Universidad de La Habana
Facultad de Matemática y Computación

Metodología para la predicción del desempeño de un estudiante en un curso dentro de un Entorno Virtual de Aprendizaje

Autor:
Jorge Alejandro Soler González

Tutores:

MSc. Carmen Fernández Montoto[1]
MSc. Jorge Miguel Soler McCook[2]

Trabajo de Diploma
presentado en opción al título de
Licenciado en Ciencia de la Computación

Enero 2024

[1]Facultad de Matemática y Computación, Universidad de la Habana, Cuba
[2]Universidad Metropolitana del Ecuador, Ecuador

A mi abuela Cirenia ...

Opinión del tutor

En los últimos años, en las instituciones educacionales existe una marcada tendencia hacia el uso de los Entornos Virtuales de Aprendizaje, los cuales permiten la propagación de su instrucción a mayor número de personas, apoyándose en las facilidades que ofrecen estas plataformas.

En estos entornos se produce una gran cantidad de información valiosa referente al aprendizaje de los alumnos que no siempre es objeto de análisis por parte de la institución, el profesorado, incluso del estudiantado, en aras de lograr mejores resultados en su proceso de enseñanza-aprendizaje.

La Minería de Datos Educacionales (MDE), es un campo de estudio que aporta un conjunto de técnicas que facilitan el análisis de grandes repositorios de datos generados o relacionados con las actividades de aprendizaje en los centros educativos, con el objetivo de orientar mejor el proceso de instrucción, evaluar el comportamiento del desempeño de sus estudiantes, desarrollar un mejor trabajo colaborativo en los educandos, entre otras muchas acciones proveniente del análisis efectuado.

El trabajo de investigación que se aborda en esta tesis, parte del análisis de las técnicas de minería de datos que faciliten la extracción de la información educacional y a partir de ella poder realizar predicciones del comportamiento del desempeño que llevan los estudiantes de un determinado curso, en aras de facilitar el proceso de enseñanza-aprendizaje.

El autor primeramente ha tenido que dedicar un tiempo considerable a profundizar en la plataforma Moodle desde el punto de vista usuario y desarrollador, profundizar en los registros necesarios para obtener la información necesaria para aplicar las técnicas de aprendizaje de máquina, determinar aquellas que se fueran mejores para establecer las posibles predicciones y brindarla a la institución para su posterior análisis para la toma de decisiones.

Ha tenido que realizar un estudio del estado del arte para determinar el camino a seguir, así como profundizar en las estructuras de la plataforma para poder insertarse en el contexto de esta. Para ello, ha tenido que consultar diferentes bibliografías, en su mayoría en idioma inglés.

En el desarrollo de esta tesis el estudiante evidenció su motivación por la investigación, la cual ha desarrollado de manera independiente, con profesionalidad,

aportando sugerencias válidas para el desarrollo de la misma. Demostró el dominio alcanzado en los contenidos estudiados y los resultados alcanzados avalan la calidad de la plataforma instrumentada.

Es de destacar la constancia y dedicación mostrada en el desarrollo de la investigación, así como las acciones que tuvo que desarrollar para recopilar los datos de los casos de estudios presentado, que no solo se redujo a los presentados en la memoria escrita.

Consideramos que el autor posee las habilidades necesarias de un profesional en Ciencia de la Computación, evidenciándose en la aplicación de los conocimientos adquiridos al desarrollo de otras áreas del saber, por tales razones proponemos al tribunal se le otorgue la calificación de excelente (5).

MSc. Carmen Fernández Montoto

MSc. Jorge Miguel Soler McCook

Resumen

Durante la pandemia de COVID-19, muchas universidades estuvieron en su mayor parte cerradas y sus aulas se transformaron a un formato totalmente en línea. Para los profesores era un desafío gestionar el aprendizaje virtual y, especialmente, realizar un seguimiento del comportamiento de los estudiantes, ya que no se podía establecer contacto interpersonal y por ende, el desempeño de los estudiantes no es fácil de controlar. Para aliviar este problema, una solución, que se ha vuelto cada vez más importante, es la predicción del desempeño de los estudiantes en función de sus datos históricos de registro en los entornos virtuales de aprendizaje. Este estudio, por tanto, tiene como objetivo analizar datos de comportamiento de los alumnos aplicando técnicas de aprendizaje automático a los registros de Moodle, con un total de 453941 registros. Se utilizaron cinco algoritmos de aprendizaje automático (*Random Forest*, Árbol de Decisión, Regresión Logística, Regresión Lineal y *Support Vector Machine*) para realizar la predicción académica. Además, se crearon dos conjuntos de datos con atributos distintos, los cuales se dividieron en cuatro etapas de progreso del curso (25 %, 50 %, 75 % y 100 %), a los que se le aplicó un proceso de selección de características con el algoritmo Boruta.

Los modelos de predicción podría guiar estudios futuros, motivar la autopreparación y reducir las tasas de abandono. En la investigación se evaluaron los modelos con validación cruzada quíntuple. Los resultados indicaron que en ambos conjuntos de datos el algoritmo de Regresión Lineal tuvo el mejor comportamiento en todas las etapas del curso.

Los resultados podrían aplicarse a otros cursos y en un registro más grande en entornos virtuales de aprendizaje que tengan condiciones similares de actividad de los estudiantes, en búsqueda de lograr una predicción más precisa del desempeño de los estudiantes.

Abstract

During the COVID-19 pandemic, many universities were mostly closed, and their classrooms shifted to a fully online format. For teachers, managing virtual learning posed a challenge, especially in monitoring student behavior, as interpersonal contact was not possible, making it difficult to control student performance. To alleviate this problem, an increasingly important solution has been predicting student performance based on their historical log data in virtual learning environments.

This study aims to analyze student behavior data by applying machine learning techniques to Moodle logs, totaling 453,941 records. Five machine learning algorithms (Random Forest, Decision Tree, Logistic Regression, Linear Regression, and Support Vector Machine) were used for academic prediction. Additionally, two datasets with different attributes were created, divided into four course progress stages (25 %, 50 %, 75 %, and 100 %), subjected to feature selection using the Boruta algorithm.

The predictive models could guide future studies, motivate self-preparation, and reduce dropout rates. The research evaluated the models with five-fold cross-validation. The results indicated that in both datasets, the Linear Regression algorithm performed the best at all course stages.

These findings could be applied to other courses and in a larger dataset in virtual learning environments with similar student activity conditions, aiming for a more accurate prediction of student performance.

Índice general

Índice de figuras

Introducción

La formación de competencias profesionales es un aspecto fundamental en el proceso de instrucción en las instituciones educativas, principalmente en la educación superior y tecnológica. En la actualidad, gracias a los avances tecnológicos, el aprendizaje se ha vuelto más accesible y flexible a través de los Entornos Virtuales de Aprendizaje (EVA). Recientemente, la humanidad se encuentra aún dando solución al problema global de pandemia, lo cual potenció la educación a distancia y todos sus mecanismos de enseñanza, por esta razón, la demanda de tecnología digital e innovadora para respaldar las tareas de enseñanza, administrar las clases y hacer un seguimiento de los alumnos se ha convertido en una parte crucial de la educación. [1]

En tiempos de pandemia la mayoría de los países fueron forzados a cerrar todas sus instituciones académicas y transitar al aprendizaje en línea. Aunque las clases virtuales eran portables, de fácil acceso y aumentaban las oportunidades de aprendizaje para adultos, todas las escuelas y universidades enfrentaron múltiples desafíos al requerir la adopción de programas de enseñanza no presencial.[2, 3] Muchos docentes y estudiantes, además, se encontraron con la difícil tarea de tener que continuar con sus cursos en esta modalidad, teniendo esto un impacto desfavorable en la calidad de algunos cursos.

Dentro de los problemas más típicos que se pueden citar, está el hecho de que los estudiantes se pierden el trabajo de laboratorio y reciben menos retroalimentación porque con frecuencia son demasiado tímidos para hacer preguntas durante el curso, mientras que los profesores carecen de contacto directo con sus estudiantes, no pueden observarlos ni valorarles el ritmo del aprendizaje oportunamente, y con ello notificar de inmediato a los estudiantes que pueden estar en riesgo académico. Muchos de estos factores aumentan la probabilidad de que los estudiantes reprueben, abandonen y se retiren antes de graduarse.

De igual manera, los entornos virtuales de aprendizaje han surgido como una valiosa herramienta en el ámbito educativo. Se tornan indispensables en el contexto pandémico que ha afectado a la humanidad. Estos entornos ofrecen a los estudiantes una experiencia de aprendizaje flexible y accesible, permitiéndoles adquirir conocimientos y desarrollar habilidades de manera eficiente y autónoma. Además, han demostrado su capacidad para adaptarse a las necesidades individuales de los estudiantes, brin-

dando la oportunidad de personalizar los contenidos y el ritmo de aprendizaje. A medida que la sociedad se adentra en una nueva era digital, los entornos virtuales de aprendizaje se han consolidado como una solución duradera y eficaz, que no solo ha llegado para quedarse, sino que continuará evolucionando y transformando la forma en que enseñamos y aprendemos.

Las universidades, para seguir con esta dinámica, deben proporcionar herramientas de aprendizaje en línea para apoyar la educación de manera eficiente. En este sentido [4], Moodle LMS(*Moodle Learning Management System*) ofrece un entorno de aprendizaje con software digital, de libre acceso y además, de código abierto, de amplia utilización en muchas universidades del mundo, en el cual los estudiantes obtienen acceso rápido y eficiente a los recursos y actividades de un curso, mientras los profesores pueden usarlo como una herramienta eficiente para administrar el aula a través de cuestionarios, tareas, exámenes y otras actividades. Adicionalmente esta plataforma permite recopilar datos de la actividad estudiantil y docente, generando un archivo de registros de acceso bastante extenso. Esta última ventaja se puede utilizar para el análisis y el pronóstico del rendimiento de un estudiante dentro de un curso. Por todas estas características esta investigación se centra en cursos impartidos en esta plataforma.

En estos entornos virtuales de aprendizaje como Moodle es un reto para los profesores obtener una métrica que permita analizar y predecir el desempeño de cada estudiante durante un curso en el logro de los objetivos y habilidades de estudio. Con esta investigación se busca desarrollar una herramienta que permita proveer, a tiempo, una atención personalizada a los estudiantes, logrando una intervención oportuna que permitiría conducir y afianzar la confianza de los estudiantes dentro del curso, a pesar de las brechas detectadas, en los resultados al finalizar este.

¿Qué facilita predecir el desempeño de un estudiante durante el curso?

- Entender el aprendizaje académico del estudiante en su trayectoria en cada objetivo académico dentro del entorno: esto, además, incluye el cómo está progresando hacia la consecución de los objetivos académicos específicos del curso. Esto permite a los profesores identificar áreas en las que el estudiante puede estar estancado y proporcionar intervenciones específicas para ayudarles a mejorar su rendimiento. Además, también puede ayudar a identificar las fortalezas del estudiante y permitir que se les brinden oportunidades para desarrollarlas aún más.

- Facilitar a los profesores entender las condiciones y realidades académicas en las que están sus estudiantes, esto incluye factores como la carga de trabajo del estudiante, y otros factores que pueden influir en su rendimiento académico. Al comprender mejor estas condiciones y realidades, los profesores pueden ajustar su enseñanza y proporcionar apoyo personalizado para ayudar a los estudiantes

a alcanzar las habilidades requeridas.

- Implementar planes personalizados de recomendación. Una vez que se ha realizado la predicción del rendimiento del estudiante, se pueden implementar planes personalizados de recomendación para ayudar a mejorar su rendimiento. Estos planes pueden incluir guías específicas para el estudio, la práctica y la preparación para exámenes, así como para el apoyo académico adicional, como tutorías o asesoramiento individualizado.

Una forma de detectar a los estudiantes que pudieran tener un desempeño medio o bajo dentro de un curso, es realizar predicciones tempranas de sus calificaciones en los diferentes temas que se abordan.

El uso de modelos de predicción para el análisis del aprendizaje, así como los sistemas de recomendación, presentan múltiples desafíos, especialmente el cómo obtener, procesar y usar los datos para construir el modelo de predicción apropiado. A este campo dentro de la analítica del aprendizaje se le llama: Minería de datos educacionales. En este estudio se proponen los siguientes indicadores, los cuales son extraídos del archivo de registros de Moodle:

- Cantidad de veces que el estudiante ha ingresado a las actividades y recursos de la plataforma.
- Nivel de participación en los tipos diferentes de actividades.
- Resultado de la evaluación que se otorga por el docente en cada actividad interactiva de cada objetivo académico.

Este preprocesamiento de los datos a realizarse en esta investigación tiene significativa importancia puesto que interviene directamente en la precisión y en el costo computacional del modelo predictivo a emplear.

Luego, ¿Sería posible establecer un modelo de predicción en los Entornos Virtuales de Aprendizaje que facilite el proceso de enseñanza con el aprendizaje personalizado de los estudiantes en los cursos montados en la plataforma de una institución? En el contexto de la presente tesis, el objetivo general es responder a la interrogante anterior desarrollando un modelo predictivo a partir de los datos extraídos y preprocesados de Moodle.

Como resultado de este trabajo se propone una metodología para predecir el comportamiento de un estudiante en un curso y, a partir de ella, se busca en futuros trabajos, ofrecer recomendaciones personalizadas adaptadas a las necesidades y preferencias individuales de cada estudiante, con el fin de mejorar su desempeño en el proceso de aprendizaje y aumentar su rendimiento académico.

Para lograr este objetivo se tendrán en cuenta los siguientes objetivos específicos:

1. Estudiar de la arquitectura de almacenamiento de la plataforma Moodle, dígase: Bases de datos, APIs de acceso a la información, descargas de archivos, etc., para poder realizar la extracción la información sobre las actividades realizadas por los estudiantes y sus evaluaciones.

2. Recopilar y procesar los datos pertinentes.

3. Construir el modelo predictivo: Implementación de algoritmo de clasificación de aprendizaje de máquina.

Este trabajo consta de 3 capítulos:

1. Marco teórico-conceptual: Se plantean conceptos importantes y se hace una revisión bibliográfica de soluciones similares encontradas.

2. La Propuesta: Se presenta la propuesta de solución, así como, la metodología llevada a cabo a lo largo de la investigación.

3. Detalles de Implementación y Resultados: Se expone detalladamente cada paso seguido durante la implementación y los resultados obtenidos.

Capítulo 1

Marco teórico-conceptual

1.1. Evolución de los LMS

La sociedad contemporánea tiene la transmisión de conocimientos como algo determinante para garantizar el futuro, históricamente la educación se ha tornado un elemento fundamental para el desarrollo de la sociedad por lo que se estableció como una actividad necesaria en la vida de todas las personas.

A través del tiempo la transmisión de información ha evolucionado de la mano de las tecnologías disponibles en cada época, en este apartado se expone de una de las herramientas de transmisión de la información más utilizadas en la educación en nuestra era, los Sistemas de Gestión de Aprendizaje, conocidos también como LMS, por sus siglas en inglés Learning Management System. [5]

1.1.1. ¿Qué es un LMS?

Un Sistema de Gestión de Aprendizaje (*Learning Management System*, LMS) es un tipo de software en línea que permite crear, implementar y desarrollar un programa de entrenamiento o un proceso de aprendizaje específico. Estos sistemas son utilizados por organizaciones empresariales, agencias gubernamentales; así como, instituciones educativas tradicionales, en la búsqueda de complementar o mejorar métodos educativos tradicionales mientras ahorran tiempo y recursos. Como principales características de estos entornos se puede citar:

- **Diseño responsivo**: Los usuarios pueden acceder al sistema desde cualquier dispositivo que elijan, ya sea computadora de escritorio, teléfono inteligente, tableta o laptop. El LMS desplegará automáticamente la versión más adecuada para cada uno de los dispositivos. Adicionalmente, los usuarios consiguen descargar contenido para trabajar sin tener que estar conectados a internet.

- **Interfaz amigable**: La interfaz del usuario permite a los estudiantes navegar fácilmente por la plataforma alineados a los objetivos y las habilidades, tanto del estudiante, profesor, como de la institución.
- **Reportes y analíticas**: Debe integrar herramientas de evaluación. Los profesores y administradores deben de contar con la capacidad para visualizar y revisar sus iniciativas didácticas para determinar su efectividad y realizar ajustes en caso necesario.
- **Manejo de cursos y catálogos**: Debe contener todos los cursos disponibles y accesibles para que administradores y profesores puedan manejar el catálogo y dirigirlo hacia una experiencia de aprendizaje enfocada a resultados.
- **Contenido interoperable e integrado**: El contenido creado y almacenado en un LMS debe empaquetarse de acuerdo a estándares educativos regulados por instituciones[1].
- **Servicio de soporte**: Los diferentes desarrolladores deben ofrecer distintos niveles de soporte que brinda la plataforma como foros de discusión, donde los usuarios puedan interactuar entre ellos, los *chats*, los números telefónicos de ayuda, etc.

1.1.2. Resumen histórico

El surgimiento de los software educativos parte desde 1960 con Plato (*Programmed Logic for Automated Teaching Operations*, Lógica programada para operaciones de enseñanza automatizadas) [6], considerado el primer software educativo, fue desarrollado en la Universidad de Ilinois por Donald BItzer. El sistema contaba con una interfaz táctil y los programas educativos eran desarrollados en Tutor, un lenguaje de programación orientado a la educación que contaba con herramientas como sistemas de grabación y ficheros de notas.

En 1987 NKI, universidad de Noruega desarrolla el primer sistema de educación a distancia, NKI Distance Education [7], con funcionalidades básicas de almacenamiento de recursos en línea, accesibles por todos los usuarios pertenecientes a la plataforma.

En 1990 nace FirstClass [8] un producto de Softarc una empresa tecnológica Sueca. al principio esta plataforma estuvo enfocada a solo los sistemas operativos de Macintosh(macOS). Esta plataforma incorporaba funcionalidades como foros y correos.

[1]SCORM : *Shareable Content Object Reference Model*, es uno de estos estándares. Estos empaquetados permitan crear objetos pedagógicos estructurados con objetivos fundamentales de facilitar la portabilidad de contenido de aprendizaje, poder compartir y reusarlo

Se puede decir que FirsClass fue la primera plataforma de *e-learning* que propició la interconexión entre profesores y alumnos.

Modular Object-Oriented Dynamic Learning, Moodle por sus siglas en inglés [9]. Probablemente la plataforma más conocida en la historia de los LMS fue lanzada en 2002 por Martin Dougiamas. En cuanto a las principales diferencias de esta plataforma respecto a sus antecesoras es la posibilidad de asignar roles como administrador, profesor y estudiante. Al ser una tecnología de código abierto, Moodle se ha convertido en una de los entornos preferidos en el momento de implementar proyectos académicos en entidades docentes, por su fácil adaptabilidad y personalización en base a las necesidades de cada institución. El presente estudio se centra en este LMS, en el siguiente apartado donde se profundiza acerca de sus principales características, funcionalidades y beneficios.

En 2008 los LMS comienzan a brindar sus servicios desde la nube a través del modelo SAAS(*Software as a Service*). La principal ventaja de brindar el servicio directamente desde la nube trajo un avance muy grande para el mundo de las LMS ya que facilitaba el acceso a la información educativa en cualquier lugar y desde cualquier dispositivo sin necesidad de instalaciones tediosas. El modelo SAAS se volvió el modelo más aplicado en cuanto a servicios de LMS, actualmente más del 70 % de los LMS actuales ofrecen sus servicios bajo este modelo.[10]

La actualidad de los LMS están cubiertos por tendencias que marcan el nuevo camino de estas plataformas, entre ellas se encuentran:

- **Aprendizaje Social**: El aprendizaje social o también conocido como *social learning* es una de las tendencias que las últimas plataformas LMS están incorporando, plataformas como Google Classroom o Edmodo llevan este tipo de tendencias en donde la información se presenta en un canal de comunicación y la información más antigua va quedando al final.

- **Gamificación**: La gamificacion es la aplicación de dinámicas propias de los juegos a los procesos de aprendizaje, aspectos como los puntos o los premios forman parte de esta herramienta, muchas plataformas han comenzado a incorporar estas funcionalidades para poder mejorar la interacción y el compromiso de los estudiantes.

- **Experiencia de Usuario**: La experiencia de usuario se ha tornado un factor determinante para la creación de experiencias de aprendizajes, el desarrollo de interfaces y el estudio del comportamiento del usuario para refinar los detalles de la plataforma dan una ventaja para la presentación de contenidos educativos. Los videos se han vuelto los más populares en cuanto a transmisión de contenidos, tener una plataformas simple y sencilla se ha vuelto en uno de los requerimientos más básicos para los proyectos de aprendizaje en línea.

1.1.3. Futuro de los LMS

El futuro de los LMS, como software en general, estará marcado por la flexibilidad y la adaptabilidad que pueda ofrecer a las nuevas tecnologías que surjan. En este sentido es importante destacar que la inteligencia artificial y los modelos de aprendizaje automático tendrán un papel fundamental en dicho futuro, herramientas que permitan analizar a profundidad los datos históricos almacenados en un LMS con el objetivo de entender la línea de aprendizaje de cada estudiante y de esa manera poder personalizar al máximo el desempeño a cada estudiante, marcarán sin duda el camino de los LMS en unos años. En ese sentido, la presente tesis aborda, posibles soluciones al análisis de esos datos y cómo emplearlos para el beneficio del usuario.

1.2. Modular Object-Oriented Dynamic Learning: MOODLE

Moodle [4] es una plataforma de aprendizaje diseñada para proporcionarle a educadores, administradores y estudiantes un sistema integrado único, robusto y seguro para crear ambientes de aprendizaje personalizados. Soportada financieramente por una red mundial de más de 80 compañías de servicio, impulsando a cientos de miles de ambientes de aprendizaje globalmente, cuenta con la confianza de instituciones y organizaciones grandes y pequeñas. El número de usuarios a nivel mundial, ascendió a más de 200 millones de usuarios en agosto del 2020 [11], entre usuarios académicos y empresariales, lo cual la convierten en la plataforma de aprendizaje más utilizada del mundo.

1.2.1. Principales características

Este estudio se centró en la versión estable de Moodle 4.2, que al igual que las anteriores posee una variedad considerable de características, entre las que se puede destacar:

- **Interfaz moderna y fácil de usar**: Diseñada para ser responsiva y accesible, la interfaz de Moodle es fácil de navegar, tanto en computadoras de escritorio como en dispositivos móviles.
- **Actividades y Herramientas colaborativas**: Cuenta con actividades interactivas de uso muy intuitivo para el usuario; los foros, las wikis, los glosarios, actividades de bases de datos son ejemplos de herramientas que se pueden desarrollar en Moodle con el objetivo de potenciar el aprendizaje colaborativo y hacer más ameno el proceso de aprendizaje.

- **Gestión conveniente de archivos**: Cuenta con la opción de trasladar archivos desde servicios de almacenamiento en la nube, incluyendo MS OneDrive, Dropbox y Google Drive.
- **Diseño personalizable de la plataforma**: Los administradores tienen la posibilidad de personalizar y adaptar la interfaz de Moodle a las necesidades de su institución u organización.

La plataforma es de código abierto y gratuita, se puede adaptar, extender o modificar, tanto para proyectos comerciales como no-comerciales, sin pago de cuotas por licenciamiento, esto, además, significa que Moodle siempre está en proceso de mejora y actualización, cuenta con una comunidad fiel de desarrolladores y con estándares de código para su mejora. Los administradores de las instituciones se pueden apoyar en esto para flexibilizar aún más la plataforma y de esa forma cubrir, de una manera más eficiente, las necesidades de dichas instituciones.

Su configuración modular y diseño inter-operable les permite a los desarrolladores el crear *plugins*[12] e integrar aplicaciones externas para lograr funcionalidades específicas, además, cuenta con una base de datos relacional (SQL) perfectamente escalable para soportar, desde clases pequeñas con pocos estudiantes hasta grandes organizaciones con millones de usuarios.

Para la comunicación con esta base de datos desde la perspectiva de un programador, Moodle cuenta con un API (del inglés, *Application Programming Interface*, en español, Interfaz de Programación de Aplicaciones), con todas las funciones necesarias para su modificación desde el código, permitiendo de esta manera extender MOODLE a aplicaciones externas o herramientas que monitoreen la plataforma en cuestión.

En esta investigación se utiliza esta API como principal forma de extracción de datos de la plataforma, en el siguiente capítulo se profundiza acerca de las diferentes funciones utilizadas así cómo la correcta forma de usarlas.

Otra de las características a tener en cuenta es que Moodle almacena, en su extensa base de datos, cada acción realizada por el usuario con fecha y hora, administradores y profesores tienen acceso a este registro de actividad realizada por los usuarios en sus respectivos cursos, o en la plataforma viéndolo de una manera más general, Moodle posibilita esto a través de reportes, accesibles en la misma plataforma y con posibilidad de descarga para su posterior análisis. A estas acciones se les llama "*logs*", el problema principal de estos reporte es que son muy tediosos de analizar por una persona, por lo que varios estudios y herramientas se han centrado en el análisis de estos, para de alguna manera "traducirlos" a un lenguaje más entendible por la persona interesada.

En este sentido Brian Sal Sarria, de la Universidad de Cantabria en España, en su tesis de grado de Ingeniería Informática [13] plantea una guía para la traducción de estos *logs*, en donde para acceder a estos, se debe descargar el reporte mencionado

anteriormente. Por otra parte en el estudio realizado por Cristóbal Romero Morales, del departamento de Ciencias de la Computación y Análisis Numérico de la Universidad de Córdoba en España, titulado "Aplicando Minería de Datos en Moodle" [14], utiliza un enfoque de trabajo directo con la base de datos de Moodle, emplea el lenguaje SQL para acceder a la información de las tablas de dicha base de datos. En el capítulo 2 se plantea la propuesta llevada a cabo en la presente tesis.

1.2.2. Futuro

El futuro de esta plataforma está altamente influenciado por la comunidad de desarrolladores que posee, así como por los usuarios que lo utilizan. Cualquier idea novedosa es bienvenida siempre que cumpla con los estándares de la plataforma. Como mismo el futuro de los LMS en general está marcado por la inteligencia artificial, Moodle no es una excepción, ya se trabaja en modelos de aprendizaje automático que se integrarán a la plataforma, con el objetivo de potenciar el aprendizaje personalizado y hacer de esta un lugar mucho más amigable, eficiente y productivo para estudiantes y profesores, también se valora la posibilidad de que cualquier desarrollador que proponga un modelo o algoritmo para resolver una problemática determinada, tenga su espacio en la plataforma, ya sea integrando el modelo o como una herramienta externa en forma de *plugin*.

1.3. Minería de datos educacionales

Las instituciones docentes disponen de una gran cantidad de información, la cual posee un alto valor pedagógico y una gran importancia para la evolución del proceso de aprendizaje. Este tipo de información puede utilizarse en la toma de decisiones para mejorar sus estrategias y políticas docentes educativas. Además, las estadísticas y el flujo de información que se produce durante el análisis de los datos educativos, favorecen al profesor para profundizar en sus investigaciones. No obstante, en la mayoría de las instituciones educativas, en sus entornos virtuales de aprendizaje, el gran volumen de información disponible no es aprovechado al máximo.

Para lidiar con esta situación, se han llevado a cabo múltiples esfuerzos en el diseño de herramientas que utilizan técnicas de extracción de conocimiento para procesar los datos obtenidos en los entornos educativos [15]. La disciplina que engloba este grupo de técnicas y metodologías se conoce como Minería de Datos Educacionales (*Educational Data Mining* - EDM) y recientemente se ha evidenciado un significativo avance en esta línea de investigación [16, 17].

Los objetivos que se persiguen al aplicar esas técnicas dependen de a quién va dirigido el conocimiento extraído. Los distintos roles que utilizan en EDM se pueden

clasificar en tres categorías principales, estudiantes, profesores e instituciones académicas [15].

Desde el punto de vista del estudiante tiene como objetivos establecer qué actividades, recursos y tareas podrían mejorar su rendimiento académico y su motivación. También resulta importante determinar qué actividades se ajustan mejor a su perfil y fijar qué camino recorrer para obtener un resultado concreto, basado en el desempeño del alumno en el camino recorrido hasta el momento, así como, por comparación con lo realizado por otros estudiantes de características similares [15].

Desde el punto de vista del profesor, con la EDM se pueden resolver múltiples problemas y tareas de gran incidencia en la efectividad de los métodos de aprendizaje [15].

Uno de los objetivos primarios sería evaluar la eficacia del proceso de enseñanza y en función de esta corregir el contenido de los cursos para mejorar la estructura de los mismos. Para llevar a cabo esta tarea, resulta indispensable monitorizar cada actividad y determinar su grado de dificultad y los errores más frecuentes cometidos en su ejecución.

Otro aspecto de gran importancia para los profesores es identificar las relaciones existentes entre los usuarios para organizarlos en grupos homogéneos. Este tipo de organización permitiría encausar los intereses comunes de los estudiantes para fomentar el desarrollo de grupos participativos y líneas de investigación, colaborando con personas dedicadas a temáticas afines.

Además, se puede incrementar la motivación de los estudiantes al personalizar y adaptar el contenido de cursos diseñando los planes de instrucción a partir de las características de cada grupo. Otra tarea consiste en investigar el comportamiento de los alumnos y su desempeño en los cursos para buscar patrones generales y patrones anómalos en su rendimiento.

Finalmente desde el punto de vista de las instituciones académicas los objetivos de la EDM son mejorar la eficiencia de los sitios web educativos así como adaptarlo a los hábitos de sus usuarios teniendo en cuenta el tamaño de servidor óptimo y la distribución del tráfico en la red [15].

El proceso de minería de datos en entornos de aprendizaje virtual, como Moodle, involucra varios pasos fundamentales. Moodle, una plataforma de aprendizaje gestionada, proporciona un amplio rango de datos que pueden ser utilizados para este fin, incluyendo información sobre la interacción de los estudiantes con los materiales del curso, sus patrones de navegación, resultados de evaluaciones, participaciones en foros y muchos otros indicadores.

Pasos en la Minería de Datos Educativos en Moodle:

1. **Recopilación de Datos**: En Moodle, la recopilación de datos puede realizarse manualmente o de forma automatizada.

- **Manualmente**: El administrador o el profesor puede descargar registros de actividades, calificaciones, y contribuciones en foros directamente desde la interfaz de Moodle en formatos como .csv o .xls. Estos registros contienen todo lo que ha hecho un usuario en la plataforma, contiene campos como: Componente Afectada, Nombre del evento, Descripción, Fecha y Hora, que son fundamentales para el análisis del comportamiento del usuario [13]. En este estudio se utilizan todos los campos anteriormente mencionados.
- **Automatizada**: Se puede utilizar API's de Moodle o *plugins* específicos que recolectan datos continuamente y los almacenan en una base de datos externa más robusta para su posterior análisis, además que se pueden realizar peticiones a la base de datos para extraer todos estos registros, que tienen los mismos campos mencionados anteriormente.[14]

En la presente tesis se utilizó un enfoque combinado, manejando el excel extraído manualmente y la obtención de datos necesarios a través de la API que brinda Moodle, todo esto se verá en el siguiente capítulo.

2. **Preprocesamiento**: Los datos extraídos deben ser limpiados y transformados para ser utilizados eficientemente. Esto puede implicar la eliminación de *outliers*, la corrección de errores y la transformación de datos en formatos que los algoritmos de minería de datos puedan procesar. En este sentido el campo de Descripción que viene en esos datos extraídos juega un papel esencial ya que tiene la mayor información a extraer, vienen datos como el id del usuario en cuestión, el id del curso afectado, así como el módulo utilizado. Como este campo es un texto la mejor técnica a emplear son las expresiones regulares,[13] de esta forma se pueden obtener los datos necesarios para el posterior análisis. En este paso también es importante escoger las características de esos datos que se van a emplear.

   En la bibliografía revisada hay mucha variedad en cuanto a los campos que se van a utilizar, además esto tiene mucho que ver con la problemática a resolver. En el caso de la predicción académica es común el uso de datos cuantitativos sobre la utilización de un curso por un estudiante, esto incluye por ejemplo: la cantidad de veces que el estudiante accedió a un recurso, la cantidad de veces que un estudiante participó en un forum, la cantidad de intentos realizados en un cuestionario.[18] Otros estudios además incluyen factores demográficos como el sexo, el lugar de residencia, el horario en el que acceden a la plataforma.[19] En este paso se debe ser capaz de transformar los datos extraídos en un formato que cumpla con esas características por cada estudiante.

3. **Análisis de Datos**: Utilizando algoritmos de minería de datos, se busca identificar patrones y tendencias en los datos. Esto implica técnicas como la agrupa-

ción, clasificación, reglas de asociación y análisis de secuencias. Dentro de este aspecto hay varios enfoques para el análisis en dependencia de la problemática en cuestión. Las principales técnicas de minería de datos que se emplean en este campo son:

a) **Agrupamiento(Clustering)**: Se utiliza principalmente para identificar grupos de estudiantes con patrones de comportamiento similares.[14]

b) **Reglas de asociación**: dentro de estas reglas el algoritmo Apriori ayuda a encontrar qué recursos o actividades están comúnmente están asociados con altas calificaciones.[14]

c) **Algoritmos de clasificación**: dentro de estos algoritmos se encuentran, los árboles de decisión, el *Random Forest*, regresión lineal, regresión logística, *Support Vector Machine*, redes neuronales, etc. Los cuales a partir de parámetros o atributos y una o varias clases de salida dentro de un conjunto de datos, realiza la clasificación.

En este estudio se utilizaron los algoritmos de clasificación para resolver la problemática de predecir el desempeño de un estudiante dentro de un curso virtual. Específicamente los algoritmos: Arboles de decisión, *Random Forest*, regresión lineal, regresión logística y *Support Vector Machine*.

4. **Resultados**: Los resultados deben ser interpretados y presentados de tal manera que los educadores puedan entenderlos fácilmente y tomar decisiones informadas para mejorar las estrategias de enseñanza. Esta etapa se divide a su vez en 3 partes:

a) **Visualización de Datos**: Representación gráfica de los resultados para facilitar su interpretación.

b) **Evaluación estadística**: Deben realizarse pruebas estadísticas para determinar la importancia de los patrones encontrados.

c) **Validación**: Comprobar los resultados obtenidos con expertos o verificar que se siguen cumpliendo estos patrones en cursos venideros.

Esta metodología de trabajo dividida en estos 4 pasos es la seguida por este estudio la cual se abundará en el siguiente capítulo.

En el siguiente epígrafe se verá las diferentes formas en que se puede aplicar el aprendizaje automático para resolver el problema de predicción académica.

1.4. Modelos de Aprendizaje Automático

El aprendizaje automático o aprendizaje de máquinas (del inglés, ML: *Machine Learning*) es una rama de la Inteligencia Artificial, cuyo objetivo es desarrollar

técnicas que permitan a los sistemas computacionales aprender mediante un modelo basado en los datos históricos de una entidad, que facilite el análisis y deducción del comportamiento de indicadores de la gestión, en aras de mejoras en la mencionada entidad.

El aprendizaje automático también está estrechamente relacionado con el reconocimiento de patrones puede ser visto como un intento de automatizar métodos matemáticos que faciliten el proceso de inducción del conocimiento.

El aprendizaje automático tiene una amplia gama de aplicaciones, incluyendo motores de búsqueda, diagnósticos médicos, detección de fraude en el uso de tarjetas de crédito, análisis de mercado para los diferentes sectores de actividad, clasificación de secuencias de ADN, reconocimiento del habla y del lenguaje escrito, juegos y robótica, evidentemente puede ser utilizado para el comportamiento del desempeño de los estudiantes en un determinado curso académico.

El aprendizaje automático tiene como resultado un modelo para resolver determinado problema. Entre los modelos se distinguen: los modelos geométricos, los probabilísticos, los lógicos, de agrupamiento y de clasificación.

Los algoritmos de aprendizaje automático conforman un conjunto de datos para crear un modelo. A medida que se introducen nuevos datos de entrada en el algoritmo de aprendizaje automático, se utiliza el modelo desarrollado para realizar una predicción.

Los pasos clave para crear un modelo de aprendizaje automático son:

1. Recopilación de datos: Compilación de información confiable para informar al modelo predictivo.

2. Preparación de datos: Realizar la preparación de los datos, como agrupar y seleccionar los datos relevantes. Los datos se dividen en dos conjuntos: los datos de entrenamiento que se utilizan en el aprendizaje y los datos de evaluación que sirven para medir la efectividad del modelo una vez entenado.

3. Elegir un modelo: Existen muchos prototipos de aprendizaje automático, y algunos se adaptan mejor a casos de uso específicos que otros, por ende se debe seleccionar aquel que garantice mejorar la eficacia y precisión con el tiempo.

4. Entrenamiento: Los datos refinados son elegidos para mejorar la capacidad predictiva del modelo.

5. Evaluación: Introducir de nuevos datos para comprobar la efectividad de sus capacidades predictivas.

6. Ajuste de parámetros: Ajustar los parámetros de prueba específicos que puedan amoldarse para producir mejores resultados.

Existen varios tipos de algoritmos de aprendizajes:

- Aprendizaje supervisado: Los algoritmos de aprendizaje supervisado basan su aprendizaje en un juego de datos de entrenamiento previamente etiquetados.
- Aprendizaje no supervisado: Los métodos no supervisados (*unsupervised methods*) son algoritmos que basan su proceso de entrenamiento en un juego de datos sin etiquetas o clases previamente definidas. Está dedicado a las tareas de agrupamiento, también llamadas *clustering* o segmentación, donde su objetivo es encontrar grupos similares en el conjunto de datos.
- Aprendizaje semi-supervisado: es una clase de técnicas de aprendizaje automático que utiliza datos de entrenamiento tanto etiquetados como no etiquetados: normalmente una pequeña cantidad de datos etiquetados junto a una gran cantidad de datos no etiquetados.
- Aprendizaje por refuerzo: es un área del aprendizaje automático inspirada en la psicología conductista, cuya ocupación es determinar qué acciones debe escoger un agente de software en un entorno dado con el fin de maximizar alguna noción de "recompensa".

Para este estudio se utilizaron algoritmos de aprendizaje supervisado (Árboles de decisión, *Random Forest*, regresión lineal, regresión logística, *Support Vector Machine*), ya que los datos recopilados están previamente etiquetados.

El aprendizaje automático se ha utilizado en el proceso de enseñanza-aprendizaje haciéndose énfasis en los entornos virtuales de aprendizaje convirtiéndose en una dirección importante de las aplicaciones de minería de datos educativos [20], en aras de lograr mejoras en la identificación del comportamiento de los estudiantes, la evaluación de calificaciones de aprobado, reprobado y, especialmente en la predicción del desempeño de los estudiantes. Por tanto, el desarrollo y combinación de educación y aprendizaje automático es la tendencia actual en los centros educacionales [21].

La mayoría de los trabajos publicados se concentran en el uso de un único algoritmo de clasificación. Por ejemplo, en [22] se utilizó la regresión logística para analizar cursos de variables predictivas de LMS y predecir las calificaciones de los estudiantes durante 10 semanas. Los resultados indicaron que las calificaciones de las evaluaciones se relacionan con las calificaciones finales, mientras que los eventos como debates, foros o uso de wiki son un predictor menos confiable de la calificación final.

En [23], se aplicó *Random Forest* para construir un modelo utilizando eventos de registro (conferencias, cuestionarios, laboratorios, videos) para predecir el fracaso de los estudiantes dentro de un curso, la precisión fue de 96.3 %. Revelaron que los resultados de las puntuaciones de laboratorio son el predictor más sólido en este estudio.

Asimismo, [24] utiliza datos de cuatro cursos, como material semanal, videos conferencias, cuestionarios y ejercicios, de la Universidad Van Lang en Vietnam, aplicando un clasificador de regresión lineal para pronosticar el riesgo de reprobar el curso. Se encontró que los estudiantes con menos del 37 % de interacción estaban en riesgo de reprobar. Se realizó un análisis de 30000 expedientes estudiantiles [25] que incluyó cinco indicadores como rendimiento académico, tareas, acceso, aspectos sociales y cuestionarios, los cuales revelaron que luego de aplicar el árbol de regresión, el modelo implementado tuvo una tasa de precisión de 89.70 %.

En [26] se propone un árbol de decisión para predecir los resultados de los estudiantes en riesgo en tres niveles (alto, promedio y bajo). Se recopilaron muchas características (género, sesión, duración de la clase, GPA, especialización título, año, asistencia y puntuación de mitad de período) de los estudiantes del curso de Introducción a la Programación en el Buraimi University College en Omán. La precisión del modelo fue del 87.88 %, demostrando su efectividad.

De manera similar se propuso el modelo de árbol de decisión para predecir las tasas de abandono estudiantil en la Universidad Phayao en Tailandia [27]. Se analizó un conjunto de datos de 397 estudiantes y se consideraron las causas de deserción. El resultado produjo una precisión general de alrededor del 87.21 % de precisión.

Aunque el rendimiento de la predicción funciona con un único algoritmo, en este estudio se exploran y comparan diferentes modelos con el objetivo de encontrar el de mejor precisión. Además, ninguno ofrece una comparación de la predicción en las diferentes etapas del curso, y tampoco se utilizan diversos predictores de entrada en términos de predicción temprana del desempeño de los estudiantes. En el siguiente capítulo se abordará con más profundidad este tema.

Capítulo 2

Propuesta

2.1. Antecedentes

Como ya se ha mencionado la educación a distancia ha experimentado un crecimiento significativo en los últimos años, y con ese crecimiento la necesidad de soluciones innovadores a los diferentes problemas que se presentan dentro de este campo, uno de los problemas más comunes es el abandono y el bajo rendimiento de los estudiantes en los cursos virtuales, este problema puede estar influenciado por varios factores, pero sin duda uno de ellos es la poca enseñanza personalizada que existe hoy en día en esa modalidad de enseñanza, lo que conlleva a un desinterés por parte del estudiante a la hora de abordar un curso.

Históricamente la forma de lograr una enseñanza personalizada en la educación tradicional, es basada en la percepción que puede tener el docente dentro de un aula, pero en la educación distancia cambia porque los profesores no ven continuamente a sus estudiantes. Por lo tanto, lograr una predicción temprana del rendimiento que pueda tener un estudiante dentro de un curso virtual pasa a ser una parte crucial en el camino de evitar el abandono o el desinterés en el curso por parte de los estudiantes, ya que esto permite una toma de medidas temprana para ayudar a ese grupo de estudiantes que están potencialmente en riesgo.

La educación virtual, y su influencia en el rendimiento académico de los estudiantes puede ser tanto positiva como negativa, dependiendo de la calidad de la instrucción y la adaptación a la modalidad virtual. Moodle puede proporcionar una experiencia de aprendizaje enriquecedora y efectiva si se utiliza adecuadamente, pero también puede ser un obstáculo si no se adapta correctamente a las necesidades de los estudiantes. Por lo tanto, es importante lograr tener cursos atractivos y de fácil manejo por los estudiantes, utilizando las diferentes herramientas de Moodle, logrando así un curso verdaderamente interactivo donde el estudiante se sienta motivado, ya que esto influye directamente en el rendimiento académico del estudiante.

Identificar a los estudiantes que tienen más probabilidades de tener dificultades, se pueden tomar medidas preventivas para mejorar su rendimiento y reducir la tasa de deserción. De similar manera, si se identifica las mejores prácticas para perfeccionar el rendimiento académico de los estudiantes, estas se pueden implementar en otros cursos para mejorar la calidad de la enseñanza-apredizaje, la cual se logra mediante la integración de *plugins* desarrollados por los programadores, con el objetivo de poder utilizar la herramienta en los cursos venideros.

Para analizar y predecir el desempeño de los estudiantes en cursos virtuales, es necesario desarrollar modelos predictivos de rendimiento académico, utilizando datos generados en la gestión educativa y aplicando técnicas de aprendizaje automático. Estos modelos pueden ayudar a identificar a los estudiantes que tienen más probabilidades de tener dificultades y proporcionar apoyo adicional, por parte del docente, para mejorar su rendimiento; así como identificar las mejores prácticas para elevar el rendimiento académico de los estudiantes más sobresalientes y con ello mejorar la calidad del proceso enseñanza-aprendizaje, en los EVA.

2.2. Contexto del estudio

A partir del 2021 los procesos educacionales se vieron afectados por la pandemia de COVID-19 producto del aislamiento interpersonal que sufrió la humanidad por casi dos años, a partir de ello las instituciones educacionales tuvieron que cambiar drásticamente su modalidad hacia la forma de enseñanza en línea o virtual, la cual ha llegado para quedarse. Por esta razón la demanda de tecnología digital innovadora se ha convertido en una parte crucial de la educación para respaldar las tareas docentes, educativas, el aprendizaje auto-gestionado y colaborativo, gestionar el proceso enseñanza aprendizaje mediado por las tecnologías, así como y realizar un seguimiento personalizado de los alumnos en el proceso de compresión y asimilación de los contenidos y las habilidades alcanzadas en cada curso matriculado.

La tendencia hacia un aprendizaje centrado en el estudiante y que responda a sus necesidades ha aumentado de manera significativa el uso de entornos virtuales de aprendizaje. En este sentido, como se explicó en el capítulo anterior, Moodle es una plataforma que ofrece un ambiente ideal para llevar a cabo este enfoque, debido a sus potencialidades para la gestión de contenidos y la interrelación entre docentes y estudiantes, lo cual ha provocada su amplia utilización en los centros educacionales.

Ejemplos del uso expansivo de esta plataforma se pueden citar los centros de altos de estudio asociados al Ministerio de Educación Superior de Cuba, La Universidad Metropolitana del Ecuador, Universidad de Barcelona, Universidad de Harvard, entre otras muchas a nivel internacional.

Para este estudio se utilizaron datos brindados por la Universidad Metropolitana del Ecuador, que posee un sistema de educación a distancia en Moodle desde el 2018,

y que a partir de la pandemia se hizo mucho énfasis en esta modalidad de enseñanza y desde el 2020 se implementa un calendario por períodos, donde se imparten todo tipo de cursos. Los datos a los que se tuvieron acceso pertenecen a los cursos de grado de aprobación regular del primer semestre del 2023, en total son 401 cursos, con 453941 registros de actividad desde el inicio del período el 2 de abril del 2023 hasta el 1ro de julio del 2023. El archivo de registros proporcionado se muestran en la Figura 2.1

Hora	Contexto del evento	Componente	Nombre evento	Descripción
1/07/2023 23:53	Tarea: Actividad eva	Archivos enviados	Un fichero ha sido subid	The user with id '5318' has uploaded a file to the submission with id '573144' in the
1/07/2023 23:45	Tarea: Actividad No.	Tarea	Se ha enviado una entre	The user with id '5671' has submitted the submission with id '573007' for the assign
1/07/2023 23:45	Tarea: Actividad No.	Archivos enviados	Entrega creada.	The user with id '5671' created a file submission and uploaded '1' file/s in the assign
1/07/2023 23:45	Tarea: Actividad No.	Archivos enviados	Un fichero ha sido subid	The user with id '5671' has uploaded a file to the submission with id '573007' in the
1/07/2023 23:43	Tarea: Actividad No.	Tarea	Se ha enviado una entre	The user with id '5610' has submitted the submission with id '573941' for the assign
1/07/2023 23:43	Tarea: Actividad No.	Archivos enviados	Entrega creada.	The user with id '5610' created a file submission and uploaded '1' file/s in the assign
1/07/2023 23:43	Tarea: Actividad No.	Archivos enviados	Un fichero ha sido subid	The user with id '5610' has uploaded a file to the submission with id '573941' in the
1/07/2023 23:42	Tarea: Actividad No.	Tarea	Se ha enviado una entre	The user with id '4388' has submitted the submission with id '568340' for the assign
1/07/2023 23:42	Tarea: Actividad No.	Archivos enviados	Entrega creada.	The user with id '4388' created a file submission and uploaded '1' file/s in the assign
1/07/2023 23:42	Tarea: Actividad No.	Archivos enviados	Un fichero ha sido subid	The user with id '4388' has uploaded a file to the submission with id '568340' in the
1/07/2023 23:40	Tarea: Actividad No.	Tarea	Se ha enviado una entre	The user with id '4642' has submitted the submission with id '573944' for the assign
1/07/2023 23:40	Tarea: Actividad No.	Archivos enviados	Envío actualizado.	The user with id '4642' updated a file submission and uploaded '1' file/s in the assig
1/07/2023 23:40	Tarea: Actividad No.	Archivos enviados	Un fichero ha sido subid	The user with id '4642' has uploaded a file to the submission with id '573944' in the
1/07/2023 23:39	Tarea: Actividad No.	Tarea	Se ha enviado una entre	The user with id '4363' has submitted the submission with id '572363' for the assign
1/07/2023 23:39	Tarea: Actividad No.	Archivos enviados	Entrega creada.	The user with id '4363' created a file submission and uploaded '1' file/s in the assign
1/07/2023 23:39	Tarea: Actividad No.	Archivos enviados	Un fichero ha sido subid	The user with id '4363' has uploaded a file to the submission with id '572363' in the
1/07/2023 23:33	Tarea: Actividad No.	Tarea	Se ha enviado una entre	The user with id '4642' has submitted the submission with id '573944' for the assign
1/07/2023 23:33	Tarea: Actividad No.	Archivos enviados	Entrega creada.	The user with id '4642' created a file submission and uploaded '1' file/s in the assign
1/07/2023 23:33	Tarea: Actividad No.	Archivos enviados	Un fichero ha sido subid	The user with id '4642' has uploaded a file to the submission with id '573944' in the
1/07/2023 23:32	Tarea: Actividad No.	Tarea	Se ha enviado una entre	The user with id '4608' has submitted the submission with id '572477' for the assign
1/07/2023 23:32	Tarea: Actividad No.	Archivos enviados	Entrega creada.	The user with id '4608' created a file submission and uploaded '1' file/s in the assign

Figura 2.1: Muestra del archivo de registros proporcionado por la mencionada universidad

En estos registros también se encuentran eventos realizados por los profesores, pero para la finalidad de esta investigación solamente se tienen en cuenta las actividades relacionadas con los estudiantes.

Los 401 cursos se filtraron por los cursos que tuvieran bien establecido el libro de calificaciones de Moodle, que es lo que proporciona la nota final del estudiante dentro del curso. Se asume que el estudiante que tenga una nota por debajo de los 70 puntos su desempeño fue insatisfactorio. Finalmente se utilizaron 271 cursos, con un total de 2885 estudiantes.

Las actividades y recursos dentro de un curso se les llama módulos, atendiendo a los datos facilitados de la universidad antes mencionada, se tuvo en cuenta los módulos más utilizados por los estudiantes, se llega a la conclusión de utilizar los registros relacionados con los siguientes tipos de módulos:

- Tarea
- Glosario
- Foro

- Cuestionario
- Carpeta
- Recurso
- URL

En el epígrafe 2.3 se verá la propuesta de procesamiento de datos llevada a cabo.

2.3. Propuesta de procesamiento de datos

En esta tesis se presentan técnicas de análisis predictivo para la investigación de los datos referentes al comportamiento estudiantil. Se cuentan con 453.941 registros obtenidos de la plataforma Moodle de la mencionada institución universitaria. El objetivo es lograr predecir el desempeño académico de los estudiantes en las etapas del curso y hacer una comparación en el rendimiento de los diferentes modelos de aprendizaje automático utilizados (*Random Forest*, Árboles de Decisión, regresión logística, regresión lineal y *Support Vector Machine*).

El diseño para la predicción se dividió en cuatro etapas fundamentales como se muestra en la figura 2.2:

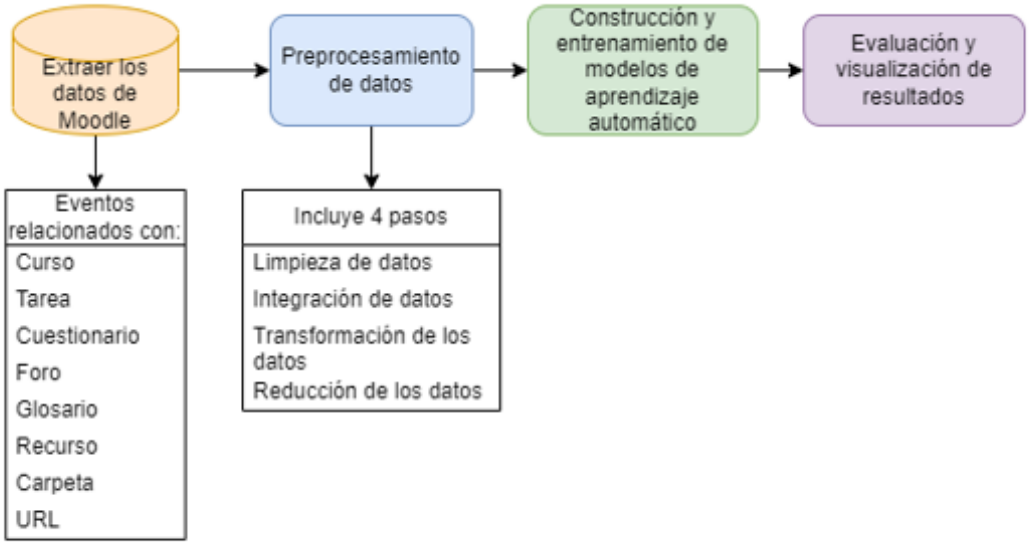


Figura 2.2: Etapas de la investigación

A continuación se expone la propuesta de solución en cada etapa.

2.3.1. Recopilación de datos de Moodle

Para el proceso de predicción se necesitan los datos de interacción de los estudiantes con la plataforma, así como la calificación del estudiante en cuestión, dentro del curso donde ha matriculado. Para esto Moodle cuenta con la opción de exportar un archivo de registros donde se encuentran todas las acciones realizadas dentro de los cursos que se quieran analizar. Este archivo de registros cuenta con la información tanto de estudiantes como de docentes. En el presente trabajo solo se abordan usuarios con rol de estudiante.

Una vez obtenidos los registros, se necesita contrastar las acciones llevadas a cabo por los estudiantes con el desempeño obtenido en el curso con la nota final alcanzada. Las calificaciones obtenidas por un estudiante en un curso están organizadas en una estructura de categorías evaluativas a las que pertenecen las diferentes actividades. La organización de estas categorías y la relación y la formulación entre los resultados alcanzados en las mismas es lo que se denomina: el libro de calificaciones en Moodle.

Luego ¿ cómo funciona el libro de calificaciones en Moodle ? Dentro de un curso existen varias etapas con objetivos a cumplir. Esas etapas se pueden entender como cortes parciales dentro de un curso. Dentro del libro de calificaciones estos cortes parciales se representan por categorías evaluativas, a estas categorías se le definen el peso que tendrá la nota alcanzada por el estudiante en ese corte parcial, con respecto a la nota total. Toda actividad dentro de una categoría de evaluación tiene un peso dentro de esa categoría.

$$NT = \sum_{i=1}^{m} PC_i NC_i$$

Donde:

- NT = Nota total
- m = Número de cortes parciales
- PC_i = Peso asignado a cada corte parcial (El examen final debe estar categorizado como un corte parcial, para que tenga su peso asignado)
- $NC_i = \sum_{j=1}^{k} PA_j NA_j$ donde:
  - PA_j = Peso asignado a cada actividad dentro del corte parcial i
  - NA_j = Nota alcanzada en cada actividad

Para este trabajo se analizaron los cursos que cumplen con la fórmula anterior, es decir, todas sus actividades están categorizadas y es posible calcular la nota final del estudiante mediante esta fórmula.

Para extraer estos datos se utiliza la API proporcionada por Moodle que permite el acceso a toda la información que se registra en esta plataforma. Como mecanismo de seguridad para realizar los llamados a las funciones que permiten obtener estos datos, se utiliza un *token* de acceso que se gestiona con el rol de administrador de la plataforma. Estas funciones gestionan (obtener, modificar, crear o eliminar) cualquier módulo, curso o usuario de Moodle. En este estudio se utilizaron las siguientes funciones:

- `gradereport_overview_get_course_grades` : A partir del id de un usuario, devuelve las calificaciones finales de todos los cursos en los que ha participado.
- `gradereport_user_get_grade_items` : A partir del id de un curso y el id de un usuario, devuelve el libro de calificaciones asociado.
- `core_enrol_get_enrolled_users` : A partir del id de un curso, devuelve los usuarios matriculados en dicho curso y el rol con asignado.
- `core_course_get_contents` : A partir del id de un curso, devuelve el contenido del curso, dividido por secciones y módulos.
- `core_course_get_courses_by_field` : A partir de un campo determinado, se obtiene el id de un curso.

Con el acceso al API de Moodle, para cada curso se extrae toda la información relativa a su contenido (todo tipo de módulos a través de los cuales el docente establece un recurso o actividad en su curso), matrícula y las calificaciones otorgadas a cada estudiante. Las funciones utilizadas para obtener este tipo de información son de solo lectura garantizando la posibilidad de construir una herramienta, que incluso pueda repetir varias veces el proceso de investigación, sin correr el riesgo de que se modifique ningún registro del LMS. El resultado que que se obtiene en cada función del API se recibe en el formato JSON y esa respuesta se interpreta y se almacena para su posterior integración con el archivo de registros extraído de Moodle y de conjunto crear el *dataset*.

La cantidad de cursos influye en el tiempo de duración para el proceso de extracción de los datos, ya que por cada curso hay que hacer un conjunto de llamados a la API. Cada llamado tiene un costo de conexión a una base de datos remota y a la obtención de información mediante consultas de un volumen importante y creciente de información. Para resolver esta problemática se utilizaron técnicas de programación en paralelo, de manera que se puedan realizar las mismas operaciones en diferentes

hilos. Con estas acciones se optimizó el tiempo de prospección de datos de 45 minutos en modo secuencial a 12 minutos en paralelo.

2.3.2. Preprocesamiento de los datos

El propósito de preprocesar los datos es filtrarlos para su posterior análisis y modelado. El proceso de limpieza y conversión de datos sin procesar que conducen al procesamiento y al análisis se conoce como "preparación de datos". Es una fase vital antes de procesar los datos que normalmente implica reformatearlos, realizar correcciones e integrar fuentes para aclararlos. En este estudio se crearon dos *datasets* con atributos diferentes y se analizaron todos los modelos con ambos conjuntos de datos.

Dentro del archivo de registros de Moodle vienen, entre otras cosas, la fecha en que ocurrió la interacción. Conociendo la duración del curso (en este caso todos tienen la misma duración, que coincide con la duración del período académico) , se dividieron ambos *datasets* en cuatro nuevos conjuntos, uno con el 25 % del curso, otro con el 50 %, otro con el 75 % y otro con el 100 % del curso, se analizaron todos los algoritmos en cada uno de los *datasets* con el objetivo de encontrar el mejor modelo de predicción para cada etapa del curso.

El primer conjunto de datos está hecho desde un punto de vista más general del estudiante. Se reunieron todas las interacciones por cada uno de los eventos relacionados con los módulos mencionados en la figura 2.2. Cada estudiante es un vector, donde cada componente es un número natural que expresa el total de interacciones correspondiente al atributo de esa componente, los atributos son los siguientes:

- Tarea
- Glosario
- Cuestionario
- Foro
- Carpeta
- Recurso
- URL
- Estado : Si está aprobado o no.

Este conteo se logra a partir del campo **Componente** dentro del archivo de registros de Moodle, el cual asocia cada interacción con el módulo correspondiente. En este *dataset* no se tiene en cuenta el evento relacionado con este módulo.

El segundo conjunto está hecho teniendo en cuenta aspectos mucho más específicos de las interacciones. Los registros de cada uno de los atributos vistos en el *dataset* anterior se dividen en nuevos atributos relacionados con el módulo en cuestión, logrando así una visión más detallada del comportamiento del estudiante. En este *dataset* se tendrá en cuenta el campo **Evento** del archivo de registros. Se realiza de la siguiente manera:

- Tarea : Se contabilizan por separado los eventos de vista y entrega de una tarea, añadiendo los siguientes atributos:
  - Vista de Tarea: Asociado al módulo **Tarea** y al evento **Módulo de curso visto**.
  - Entrega de Tarea: Asociado al módulo **Tarea** y al evento **Se ha enviado una entrega**.
- Cuestionario: Se contabilizan por separado los eventos de vista, envío e intento, añadiendo los siguientes atributos:
  - Vista de Cuestionario: Asociado al módulo **Cuestionario** y al evento **Módulo de curso visto**.
  - Intento de Cuestionario: Asociado al módulo **Cuestionario** y al evento **Ha comenzado el intento**.
  - Entrega de Cuestionario: Asociado al módulo **Cuestionario** y al evento **Intento enviado**.
- Foro: Se contabilizan por separado las vistas del foro y la participación en este, añadiendo los siguientes atributos:
  - Vista de Foro: Asociado al módulo **Foro** y al evento **Módulo de curso visto**.
  - Participación en Foro: Asociado al módulo **Foro** y a los eventos **Tema creado**, **Algún contenido ha sido publicado**, **Mensaje actualizado**, **Mensaje creado**, **Suscripción activada**.

Lo relacionado con los demás módulos (Recurso, Carpeta y URL) se mantiene igual al *dataset* anterior, porque el único evento asociado con ellos es **Módulo de curso visto**, por lo tanto lo que se contabiliza en cada uno son las vistas. Además, se añadieron 2 nuevos atributos al *dataset* que tienen que ver con las fechas de acceso por los estudiantes:

- TAD: *total access days* por sus siglas en inglés, se cuenta cada día (único) en el que el estudiante realizó una acción en la plataforma.
- ADS: *access density score* por sus siglas en inglés, se calcula a partir de la división entre el TAD y el total de días que dura el curso, logrando de esta manera una métrica del esfuerzo hecho por el estudiante.

Esta idea viene del estudio [28] donde utilizan esas variables en su *dataset*, logrando un mejor análisis de lo que sucede con el estudiante en la plataforma. Además, por cuestiones de experimentación se añadieron 4 nuevas variables que describen el horario del día en el que los estudiantes realizan las acciones:

- AM+: contabiliza las acciones realizadas de 00:00 hasta las 6:00.
- AM- : contabiliza las acciones realizadas de 6:01 has las 12:00.
- PM+ : contabiliza las acciones realizadas de 12:01 has las 18:00.
- PM- : contabiliza las acciones realizadas de 18:01 has las 23:59.

Para lograr obtener ambos *datasets* se siguió una metodología dividida en 5 pasos:

1. **Limpieza da datos**: es el proceso de corregir o eliminar datos incorrectos, corruptos, con mal formato, duplicados o incompletos de un conjunto. Es un proceso necesario para remover imperfecciones, inexactitudes y datos distorsionados. En este caso, para limpiar los registros a los que se accedieron, se eliminaron las interacciones de los profesores y del administrador ya que no tienen ningún impacto en este proceso. Además, una vez obtenidas las calificaciones, el archivo de registros pasa por otro filtro donde se eliminan todas las interacciones de los cursos que no cuentan con calificación final, ya sea porque no la tienen, o porque su libro de calificación es incorrecto.
2. **Integración de datos**: es el proceso de fusionar datos de numerosos sistemas fuente para crear conjuntos unificados de información con fines analíticos. Su propósito es crear conjuntos de datos limpios y consistentes que satisfagan las necesidades de información de los usuarios. En este caso, se fusionan las calificaciones de los estudiantes con los registros de Moodle.

Descripción
The user with id '5318' has uploaded a file to the submission with id '573144' in the assignment activity with course module id '657652'.
The user with id '5671' has submitted the submission with id '573007' for the assignment with course module id '610894'.
The user with id '5671' created a file submission and uploaded '1' file/s in the assignment with course module id '610894'.
The user with id '5671' has uploaded a file to the submission with id '573007' in the assignment activity with course module id '610894'.
The user with id '5610' has submitted the submission with id '573941' for the assignment with course module id '595468'.
The user with id '5610' created a file submission and uploaded '1' file/s in the assignment with course module id '595468'.
The user with id '5610' has uploaded a file to the submission with id '573941' in the assignment activity with course module id '595468'.
The user with id '4388' has submitted the submission with id '568340' for the assignment with course module id '603663'.
The user with id '4388' created a file submission and uploaded '1' file/s in the assignment with course module id '603663'.
The user with id '4388' has uploaded a file to the submission with id '568340' in the assignment activity with course module id '603663'.
The user with id '4642' has submitted the submission with id '573944' for the assignment with course module id '603663'.
The user with id '4642' updated a file submission and uploaded '1' file/s in the assignment with course module id '603663'.
The user with id '4642' has uploaded a file to the submission with id '573944' in the assignment activity with course module id '603663'.
The user with id '4363' has submitted the submission with id '572363' for the assignment with course module id '606563'.
The user with id '4363' created a file submission and uploaded '1' file/s in the assignment with course module id '606563'.
The user with id '4363' has uploaded a file to the submission with id '572363' in the assignment activity with course module id '606563'.
The user with id '4642' has submitted the submission with id '573944' for the assignment with course module id '603663'.
The user with id '4642' created a file submission and uploaded '1' file/s in the assignment with course module id '603663'.
The user with id '4642' has uploaded a file to the submission with id '573944' in the assignment activity with course module id '603663'.
The user with id '4608' has submitted the submission with id '572477' for the assignment with course module id '605725'.
The user with id '4608' created a file submission and uploaded '1' file/s in the assignment with course module id '605725'.
The user with id '4608' has uploaded a file to the submission with id '572477' in the assignment activity with course module id '605725'.
The user with id '4403' has submitted the submission with id '570520' for the assignment with course module id '591314'.

Figura 2.3: Campo descripción del archivo de registros de Moodle

3. **Transformación de datos**: La transformación de datos es el método de convertir datos de un formato a otro, generalmente del formato de un sistema origen al formato de un sistema destino. En este caso, se debe transformar el archivo de registros de Moodle a un formato que se adapte a las necesidades de los modelos de clasificación que se emplea. Un primer ejemplo, radica en el campo "Descripción" de este archivo de registros, que muestra su contenido como en la figura 2.3. En la descripción aparece el id del estudiante, el id del módulo y en ocasiones, también el id del curso donde se está realizando la acción. Todo esto es información fundamental a la hora de organizar el *dataset* que se desea obtener. El id del usuario se debe extraer para encontrar la calificación del mismo dentro del curso. Por otro lado, el id del módulo se extrae para saber a qué curso pertenece dicha acción. Por lo tanto, este campo se transforma en tres campos distintos por cada fila: el id del usuario, el id del módulo y el id del curso. Los mismos se obtuvieron con la utilización de expresiones regulares, que son cadenas de caracteres que permiten, a partir de un patrón en el texto, identificarlo y extraer la correspondiente información.

   Una vez obtenidas estas nuevas columnas, agrupando por estudiante y luego por el curso, se obtienen todas las interacciones que posee un estudiante dentro de un curso determinado, y se empieza a contabilizar por cada uno de los atributos mencionados anteriormente correspondientes al *dataset* que se desea formar. Además, para la predicción se convierten los valores numéricos de las califica-

ciones en valores binarios, de aprobado o no en el curso. Como se mencionó en el epígrafe 2.1, en este estudio se asume que un estudiante está desaprobado cuando tiene menos de 70 puntos y en caso contrario está aprobado.

4. **Reducción de los datos**: La reducción de datos es la técnica que se emplea para obtener una representación reducida del conjunto manteniendo la integridad de los original de los mismos. En este estudio, por la cantidad de datos con los que se cuenta, no fue necesario realizar este paso.

5. **Selección de características**: Una vez obtenidos los *datasets*, se realizó un trabajo de selección de características dentro de los mismos. Con el objetivo de encontrar los atributos que más inciden en la predicción académica de un estudiante.

   La selección de características es el proceso de escoger un subconjunto de atributos relevantes para construir modelos de aprendizaje robustos. Esta selección se clasifica en tres tipos: *Wrapper*, *Filter* e Híbridos. El primero emplea un algoritmo de aprendizaje automático para evaluar la fiabilidad de un conjunto de características; entre estos destacan dos algoritmos a mencionar: Boruta e Importancia de permutación (*Permutation importance*). Mientras que el método *Filter* utiliza las características de los datos para evaluar su importancia o rango por medida de distancia, medidas de correlación, medidas de consistencia y medida de información. Finalmente, los métodos Híbridos son una combinación de los dos anteriores, siendo útiles cuando existe un gran número de atributos para usar algoritmos del tipo *Wrapper* y el rendimiento del enfoque basado en filtros no es satisfactorio. Este tipo de algoritmo pone en marcha un filtrado para medir la importancia de cada atributo.

   Posteriormente, dentro del conjunto total se seleccionan diferentes subconjuntos de atributos y se evalúan desde el punto de vista de *Wrapper*. En este estudio se utilizó el algoritmo Boruta [29] por su fácil manera de implementar. Es un algoritmo de tipo wrapper, el cual funciona extendiendo con otras acciones el *Random Forest* y es capaz de trabajar con cualquier método de clasificación que pueda aplicar medidas de importancia de la variable. A continuación, se describe el funcionamiento del algoritmo:

   a) En la figura 2.4 se observa que el algoritmo añade aleatoriedad al conjunto de datos al crear copias barajadas de un número determinado de características (denominadas características sombra u ocultas), por ejemplo, el atributo 1 con valores originales de 1 y 9, son asignados con valores sombra de 2 y 7.

CONJUNTO ORIGINAL			
Atributo 1	**Atributo 2**	**Atributo 3**	**Atributo 4**
1	2	3	4
9	7	5	4

CONJUNTO SOMBRA			
Atributo 1	**Atributo 2**	**Atributo 3**	**Atributo 4**
2	1	4	3
7	9	4	5

Figura 2.4: Conjunto de datos sombra y original (Descripción del algoritmo Boruta)

b) El siguiente paso es entrenar un clasificador de *Random Forest* en el conjunto de datos extendido y aplica una medida de importancia de la característica (el valor predeterminado es Precisión de disminución media) para evaluar la importancia de cada característica, cuando más reducción exista mayor significancia tendrá.

c) En cada iteración, verifica si una característica real tiene mayor importancia que la mejor de sus características sombra (es decir, si la característica tiene una puntuación Z más alta que la puntuación Z máxima de sus características sombra) y elimina constantemente las características que se consideran poco importantes.

d) Finalmente, el algoritmo se detiene cuando se confirman o rechazan todas las características o cuando alcanza un límite especificado de *Decision Tree*.

En este estudio se analizan el comportamiento de los modelos con los datasets originales y luego con los *datasets* filtrados por las características relevantes encontradas por el algoritmo Boruta.

En el siguiente epígrafe se verán los modelos utilizados a profundidad.

2.4. Análisis de modelos para la predicción

Como se mencionó en el epígrafe anterior el proceso de investigación constó de 4 etapas, en este epígrafe se verá la propuesta de solución para la tercera y cuarta etapa.

2.4.1. Construcción y entrenamiento de modelos de aprendizaje automático

En este estudio se tuvo en cuenta cinco algoritmos de aprendizaje automático:

- **Regresión Lineal**: es un método estadístico muy utilizado en técnicas de *Machine Learning*. Estas estimaciones de regresión explican la correlación entre las variables dependientes e independientes. La forma más simple de la regresión se define mediante la siguiente ecuación:

  $$Y = \beta_0 + \beta_1 X + \epsilon$$

  donde $\beta_0, \beta_1 \in \mathbb{R}$ son constantes desconocidas llamadas coeficientes de regresión, las cuales se pueden calcular como estimaciones de algunos parámetros del modelo, definiendo la relación entre dos entidades (valor del predictor y respuesta). Hay tres tipos principales de análisis de regresión, que incluyen pronosticar un efecto, determinar la fuerza de los predictores y pronosticar tendencias. [30]

- **Regresión logística**: es un tipo de análisis de clasificación utilizado para predecir el resultado de una variable categórica (una variable que puede adoptar un número limitado de categorías) en función de las variables independientes o predictoras. Es útil para modelar la probabilidad de un evento ocurriendo en función de otros factores, en el caso que ocupa este trabajo puede ser: si un estudiante aprobó o no. Este algoritmo puede manejar no sólo el ajuste de una línea de regresión sino también el ajuste de un método logístico en forma de "S" que predice dos estados máximos (0 o 1). [31]

- **Árbol de decisión**: es una de las técnicas de aprendizaje más efectivas y ampliamente utilizada en diversas áreas como educación, estadística, banca y reconocimiento, entre otras. La estructura del árbol se construye a partir de un nodo raíz, nodos internos y nodos hoja. Cada nodo se puede definir como el momento en el que se ha de tomar una decisión de entre varias posibles, lo que va haciendo que a medida que aumenta el número de nodos aumente el número de posibles finales a los que se puede llegar. El algoritmo prueba un atributo en todos los nodos internos, la salida de la prueba está en una rama, mientras que

a cada nodo hoja se le asigna una etiqueta de clase, es decir, la clasificación a la que se desea llegar. [30]

- ***Random Forest***: es un método de aprendizaje automático supervisado de uso común que desempeña un papel importante en los problemas de clasificación y regresión. Es un grupo de árboles de decisión, pero en este caso solo se seleccionan un subconjunto de características, mientras que el árbol de decisión considera todas las posibles divisiones de atributos. El algoritmo presenta características claves, como producir una predicción razonable sin ajuste de hiperparámetros, reduce el riesgo de sobre ajuste, proporciona flexibilidad y determina fácilmente la importancia de las características. [30]
- ***Support Vector Machine (SVM)***: es un método de aprendizaje supervisado adaptable para problemas que involucran clasificación y regresión. Cada punto de datos en un espacio de n dimensiones corresponde al valor de cada atributo. Luego, la clasificación se realiza seleccionando el hiperplano que mejor distingue las dos clases. SVM puede realizar una clasificación lineal de manera eficiente al *mapear* el conjunto de entrada dado en espacios de dimensiones superiores. Se agrupa en dos tipos diferentes, SVM lineal y no lineal. [31]

El proceso de construcción y entrenamiento se dividió en tres etapas:

1. Los datos recopilados y ya procesados se dividieron en dos conjuntos, con el 80 % el primer conjunto sirviendo como datos de entrenamiento y el 20 % restante como datos de prueba.
2. Se seleccionaron los 5 algoritmos basados en los más utilizados para la clasificación (*Random Forest*, Árbol de decisión, Regresión lineal, Regresión logística, *Support Vector Machine*).
3. Entrenamiento y evaluación de los modelos seleccionados.

2.4.2. Evaluación

El experimento se dividió en dos pasos para la fase de evaluación del modelo: validación cruzada de k veces y evaluación del modelo.

1. Se evaluaron los 5 clasificadores mediante validación cruzada quíntuple, una de las técnicas más utilizadas para medir un modelo en *Machine Learining* (Aprendizaje de máquina). En este paso el conjunto de entrenamiento en cada iteración de la validación cruzada se dividió nuevamente en el 80 % para el entrenamiento y el 20 % para la validación. Por lo tanto se construyó un modelo compuesto por cinco iteraciones, donde, en cada iteración se utilizaron conjuntos de validación y entrenamiento diferentes.

2. Se evaluó el rendimiento de los modelos calculando la puntuación de rendimiento promedio de todos los k pasos. Cada etapa del curso (25 %, 50 %, 75 % y 100 %), se le aplicó el enfoque mencionado de la misma manera.

La evaluación del desempeño se realizó con la Matriz de Confusión con dos etiquetas de clase, ampliamente utilizada para la evaluación de la calidad de la clasificación [30]. Se utilizaron una variedad de cuatro medidas comunes: exactitud, precisión, recuperación y medida F1 [32].

Sean:

- TP = Verdadero Positivo (predicho verdadero y verdadero en la realidad)
- TN = Verdadero Negativo (predicho falso y falso en realidad)
- FP = Falso Positivo (predicho verdadero y falso en realidad)
- FN = Falso Negativo (predicho falso y verdadero en la realidad)

Cada métrica está definida de la siguiente manera:

- **Exactitud**: se define como la proporción entre el número total de clasificaciones correctas y el total de clasificaciones.

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

- **Precisión**: es la proporción de casos de predicción positivos que se clasifican correctamente.

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

- **Recuperación**: es la proporción de casos de predicción realmente positivos que se clasificaron correctamente.

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

- **Medida F1**: es la media armónica entre la precisión y la recuperación. Sus valores oscilan entre 0 y 1, mientras más cercano a 1 sea su valor mejor es el clasificador.

$$F1 = \frac{2 * Precision * Recuperacion}{Precision + Recuperacion}$$

En el siguiente capítulo se aborda con más detalle la implementación y los resultados de todo el proceso de investigación.

Capítulo 3

Detalles de Implementación y Resultados

Siguiendo el proceso de investigación descrito en el capítulo anterior, en el presente se detallará la implementation de dicho proceso.

3.1. Construcción del conjunto de datos

Para la creación de los *datasets* se utilizó Python 3.10 como lenguaje de programación. Este lenguaje cuenta con la característica de poseer muchas librerías de gran utilidad en el análisis y manejo de datos. Las utilizadas en este estudio para la recopilación y el procesamiento de los datos son: `pandas` y `requests`.

La librería `pandas` es ampliamente utilizada cuando se trabaja con grandes volúmenes de datos, también para el manejo de archivos .xlsx. Los principales tipos de datos en esta librería son: `DataFrame` y `Series`. Para esta investigación se empleó fundamentalmente el tipo de dato `DataFrame` el cual modela el funcionamiento de una hoja de trabajo en excel. El archivo de registros de Moodle se procesó con esta librería, mediante el método `extract()` se transformó la columna de "Descripción" por tres nuevas columnas (ID_USUARIO, ID_MODULO, ID_CURSO).

La librería `requests` se utiliza para hacer las peticiones al API de Moodle. Estas se realizan a partir de un *token* y una url proporcionadas por la institución en cuestión.

En la figura 3.1 se presenta un ejemplo de un método que obtiene el libro de calificaciones de un curso en Moodle a través del API.

```python
def get_grade_items(courseID : int):
    params = {
        "wstoken" : MOODLE_TOKEN,
        "moodlewsrestformat" : "json",
        "wsfunction" : "gradereport_user_get_grade_items",
        "courseid" : courseID,
    }
    try:
        moodle_grade_items = requests.get(MOODLE_URL, params=params).json()
    except Exception as e:
        raise e

    return moodle_grade_items
```

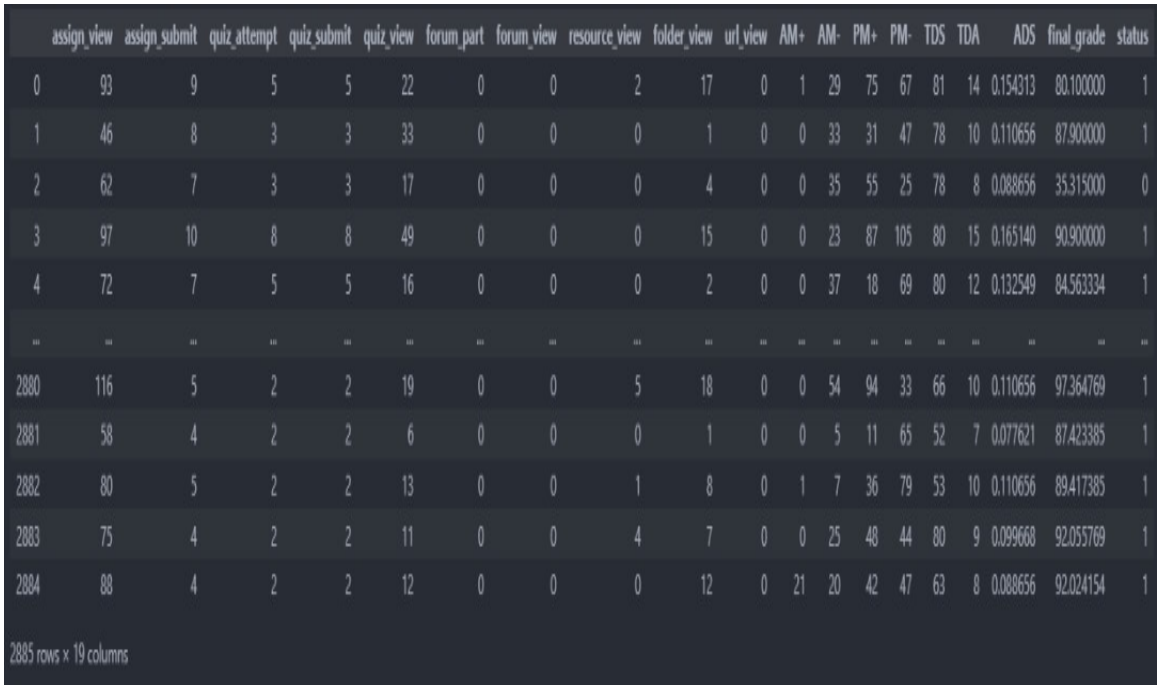
Figura 3.1: Ejemplo de código para el llamado al API de Moodle

Una vez procesado el archivo de registros y fusionado con las calificaciones de los estudiantes, se obtienen los *datasets* que se presentan en las figuras 3.2 y 3.3.

	Tarea	Glosario	Cuestionario	Foro	Carpeta	Recurso	URL	Nota	status
0	103	0	32	0	17	2	0	80.100000	1
1	55	0	39	0	1	0	0	87.900000	1
2	70	1	23	0	4	0	0	35.315000	0
3	109	6	65	0	15	0	0	90.900000	1
4	79	3	26	0	2	0	0	84.563334	1
...	...	...	...	...	...	...	...	...	...
2880	121	0	23	0	18	5	0	97.364769	1
2881	62	0	10	0	1	0	0	87.423385	1
2882	85	0	17	0	8	1	0	89.417385	1
2883	79	0	15	0	7	4	0	92.055769	1
2884	92	0	16	0	12	0	0	92.024154	1

2885 rows × 9 columns

Figura 3.2: Muestra del *Dataset* 1



	assign_view	assign_submit	quiz_attempt	quiz_submit	quiz_view	forum_part	forum_view	resource_view	folder_view	url_view	AM+	AM-	PM+	PM-	TDS	TDA	ADS	final_grade	status
0	93	9	5	5	22	0	0	2	17	0	1	29	75	67	81	14	0.154313	80.100000	1
1	46	8	3	3	33	0	0	0	1	0	0	33	31	47	78	10	0.110656	87.900000	1
2	62	7	3	3	17	0	0	0	4	0	0	35	55	25	78	8	0.088656	35.315000	0
3	97	10	8	8	49	0	0	0	15	0	0	23	87	105	80	15	0.165140	90.900000	1
4	72	7	5	5	16	0	0	0	2	0	0	37	18	69	80	12	0.132549	84.563334	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2880	116	5	2	2	19	0	0	5	18	0	0	54	94	33	66	10	0.110656	97.364769	1
2881	58	4	2	2	6	0	0	0	1	0	0	5	11	65	52	7	0.077621	87.423385	1
2882	80	5	2	2	13	0	0	1	8	0	1	7	36	79	53	10	0.110656	89.417385	1
2883	75	4	2	2	11	0	0	4	7	0	0	25	48	44	80	9	0.099668	92.055769	1
2884	88	4	2	2	12	0	0	0	12	0	21	20	42	47	63	8	0.088656	92.024154	1

2885 rows × 19 columns

Figura 3.3: Muestra del *Dataset* 2

A partir del campo **Fecha**, que se encuentra en el archivo de registros de Moodle, se dividieron ambos *datasets* en 4 (25 %, 50 %, 75 %, 100 %). Para la selección de características se empleó el conjunto de datos que contiene el 100 % en cada caso. Para tener una mejor apreciación de cómo se comportan las características seleccionadas en ambos *datasets* se calculó la Matriz de Correlación, y se realizó un proceso de selección de características con el algoritmo **Boruta**.

La Matriz de Correlación es una herramienta estadística que muestra la intensidad y la dirección de la relación entre dos o más variables. Muestra cómo se relacionan entre sí todos los posibles pares de valores de una tabla con el coeficiente de correlación. Este oscila entre "-1" y "1", donde "-1" significa una correlación negativa perfecta, "1" significa una correlación positiva perfecta y "0" significa que no hay correlación entre las variables. El coeficiente de correlación se calcula a partir de la siguiente fórmula:

$$r = \frac{n\sum XY}{\sqrt{(n\sum X^2 - (\sum X)^2)(n\sum Y^2 - (\sum Y)^2)}}$$

Donde :

- r = coeficiente de correlación
- n = número de observaciones

- $\sum XY$ = suma del producto de cada par de observaciones correspondientes de las dos variables
- $\sum X$ = suma de las observaciones de la primera variable
- $\sum X^2$ = suma de los cuadrados de las observaciones de la primera variable
- $\sum Y^2$ = suma de los cuadrados de las observaciones de la segunda variable

Esta matriz hace que sea fácil y rápido observar cómo están relacionadas las distintas variables y de esta forma encontrar patrones en ellas.

Se construyeron Modelos de Aprendizaje automático con este *dataset* en las cuatro etapas del curso. El lenguaje de programación utilizado fue con la librería *scikit-learn* en su versión 1.3.2, a través de Google Colab, una plataforma de código abierto que admite muchas bibliotecas populares de aprendizaje automático, entre ellas *scikit-learn*. [31].

Se crearon y ejecutaron 5 clasificadores binarios con parámetros predeterminados, como Árbol de Decisión, *Random Forest*, Regresión Logística, Regresión Lineal, *Support Vector Machine*. Estos modelos se entrenan en las diferentes etapas de un curso, una primera vez con todas los atributos y luego con un filtro de características a partir del resultado del algoritmo Boruta.

3.1.1. Análisis del primer conjunto de datos

En la figura 3.4 se muestra que las variables implicadas tienen muy poca relación entre ellas. El vínculo más importante es el que se establece entre cada una de las características y el estado final, ya que esto da una métrica de cuáles son los atributos que tienen mayor implicación en el desempeño del estudiante que es el estado final. El mayor coeficiente 0,090817 perteneciente a la relación entre la variable **Tarea** y el estado final da una muestra de la poca conexión existente en las variables implicadas. Este resultado implica que utilización de la plataforma no es la correcta por parte de los estudiantes, así como existe poca rigurosidad por parte de los profesores a la hora de establecer una enseñanza en línea.

	Tarea	Glosario	Cuestionario	Foro	Carpeta	Recurso	URL	Nota	status
Tarea	1.000000	0.077959	-0.027072	0.095802	0.159103	0.105535	0.083060	0.206169	0.090817
Glosario	0.077959	1.000000	-0.044135	-0.014757	-0.068563	-0.007688	0.117346	0.024676	0.031830
Cuestionario	-0.027072	-0.044135	1.000000	0.061370	0.245832	0.013647	0.184564	0.072772	0.043508
Foro	0.095802	-0.014757	0.061370	1.000000	0.144319	0.058370	0.050913	0.001103	0.019954
Carpeta	0.159103	-0.068563	0.245832	0.144319	1.000000	0.200012	0.144430	0.110297	0.068718
Recurso	0.105535	-0.007688	0.013647	0.058370	0.200012	1.000000	0.138477	0.048369	0.023546
URL	0.083060	0.117346	0.184564	0.050913	0.144430	0.138477	1.000000	-0.005965	-0.027468
Nota	0.206169	0.024676	0.072772	0.001103	0.110297	0.048369	-0.005965	1.000000	0.794484
status	0.090817	0.031830	0.043508	0.019954	0.068718	0.023546	-0.027468	0.794484	1.000000

Figura 3.4: Matriz de correlación del *Dataset* 1

A este *dataset* se le aplicó 30 iteraciones del algoritmo Boruta en busca de encontrar las variables más importantes en la clasificación. El resultado fue el siguiente:

Feature	Rank	Keep
Tarea	1	True
Glosario	2	False
Cuestionario	1	True
Foro	3	False
Carpeta	1	True
Recurso	3	False
URL	5	False
status	1	True

Figura 3.5: Resultado del algoritmo Boruta en el primer *Dataset*

En la figura 3.5 da como respuesta que se deben mantener los atributos: **Tarea**, **Cuestionario**, **Carpeta**, **URL**. Este resultado coincide con los mayores coeficientes de correlación de los atributos con el estado final del estudiante (figura 3.4). Además se muestra que el uso de los foros, glosarios y recursos no es bueno en este contexto.

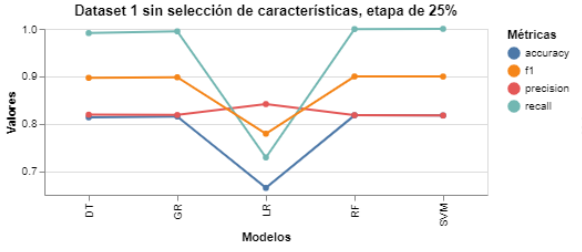


Figura 3.6: Resultados del *dataset* 1, etapa 25 %

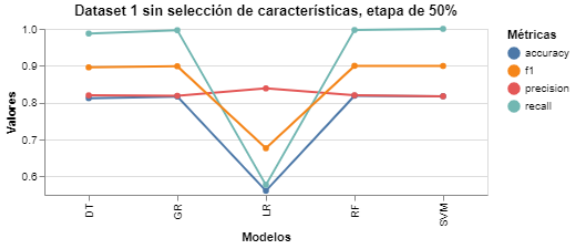


Figura 3.7: Resultados del *dataset* 1, etapa 50 %

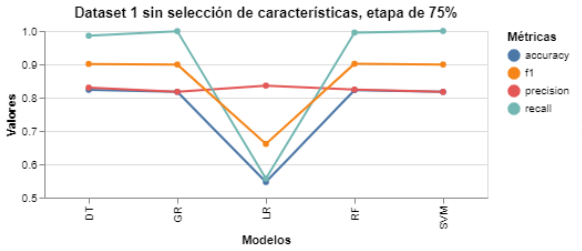


Figura 3.8: Resultados del *dataset* 1, etapa 75 %

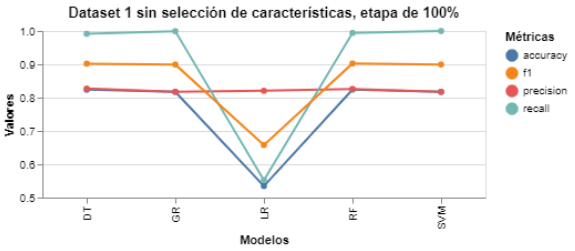


Figura 3.9: Resultados del *dataset* 1, etapa 100 %

En las figuras [3.6-3.9] se muestra una perspectiva diferente de los 5 modelos predictivos utilizados sin el filtrado de características en términos de 4 medidas: exactitud (*accuracy*), precisión (*precision*), recuperación (*recall*) y la medida f1.

De estos resultados se concluye:

- Primero, con un 25 % de progreso del curso, como se muestra en la figura 3.6, la medida f1 más alta para esta etapa promedió 0.899577 (89.9 %) por parte del algoritmo *Random Forest*, sin embargo el modelo de Regresión Lineal presente la mayor precisión con 0.841331, además de ser el modelo mejor balanceado.

- En segundo lugar, en la figura 3.7 se muestra la etapa intermedia del 50 % del curso, no se muestran cambios relevantes en cuanto a las medidas en ninguno de los modelos, en este caso el Árbol de decisión logró la mejor medida f1 de 0.895660 y el algoritmo de Regresión Lineal la mayor precisión con 0.838568. .

- De manera similar en el 75 % de progreso como se observa en la figura 3.8 el modelo Regresión Lineal obtuvo la mejor precisión de 0.836072 y y el Árbol de Decisión la mayor medida f1 de 0.901137.

- Por último, cuando se completó el curso como se muestra en la figura 3.9, el Árbol de Decisión superó a los demás con una precisión de 0.827242 y una medida f1 de 0.902335.

También se observa que en todas las etapas la recuperación es alta, esto significa que todos los modelos tienen la capacidad de encontrar de manera efectiva los verdaderos positivos en la predicción, cosa que puede ser perjudicial porque esto puede ser a expensas de aumentar el número de falsos positivos. Se propone aumentar el conjunto de datos, y añadir más estudiantes suspensos para mejorar este aspecto.

En las figuras [3.10 - 3.13] se observa el comportamiento de todos los modelos con un filtrado de características luego de la selección con el algoritmo Boruta.

Como se puede apreciar no hubo cambios significativos con respecto al entrenamiento con todas las características. Siendo el modelo Regresión Lineal nuevamente

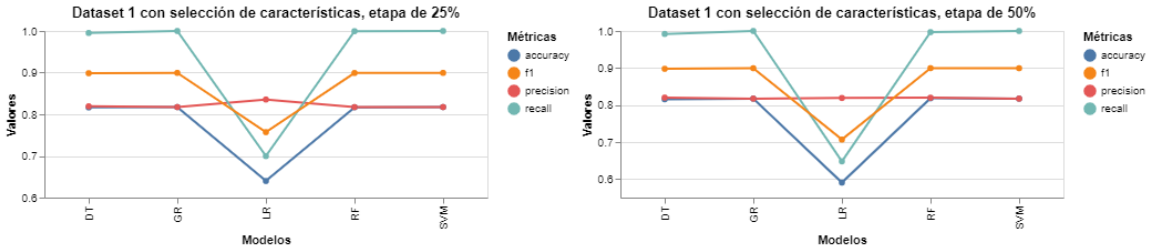


Figura 3.10: Resultados del *dataset* 1 con filtro de características, etapa 25 %

Figura 3.11: Resultados del *dataset* 1 con filtro de características, etapa 50 %

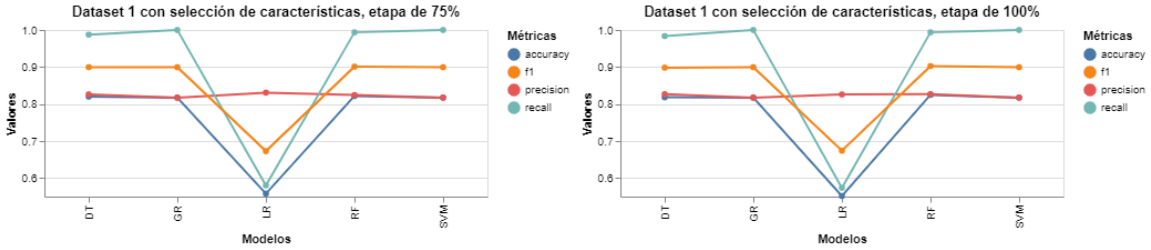


Figura 3.12: Resultados del *dataset* 1 con filtro de características, etapa 75 %

Figura 3.13: Resultados del *dataset* 1 con filtro de características, etapa 100 %

el de mejor precisión con respecto a los demás y el *Random Forest* el de mejor medida f1.

3.1.2. Análisis del segundo conjunto de datos

En este conjunto de datos se tuvieron en cuenta otras variables como se muestra en la figura 3.3. Con este cambio los coeficientes de correlación mejoraron con respecto al *dataset* anterior, como se observa en la figura 3.14 la mayor conexión con el estado final la alcanzan los atributos: TAD (*total access days*, cantidad de días distintos en que el estudiante accedió al sitio) y ADS (*access density score*, proporción entre TAD y el total del días de duración del curso), con 0.151449 y 0.151999 respectivamente. En general los coeficientes se mantienen bajos, pero este *dataset* da una perspectiva más amplia de lo que realiza un estudiante dentro de un curso.

	assign_view	assign_submit	quiz_attempt	quiz_submit	quiz_view	forum_part	forum_view	resource_view	folder_view	url_view	AM+	AM-	PM+	PM-	TDS	TDA	ADS	final_grade	status
assign_view	1.000000	0.791445	-0.089560	-0.089638	-0.003702	0.060765	0.100519	0.106757	0.162285	0.084689	0.275107	0.532908	0.634744	0.662070	0.450732	0.645908	0.646475	0.200901	0.088197
assign_submit	0.791445	1.000000	-0.094984	-0.095211	-0.073963	0.043392	0.056759	0.076041	0.095528	0.062966	0.234253	0.439722	0.483188	0.524217	0.408193	0.664234	0.665134	0.226363	0.098634
quiz_attempt	-0.089560	-0.094984	1.000000	0.999927	0.713261	0.055949	0.043498	-0.042911	0.201553	0.165747	-0.064353	0.135794	0.094423	0.206911	0.232147	0.316674	0.317371	0.129814	0.110221
quiz_submit	-0.089638	-0.095211	0.999927	1.000000	0.712815	0.055994	0.043545	-0.042846	0.201258	0.165844	-0.064274	0.135267	0.094308	0.206964	0.232019	0.316626	0.317322	0.129683	0.110144
quiz_view	-0.003702	-0.073963	0.713261	0.712815	1.000000	0.047543	0.061069	0.026801	0.242499	0.178462	-0.037222	0.274868	0.196826	0.270079	0.215504	0.238470	0.239126	0.054481	0.024526
forum_part	0.060765	0.043392	0.055949	0.055994	0.047543	1.000000	0.724811	0.043827	0.100125	0.055588	0.007442	0.154999	0.084000	0.056244	0.072178	0.104957	0.104436	0.007420	0.025738
forum_view	0.100519	0.056759	0.043498	0.043545	0.061069	0.724811	1.000000	0.058598	0.146659	0.047366	0.014260	0.191373	0.121218	0.095194	0.099146	0.139538	0.139001	-0.000315	0.017711
resource_view	0.106757	0.076041	-0.042911	-0.042846	0.026801	0.043827	0.058598	1.000000	0.200012	0.138477	0.033521	0.194661	0.208425	0.102455	0.107777	0.176283	0.175256	0.048369	0.023546
folder_view	0.162285	0.095528	0.201553	0.201258	0.242499	0.100125	0.146659	0.200012	1.000000	0.144430	-0.034851	0.277967	0.319885	0.256031	0.264351	0.313716	0.312824	0.110297	0.068718
url_view	0.084689	0.062966	0.165747	0.165844	0.178462	0.055588	0.047366	0.138477	0.144430	1.000000	0.058589	0.131425	0.137394	0.139996	0.127396	0.119353	0.119430	-0.005965	-0.027468
AM+	0.275107	0.234253	-0.064353	-0.064274	-0.037222	0.007442	0.014260	0.033521	-0.034851	0.058589	1.000000	-0.026583	0.018990	0.128648	0.062563	0.164463	0.164569	0.058306	0.052068
AM-	0.532908	0.439722	0.135794	0.135267	0.274868	0.154999	0.191373	0.194661	0.277967	0.131425	-0.026583	1.000000	0.330930	0.120097	0.329147	0.459546	0.459722	0.141383	0.068462
PM+	0.634744	0.483188	0.094423	0.094308	0.196826	0.084000	0.121218	0.208425	0.319885	0.137394	0.018990	0.330930	1.000000	0.208084	0.335761	0.465171	0.465510	0.164980	0.097661
PM-	0.662070	0.524217	0.206911	0.206964	0.270079	0.056244	0.095194	0.102455	0.256031	0.139996	0.128648	0.120097	0.208084	1.000000	0.421399	0.563433	0.563883	0.151226	0.046886
TDS	0.450732	0.408193	0.232147	0.232019	0.215504	0.072178	0.099146	0.107777	0.264351	0.127396	0.062563	0.329147	0.335761	0.421399	1.000000	0.595009	0.596702	0.376494	0.199486
TDA	0.645908	0.664234	0.316674	0.316626	0.238470	0.104957	0.139538	0.176283	0.313716	0.119353	0.164463	0.459546	0.465171	0.563433	0.595009	1.000000	0.999974	0.278652	0.151449
ADS	0.646475	0.665134	0.317371	0.317322	0.239126	0.104436	0.139001	0.175256	0.312824	0.119430	0.164569	0.459722	0.465510	0.563883	0.596702	0.999974	1.000000	0.279691	0.151999
final_grade	0.200901	0.226363	0.129814	0.129683	0.054481	0.007420	-0.000315	0.048369	0.110297	-0.005965	0.058306	0.141383	0.164980	0.151226	0.376494	0.278652	0.279691	1.000000	0.794484
status	0.088197	0.098634	0.110221	0.110144	0.024526	0.025738	0.017711	0.023546	0.068718	-0.027468	0.052068	0.068462	0.097661	0.046886	0.199486	0.151449	0.151999	0.794484	1.000000

Figura 3.14: Matriz de correlación del *Dataset* 2

Al aplicar el algoritmo Boruta nuevamente con 30 iteraciones, se produjo el siguiente resultado:

Feature	Rank	Keep
assign_view	1	True
assign_submit	1	True
quiz_attempt	1	True
quiz_submit	1	True
quiz_view	1	True
forum_part	8	False
forum_view	7	False
resource_view	6	False
folder_view	2	False
url_view	5	False
AM+	4	False
AM-	3	False
PM+	1	True
PM-	1	True
TDS	1	True
TDA	1	True
ADS	1	True
status	1	True

Figura 3.15: Resultado del algoritmo Boruta en el segundo *Dataset*

Como se muestra en la figura 3.15 las características más importantes coinciden con los atributos de mayor coeficiente de correlación con respecto al estado final. Además se vuelve a evidenciar que el foro y los recursos no son relevantes dentro de este contexto.

Como con el conjunto de datos anterior, se entrenaron todos los modelos primeramente sin el filtrado de características y luego con los atributos relevantes.

En las figuras [3.16-3.19] se muestran los resultados de cada modelo en cada etapa. De estos resultados se concluye:

- En el 25 % de progreso del curso, como se muestra en la figura 3.16, la medida f1 más alta para esta etapa promedió 0.903166 (90.3 %) por parte del algoritmo

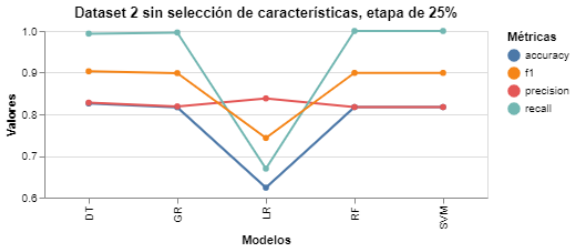


Figura 3.16: Resultados del *dataset* 2, etapa 25 %

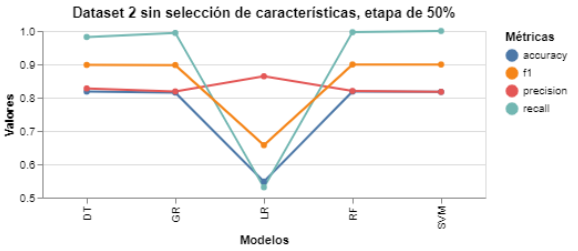


Figura 3.17: Resultados del *dataset* 2, etapa 50 %

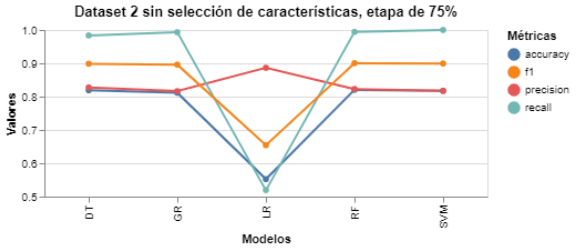


Figura 3.18: Resultados del *dataset* 2, etapa 75 %

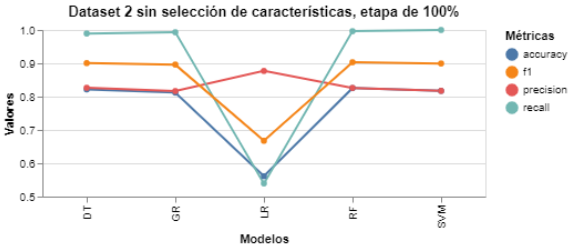


Figura 3.19: Resultados del *dataset* 2, etapa 100 %

Árbol de Decisión, sin embargo el modelo de Regresión Lineal presenta la mayor precisión con 0.841331.

- En segundo lugar, en la figura 3.17 se muestra la etapa intermedia del 50 % del curso, no se muestran cambios relevantes, en este caso el *Random Forest* logró la mejor medida f1 de 0.899510 y el algoritmo de Regresión Lineal la mayor precisión con 0.864092.
- De manera similar en el 75 % de progreso como se observa en la figura 3.18 el modelo Regresión Lineal alcanzó la mejor precisión de 0.886431 y el *Random Forest* la mayor medida f1 de 0.900394.
- Por último, cuando se completó el curso como se muestra en la figura 3.19, el algoritmo Regresión Lineal superó a los demás en cuanto a la precisión con una puntuación de 0.877038 y el *Random Forest* tuvo la mejor medida f1 con 0.903177.

Al igual que en el *dataset* anterior la recuperación en todos los modelos excepto en el de Regresión Lineal, es muy cercana a 1, lo que vuelve a poner de manifiesto el mismo problema.

En las figuras [3.20 - 3.23] se observa el comportamiento de todos los modelos con un filtrado de características luego de la selección con el algoritmo Boruta en el segundo conjunto de datos.

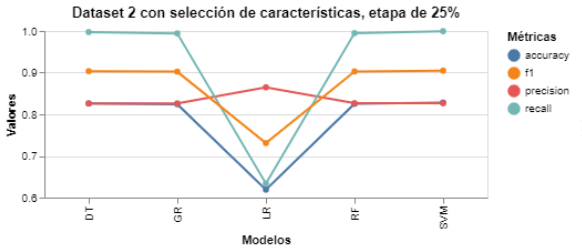


Figura 3.20: Resultados del *dataset* 2 con filtro de características, etapa 25 %

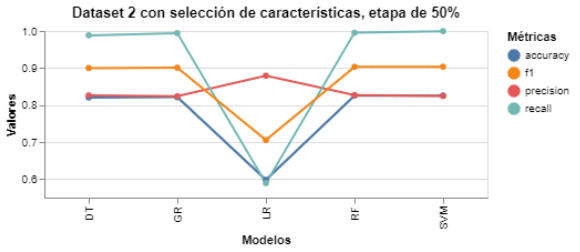


Figura 3.21: Resultados del *dataset* 2 con filtro de características, etapa 50 %

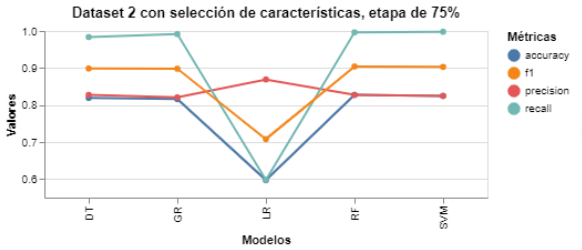


Figura 3.22: Resultados del *dataset* 2 con filtro de características, etapa 75 %

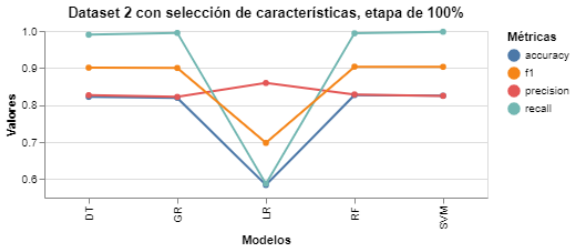


Figura 3.23: Resultados del *dataset* 2 con filtro de características, etapa 100 %

Como se puede apreciar no hubo cambios significativos con respecto al entrenamiento con todas las características. Siendo el modelo Regresión Lineal nuevamente el de mejor precisión con respecto a los demás y el *Random Forest* el de mejor medida f1.

De manera general el algoritmo Regresión Lineal es el más balanceado en cada una de las etapas para ambos *datasets*. Dentro del contexto de esta investigación este modelo es el mejor dado que a diferencia de los demás. Presenta valores de recuperación que oscilan entre los 0.5 y 0.7, que en este problema en particular de predicción académica es bueno, ya que eso significa que el modelo no hace énfasis en los estudiantes aprobados, y aún así logra una buena precisión a la hora de la evaluación.

Conclusiones

Como resultado de la investigación se logró la creación de una solución computacional que responde a la problemática de predicción académica en Entornos Virtuales de Aprendizaje, en la que se integran técnicas estadísticas, de análisis de datos y de aprendizaje automático. La solución se caracteriza por la construcción de modelos de datos educacionales y de aprendizaje automático con vista a favorecer el análisis descriptivo del comportamiento de un estudiante dentro de un curso virtual, así como el uso de las potencialidades que brinda una LMS como Moodle.

A partir de la profundización en las áreas de conocimiento asociadas, el acercamiento al contexto educativo y el estudio de soluciones similares a la predicción del desempeño académico. Se logró el diseñó de una metodología general para el procesamiento de datos educacionales de la plataforma Moodle, a partir de la cual, se implementó un prototipo sobre cuya base se aplicó un conjunto de experimentos que permitió establecer la validez de la concepción global.

La creación de una solución computacional basada en los avances científicos y tecnológicos para el análisis de los datos históricos de los estudiantes, la construcción de diferentes conjuntos de datos con distintos atributos implicados, la selección de características dentro de estos, el procesamiento del lenguaje natural a la hora de limpiar los datos, así como realizar la predicción en las diferentes etapas del curso constituyen aportes con respecto a las herramientas o soluciones computacionales implementadas con anterioridad en el ámbito de la analítica del aprendizaje.

Los resultados de la investigación permiten responder afirmativamente a la pregunta científica, ya que ha sido posible la predicción del desempeño estudiantil a partir de modelos de aprendizaje automático, incorporando nuevas formas del procesamiento de los datos de Moodle.

Recomendaciones

- Contribuir a la evolución de esta investigación con una herramienta que automatice todos los procesos: preprocesamiento de datos de Moodle para la posterior predicción.
- Realizar el entrenamiento de los modelos con un cojunto de datos más grande con el objetivo de lograr mayor robustez en los algoritmos empleados.
- Incorporar nuevos modelos (redes neuronales), así como, otros enfoques, como el de aprendizaje no supervisado.
- Mejorar el empleo de las plataformas por parte de los profesores y estudiantes, con el objetivo de obtener datos más correspondientes con la realidad del curso virtual en cuestión.
- Implementar un sistema de recomendaciones en el Entorno Virtual de Aprendizaje que facilite la retroalimentación de los estudiantes en aras de mejorar su desempeño partiendo de las predicciones que se tengan de cada estudiante o de cada curso.

Bibliografía

[1] M. Zabolotniaia, Z. Cheng, E. Dorozhkin y A. Lyzhin, «Use of the LMS Moodle for an effective implementation of an innovative policy in higher educational institutions,» *International Journal of Emerging Technologies in Learning*, vol. 15, n.º 13, págs. 172-189, 2020. DOI: 10.3991/ijet.v15i13.14945. dirección: https://doi.org/10.3991/ijet.v15i13.14945 (vid. pág. 1).

[2] P. Wattanakasiwich et al., «Investigating challenges of student centered learning in Thai higher education during the COVID-19 pandemic,» en *2021 IEEE Frontiers in Education Conference (FIE)*, Lincoln, NE, USA, 2021, págs. 1-7. DOI: 10.1109/FIE49875.2021.9637298. dirección: https://doi.org/10.1109/FIE49875.2021.9637298 (vid. pág. 1).

[3] P. Nuankaew, «Dropout situation of business computer students, University of Phayao,» *International Journal of Emerging Technologies in Learning*, vol. 14, n.º 19, págs. 115-131, 2019. DOI: 10.3991/ijet.v14i19.11177. dirección: https://doi.org/10.3991/ijet.v14i19.11177 (vid. pág. 1).

[4] Moodle. «Moodle.» Consultado el 28 de octubre de 2023. (), dirección: https://moodle.com/es/solutions/lms/ (vid. págs. 2, 8).

[5] «Bit4Learn.» (), dirección: https://bit4learn.com/es/lms/ (vid. pág. 5).

[6] «Wikipedia.» (), dirección: https://es.wikipedia.org/wiki/Programmed_Logic_Automated_Teaching_Operations (vid. pág. 6).

[7] «virtualeducation.» (), dirección: http://www.virtualeducation.wiki/index.php/NKI_Nettstudier (vid. pág. 6).

[8] «comparasoftware.» (), dirección: https://www.comparasoftware.com/firstclass-lms (vid. pág. 6).

[9] «Moodle.» (), dirección: https://docs.moodle.org/all/es/Acerca_de_Moodle (vid. pág. 7).

[10] easy-lms. «¿Qué es un sistema de gestión del aprendizaje basado en la nube (SaaS)?» (), dirección: https://www.easy-lms.com/es/centro-de-conocimiento/centro-lms/que-es-un-lms-basado-en-la-nube-saas/item12782 (vid. pág. 7).

[11] «Moodle.» (), dirección: https://moodle.org/stats/ (vid. pág. 8).

[12] «Moodle.» (), dirección: https://docs.moodle.org/dev/Plugins (vid. pág. 9).

[13] B. S. Sarria, «Moodle's Events Log Processing for the Generation of Users' Activity Reports,» Tesis doct., Universidad de Cantabria, 2020 (vid. págs. 9, 12).

[14] C. R. Morales, «Aplicando Minería de datos en Moodle,» Tesis doct., Universidad de Córdoba', 2021 (vid. págs. 10, 12, 13).

[15] S. Ventura, «Minería de Datos en Sistemas Educativos,» 2008 (vid. págs. 10, 11).

[16] V. P. Bresfelean, «Data Mining Applications in Higher Education and Academic Inteligence Management,» 2008 (vid. pág. 10).

[17] T. A. R. Sherine Dominik, «Analizing the student performance using classification techniques to the better suited classifer,» *International Journal of Computer Applications*, vol. 104, n.º 4, págs. 1-3, 2014 (vid. pág. 10).

[18] R. C. R. J. Rosalina Rebucas Estacio, «Analyzing students online learning behavior in blended courses using Moodle,» *Asian Association of Open Universities Journal*, vol. 12, n.º 1, 2017. DOI: 10.1108/AAOUJ-01-2017-0016 (vid. pág. 12).

[19] M. P. Sushil Shrestha, «Educational data mining in moodle data,» *International Journal of Informatics and Communication Technology*, vol. 10, n.º 1, 2021. DOI: 10.11591/ijict.v10i1.pp9-18 (vid. pág. 12).

[20] D. F. Murad, Y. Heryadi, B. D. Wijanarko, S. M. Isa y W. Budiharto, «Recommendation system for smart LMS using machine learning: a literature review,» en *2018 4th International Conference on Computing, Engineering and Design (ICCED)*, Bangkok, Thailand, 2018, págs. 113-118. DOI: 10.1109/ICCED.2018.00031. dirección: https://doi.org/10.1109/ICCED.2018.00031 (vid. pág. 15).

[21] A. Suleiman, «Recommen- dation system for smart LMS using machine learning: a literature review,» *International Conference on Computer Supported Cooperative Work in Design*, vol. 10, n.º 1, 2022. DOI: 10.1109/CSCWD54268.2022.9776102. dirección: https://doi.org/10.1109/CSCWD54268.2022.9776102 (vid. pág. 15).

[22] R. Conijn, C. Snijders, A. Kleingeld y U. Matzat, «Predicting student performance from LMS data: A comparison of 17 blended courses using Moodle LMS,» *IEEE Transactions on Learning Technologies*, vol. 10, n.º 1, págs. 17-29, 2016. DOI: 10.1109/TLT.2016.2616312. dirección: https://doi.org/10.1109/TLT.2016.2616312 (vid. pág. 15).

[23] D. Ljubobratović y M. Matetić, «Using LMS activity logs to predict student failure with random forest algorithm,» en *2019 7th International Conference on Future of Information Sciences: Knowledge in the Digital*, 2019, págs. 113-119. DOI: 10.17234/INFUTURE.2019.14. dirección: https://doi.org/10.17234/INFUTURE.2019.14 (vid. pág. 15).

[24] A. Thi-Diem Nguyen, «Using machine learning to predict the low grade risk for students based on log file in Moodle learning management system,» *International Journal of Computing & Digital System*, vol. 10, n.º 1, págs. 1134-1140, 2021. DOI: 10.12785/ijcds/110191. dirección: https://doi.org/10.12785/ijcds/110191 (vid. pág. 16).

[25] B. Maraza-Quispe, E. Damian Valderrama-Chauca, L. Henry Cari-Mogrovejo y J. Milton Apaza-Huanca, «Predictive model of student academic performance from LMS data based on learning analytics,» en *2021 13th International Conference on Education Technology and Computers (ICETC)*, 2021, págs. 13-19. DOI: 10.1145/3498765.3498768. dirección: https://doi.org/10.1145/3498765.3498768 (vid. pág. 16).

[26] H. Mi, Z. Gao, Q. Zhang e Y. Zheng, «Research on constructing online learning performance prediction model combining feature selection and neural network,» *International Journal of Emerging Technologies in Learning*, vol. 17, n.º 7, págs. 94-111, 2022. DOI: 10.3991/ijet.v17i07.25587. dirección: https://doi.org/10.3991/ijet.v17i07.25587 (vid. pág. 16).

[27] I. Khan, A. R. Ahmad, N. Jabeur y M. N. Mahdi, «Machine learning prediction and recommendation framework to support introductory programming course,» *International Journal of Emerging Technologies in Learning*, vol. 16, n.º 17, págs. 42-59, 2021. DOI: 10.3991/ijet.v16i17.18995. dirección: https://doi.org/10.3991/ijet.v16i17.18995 (vid. pág. 16).

[28] J. D. R. Rodolfo C. Raga Jr, «Monitoring Class Activity and Predicting Student Performance Using Moodle Action Log Data,» *International Journal of Computing Sciences Research*, vol. 1, n.º 3, págs. 1-16, 2017. DOI: 10.25147/ijcsr.2017.001.1.09. dirección: https://doi.org/10.25147/ijcsr.2017.001.1.09 (vid. pág. 25).

[29] «andreaperlato.» (), dirección: https://www.andreaperlato.com/mlpost/feature-selection-using-boruta-algorithm/ (vid. pág. 27).

[30] A. C. Faul, *A Concise Introduction to Machine Learning*. Boca Raton, FL: CRC Press, 2019. DOI: 10.1201/9781351204750. dirección: https://doi.org/10.1201/9781351204750 (vid. págs. 29-31).

[31] L. Wei-Meng, *Python Machine Learning*. Hoboken, NJ: Wiley Press, 2019 (vid. págs. 29, 30, 35).

[32] A. Rehman, S. Naz, M. I. Razzak e I. A. Hameed, «Automatic Visual Features for Writer Identification: A Deep Learning Approach,» *IEEE Access*, vol. 7, págs. 17 149-17 157, 2019. DOI: 10.1109/ACCESS.2018.2890810. dirección: https://doi.org/10.1109/ACCESS.2018.2890810 (vid. pág. 31).