipación, se podrán implementar estrategias de apoyo y tutoría personalizada para mejorar su rendimiento y prevenir la reprobación.
- Optimización de recursos académicos: la disminución de la reprobación implicaría una mejor utilización de los recursos docentes y de tutoría, enfocándose en aquellos estudiantes que realmente los necesitan.

- Mejora del rendimiento académico general: un mejor desempeño en Cálculo I sentaría las bases para un mejor rendimiento académico en cursos posteriores, impactando positivamente en el índice académico general de la universidad.

II. Objetivos

A. General

- Desarrollar un sistema predictivo del rendimiento académico en Cálculo I, utilizando modelos de aprendizaje automático, que permita identificar a los estudiantes en riesgo de reprobación.

A. Específicos

- Realizar una limpieza de los datos iniciales para obtener un conjunto de datos que permita determinar si un estudiante aprobará o no la asignatura.

- Realizar un análisis exploratorio para definir las variables de mayor impacto para la predicción de los estudiantes en riesgo.

- Desarrollar al menos tres modelos de aprendizaje automático usando diferentes algoritmos.

- Seleccionar el modelo que presente un mejor desempeño de acuerdo a las métricas más adecuadas para modelos de clasificación.

- Desarrollar una interfaz de usuario que permita utilizar el modelo con mejor desempeño.

III. Justificación

Cálculo I, una de las materias con mayor índice de reprobación en la Universidad del Valle, según estudios, presenta un panorama desafiante para los estudiantes. Con el objetivo de mitigar este problema y apoyar a los alumnos en su trayectoria académica, se propone este proyecto de utilizar modelos de regresión y de aprendizaje de máquina.

Este proyecto busca desarrollar una herramienta innovadora que permita predecir la probabilidad de que un estudiante repruebe Cálculo I en cualquier momento del ciclo académico, con base en sus calificaciones acumuladas hasta ese momento. De esta manera, se espera brindar información oportuna a los estudiantes para que puedan tomar decisiones acertadas que mejoren su rendimiento académico y prevengan la reprobación.

La implementación de esta herramienta no solo beneficiará directamente a los estudiantes, permitiéndoles identificar su situación académica y tomar medidas oportunas para mejorar su desempeño, sino que también contribuirá a reducir la tasa de reprobación en Cálculo I. Esto, a su vez, optimizará los recursos académicos de la Universidad del Valle y fomentará un mejor rendimiento académico en general.

En definitiva, este estudio, a través del uso de modelos de regresión y clasificación, tiene el potencial de convertirse en una herramienta de valor para apoyar a los estudiantes en su proceso de aprendizaje en Cálculo I. Al prevenir la reprobación y contribuir al éxito académico de la comunidad universitaria, este proyecto representa un avance significativo en la búsqueda de mejorar la calidad educativa en la Universidad del Valle.

IV. Marco teórico

A. Reprobación de cursos clave durante la educación superior

La reprobación de cursos clave, como Cálculo I, en la Universidad del Valle de Guatemala tiene efectos significativos en la trayectoria académica de los estudiantes. En muchas otras universidades, estos cursos actúan como filtros académicos que determinarán la continuidad o deserción de los estudiantes en sus carreras. Un alto índice de reprobación puede llevar a una mayor tasa de deserción, prolongación del tiempo de estudio y un uso ineficiente de los recursos académicos disponibles en la institución.

1. Prolongación de estudios universitarios

Cálculo I es un curso fundamental para muchos de los programas universitarios en áreas de STEM (Ciencia, Tecnología, Ingeniería, Matemáticas). Este curso no solo introduce conceptos matemáticos esenciales, sino que también es un requisito para llevar cursos más avanzados incluso en la Universidad del Valle. La reprobación de Cálculo I puede retrasar significativamente la capacidad de un estudiante para inscribirse a cursos subsecuentes que dependen de este conocimiento, lo que lleva a un retraso en la graduación.

Diversos estudios han demostrado que la reprobación en cursos clave está fuertemente correlacionada con la prolongación de los estudios. Esto se debe a varios factores, incluyendo la necesidad de repetir el curso lo que llega a afectar el progreso en otras áreas del mapa curricular que tienen Cálculo I como prerrequisito. Además, la repetición del curso no solo consume tiempo adicional, sino que también puede impactar la moral y motivación del estudiante, aumentando la probabilidad de abandono o prolongación del tiempo para completar su grado (Yonghong L. y McDowell, 2021).

2. Estudios de caso y datos institucionales

Algunas universidades del mundo han documentado el impacto de la reprobación de cursos en el rendimiento general y el uso de sus recursos. Estudios realizados en la Universidad Nacional Autónoma de México (UNAM), han revelado que la reprobación de asignaturas de matemáticas básicas está fuertemente correlacionada con la deserción en carreras de ingeniería y ciencias aplicadas.

La UNAM, en particular, encontró que la alta tasa de reprobación en estos cursos contribuye directamente al bajo índice de graduación y al aumento en los tiempos de permanencia en la universidad (Sánchez Mendiola y Galindo Sontheimer, 2019).

B. Fundamentos de la minería de datos y la ciencia de los datos.

1. Conceptos clave de la minería y la ciencia de los datos

a) Preprocesamiento de datos:

El preprocesamiento de datos incluye una serie de procedimientos destinados a transformar datos crudos o no estructurados en información de alta calidad para un análisis estadístico o de aprendizaje automático. El preprocesamiento se asegura de:

- Unificar los datos en tipos que sean compatibles.
- Ajustar los datos a una escala estándar para normalizar las condiciones de los datos.
- Hacer comparaciones válidas y para transformar variables categóricas en numéricas (Hastie et al., 2009).

b) Limpieza de datos

Identifica y corrige errores o inconsistencias en los conjuntos utilizando técnicas como:

- Eliminación o imputación de valores faltantes: los valores ausentes pueden afectar el análisis y modelos predictivos, por lo que se pueden eliminar registros incompletos o imputar datos mediante técnicas estadísticas.
- Detección y tratamiento de datos atípicos: datos atípicos o errores de entrada pueden sesgar los resultados, por lo que es importante identificarlos y decidir si eliminarlos o ajustarlos.
- Corrección de duplicados: los registros repetidos pueden inflar la información, sesgando análisis o modelos predictivos, especialmente en algoritmos sensibles al volumen de datos (Hastie et al., 2009).

c) Aprendizaje automático:

El aprendizaje automático o machine learning, es una rama de la inteligencia artificial que se centra en el desarrollo de algoritmos y modelos que permiten a las computadoras aprender y hacer predicciones o decisiones basadas en datos previos. Estos algoritmos mejoran automáticamente su desempeño a medida que se exponen a más datos para su entrenamiento.

(1) Tipos de aprendizaje:

Incluye el aprendizaje supervisado, donde el modelo es entrenado con datos etiquetados; aprendizaje no supervisado, donde el modelo descubre patrones en datos sin etiquetas; y aprendizaje por refuerzo, donde el modelo aprende a tomar decisiones a través de la recompensa y el castigo (Kotsiantis, 2007).

d) Boosting

El *boosting* es una técnica de aprendizaje automático enfocada en mejorar el rendimiento general de los modelos de predicción, combinando varios modelos cuyo rendimiento es apenas mejor que adivinar al azar y crear con ellos un modelo fuerte, que tiene una mayor precisión, que sea mucho más robusto y que se enfoca en los errores que comete cada uno de los modelos anteriores.

Cada uno de los modelos aprende en función de los errores que cometen modelos anteriores de manera que trata de corregir los errores pasados dándole más peso a los datos mal clasificados por los modelos menores anteriores.

El boosting es un campo útil en problemas de clasificación y regresión, y suele usarse en la ciencia de los datos debido a su alta precisión para clasificar (Hastie et al., 2009).

e) SMOTE

SMOTE o técnica de sobremuestreo sintético para minorías es una técnica de preprocesamiento para el aprendizaje automático, se enfoca en balancear la cantidad de muestras de cada una de las clasificaciones o categorías posibles.

Funciona creando muestras sintéticas de la clase minoritaria para equilibrar el conjunto de datos seleccionando una muestra de la de la clase minoritaria y, a partir de ella, elegir a sus valores más cercanos de otros registros. Esto quiere

decir que esta técnica de preprocesamiento no solo duplica datos sino que genera nuevos a partir de los ya existentes (Chawla et al., 2002).

f) Herramientas para el desarrollo de ciencia de datos

(1) Python

Es un lenguaje de programación que puede utilizarse para la ciencia de los datos y la minería de datos gracias a su versatilidad para procesar y modelar datos y a su sintaxis accesible, así como su amplia disponibilidad de librerías que facilitan cada una de las etapas de un proyecto de ciencia de datos, desde la recolección y la limpieza de los datos hasta la modelación y la visualización de los resultados (Vanderplas y VanderPlas, 2016).

(a) Uso de librerías para desarrollo de interfaz gráfica

Algunas de las librerías que se pueden utilizar con Python pueden ser utilizadas para el desarrollo de una interfaz gráfica que permita interactuar con los datos y los modelos.

Otras librerías permiten a los usuarios proveer los datos para generar gráficos sobre los mismos. Algunas de estas librerías permiten hacer más de un solo tipo de gráfico por lo que resultan ser muy útiles para graficar cualquier tipo de dato.

- Tkinter es una librería que permite desarrollar interfaces gráficas con cuadros de diálogo o tablas, así como esperar una entrada de datos de un archivo.

- Matplotlib es una librería que permite generar múltiples tipos de gráficos a partir de datos como los datos procesados, los datos limpios y balanceados y resultados de las predicciones.

C. Aplicaciones de la minería y la ciencia de los datos en la educación.

1. Minería de datos educativos

La minería de datos educativos es una rama de la ciencia de los datos que se especializa en analizar grandes volúmenes de datos provenientes de los sistemas de aprendizaje que se utilizan en las distintas instituciones educativas. Se enfoca

principalmente en:

- Analizar patrones de aprendizaje para entender las preferencias de estudio y los factores que afectan el rendimiento de los alumnos.

- Detección de casos de alto riesgo de reprobación de cursos para optimizar el desempeño de los estudiantes en diferentes temas (Baker y Yacef, 2009).

2. Predicción del rendimiento académico

La minería de datos y la ciencia de los datos permiten analizar grandes volúmenes de datos educativos para predecir el rendimiento académico de los estudiantes. Utilizando modelos predictivos, es posible identificar a los estudiantes en riesgo y proporcionar intervenciones personalizadas para mejorar sus resultados (Baker y Yacef, 2009).

3. Análisis del aprendizaje

El análisis del aprendizaje (learning analytics) utiliza datos de estudiantes para comprender y optimizar los procesos de aprendizaje. Se enfoca en monitorear el progreso de los estudiantes, identificar brechas de conocimiento, y mejorar la toma de decisiones educativas (Siemens y Baker, 2012).

4. Personalización del aprendizaje

La personalización del aprendizaje se refiere a adaptar el proceso educativo a las necesidades individuales de los estudiantes mediante el análisis de datos. Esto incluye adaptar los materiales de enseñanza, el ritmo de aprendizaje y las estrategias pedagógicas para maximizar el potencial de cada estudiante (Zhang, 2019).

D. Modelos predictivos

1. Tipos de modelos predictivos

a) Regresión lineal

Adecuada para predecir un valor numérico continuo. La regresión lineal es una herramienta que se puede utilizar para observar la evolución de datos numéricos

durante en función del tiempo y de otros factores que dependen de cada uno de los casos de uso.

La regresión lineal se ajusta a los datos históricos por lo que es capaz de predecir el valor de una variable en el futuro basándose en la tendencia observada (Hernández Sampieri et al., 2014).

b) Regresión logística

Se enfoca en problemas de clasificación binaria y es útil para predecir la probabilidad de ocurrencia de un evento particular.

La regresión logística también se basa en datos históricos y estima la probabilidad de ocurrencia de un evento basándose en la tendencia de los datos iniciales (Hernández Sampieri et al., 2014).

2. Algoritmos de aprendizaje automático para modelos predictivos

a) Support vector machines (SVM)

Las máquinas de vectores de soporte (SVM) son un poderoso algoritmo de aprendizaje supervisado utilizado principalmente para tareas de clasificación. La idea central de un SVM es encontrar el hiperplano óptimo que separe las diferentes clases de datos. Imaginemos un hiperplano como una línea recta en un espacio bidimensional o un plano en un espacio tridimensional. En dimensiones superiores, este concepto se generaliza a un hiperplano.

El objetivo del SVM es encontrar el hiperplano que maximice la distancia entre él y los puntos de datos más cercanos de cada clase, conocidos como vectores de soporte. Esta distancia se denomina margen. Un margen máximo garantiza una mejor generalización del modelo a nuevos datos.

Sin embargo, no todos los conjuntos de datos son linealmente separables. Es decir, no siempre es posible encontrar un hiperplano que separe perfectamente las clases. Para abordar este problema, se utilizan las funciones *kernel*. Estas funciones transforman los datos originales a un espacio de mayor dimensión donde es más probable encontrar un hiperplano que separe las clases. La elección de la función *kernel* adecuada es crucial para el rendimiento del SVM. Algunos ejemplos comunes de *kernels* incluyen el lineal, polinomial y RBF (Páez, 2022).

b) Light gradient boosting machine (LGBM)

LightGBM es un algoritmo de aprendizaje automático que ha revolucionado el campo del *boosting*. Se basa en la construcción secuencial de árboles de decisión donde cada nuevo árbol se enfoca en corregir los errores del modelo anterior. A diferencia de otros modelos, LightGBM prioriza las divisiones que maximizan la ganancia de información lo que resulta en modelos más precisos y eficientes. La teoría detrás de LightGBM se fundamenta en la optimización de una función de pérdida lo que permite ajustar los parámetros del modelo de manera iterativa para minimizar el error entre las predicciones del modelo y los valores reales.

Una de las principales ventajas de este modelo es su capacidad para manejar grandes conjuntos de datos y características de dimensiones altas gracias a diversas optimizaciones como la utilización de histogramas para discretizar los valores de las características y acelerar el proceso de entrenamiento. Además, LightGBM emplea una técnica de crecimiento llamada *Leaf-Wise* la cual divide las hojas con la mayor ganancia de información evitando el sobreajuste.

LightGBM puede ser utilizada en el ámbito educativo para predecir el rendimiento académico de los estudiantes en función de sus calificaciones anteriores o factores como la cantidad de actividades que no aportan a la calificación final que fueron entregadas por los estudiantes. También podría ayudar en la detección temprana de estudiantes en riesgo de abandonar el curso y la optimización de recursos educativos como el programa de tutorías (Páez, 2022).

3. Evaluación del rendimiento de un modelo de clasificación

La evaluación del rendimiento de un modelo de clasificación es un proceso con el que se mide la capacidad de un modelo para realizar predicciones correctas en las etiquetas de una clase en comparación con los datos observados.

Esta evaluación es fundamental para determinar la eficacia de un modelo y es esencial para garantizar que un modelo de clasificación se ajuste bien a los datos con los que fue entrenado así como para asegurar su generalización y efectividad con nuevos datos.

Hay varias métricas que nos ayudan evaluar el rendimiento de un modelo de clasificación como las siguientes:

a) Matriz de confusión

La matriz de confusión es una tabla que permite visualizar el rendimiento de un modelo de clasificación. Generalmente se utiliza para modelos de clasificación binaria.

Cada celda de la matriz representa una combinación de las predicciones del modelo y la etiqueta real de los datos siendo cada una de las celdas, las siguientes:

- Verdaderos positivos: casos en los que el modelo ha predicho correctamente la clase positiva.
- Falsos positivos: casos donde el modelo predice Positivos pero la etiqueta real es negativa.
- Verdaderos negativos: casos en los que el modelo ha predicho correctamente la clase negativa.
- Falsos negativos: casos donde el modelo predice negativo pero la etiqueta real es positiva (Hernández Sampieri, 2018).

b) Precisión (*accuracy*)

Es una de las métricas más comunes y sencillas de interpretar. Representa la proporción de predicciones correctas sobre el total de predicciones realizadas.

$$precisión = \frac{Predicciones\ correctas}{Total\ de\ predicciones}$$

La precisión da una idea general del porcentaje de casos correctamente clasificados, sin embargo, en conjuntos donde una clase es mucho más común que otras, la precisión puede ser engañosa (Hernández Sampieri, 2018).

c) Sensibilidad (*recall*)

La sensibilidad mide la capacidad de un modelo para identificar correctamente los casos positivos. Esta es un poco más compleja ya que se define como la proporción de verdaderos positivos sobre el total de positivos.

$$sensibilidad = \frac{Verdaderos\ Positivos}{(Verdaderos\ Positivos) + (Falsos\ Negativos)}$$

Un modelo con alta sensibilidad detectará la mayor parte de los casos en los que la clasificación pertenezca a una clase aunque también puede dar falsos positivos (Hernández Sampieri, 2018).

d) Especificidad (true negative rate)

Mide la proporción de verdaderos negativos sobre el total de negativos reales. Define qué tan bien puede un modelo identificar los casos negativos.

$$especificidad = \frac{Verdaderos\ negativos}{(Verdaderos\ Negativos) + (Falsos\ Positivos)}$$

Una alta especificidad significa que el modelo minimiza las falsas alarmas, dando a entender que la clasificación es correcta.

e) Puntuación F1

La puntuación o medida F1 es una media armónica entre la precisión y la sensibilidad. Esta métrica se utiliza especialmente cuando el interés está en equilibrar ambas métricas, en especial con clases desbalanceadas.

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

La puntuación F1 proporciona una única medida que captura la precisión y la sensibilidad. Esta resulta ser particularmente útil cuando ambas métricas son igualmente importantes, en especial en los casos donde clasificar incorrectamente en cualquiera de las clases disponibles, puede traer consecuencias graves.

Una alta puntuación F1 indica que el modelo tiene un buen equilibrio entre la identificación correcta de cada una de las clases y la precisión en evitar clasificaciones erróneas (Hastie et al., 2009).

4. Consideraciones éticas y legales

a) Sesgo

El sesgo en los modelos predictivos es una preocupación ética recurrente, ya que los algoritmos pueden perpetuar o incluso amplificar desigualdades existentes si los datos de entrenamiento no son representativos o si reflejan prejuicios históricos

Por ello, es necesario diseñar los modelos con un enfoque de “equidad algorítmica,” que implica revisar los datos de entrenamiento y validar las predicciones en subgrupos diversos. De esta manera, se pueden detectar posibles fuentes de sesgo y hacer ajustes que promuevan resultados equitativos (Hernández Sampieri, 2018).

V. Metodología

A. Obtención de datos

Se obtuvieron datos anonimizados de la base de datos de la Universidad del Valle de Guatemala, con la colaboración de la Inga. Lynette García. Las variables recolectadas inicialmente incluyen: año del curso, ciclo académico, sección, tipo de calificación, puntaje obtenido, puntaje máximo, fecha de entrega, fecha de calificación, entre otras.

B. Preprocesamiento y limpieza de datos

1. Limpieza de valores faltantes y formateo

Se eliminaron los datos faltantes o no válidos representados por los caracteres \N en el campo de Nota, y se normalizaron los valores numéricos convirtiéndolos al formato adecuado para su procesamiento. Además, se aseguraron tipos de datos correctos para cada una de las variables, este proceso incluyó convertir fechas a formatos de tiempo y convertir los números a valores decimales de precisión para cálculos.

2. Separación de actividades regulares y extras

Las actividades que no aportan a zona fueron llamadas actividades extra. Se identificó primero que todas las actividades cuyo tipo de calificación fuera *aprobado/reprobado*, sin contar la actividad que describe el porcentaje de asistencia a la clase, eran las que no aportan a zona.

Se detectaron casos de algunas actividades extra las cuales tenían tipo de calificación en porcentaje con calificación posible de hasta 200 puntos por lo que fueron tomadas en cuenta como parte de las actividades extra que no aportan a la calificación final.

Realizar esta separación permitió contar la cantidad de actividades que no aportan a la zona que entregó cada uno de los estudiantes.

C. Cálculo del porcentaje de asistencia

Se implementó un cálculo detallado para medir la asistencia de los estudiantes durante el tiempo que duró su curso. Para este cálculo se tomaron en cuenta tres escenarios clave los cuales definieron el porcentaje de asistencia de cada uno de los estudiantes:

1. Los registros de asistencia cuyo tipo de calificación era puntos netos fueron tomados como porcentaje tomando en cuenta que el valor máximo que se podía obtener en estas actividades era de 100 puntos.

2. Los registros de asistencia cuyo tipo de calificación estaba registrado como *aprobado/reprobado* fueron tomados como 100 % para aprobado y como 79 % para reprobado, debido a que el mínimo de asistencias según el reglamento de la universidad es del 80 %.

3. Se detectaron casos en los que los alumnos tenían hasta 3 registros de asistencia, donde solo el que tenía la calificación obtenida por el estudiante, fue tomada en cuenta ya que los otros valores eran inválidos.

D. Identificación y cálculo de intentos por estudiante

Para analizar el desempeño de cada uno de los estudiantes a través de cada uno de los intentos que realizó de llevar la clase, fue necesario identificar cada vez que un estudiante repetía la materia tomando en cuenta lo siguiente:

- Cada estudiante fue agrupado utilizando su identificador, el año y la sección a la que pertenecía. De este modo, se identificaron los registros correspondientes a cada uno de los intentos realizados por el estudiante.

- A cada combinación única de estudiante y sección del curso, se le asignó un número de intento secuencial mediante una función de conteo acumulativo. Este número representa el orden en el que el estudiante tomó el curso a lo largo de los distintos ciclos y años.

E. Transformación y cálculo de métricas de rendimiento

A partir de las actividades regulares o actividades de zona, se realizaron transformaciones adicionales para calcular métricas de rendimiento acumulado por cada uno de los estudiantes.

Se definieron funciones para calcular la calificación en porcentaje y la calificación neta de cada una de las actividades entregadas por cada uno de los

estudiantes con el objetivo de obtener su calificación real tomando en cuenta las actividades cuya calificación era en puntos netos o en porcentaje.

Para reflejar la estructura de los ciclos académicos de cinco meses que se hacen en la institución, se estandarizaron los valores de mes en un formato de meses numerados del 1 al 5.

Se aplicaron funciones de acumulación de puntos posibles y de puntos obtenidos por cada uno de los estudiantes, en cada uno de los meses del ciclo para posteriormente calcular el porcentaje de rendimiento acumulado por mes, excluyendo actividades extra y redondeando a dos decimales para tener una representación más precisa en el análisis.

F. Construcción del conjunto de datos final

Una vez procesados los datos, se consolidó toda la información en un nuevo conjunto de datos por estudiante el cual incluye cada uno de sus intentos. Este nuevo conjunto de datos incluye las métricas de las calificaciones obtenidas por el estudiante al finalizar el curso y durante cada uno de los meses en los que el curso se desarrolla, la cantidad de actividades extra que entregó el estudiante, su porcentaje de asistencia y cada uno de los intentos realizados por el estudiante para analizar mejor las tendencias de los alumnos que reprueban la clase y la vuelven a llevar.

G. Análisis exploratorio de los nuevos datos

Para obtener una mejor visión general del comportamiento académico de los estudiantes y los posibles factores que afecten a su rendimiento, se inició por generar un resumen estadístico para evaluar las estadísticas clave de las numéricas principales como la calificación neta final y el porcentaje de asistencia, se exploraron las distribuciones de variables de interés para visualizar la dispersión y el comportamiento general de las variables y se analizó posibles agrupamientos que señalan diferencias significativas en los resultados finales.

Se construyó una matriz de correlación utilizando un mapa de calor para analizar la relación entre variables continuas y poder examinar dependencias lineales entre sí.

Se emplearon gráficos de caja para evaluar la relación entre variables potencialmente de impacto en el rendimiento académico final del estudiante y explorar la influencia de estas variables.

H. Balanceo de datos

Para mejorar la equidad en el desempeño de los modelos predictivos se empleó la técnica *SMOTE* la cual permite generar datos sintéticos de la clase minoritaria para evitar sobre ajustes en los modelos que serían utilizados posteriormente.

Para aplicar un balanceo de datos de *SMOTE*, se requiere primero definir la variable categórica sobre la cual el algoritmo reconoce la clase minoritaria y de la cual se sintetizan los nuevos datos logrando que los datos estén proporcionados.

I. Implementación de los modelos

Una vez realizado el balance de los datos, se deben separar en proporciones diferentes para entrenar los modelos y para probarlos también. Una vez que los datos han sido separados, se realiza el entrenamiento del modelo utilizando la mayor cantidad de los datos y se pasa a hacer predicciones con la cantidad restante de datos. Las pruebas serán hechas usando de 1 a 5 meses de calificaciones para intentar predecir el resultado del estudiante desde el inicio del ciclo hasta en los meses críticos.

A su vez, los datos procesados y separados también ayudaron a entrenar un nuevo modelo de regresión múltiple el cual realiza una proyección de las calificaciones de los estudiantes al término de cada uno de los meses del ciclo. Este modelo de regresión múltiple también fue entrenado con cantidades de 1 hasta 4 meses de calificaciones para poder realizar la proyección de los meses faltantes para el término del curso.

J. Medición de rendimiento de los modelos

Una vez que los modelos han realizado las predicciones se puede obtener la matriz de confusión de cada uno de los modelos junto a sus métricas de precisión, sensibilidad, especificidad y puntuación F1 aplicando las ecuaciones descritas con anterioridad. Ya que se realizarán pruebas utilizando múltiples meses, se utilizará un promedio de todas las métricas obtenidas.

K. Selección del modelo con el mejor rendimiento

Para seleccionar el modelo, se hizo una comparación de todas las métricas descritas anteriormente, pero se profundizará en el análisis de la matriz de confusión de cada uno de los modelos, especialmente en los falsos positivos.

L. Desarrollo de interfaz gráfica para hacer uso del modelo

Se desarrolló una interfaz gráfica utilizando la librería Tkinter de Python para realizar predicciones sobre las calificaciones de una o varias secciones. La aplicación permite al usuario cargar archivos CSV con los datos de entrada, procesarlos y visualizar las predicciones del modelo. Además, la interfaz incluye un gráfico de progreso que muestra la evolución de las calificaciones acumuladas durante el ciclo académico.

VI. Resultados

A. Transformación del conjunto de datos inicial

El conjunto de datos inicial incluía un registro por cada una de las actividades que cada estudiante realizaba, con variables como el año, la sección, el tipo de calificación de la actividad, el puntaje obtenido en la actividad, el puntaje máximo obtenible en la actividad, la actividad de asistencia, entre otras.

Seccion	Anio	Actividad	TipoCalifi	PuntosPo	RevisionP	FechaTod	FechaVen	TipoEntre	Nota
90	2019	Ejercitaci	points	0.5	FALSE	7/9/2019	59:59.0	on_paper	\N
90	2019	Ejercitaci	points	0.5	FALSE	7/9/2019	59:59.0	on_paper	0
90	2019	Ejercitaci	points	0.5	FALSE	7/9/2019	59:59.0	on_paper	0.345
90	2019	Ejercitaci	points	0.5	FALSE	7/9/2019	59:59.0	on_paper	0.345
90	2019	Ejercitaci	points	0.5	FALSE	7/9/2019	59:59.0	on_paper	0.345
90	2019	Ejercitaci	points	0.5	FALSE	7/9/2019	59:59.0	on_paper	0.37
90	2019	Ejercitaci	points	0.5	FALSE	7/9/2019	59:59.0	on_paper	0.37
90	2019	Ejercitaci	points	0.5	FALSE	7/9/2019	59:59.0	on_paper	0.445
90	2019	Ejercitaci	points	0.5	FALSE	7/9/2019	59:59.0	on_paper	0.45
90	2019	Ejercitaci	points	0.5	FALSE	7/9/2019	59:59.0	on_paper	0.45
90	2019	Ejercitaci	points	0.5	FALSE	7/9/2019	59:59.0	on_paper	0.45

Cuadro 1: Algunos de los datos iniciales

Luego de todo el procesamiento anterior, se genera un nuevo conjunto de datos el cual unifica las calificaciones de cada uno de los estudiantes en únicamente una fila por cada intento realizado por cada uno de los estudiantes, su calificación al terminar el curso, el porcentaje de puntos acumulados al finalizar cada mes del ciclo, su porcentaje de asistencia y la cantidad de actividades extra entregadas durante todo el ciclo.

Est_ID	Anio	Seccion	NotaNeta	Porcentaj	Porcentaj	Porcentaj	Porcentaj	Porcentaj	Porcentaj	Actividade	Intento	Asistencia
1	2021	110	74.85	79.94	73.24	72.84	74.03	74.77	74.77	0	1	80
2	2022	100	51.28	37.36	48.62	52.58	55.49	56.23	56.23	0	1	80
3	2020	20	69.49	74.05	69.55	66.79	68.33	69.42	69.49	6	1	100
5	2020	60	89.76	92.79	83.47	87.42	89.45	89.76	89.76	0	1	80
6	2021	10	78.43	88	76.19	79.35	76.97	78.28	78.28	0	1	80
7	2020	20	67.1	79.44	65.5	66.78	65.17	67.1	67.1	0	1	80
8	2020	20	88.15	89.21	87.09	86.8	88.48	88.15	88.15	0	1	80
9	2022	10	85	95	95	95	91.43	92.5	83.75	0	1	80
10	2021	20	74.78	77.98	79.42	80.76	81.7	79.88	79.88	0	1	80
11	2019	40	61.27	50.7	54.04	56.17	53.47	55.32	55.32	0	1	80
12	2021	30	71.24	67.89	75.64	74.78	71.16	71	71	0	1	80

Cuadro 2: Datos finales después de la limpieza

B. Análisis exploratorio del conjunto de datos generado

Durante el análisis exploratorio se observaron diversos patrones y características clave que ayudaron definir el enfoque predictivo, tales como:

- El análisis de la nota neta reveló que la mayoría de los estudiantes alcanza entre 60 y 80 puntos, cumpliendo con el requisito de aprobación. Sin embargo, también se observan estudiantes con notas muy bajas, lo que indica la existencia de un grupo en riesgo de reprobación. Esto sugiere que, aunque la mayoría logra un desempeño satisfactorio, un grupo requiere apoyo adicional o estrategias específicas de intervención. Esta información es crucial para la construcción de un sistema predictivo, ya que un modelo adecuado podría identificar y orientar a este grupo en riesgo a tiempo.

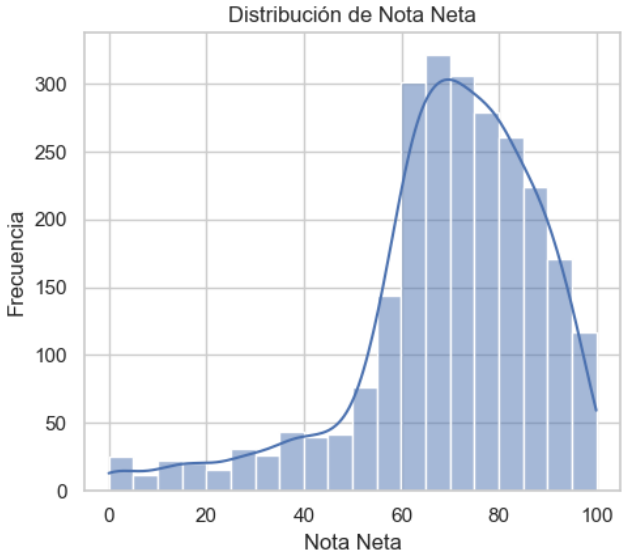


Figura 1: Histograma de la nota neta al finalizar el curso

- La progresión mensual en el porcentaje acumulado de calificaciones muestra un incremento sostenido, especialmente durante los últimos meses del ciclo académico (Figura 2). Con promedios que oscilan entre el 70 % y el 80 %, esta tendencia sugiere que los estudiantes logran mejorar sus resultados con el tiempo. Esto puede estar relacionado con un mayor dominio de los contenidos o con estrategias de evaluación que permiten recuperación académica. Este patrón justifica el enfoque mensual del análisis predictivo, ya que el seguimiento acumulativo ofrece una forma de medir el progreso individual en puntos clave del ciclo.

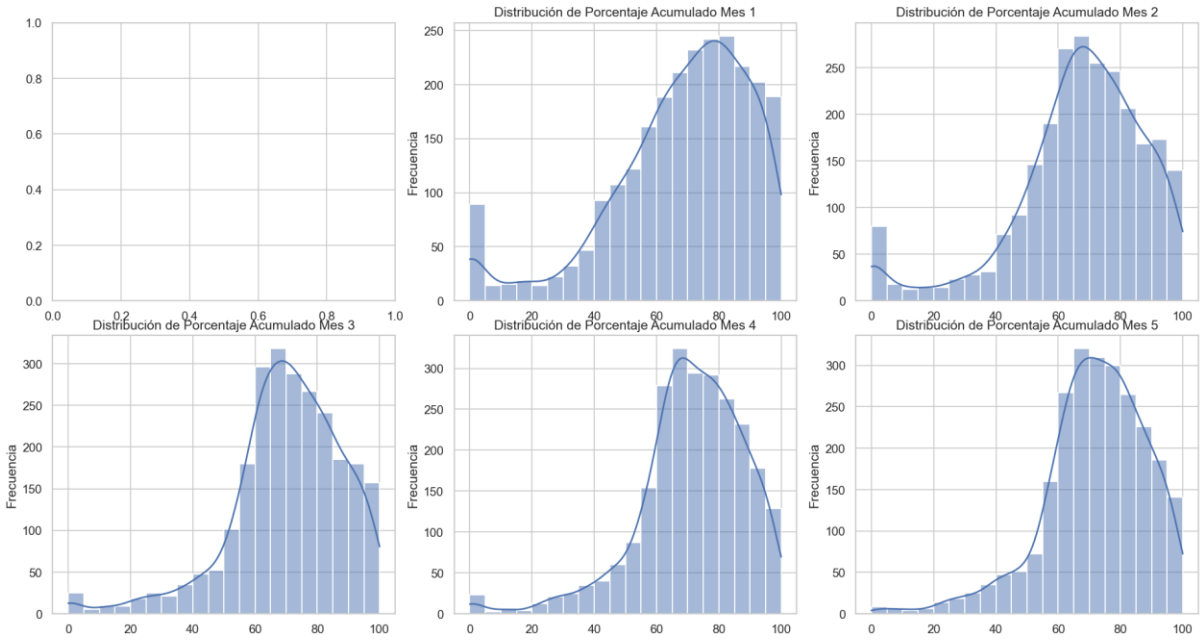


Figura 2: Histogramas de porcentajes acumulados durante cada mes del semestre

- Los datos muestran una alta correlación positiva entre la nota neta y los porcentajes acumulados mensuales, con una relación más fuerte en los últimos meses del ciclo académico, esto puede observarse en la Figura 3. Este hallazgo es importante para el modelo predictivo, ya que sugiere que un rendimiento acumulado sostenido, especialmente en los meses finales, es un buen indicador de la calificación final. Además, el análisis de esta correlación es útil para ajustar las ponderaciones en los modelos de predicción, enfatizando las métricas de rendimiento en los meses críticos.

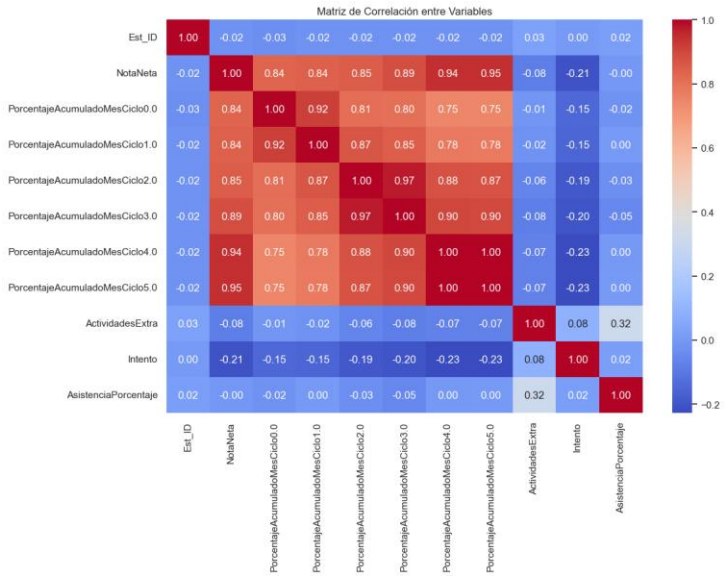


Figura 3: Mapa de calor de la matriz de correlación

- De manera sorprendente, la correlación entre la asistencia y las actividades extra con la nota neta es ligeramente negativa. Esto indica que los estudiantes que realizan más actividades adicionales o que cumplen estrictamente con la asistencia no necesariamente obtienen mejores resultados. Esta tendencia sugiere que el éxito académico en este curso podría depender de factores más allá de la asistencia y la realización de actividades extra, como la comprensión de los contenidos o el desempeño en exámenes. Este hallazgo justifica la inclusión de métricas de rendimiento mensual en el modelo, por encima de la asistencia o de las actividades adicionales.

- El análisis de los intentos de los estudiantes muestra una correlación negativa entre el número de intentos y la nota final.
  Los datos demuestran que, 1854 estudiantes aprobaron el curso en su primer intento, 66 estudiantes aprobaron en el segundo intento, 11 estudiantes en el tercer intento y una gran cantidad de estudiante (393) no llegan a aprobar el curso en las 3 oportunidades que permite la Universidad.

- La Figura 4 muestra una mayor variabilidad en las notas en los intentos 2 y 3. Es interesante notar que en los intentos 2 y 3 las notas bajas no se muestran como puntos atípicos, no siendo así el caso del primer intento. Aunque muchos estudiantes logran aprobar el curso en intentos sucesivos, sus calificaciones no necesariamente mejoran. Esto podría deberse a factores externos, como la desmotivación o la acumulación de problemas de aprendizaje. Estos resultados sugieren la necesidad de estrategias de apoyo especializadas para estudiantes que repiten el curso, y también indican que la repetición de una asignatura no garantiza un dominio completo de los contenidos.

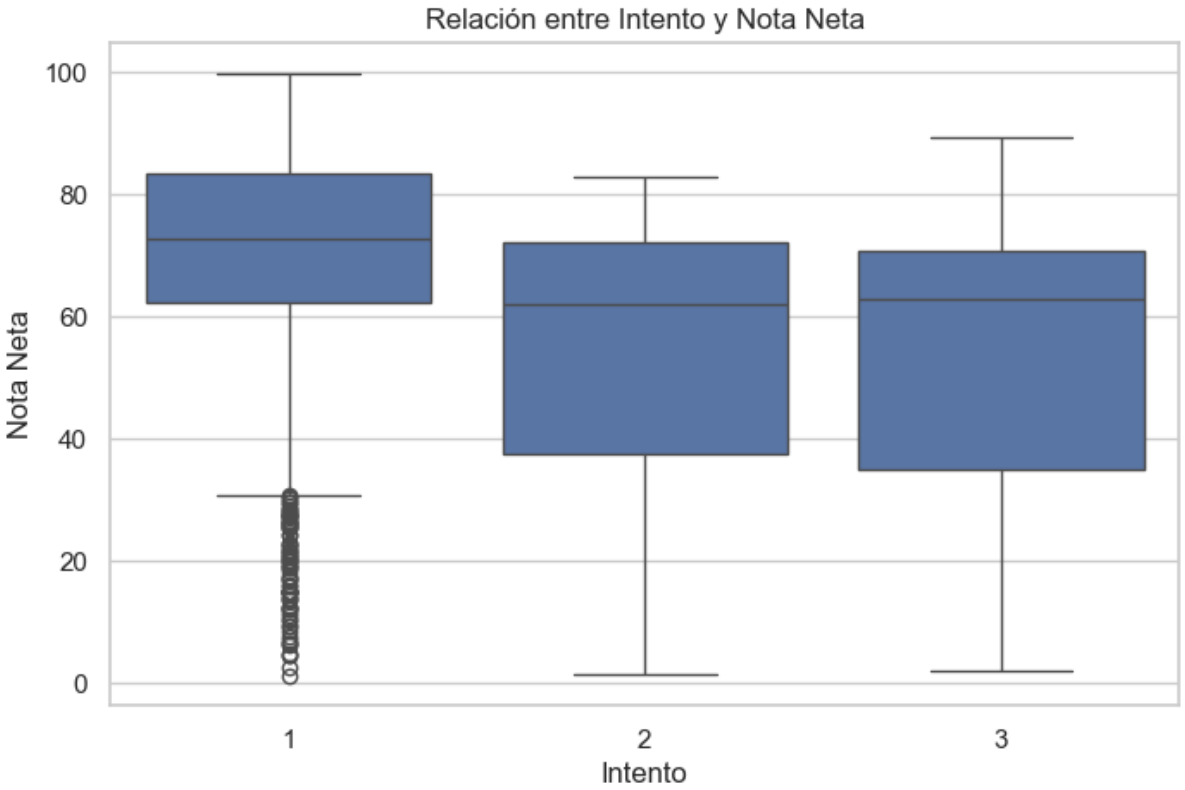


Figura 4: Diagrama de caja de la relación entre intento y calificación final

C. Balanceo del conjunto de datos generado

El balanceo utilizando la técnica SMOTE permitió equilibrar y mejorar la equidad en la representación de clases. Este proceso generó un nuevo conjunto de datos donde se duplicaron algunas entradas de la clase minoritaria, la cual era la de los estudiantes reprobados.

SMOTE permitió tener un dataset con una proporción de 50:50 entre estudiantes aprobados y reprobados, le proporcionó una base equitativa para la etapa de modelado permitiendo que los modelos tuvieran una menor tendencia a sobre ajustarse.

D. Implementación, medición y selección de los modelos

Los modelos seleccionados fueron la light gradient boosting machine (LGBM), la support vector machines (SVM) y la regresión logística, modelos que fueron entrenados utilizando datos correspondientes a diferentes números de meses (1 a 5) para predecir cómo progresaban los estudiantes en el curso de Cálculo 1,

por lo que los primeros meses mejoraron el rendimiento desde el análisis de fin de año.

Cada modelo se entrenó con datos recopilados mensualmente y proporcionó predicciones basadas en el progreso parcial de un estudiante. Esto significa que se intentó predecir si un estudiante reprobaría desde el primer mes, con la idea de identificar lo antes posible a los estudiantes en riesgo.

Modelo	Meses	Precision	Especificidad	sensibilidad	F1	Falsos positivos
LGBM	1	0.881	0.679	0.936	0.925	51
LGBM	2	0.912	0.698	0.971	0.95	48
LGBM	3	0.932	0.78	0.97	0.96	34
LGBM	4	0.958	0.86	0.98	0.97	21
LGBM	5	0.961	0.88	0.98	0.98	19
Logistica	1	0.889	0.597	0.969	0.932	64
Logistica	2	0.909	0.667	0.976	0.944	53
Logistica	3	0.934	0.736	0.988	0.959	42
Logistica	4	0.961	0.868	0.986	0.975	21
Logistica	5	0.963	0.874	0.988	0.977	20
SVM	1	0.892	0.967	0.616	0.933	61
SVM	2	0.907	0.679	0.969	0.942	51
SVM	3	0.934	0.748	0.984	0.959	40
SVM	4	0.959	0.868	0.984	0.974	21
SVM	5	0.962	0.874	0.986	0.976	20

Cuadro 3: Tabla de métricas de cada modelo en cada mes

Una vez obtenidas las métricas, se realizó un promedio de cada de una de ellas y también se realizó un análisis individual por cada uno de los modelos en cada mes lo cual dio como resultado que el modelo de LGBM es el que mejor desempeño general tiene respecto al resto de los modelos en cada uno de los meses y en promedio también. Siendo este el que tiene una menor cantidad de falsos positivos también en cada uno de los meses.

Modelo	precision	especificidad	sensibilidad	F1	falsos positivos
LGBM	0.929	0.779	0.967	0.957	34.6
SVM	0.931	0.827	0.908	0.957	38.6
Logistica	0.932	0.822	0.911	0.958	39.4

Cuadro 4: Tabla de promedios de métricas de cada modelo

Los modelos de regresión múltiple lograron proyectar con éxito el desempeño académico de los estudiantes en meses futuros, mostrando un ajuste satisfactorio a los datos de entrenamiento. Para cada modelo, las características utilizadas

como entrada y los meses proyectados se ajustaron al número de meses de datos disponibles:

- Modelo con 1 mes de datos: utilizó las calificaciones del primer mes para predecir los porcentajes acumulados de los meses 2, 3, 4 y 5.
- Modelo con 2 meses de datos: incorporó las calificaciones de los meses 1 y 2 para predecir los acumulados de los meses 3, 4 y 5.
- Modelo con 3 meses de datos: basado en los porcentajes acumulados hasta el tercer mes, proyectó los acumulados de los meses 4 y 5.
- Modelo con 4 meses de datos: predijo únicamente el acumulado del quinto mes.

Cada modelo demostró una capacidad adecuada para identificar patrones en los datos históricos y generar proyecciones coherentes. Estos resultados fueron fundamentales para incluir las proyecciones en la interfaz gráfica desarrollada, permitiendo a los usuarios observar tanto los datos históricos como las predicciones futuras para cada estudiante.

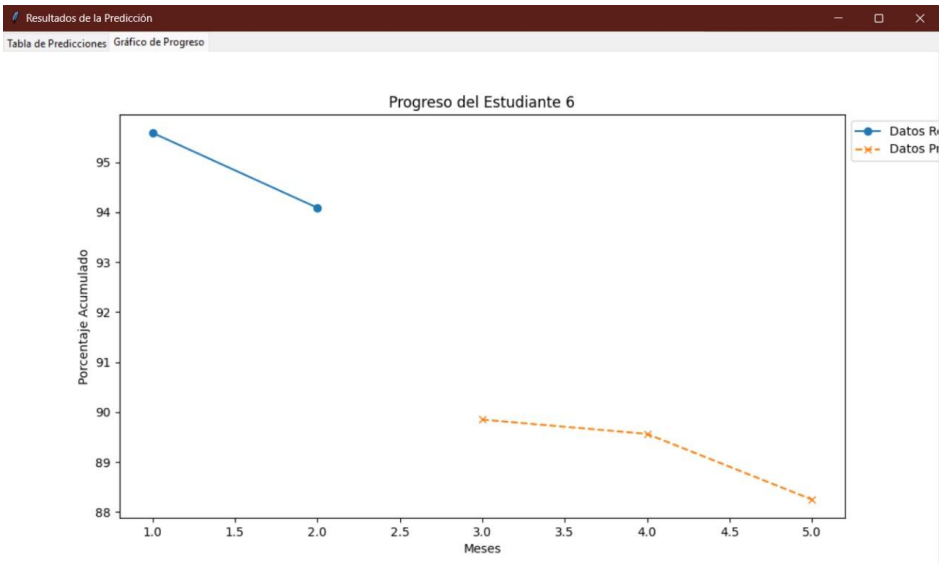


Figura 5: Gráfica de proyección de calificaciones para un estudiante

E. Interfaz gráfica de usuario

La implementación de la interfaz gráfica dio como resultado una herramienta funcional que permite a los usuarios cargar archivos CSV y recibir predicciones instantáneas sobre el rendimiento académico de sus estudiantes.

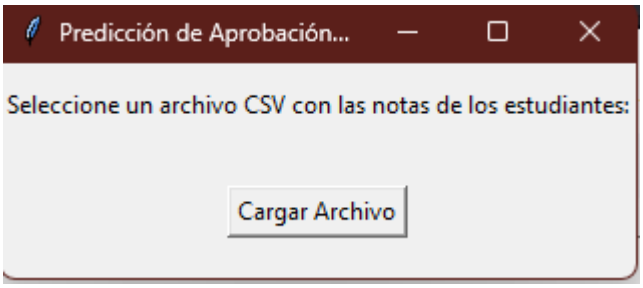


Figura 6: Muestra del cuadro de entrada de los datos

Los resultados de la predicción se muestran en forma de tabla e indican si el estudiante aprobará o reprobará el curso. Esta tabla permite seleccionar a uno de los estudiantes para observar el gráfico del progreso del estudiante durante los meses detectados.

Resultados de la Predicción

Est_ID	Predicción
26	Aprobado
43	Reprobado
107	Aprobado
313	Aprobado
333	Aprobado
447	Aprobado
463	Aprobado
496	Aprobado
500	Aprobado
555	Aprobado
582	Reprobado

Figura 7: Muestra de la tabla de estudiante mostrada por la interfaz gráfica

Además, crea un gráfico de progreso que muestra el desempeño de los estudiantes durante cada mes del año escolar, lo que facilita la identificación de tendencias y patrones en el desempeño.

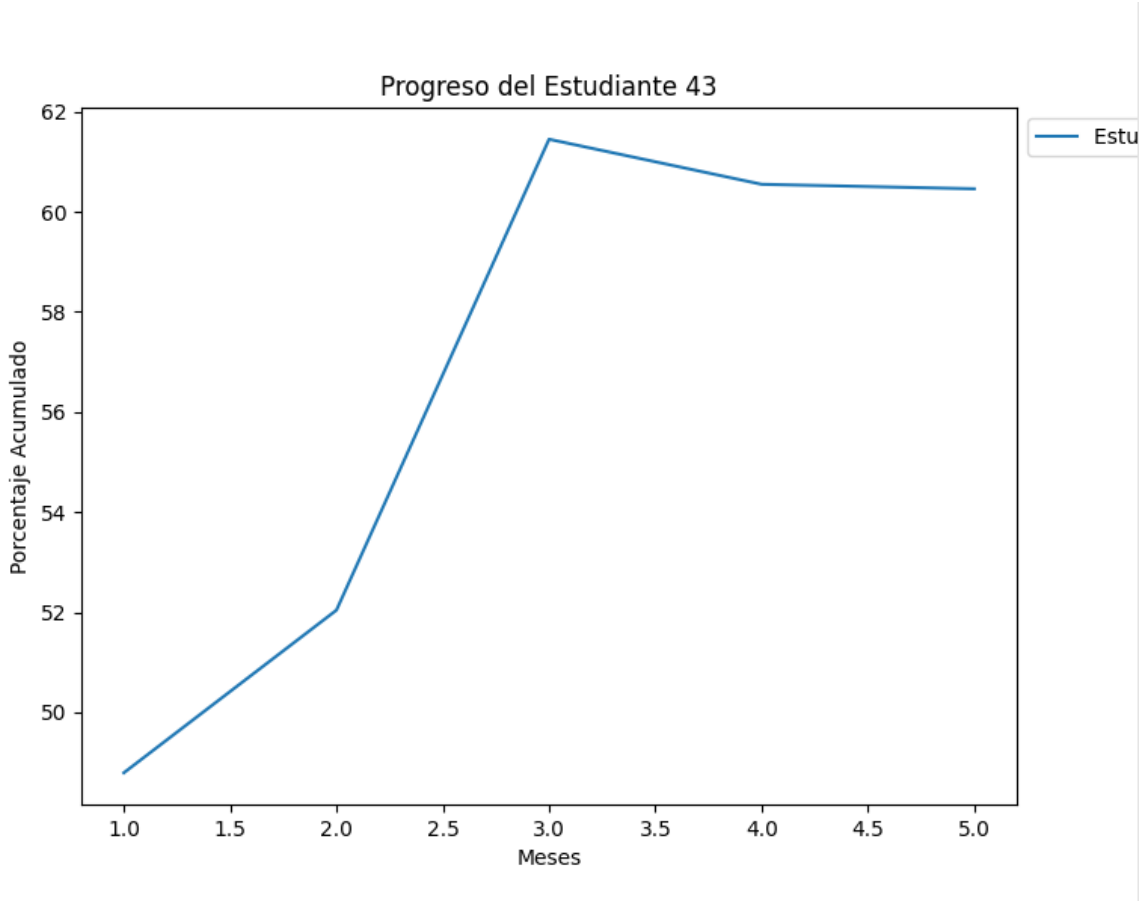


Figura 8: Muestra del gráfico de progreso del estudiante

El modelo LGBM calibrado mensualmente mostró un rendimiento estable con alta precisión y sensibilidad. y el análisis de la matriz de confusión revela que este modelo minimiza efectivamente los falsos positivos. Esta fue una prioridad para evitar predicciones falsas que pudieran impactar negativamente en las decisiones académicas de los estudiantes.

VII. Discusión

A. Retos en la limpieza de datos del conjunto de datos inicial

El análisis exploratorio permitió identificar tanto patrones esperados como aspectos inesperados que definieron el estudio y varios de los supuestos del mismo. Algunas de las secciones y estudiantes presentaron datos curiosos que añadieron complejidad a la capa inicial de limpieza de datos. Otro de los retos más difíciles a superar fue el cálculo de la asistencia ya que algunas de las secciones contaban con más de una calificación de esta actividad y otras solo la marcaban como completa o incompleta.

B. Relación entre intentos y nota neta

La expectativa era que los estudiantes que repiten el curso tendrán una probabilidad mayor de mejorar sus calificaciones en el segundo intento, dado que habrían tenido una exposición previa al contenido. Sin embargo, los datos indican que las calificaciones tienden a ser incluso más bajas en los intentos subsiguientes. Este patrón sugiere factores adicionales que podrían estar afectando el rendimiento, como la acumulación de dificultades en el aprendizaje o la falta de motivación en los estudiantes que repiten el curso. Adicionalmente, esto plantea la necesidad de evaluar las estrategias alternativas de apoyo académico para estudiantes en riesgo de reprobación para abordar las dificultades que el estudiante pueda presentar durante su segundo intento del curso.

C. Importancia de los falsos positivos

En este estudio se le dió importancia en la selección de los modelos, la cantidad de falsos positivos ya que estos son los que podrían llegar a manchar el récord académico de los estudiantes si deciden confiar en lo que dice el modelo sobre seguir cursando la clase y terminan reprobando. Es por esto que era de suma importancia que los modelos dieran la menor cantidad de falsos positivos posible.

D. Selección del modelo

Tomando en cuenta lo descrito anteriormente sobre la importancia de los falsos positivos en este estudio, se decidió continuar con la implementación de LGBM ya que fue el modelo que se desempeñó mejor en cada uno de los meses, siendo siempre el que presentó menor cantidad de falsos positivos y una mejor

sensibilidad lo que indica que detectará la mayor parte de los casos en los que la clasificación pertenezca a una clase aunque también puede dar falsos positivos.

E. Implementación de una interfaz gráfica de usuario

- Al crear modelos específicos para cada número de meses, puede ajustar su sistema de predicción al progreso real de sus estudiantes durante el año escolar.

- Esto permite a los estudiantes en el primer mes del curso recibir predicciones precisas basadas en los datos disponibles hasta ese momento, aumentando la flexibilidad y aplicabilidad de la herramienta.

- El desarrollo de una interfaz gráfica facilita el uso de modelos en un entorno accesible y fácil de usar.

- Esto es esencial para que incluso los usuarios no técnicos puedan beneficiarse de la herramienta.

- Además, la capacidad de graficar el progreso mensual proporcionó valor agregado, permitiendo a profesores y estudiantes identificar visualmente áreas de mejora y momentos clave en el ciclo académico.

- En resumen, la combinación de un modelo LGBTM especialmente capacitado y una interfaz gráfica robusta proporciona una herramienta poderosa para apoyar la toma de decisiones académicas, reduciendo el riesgo de fracaso y mejorando el uso de los recursos educativos.

VIII. Hallazgos

A. Estudiantes de prueba de la plataforma de calificaciones

Se identificó un caso de un estudiante que estaba presente en más de 10 secciones en todos los datos. Ya que se conoce que la plataforma de calificaciones tiene estudiantes de prueba para los profesores calificadores de las asignaturas.

B. Desbalanceo de calificaciones

Se identificaron casos en los que las secciones tenían discrepancias en sus calificaciones ya que sumaban más de 100 puntos, algunas de ellas, 100 puntos por cada una de las actividades.

C. Impacto de la asistencia en el rendimiento académico

Aunque el análisis inicial mostró una relación negativa entre la asistencia y las calificaciones finales, lo cual se puede interpretar como que un porcentaje alto de asistencia, no asegura mejores calificaciones, lo que sugiere que sólo la asistencia no es un factor determinante en la calificación final de la clase.

D. Frecuencia de repetición del curso y efecto en el desempeño

El análisis de estudiantes que repitieron la materia sugiere que la probabilidad de reprobar no disminuye en intentos subsecuentes lo que no se alinea con la expectativa inicial de una mejora en el rendimiento, siendo otro de los hallazgos más sorprendentes de este estudio.

IX. Conclusiones

- La limpieza profunda de los datos crudos realizada fue crucial para poder obtener un conjunto de datos con el cual es posible determinar si un estudiante está en riesgo de reprobación o no.

- Es importante realizar un profundo análisis sobre las variables que pueden llegar a ser de impacto para el estudio y la predicción del rendimiento académico de los estudiantes. Se concluye que realizar un análisis exploratorio sobre las correlaciones de las variables y su impacto sobre el rendimiento académico es de suma importancia para poder realizar predicciones sobre el mismo.

- El modelo basado en *light gradient boosting machine (LGBM)* presentó el mejor desempeño en la predicción de los estudiantes en riesgo de reprobación del curso. Este modelo sobresale en métricas de precisión con especial énfasis en minimizar los falsos positivos, lo cual es crucial para guiar a los estudiantes en cuanto a sus decisiones académicas.

- A pesar de las expectativas iniciales, se concluye que la asistencia por sí sola no es un factor determinante en la calificación final de los estudiantes y que los estudiantes que repiten el curso no necesariamente mejoran sus calificaciones en intentos subsecuentes al mismo.

- Se concluye que el desarrollo de una interfaz gráfica para poder utilizar los modelos con datos de secciones más recientes y poder hacer uso del modelo, es posible y que es útil para hacer sencillo el uso del modelo con el mejor desempeño, logrando que esta sea una herramienta de fácil implementación en la institución académica.

X. Recomendaciones

- Se recomienda realizar monitoreo constante del desempeño del modelo y de realizar actualizaciones periódicas con las calificaciones de nuevos estudiantes para reflejar posibles cambios en las características de los estudiantes y los patrones de rendimiento académico, realizando ajustes al modelo para mantener su precisión y su relevancia.

- Se recomienda ampliar este tipo de estudios a otras materias de alta dificultad como puede llegar a ser Ecuaciones Diferenciales, donde, la tasa de reprobación de estudiantes también llega a ser bastante alta. Esto ayudaría a optimizar el uso de los recursos académicos y a mejorar el rendimiento general de los estudiantes

- Se recomienda implementar medidas estrictas de protección de datos y realizar auditorías periódicas para evitar sesgos en las predicciones y garantizar la aceptación y la confianza del sistema.

- Se recomienda a la Universidad del Valle de Guatemala considerar la integración temprana del sistema predictivo en sus procesos académicos. Esto permitirá identificar y apoyar a los estudiantes en riesgo de reprobar desde las fases tempranas del curso y aumentar las posibilidades de éxito de los estudiantes reduciendo la tasa de reprobación.

XI. Bibliografía

Baker, y Yacef. (2009, October 1). *The State of Educational Data Mining in 2009: A Review and Future Visions*. Journal of Educational Data Mining. Retrieved August 26, 2024, from https://jedm.educationaldatamining.org/index.php/JEDM/article/view/8

Chawla, N. V., Bowyer, K. W., Hall, L. O., y Kegelmeyer, W. P. (2011, June 9). *SMOTE: Synthetic Minority Over-sampling Technique*. arXiv. https://arxiv.org/abs/1106.1813

Contreras, L. E., Fuentes, H. J., y Rodríguez3, J. I. (2020, Oct). Predicción del rendimiento académico como indicador de éxito/fracaso de los estudiantes de ingeniería, mediante aprendizaje automático. *SciELO*. https://www.scielo.cl/scielo.php?pid=S0718-50062020000500233yscript=sci_arttext

Hastie, T., Tibshirani, R., y Friedman, J. H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer.

Hernández Sampieri, R. (2018). *METODOLOGÍA DE LA INVESTIGACIÓN: LAS RUTAS CUANTITATIVA, CUALITATIVA Y MIXTA*. McGraw-Hill Interamericana.

Hernández Sampieri, R., Fernández Collado, C., y Baptista Lucio, P. (2014). *Metodología de la investigación* (P. Baptista Lucio, Ed.). McGraw-Hill Education.

Kotsiantis. (2007). *Supervised Machine Learning: A Review of Classification Techniques*. DataJobs.com. Retrieved August 26, 2024, from

https://datajobs.com/data-science-repo/Supervised-Learning-%5BSB-Kotsiantis%5D.pdf

Páez, A. R. (2022, Jul 25). *Modelos predictivos progresivos del rendimiento académico de estudiantes universitarios*. SciELO México. https://www.scielo.org.mx/scielo.php?pid=S2007-74672022000100044yscript=sci_arttext

Sánchez Mendiola, M., y Galindo Sontheimer, D. A. (2019). *Perspectivas de la Innovación Educativa en Universidades de México: Experiencias y reflexiones de la RIE 360*. Universidad Nacional Autónoma de México (UNAM).

Siemens, y Baker. (2012). *American Behavioral Scientist*. Retrieved 2024, from https://iu.instructure.com/files/56153619/download?download_frd=1

Vanderplas, J. T., y VanderPlas, J. (2016). *Python Data Science Handbook: Essential Tools for Working with Data*. O'Reilly.

Yonghong L., y McDowell. (2021, 06 26). *Calculus Misconceptions of Undergraduate Students* [Columbia University Libraries]. Columbia Academic Commons. https://academiccommons.columbia.edu/doi/10.7916/d8-vz70-4569

Zhang, L. (2019, July 6). *(PDF) A systematic review of learning computational thinking through Scratch in K-9*. ResearchGate. Retrieved August 26, 2024, from https://www.researchgate.net/publication/333944299_A_systematic_review_of_learning_computational_thinking_through_Scratch_in_K-9