Predicción del rendimiento académico de estudiantes universitarios usando redes neuronales

Bill B. Chávez Arias

*Depto. Ciencias de la Computación*

*Universidad Peruana de Ciencias Aplicadas*

Lima, Perú

U20171C042@upc.edu.pe

Sebastian W. Contreras Rosas

*Depto. Ciencias de la Computación*

*Universidad Peruana de Ciencias Aplicadas*

Lima, Perú

U20171D516@upc.edu.pe

Carlos V. Tarazona Hurtado

*Depto. Ciencias de la Computación*

*Universidad Peruana de Ciencias Aplicadas*

Lima, Perú

U201712780@upc.edu.pe

*Abstract*—

Keywords—

# Introdución

El uso de la inteligencia artificial tiene un gran impacto en la mayoría de las industrias, no obstante, la industria educativa no es ajena a esta, ya que el uso de estas tecnologías te brinda conocimiento que beneficiará el sector, tales como entender el proceso de aprendizaje de los estudiantes o mejorar los resultados del aprendizaje.

El bajo rendimiento académico es uno de los principales problemas que enfrenta la educación superior en el mundo. Una de las causas de que sea tan difícil solucionar dicho problema, es porque hay varios factores que influyen el bajo rendimiento académico. El estudio [1] menciona que hay 5 factores: identificación, psicológica, académicas, pedagógicas y sociofamiliares. Asimismo, las consecuencias del bajo rendimiento pueden ser muchas entre ellas la deserción académica. Según el estudio [2] afirma que uno de los principales factores para la deserción académica es el bajo rendimiento académico. En la misma línea, [3] sostiene que el porcentaje de alumnos que desaprueban asignaturas el primer ciclo es elevado y cuando vuelven a llevar el curso y lo desaprueban, muchos deciden abandonar la universidad. También, hay que considerar que el bajo rendimiento académico es un indicador de calidad para las instituciones educativas, por lo que su estudio es esencial para dichas instituciones. En Perú, [4] menciona que el seguimiento del estudiante y la nivelación de las deficiencias para evitar el bajo rendimiento académico de los alumnos es considerado un estándar de calidad para los programas universitarios y ser acreditada.

En las investigaciones [5] y [6], se encontró que realizan la predicción del rendimiento académico de los alumnos con diferentes algoritmos, el [5] propone una red neuronal para la predicción del rendimiento académico consiguiendo una precisión de 78%, mientras que el [6] menciona varios modelos como KNN y Bosque Aleatorio consiguiendo una precisión de 80%. Sin embargo, dichos trabajos no justifican porqué están realizando la predicción del rendimiento académico de los alumnos y donde lo van a utilizar o a quien van a ayudar con dicha predicción.

El objetivo de esta investigación es proponer y desarrollar un modelo de predicción del rendimiento académico usando redes neuronales. De esta forma, solucionamos el problema de bajo rendimiento académico en las universidades, ya que predecir el rendimiento académico del alumno permite a la institución educativa identificar a los alumnos en riesgo y poder tomar medidas para apoyar al alumno, consiguiendo que el estudiante culmine con éxito su ciclo académico, y que la institución, por consecuencia, obtenga más prestigio.

Los datos que se utilizarán para el entrenamiento del modelo son proporcionados por [7], donde se puede conseguir información como: cursos, exámenes, información personal de estudiantes y la interacción del estudiante con los recursos de la plataforma educativa online de The Open University. Primero, se realizará técnicas de preprocesamiento para preparar la data con datos que nos van a servir para el entrenamiento. Después, clasificaremos a los alumnos en aprobados y desaprobados. Luego, dividiremos la data en datos de entrenamiento y datos de prueba. Finalmente, entrenaremos al modelo con el dataset. Como resultado se mostrará, la precisión que tuvo nuestro modelo de predicción y se comparará con otros trabajos similares.

# Estado del arte

## Empleo de la teoría del algoritmo Naive Bayes para la predicción del rendimiento académico de los estudiantes

En los estudios encontramos variedad de aplicaciones matemáticas, de las cuales el mayor grupo que resalta por sus porcentajes de aplicaciones fue la teoría del algoritmo Naive Bayes. En esta sección mencionaremos los porcentajes de éxito de cada estudio que desarrolló el algoritmo Naive Bayes, así también características del mismo como cantidad de información requerida y bajo qué circunstancias se realizaron.

En la investigación [8] se desarrolló la evaluación del rendimiento de los alumnos en los exámenes técnicos de procesos de reclutamiento, ya que estos son importantes para las universidades, la técnica para abordar el problema fue predecir si un estudiante aprueba o desaprueba un examen técnico a partir de su rendimiento en algunos cursos de carrera. Se usaron datos de 200 estudiantes de K L University entre 2013-2017. Luego de la clasificación de los estudiantes en la que se usaron dos algoritmos: K-means Clustering y Hierarchical Clustering, una vez clasificado se empleó para la predicción el algoritmo de Naive Bayes el cual tuvo una precisión del 72%.

En el estudio [9] refiere al abandono de los estudiantes en las primeras etapas del programa, para esto se calculó los factores que más influyen en el rendimiento académico del estudiante. Se utilizaron 887 instancias y 19 variables como datos para la predicción. Se empleó el algoritmo Naive Bayes el cual tiene un 85,7% de precisión en el estudio.

En la investigación [10] se desarrolló en base al análisis de que muchos estudiantes dejan su carrera, para esto la investigación realizada tiene como objetivo predecir los factores personales, familiares, económicos que más influyen en el rendimiento del estudiante para saber si el estudiante completa los estudios o lo deja. Para las pruebas, se realiza data cleaning en el dataset de tal forma que, se quedan con 50 estudiantes que terminaron los estudios y otros 50 que lo dejaron. Como resultado, el algoritmo Naive Bayes consiguió un porcentaje de 84,8% en F1-Score.

En el estudio [11] se desarrolló en base a saber el desempeño académico de los estudiantes del primer año, para esto trabajaron un enfoque de aprendizaje semi-supervisado para clasificar el desempeño de los estudiantes. En el estudio se utilizó el clasificador Naive Bayes como metodología y el porcentaje de precisión fue del 96%

En la investigación [12] se abordó el tema del abandono de los estudiantes en los estudios superiores, el sistema propuesto es una aplicación basada en web que hace uso de la técnica de minería Naive Bayesiana para la extracción de información útil. Según el proyecto en este rubro demuestra que el algoritmo básico proporciona más precisión que otros métodos como regresión, árbol de decisión y redes neuronales.

En el estudio [13] se enfocan en brindar proyecciones de rendimiento estudiantil para mejorar la educación en la toma de decisiones. Se utilizó la minería de datos educativos para modelar el rendimiento académico de los estudiantes utilizando Naive Bayes para analizar el conjunto de datos y predecir el rendimiento académico. Los resultados muestran que el clasificador de Naive Bayes es superado por los demás algoritmos, sin embargo, tiene un pronóstico general del 88% en precisión de clasificación.

En la investigación [14] se enfocan en resolver problemas sobre diversos procesos de enseñanza y aprendizaje, para el cual diseñaron y automatizaron un modelo predictivo del desempeño académico de los estudiantes. La investigación se empleó con varios algoritmos, detallando que el mejor era el de Naive Bayes con un 73% de predicción.

El Clustering es necesario en algunos casos para un posterior análisis de datos, los artículos mencionados con anterioridad no son ajenos a esto ya que se emplea K-means en [8] [11], Hierarchical en [8] [10] [11], tipos de Clustering personalizados [14] y otros según el nivel de clasificación que desean emplear, sin embargo el más común y fácil de utilizar también mencionado por su nombre es el Naive Bayes y fueron utilizados en [9] [11] [13] cabe resaltar que es importante el tipo de dato del cual se dispondrá, ya que si este es más específico y abarca temas relacionados con el aprendizaje del estudiante, o el entorno, etc. Ya que es necesario tener los grupos de la forma más diferenciada posible como refieren [11].

La calidad de los datos como menciona [8] y la limpieza de estos son importantes. Como se explicó con anterioridad, el algoritmo Naive Bayes se utiliza para el Clustering, sin embargo, también para las predicciones, no teniendo los mejores resultados a comparación de otros algoritmos [8] [9] [10] [11] [13] [14], y en algunos casos resultando con errores porcentuales muy alejados a los demás, con un error del 12% separados de Deep Learning con 5% de error y Decision Tree con 7% de error. [13].

## Uso de Árboles de decisión/bosques aleatorios para la predicción del rendimiento académico de estudiantes

El artículo [8] también propone usar el algoritmo C5.0, un algoritmo árbol de decisión, para poder predecir el rendimiento de los alumnos. Para ello, se realizó la misma metodología que se usó para hacerlo con Naive Bayes. Como resultado se obtuvo un 81% de precisión.

La investigación [15] utiliza el algoritmo J48 del software WEKA para predecir el rendimiento académico de los estudiantes. Para ello, se realizó la siguiente metodología: (1) Diseño de la encuesta y recolección de la información;(2) Definición de la muestra; (3) Construcción de la base de datos; (4) Correlación entre atributos;(5) Procesamiento computacional; (6) Predicción del rendimiento académico; (7) Identificación de los principales atributos influyentes en el desempeño académico; y (8), Estabilidad del sistema. El resultado de la encuesta arrojó 121 estudiantes con 22 variables. Finalmente, se usaron varios tratamientos de entrenamiento y prueba (50:50, 60:40, 70:30 y 80:20), y se obtuvo 91.67% de precisión en todos. Además, concluyen que los factores que más influyen en el rendimiento académico son: pedagogía, buenos horarios, buena relación docente-alumno, calidad de los docentes, prácticas remuneradas.

El estudio [9] también propone el algoritmo Bosque Aleatorio y uno propio, el algoritmo Bosque aleatorio mejorado, con estos dos se consiguió 91% y 93% de precisión respectivamente al predecir el rendimiento académico.

En la investigación [10], otro algoritmo que usaron los investigadores para predecir si el estudiante completa los estudios o lo deja fueron el algoritmo C4.5 y el algoritmo CART. Estos algoritmos obtuvieron 76.6% y 71% en F1-score respectivamente.

En el estudio [16] se propone predecir la tasa de aprobados de estudiantes en línea usando redes neuronales profundas. Se usó data de Guo Pei online Education. Primero, se filtró las características que más afectan a la tasa de aprobados usando el algoritmo filter-type feature selection. De tal forma que resultó en 27 características representativas. Luego, se realizó preprocesamiento y estandarización en la dataset que quedó (27 características). Se utilizó el algoritmo Árbol de decisión para la predicción. Asimismo, se utilizó el algoritmo Grid Search para optimizar el Árbol de decisión. Se usaron las siguientes métricas con los algoritmos: Precision, recall, F1-score y runtime. Antes de las pruebas, los datos se dividieron en aprobados y desaprobados. Finalmente, usando el grid search algorithm y con data 50% aprobados / 50% desaprobados, Árbol de decisión logró un 96% de precisión.

En [6] los investigadores plantean usar learning analytics para predecir el rendimiento académico del alumno para ello usan una muestra de los cursos Fisica 1 y 2, dando con datos de 134 estudiantes en total. Los algoritmos usados para el modelo predictivo son: KNN y Bosque aleatorio. Para entrenar el modelo predictivo, el algoritmo se alimentó con las calificaciones y logaritmo del curso anterior de cada instructor, así como las fotografías de esos exalumnos. Una vez entrenado el modelo, las fotografías de los estudiantes actuales se utilizaron como único insumo para hacer la predicción para el primer período de evaluación. Para predecir la calificación del segundo período de evaluación, solo se utilizó Bosque aleatorio. Los resultados de los cuestionarios muestran que alrededor del 80% de los estudiantes universitarios estuvo de acuerdo y muy de acuerdo con las calificaciones que obtuvieron en las tareas, las pruebas y el examen de mitad de período. Por lo que, el algoritmo Bosque aleatorio obtuvo 80% de precisión.

En el estudio [17] presentan un modelo que trata de predecir el resultado del examen final de un estudiante. Para ello, utilizaron un dataset de 1170 estudiantes en tres cursos. Luego, realizaron el preprocesamiento en el dataset quitando columnas innecesarias como el id del estudiante. Usaron K-Neighbors Neighbours y clasificador de árbol de decisión (algoritmo ID3) para la predicción. Un 94.44% de precisión obtuvo el algoritmo clasificador de árbol de decisión.

En la investigación [18] se propone utilizar predecir el rendimiento de los alumnos utilizando datos de redes sociales como: Facebook, Twitter, Instagram y YouTube, así como juegos en línea. Se creó una entrada de formato único que sirvió como conjunto de datos de entrenamiento siguiendo el formato de archivo de relación de atributos (ARFF). Para finalizar el modelo, los investigadores consideraron usar Support Vector Machine (SVM), k-vecino más cercano (KNN) y Bosque aleatorio para la predicción. Además, para garantizar que el modelo pueda clasificar con precisión un conjunto de datos dado, se realizó una validación cruzada de 10 veces con el uso del software Weka. Se usaron tres métricas para los modelos: Precision, Recall, F-Measure. Para los resultados, se utilizó tres tipos de muestras (30, 50 y 100 instancias). Como resultado, el bosque aleatorio consiguió (en la prueba con 100 instancias), 100% en Precision, 80.6% en Recall y 89.3% en F-Measure.

En [13] los investigadores también propusieron usar un árbol de decisión para la predicción. Se utilizó la misma metodología mencionada anteriormente. Y los resultados para el árbol de decisión fueron los siguientes: 93% de precisión y 96% de F-Measure, 100% de Recall.

De todas las investigaciones en esta subsección, 6 de ellos utilizaron algoritmos árboles de decisión [8][15][10][16][17][13]. Mientras que [9][6][18] utilizaron bosques aleatorios. Mostrando que, a pesar de que los bosques aleatorios sean un conjunto de árboles de decisión, la mayoría de las investigaciones deciden a utilizar árboles de decisión para la predicción con múltiples variables como es la predicción del rendimiento académico.

De los estudios [8][15][10][16][17][13], el árbol de decisión que tuvo mayor precisión en su estudio fue de la investigación [16] con 96% de precisión. Asimismo [16] fue el segundo que consideró más variables para la predicción con 27.

El estudio [10] fue el que más variables utilizó para su predicción con 33, pero obtuvo 76.6% en el algoritmo C4.5 y 71% en el algoritmo CART, siendo el más bajo de todos los estudios que usen árbol de decisión [8][15][10][16][17][13]. Si lo comparamos con el segundo que usa más variables [16], se observa algo curioso. El [16] es el que tiene más precisión, mientras que [10] es el que tiene la menor precisión (medido con F1-score). Esto se puede deber a que las variables que usa [16] están relacionadas más con características del estudiante en lo académico, mientras que, [10], de las 33 variables, solo 6 son características del estudiante en lo académico, se enfoca más en variables económicas y familiares.

## Uso de redes neuronales para la predicción del rendimiento académico de los estudiantes

En el estudio [5] se enfoca en predecir los resultados de los estudiantes en las asignaturas Estructuras de Datos I y II, Debido a que las asignaturas seleccionadas tienen un significativo grado de dificultad, lo que exige una importante dedicación y rigurosidad por parte de los estudiantes, lo que conlleva a resultados poco eficientes. Por ello diseñaron un modelo basado en el sistema MATLAB, que permite predecir resultados académicos en las materias de interés además de dos redes neuronales, las cuales tiene la misma arquitectura, pero cada una entrenada con los datos específicos de cada asignatura. Asimismo dicha arquitectura está diseñada en base a 3 capas, donde en cada una se empleó una función de activación utilizadas en cada neurona, las cuales son las siguientes: para las capas de entrada y la oculta, la sigmoidal tangente hiperbólica; para la capa de salida, la lineal, ya que se deseaba la mayor amplitud en el rango del intervalo de salida y facilitar la interpretación de los resultados, De esta forma se logró una efectividad en la predicción superior a un 78% para el caso de la primera asignatura, mientras que para la segunda se alcanzó una efectividad superior al 75%.

La investigación [19] se desarrolló a partir de la necesidad de mejorar los sistemas basados en IA en el campo de la educación, ya que los estudiantes en las aulas o en el hogar requieren en muchas ocasiones de ayuda para poder rendir académicamente. Por lo consiguiente, emplearon dos modelos de redes neuronales , los cuales son La retropropagación y La Memoria a corto y largo plazo LSTM , donde la primera utiliza un algoritmo de descenso de gradiente durante el aprendizaje y propaga el error para actualizar los pesos y minimizar el valor de error y la segunda memoriza entradas anteriores de la red neuronal. para así dar resultados más precisos. Se emplearon los algoritmos mencionados anteriormente. Y los resultados fueron los siguientes: retropropagación (BP) (70% de varianza) Y memoria a largo y corto plazo (LSTM)(77.9% de varianza)

En el artículo [13] los investigadores también propusieron usar un modelo de aprendizaje profundo en redes neuronales para la predicción. Se utilizó la misma metodología mencionada anteriormente. Y los resultados para el aprendizaje profundo fueron los siguientes: 98% de precisión y 97% de F-Measure, 98% de Recall.

De los estudios [5][19][13], la red neuronal que obtuvo mayor precisión fue la investigación [13] con 98% en Precisión, 98% en Recall y 97% en F Measure, De la misma forma [13] es el estudio que empleó la menor cantidad de variables de entrada, usando 9 valores de entrada.

El estudio [5] es el que uso más variables, con un total de 21 valores de entrada, a pesar de ello su tasa de efectividad al predecir el rendimiento de los alumnos es considerablemente baja en comparación al rango de efectividad de [13]. con un margen de 78% y 75% de efectividad. Entrando en más detalle podemos notar que en [16] emplea pocas capas y por consiguiente pocas neuronas para su red. Lo que conlleva a un proceso más inexacto.

## Modelos de predicción de Support Vector Machine (SVM) / Support Vector Regression (SVR) para la predicción del rendimiento académico de los estudiantes

En [10], los investigadores también emplearon un modelo de aprendizaje supervisado para predecir si el estudiante completa los estudios o los abandona ,específicamente utilizaron el modelo de Máquinas de vectores de soporte , donde se obtuvo el mejor resultado con 86% en la prueba F1-Score.

En el estudio [16] también propone el modelo de aprendizaje supervisado, Máquinas de vectores de soporte (SVM) para predecir la tasa de aprobados de estudiantes en línea. Este modelo optimizado con grid search algorithm y con data 50% aprobados / 50% desaprobados obtuvo un 95% de precisión.

La investigación [19] otro modelo que usaron los investigadores para mejorar los sistemas basados en IA en el campo de la educación fue Regresión de vectores de soporte (SVR). Este modelo obtuvo 79.7% de varianza en la predicción.

En el artículo [18] otro modelo que usaron los investigadores para predecir el rendimiento de los alumnos utilizando datos de redes sociales como: Facebook, Twitter, Instagram y YouTube, fue Support Vector Machine (SVM). Este modelo obtuvo como resultado (en la prueba con 100 instancias) 100% en Precisión, 96.8% en Recall y 98.4% en F-Measure siendo así el más acertado con efectividad.

En el artículo [20] se propone predecir a los estudiantes que corren el riesgo de tener un desempeño deficiente en un curso. Esta identificación debe realizarse lo suficientemente temprano para permitir que los instructores ayuden a sus estudiantes antes de que se retrasen irremediablemente. Por ello diseñaron un modelo que predice la nota del examen final del alumno. donde el modelo utiliza un clasificador binario de máquina de vectores de soporte, el cual está entrenado con el kernel de función de base radial en base al curso elegido. Asimismo, el modelo se basa en tres parámetros que indican la gravedad de diferentes tipos de errores de clasificación. Empleando este modelo se identificó correctamente al menos el 62% de los estudiantes en riesgo.

De los estudios [10][16][19][18][20], el aprendizaje supervisado que obtuvo mayor precisión fue la investigación [18] con 100% en Precisión, 96.8% en Recall y 98.4% en F Measure, cabe mencionar que para llegar a dicha precisión se empleó la validación cruzada para evitar el sobreajuste del modelo, la cual se realizó 10 veces, de esta manera podemos ver un patrón, debido a que en [19] no se realizó ningún tipo de validación cruzada y es el que tiene menor tasa de precisión con 79.7%. de la misma forma en [10] si se realizó una validación cruzada, pero en este caso se realizó únicamente 5 veces, la mitad de la cantidad en [18]. Por ello, podemos notar que en [10] tiene una tasa de precisión de 86% la cual es considerablemente alta, pero no en comparación a las tasas de predicción de [20],[16] [18] que realizaron 10 veces la validación cruzada.

## Ética profesional

Por el artículo [21], donde se aplican distintos regímenes éticos, las decisiones de la inteligencia artificial deben de ser confiables y ser debidamente justificadas, para que los usuarios, fabricantes y legisladores puedan comprender cómo se toman estas decisiones y qué principios éticos se aplicaron al tomarlas. El artículo presenta un modelo que descompone las etapas de la toma de decisiones éticas en sus componentes elementales. De igual forma, en el proyecto planteado, se hace referencia a los temas éticos relacionados con la toma de decisiones del algoritmo para la predicción del rendimiento de los estudiantes, los temas éticos en detalle se encuentran en el punto 5, subsección B, denominado Ética.

# Estudio autodirigido

## Aprendizaje Profundo

El aprendizaje profundo o Deep learning es un tipo particular de aprendizaje automático que logra un gran poder y flexibilidad al aprender a representar el mundo como una jerarquía anidada de conceptos, con cada concepto definido en relación con conceptos más simples y representaciones más abstractas calculadas en términos de conceptos menos abstractos [22]. Una de las aplicaciones en el mundo real del Deep learning es en [23], en el cual se usa Deep learning para crear un modelo para la detección y monitoreo de eventos inesperados/accidentes en CCTV en túneles. Los investigadores clasificaron dichos accidentes en: (1) conducción en sentido contrario (WWD), (2) parada, (3) persona fuera del vehículo en el túnel (4) incendio. Luego, sobre la base del modelo, se probó el sistema de detección de accidentes CCTV (cadena televisiva china) de túnel basado en ODTS utilizando cuatro videos de accidentes que incluyen cada accidente. Como resultado, el sistema puede detectar todos los accidentes en 10 segundos como máximo.

## Análisis de sentimientos

El análisis de sentimiento, también conocido como minería de opinión (opinion mining), es un término muy discutido, pero a menudo incomprendido. Básicamente, es el proceso de determinar el tono emocional que hay detrás de una serie de palabras, y se utiliza para intentar entender las actitudes, opiniones y emociones expresadas en una mención online [24] [25]. IMBD y Amazon.com, son fuentes de datos de revisión muy famosas. IMDB es una fuente de reseñas de películas, mientras que amazon.com es una fuente de muchas reseñas de productos. Estas fuentes de datos se utilizan en tareas de Sentiment Analysis (SA), al igual que Twitter, un sitio de red social muy famoso donde sus tweets expresan la opinión de las personas y su longitud es de 140 caracteres como máximo donde también se realiza (SA), como en todos los casos anteriores, esta información reclutada con SA tiene un gran valor, no solo para las mismas empresas, ya que la información brindada por parte de los usuarios no refieren únicamente a la empresa, sino que esta es vendida a diferentes entes [26].

## Minería Web y de Textos

Consiste en buscar patrones en el texto, para así analizarlos y extraer la información que sea de utilidad para fines particulares. Asimismo, que sea web permite aprovechar la información adicional de la internet lo que a menudo, mejora los resultados al capitalizar la existencia de los directorios y de diferentes tipos de información web. La motivación para intentar extraer dicha información es debido a que, en la cultura occidental moderna, el texto es el vehículo más común para el intercambio formal de información. Por ello es conveniente incluso si el éxito es solo parcial [27]. Su aplicación se puede ver en diferentes ámbitos como [28] estudia el impacto del coronavirus en la población de París, Francia, procesando los datos recopilados de las redes sociales y utilizando tendencias relacionadas con la búsqueda. El trabajo sigue una metodología inspirada en CRISP-DM, donde primero recopilamos los datos para luego pre procesarlos y evitar información poco relevante. De esta manera el primer hallazgo es un patrón decreciente de publicaciones / interés, y el segundo está relacionado con la crisis de salud y el impacto económico generado por el coronavirus.

# Aporte

En esta sección, se presenta y explica toda la arquitectura propuesta para predecir el rendimiento académico del alumno.

## Arquitectura

Diagrama

Descripción generada automáticamente

Fig. 1. Arquitectura propuesta.

1. Definición de objetivos.

El objetivo principal del proyecto es la predicción del rendimiento académico de los estudiantes universitarios durante su desarrollo en el curso. De esta forma, permite a los profesores y la institución educativa identificar a los alumnos en riesgo y poder tomar medidas para apoyar al alumno, consiguiendo que el estudiante culmine con éxito su ciclo académico, y que la institución, por consecuencia, obtenga más prestigio.

1. Recolectar información.

El dataset utilizado para el desarrollo del proyecto proviene de Open University Learning Analytics dataset (OULAD) la cual brinda información acerca de cursos, estudiantes y sus interacciones con Virtual Learning Environment (VLE). El dataset consiste en tablas conectadas usando un único identificador. Todas las tablas están en formato csv.

1. Análisis de la información.

Posterior a la recolección de la información es necesario un análisis de este, el cual nos ayudará a segmentar la información y poder manipularla con mayor facilidad, a continuación, se detalla el conjunto de datos recolectados.

* Courses.csv guarda información acerca de los cursos.
* Assessments.csv guarda información referente a las evaluaciones de los cursos.
* Vle.csv contiene información sobre los materiales disponibles en la plataforma virtual (VLE).
* StudentInfo.csv almacena información referente al alumno en general, tanto como información demográfica como sus resultados finales.
* StudentRegistration.csv almacena información acerca del registro del estudiante a un módulo del curso.
* StudentAssessment.csv contiene información sobre el resultado de las evaluaciones de los estudiantes.
* StudentVle.csv contiene información acerca de la interacción del estudiante con los materiales en la plataforma virtual (VLE).

1. Filtro.

Es importante detallar el filtrado de la información que se ingresará en el motor de la inteligencia artificial, dado que si se entregan valores que no corresponden de una forma lógica al modelo de predicción planteado, los resultados podrían alterarse, de tal modo que las salidas y la predicción resultante serían erróneas. En este caso se han planteado los siguientes filtros:

* Aquellos alumnos que no tengan algunas notas en las evaluaciones del curso, no los consideraremos para el entrenamiento de la red neuronal.
* Los datos de alumnos que cuenten con una secuencia ilógica en sus notas se eliminan, debido a que no presentan una regularidad en ellas y, por lo tanto, puede afectar a la predicción de los resultados.

Una vez filtrados y ordenados los datos, estos están listos para pasar a la entrada del motor de la inteligencia artificial.

1. Entrada.

Las entradas para la predicción del rendimiento académico del alumno en el examen final de un curso son los siguientes:

* Número de veces que estudió el curso: Es la cantidad de veces que un alumno ha llevado el curso.
* La nota promedio de las evaluaciones: Es la nota ponderada de las evaluaciones del alumno durante el curso. No confundir con el examen final.
* Tasa de aprobación del curso: Indica la probabilidad de que un alumno apruebe el curso. Se calcula dividiendo la cantidad de evaluaciones aprobadas por el alumno sobre la cantidad de evaluaciones del curso.
* Fecha promedio de uso de materiales: promedio de las fechas de interacción del estudiante con el material medido como el número de días desde el inicio de la presentación del módulo.
* Promedio de Clicks: Promedio de interacciones que tuvo el estudiante con los materiales del curso.

1. Salida.

Las salidas serán empleadas por los maestros para poder identificar fácilmente a los alumnos con mayor riesgo de desaprobar el curso y para que puedan decidir, en su experiencia, qué tipo de metodología, en específico, pueda ayudar a los alumnos que tienen problemas en el curso. Para la predicción del rendimiento académico definimos dos clases:

* Aprueba: El alumno aprobará el curso.
* Desaprueba: El alumno podría desaprobar el curso.

La inteligencia artificial planteada, usará una de estas clases para indicar el resultado.

1. Motor de la Inteligencia Artificial.

El motor de la inteligencia artificial propuesta es una red neuronal multicapa. A continuación, se presenta la arquitectura de la red neuronal.

* 1. Descripción de las capas.

Nuestra red neuronal contiene 3 capas:

* La capa de entrada, el cual poseerá 5 neuronas que representarán a la cantidad de variables que tiene el dataset limpio. Su función de activación es “ReLU”.
* Una capa oculta, donde se elegirá el número de neuronas empleando el siguiente método empírico [29]: “El número de neuronas ocultas debe ser menos del doble del tamaño de la capa de entrada.” Por lo tanto, tendrá 8 neuronas, donde su función de activación es “ReLU”.
* Diagrama

  Descripción generada automáticamenteLa capa de salida, el cual se comprende de 1 neurona, la cuales representan el resultado de la predicción del modelo. Su función de activación es “sigmoid”.

*Fig. 2. Grafica de la Red Neuronal*

* 1. Entrenamiento y validación.

Un problema común en las redes neuronales es el Overfitting, por lo que, para evitar dicho problema realizaremos la siguiente estrategia. Dividiremos el dataset en 70/30, de tal forma que, el 70% va dirigido al entrenamiento de la red neuronal, mientras que, el 30% restante es para la validación de la red neuronal.

La cantidad de épocas para el entrenamiento será de 100 y en cada época se usará el algoritmo Backpropagation para mejores ponderaciones y sesgos de entrada para obtener un resultado más preciso o "minimizar la pérdida".

## Etica

1. Criterios de discriminación.

Dado el conjunto de variables del dataset que usamos en el proyecto, existe información personal sobre el alumno como: género. Dicha variable no la utilizaremos para el entrenamiento de la red neuronal ni para las predicciones. Consideramos que no es ético utilizar dicha información, ya que puede generar discriminación sobre qué tipo de género es más “inteligente” o tiene mayores capacidades, lo cual no consideramos que sea un punto relevante, debido a que un pequeño sector de la población no representa adecuadamente a todas las mujeres o hombres del mundo. Además, no queremos generar controversia en los salones de clase dando a entender que un sector tiene mejores capacidades que el otro cuando únicamente estamos intentando predecir su rendimiento en el curso, mas no las capacidades que puede tener una persona.

1. Exclusión de datos.

En el punto 4 de la subsección de Arquitectura llamada “Filtro” se mencionó algunos filtros que se realizarán al dataset. Uno de ellos es eliminar los datos de los alumnos que no presentan notas en varias evaluaciones del curso, debido a que puede deberse a muchos factores como: enfermedades, problemas económicos, problemas personales, etc. que no permitieron rendir la evaluación al alumno. Consideramos que es éticamente incorrecto considerar a un alumno como ineficiente si es que está pasando por inconvenientes que le impiden poder asistir a sus clases y, por lo tanto, afectan a su rendimiento académico.

1. Manipulación para fines no íntegros.

El modelo de ética empleado en el proyecto permite asegurar la información brindada por los usuarios, debido a que dicha información no será retenida y utilizada para fines no íntegros, dado a la variedad de casos donde se filtra información con malas intenciones, los datos personales no han sido requeridos para la elaboración del proyecto y, por consiguiente, no lo utilizaremos para otros fines que no sean académicos.

# Validación de resultados y discusión

En esta sección, se presentan los resultados del proceso propuesto, los cuales se explicó en la sección anterior Aporte. Asimismo, recordar que el algoritmo propuesto es una red neuronal multicapa, la cual realizará la predicción del rendimiento académico de los alumnos en algún curso.

El dataset utilizado para el entrenamiento y validación de la red neuronal fueron proporcionados por [(fuente de dataset)] y en el preprocesamiento del dataset realizamos una etiqueta a los datos, por lo que, se utilizará el aprendizaje supervisado.

## Características del dataset

El conjunto de datos obtenidos por [(fuente de dataset)] contiene la información sobre 22 cursos, 32,593 estudiantes, sus resultados de evaluación y registros de sus interacciones con el VLE representado por resúmenes diarios de los clics de los estudiantes (10,655,280 entradas). Dicha información, está ordenada en 7 tablas, las cuales fueron explicadas en la subsección Análisis de información de la anterior sección.

Para el entrenamiento, los hiper parámetros que se utilizaron fueron los siguientes:

* Epocas: 100
* Optimizador: Adam
* Función de pérdida: Binary\_CrossEntropy
* Métricas: Accuracy y Loss.

## Interpretación de resultados

Para la evaluación del entrenamiento y la validación de la red neuronal artificial se utilizaron las métricas de Accuracy y loss en cada época.

Imagen que contiene Forma

Descripción generada automáticamenteEn la Figura (2), se puede observar que se está registrando el accuracy y loss por cada época. Por un lado, el accuracy del entrenamiento, de color naranja, alcanza un cierto procentaje donde se mantiene ciertamente estable, asimismo, el accuracy de la validación, de color rojo, sigue el mismo comportamiento que el accuracy de la validación. Por otro lado, el loss del entrenamiento y la validación sigue un patrón decreciente similar llegando a un porcentaje mínimo, donde ciertamente se estabiliza.

*Fig. 2. Grafica de resultados*

|  |  |  |
| --- | --- | --- |
| **Predicción de rendimiento** | **Aprueba** | **Desaprueba** |
| Aprueba | 583 | 52 |
| Desaprueba | 158 | 692 |

*Tabla (1): Matriz de confusión de la red neuronal propuesto*

De la matriz de confusión mostrada en la Tabla (1), podemos calcular el Accuracy del modelo de red neuronal propuesto el cual es 85.85%. Además, en la Tabla (2), se muestran las siguientes métricas: Precision, Recall y F1-score. Dichas métricas se presentan para cada clase: “Aprueba” y “Desaprueba”, asimismo, en la última fila se presenta el promedio de cada métrica.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Precision** | **Recall** | **F1-score** |
| Aprueba | 93.0% | 81.0% | 87.0% |
| Desaprueba | 79.0% | 92.0% | 85.0% |
| **Promedio** | **86.0%** | **86.5%** | **86.0%** |

*Tabla (2): Reporte de clasificación de la red neuronal propuesto*

El modelo consiguió 85.85% de Accuracy, el cual indica el porcentaje de casos, en la validación,  que su predicción fue correcta. Asimismo, se obtuvo un 86.0% de Precision, el cual mide la calidad del modelo y nos muestra el número de alumnos que aprobaron el curso y que el modelo predijo que sí iban a aprobar el curso. En otras palabras, es el porcentaje de alumnos correctamente identificados, por el modelo, como aprobados del número total de alumnos identificados, por el modelo, como aprobados. También, el modelo logró 86.5% de Recall, el cual nos indica la cantidad de alumnos aprobados fue capaz de identificar el modelo, es decir, es el porcentaje de alumnos correctamente identificados como aprobados del total de alumnos aprobados. Por último, se obtuvo 86.0% F1-score, el cual es la combinación de las métricas Precision y Recall en un solo valor.

## Comparación y discusión

Con los resultados obtenidos de la predicción del rendimiento académico utilizando redes neuronales , podemos realizar comparaciones con otras investigaciones que poseen un modelo similar con diferentes algoritmos tales como Naive Bayes [6], Decision Tree [13], K-NN y SVM [18] .

Del análisis de la investigación presente, en la tabla 3 Se describen los diferentes modelos con sus respectivos porcentajes según Accuracy, Precision, Recall, F1-Mesure.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Autor** | **Técnica** | **Accuracy** | **Precision** | **Recall** | **F1 - Store** |
| **M.P** | **R.N.** | **0.858** | **0.860** | **0.865** | **0.860** |
| [18] | SVM | - | 0.981 | 1.000 | 0.990 |
| [18] | KNN | - | 0.981 | 0.990 | 0.986 |
| [9] | Naive Bayes | 0.960 | 0.857 | 0.890 | 0.873 |
| [13] | Decision Tree | 0.930 | 0.930 | 1.000 | 0.930 |

*Tabla (3): Matriz de análisis de resultados*

Luego del previo estudio realizado a las investigaciones mencionadas anteriormente y los resultado obtenido de [18], [6] y [13], se realizaron comparaciones en los resultados, como también en los datos de entrada, esta evaluación previa sirve para la posterior comparación de modelos.

Según la investigación [18], donde se evaluaron dos técnicas para la elaboración del proyecto, las cuales son SVM y k-NN. El modelo SVM da un 98% total de Precisión, esta precisión se debe al tipo de modelo empleado para la predicción de performance según los datos de entrada, la investigación [18] detalla también otros 3 modelos empleados donde se aprecia una variación en las instancias del dataset, a mayor instancias, la precisión es mayor, llegando en casos hasta el 100% de precisión. La mejora que se realizó de esta investigación [18] y la propuesta presente es que según los datos de entrada de la investigación[18], es un modelo empleado muy antiguo, a diferencia del presente que emplea modelos de entrada como, “cantidad de clicks”, “fechas”, los cuales son importantes, ya que en la actualidad se emplean más modelos educativos online a causa de la pandemia por el Covid-19.

Luego de realizar las comparaciones con la investigación [13], se emplearon 3 modelos, los cuales son Naive Bayes, Deep Learning, Decision Tree con una precisión del 96%, 98%, 93% respectivamente, a diferencia de nuestro modelo empleado, la información que requiere la investigación [13] rompe con nuestro modelo de ética, en el sentido de que se solicitan datos personales tales como “Edad”, “Address”, “Civil Status”, etc. A diferencia de nuestro modelo, que no requiere información privada, y siguiendo los lineamientos de la comparación con la investigación [18] donde el modelo de ingreso de datos era antiguo, y no regía el modelo de estudios de la actualidad (modelo de estudios online), nuestro modelo planteado devuelve un porcentaje de precisión del 86%, si bien es cierto el porcentaje de precisión es menor, sin embargo este porcentaje es obtenido sin requerir información privada, y con información requerida en la actualidad que no rompe los márgenes éticos establecidos con anterioridad.

Comparando la investigación [9] con el presente proyecto, podemos notar que en ambos casos emplean aprendizaje automático, en el caso de la investigación  [9] con minería de datos y en el de este proyecto con redes neuronales, de esta forma notamos que el porcentaje de precisión de la investigación [9] es mayor al nuestro con un 93% , sin embargo esto se debe a diferentes aspectos, ya que en la investigación [9] se centran en un conjunto de alumnos de primer año, por ello la manera en la que se formula la predicción del rendimiento académico, es diferente porque se centra más en los alumno que en el curso en general, debido a esto emplean datos más personales para así captar factores más psicológicos y cognitivos en los alumnos. A diferencia del nuestro que está desarrollado para poder predecir el rendimiento académico de un alumno mediante las notas que anteriormente dicho alumno presentó, así el programa es más flexible y permite usarlo en diferentes áreas estudiantiles.

# Conclusiones

En esta investigación, se propuso un modelo de red neuronal multicapa para la predicción de rendimiento académico del alumno en sus cursos, obteniendo un Accuracy de 85.8% en dichas predicciones.

También, la arquitectura propuesta utiliza dos datos “nuevos” para el entrenamiento del modelo de red neuronal multicapa. Dichos datos consisten en la interacción del alumno con la plataforma digital de la universidad, como la cantidad de clics que realiza el alumno a los materiales del curso y el promedio de las fechas de interacción del estudiante con el material del curso, medido como el número de días desde el inicio de la presentación del curso.

Además, en el modelo planteado fue necesario plantear métricas de ética en las cuales desarrollar el proyecto, a diferencia de las investigación comparadas fue necesario desarrollar el modelo de red neuronal con entrada de datos no personales, ya que en la actualidad se evidencia y denuncia en diversas industrias la violación y filtración de los datos personales.

# Trabajos futuros

En futuros trabajos, se podría plantear utilizar más datos referentes a la interacción del estudiante con la plataforma digital de su respectiva universidad para la predicción del rendimiento académico del alumno, ya que, por el confinamiento a causa del Covid-19, todas las actividades académicas se realizan por las plataformas digitales universitarias.

# Referencias

1. F. Tejedor y A. García-Valcárcel, "Causas del bajo rendimiento del estudiante universitario (en opinión de los profesores y alumnos). Propuestas de mejora en el marco del EEES", Revista de Educación, no. 342, pp. 443-473, 2007. Available: https://www.researchgate.net/publication/28161214\_Causas\_del\_bajo\_rendimiento\_del\_estudiante\_universitario\_en\_opinion\_de\_los\_profesores\_y\_alumnos\_Propuestas\_de\_mejora\_en\_el\_marco\_del\_EEES. [Accessed 6 May 2021].
2. "Cuando lo que se sabe nos dice cuánto no se sabe - Vicerrectorado Académico", Vicerrectorado Académico, 2017. [Online]. Available: https://vicerrectorado.pucp.edu.pe/academico/noticias/cuando-lo-que-se-sabe-nos-dice-cuanto-no-se-sabe/. [Accessed: 03- May- 2021].
3. Viale, H. "Una aproximación teórica a la deserción estudiantil", Revista Digital de Investigación en Docencia Universitaria, no. 1, 2014. Available: https://revistas.upc.edu.pe/index.php/docencia/article/view/366/346 [Accessed: 03- May- 2021].
4. SINEACE, "Modelo de Acreditación para Programas de Estudios de Educación Superior Universitaria", Lima, Perú, 2016. Available: https://www.sineace.gob.pe/wp-content/uploads/2014/08/Anexo-1-nuevo-modelo-programas-Resolucion-175.pdf. [Accessed 6 May 2021].
5. Blanco, J., Lovelle, S., Fernandez, R., Perez, E. (2016). Predicción de resultados académicos de estudiantes de informática mediante el uso de redes neuronales. Ingeniare. Revista chilena de ingeniería, vol. 24 N° 4,2016. pp. 715-727. doi:10.4067/S0718-33052016000400015
6. Rincon-Flores, E. G., Lopez-Camacho, E., Mena, J., & Lopez, O. O. (2020). Predicting academic performance with Artificial Intelligence (AI), a new tool for teachers and students. 2020 IEEE Global Engineering Education Conference (EDUCON). doi:10.1109/educon45650.2020.9125141
7. J. Kuzilek, M. Hlosta and Z. Zdrahal, "Open University Learning Analytics dataset", Scientific Data, 2017. Available: https://doi.org/10.1038/sdata.2017.171. [Accessed 6 May 2021].
8. Kumar, V. U., Krishna, A., Neelakanteswara, P., & Basha, C. Z. (2020). Advanced Prediction of Performance of a Student in an University using Machine Learning Techniques. 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC). doi:10.1109/icesc48915.2020.9155557
9. Jayaprakash, S., Krishnan, S., & V, J. (2020). Predicting Students Academic Performance using an Improved Random Forest Classifier. 2020 International Conference on Emerging Smart Computing and Informatics (ESCI). doi:10.1109/esci48226.2020.9167547
10. Daud, A., Aljohani, N. R., Abbasi, R. A., Lytras, M. D., Abbas, F., & Alowibdi, J. S. (2017). Predicting Student Performance using Advanced Learning Analytics. Proceedings of the 26th International Conference on World Wide Web Companion - WWW ’17 Companion. doi:10.1145/3041021.3054164
11. Widyaningsih, Y., Fitriani, N., & Sarwinda, D. (2019). A Semi-Supervised Learning Approach for Predicting Student’s Performance: First-Year Students Case Study. 2019 12th International Conference on Information & Communication Technology and System (ICTS). doi:10.1109/icts.2019.8850950
12. Devasia, T., Vinushree T P, & Hegde, V. (2016). Prediction of students' performance using Educational Data Mining. 2016 International Conference on Data Mining and Advanced Computing (SAPIENCE). doi:10.1109/sapience.2016.7684167
13. Amazona, M. V., & Hernandez, A. A. (2019). Modelling Student Performance Using Data Mining Techniques. Proceedings of the 2019 5th International Conference on Computing and Data Engineering - ICCDE’ 19. doi:10.1145/3330530.3330544
14. Páez, A., Guzmán, Daniel (2018) Diseño de un modelo para automatizar la predicción del rendimiento académico en estudiantes del IPN. Revista Iberoamericana para la Investigación y el Desarrollo Educativo ISSN 2007 - 7467. doi: [10.23913/ride.v8i16.340](https://doi.org/10.23913/ride.v8i16.340)
15. Castrillón, O., Sarache, W., Ruiz, S (2020). Prediction of academic performance using artificial intelligence techniques. From Univ. vol. 13 no.1 La Serena Feb 2020. Doi: 10.4067/S0718-50062020000100093
16. Ma, X., Yang, Y., & Zhou, Z. (2018). Using Machine Learning Algorithm to Predict Student Pass Rates In Online Education. Proceedings of the 3rd International Conference on Multimedia Systems and Signal Processing - ICMSSP ’18. doi:10.1145/3220162.3220188
17. Hasan, H. M. R., Rabby, A. S. A., Islam, M. T., & Hossain, S. A. (2019). Machine Learning Algorithm for Student’s Performance Prediction. 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT). doi:10.1109/icccnt45670.2019.8944629
18. Benablo, C. I. P., Sarte, E. T., Dormido, J. M. D., & Palaoag, T. (2018). Higher Education Student’s Academic Performance Analysis through Predictive Analytics. Proceedings of the 2018 7th International Conference on Software and Computer Applications - ICSCA 2018. doi:10.1145/3185089.3185102
19. Sekeroglu, B., Dimililer, K., & Tuncal, K. (2019). Student Performance Prediction and Classification Using Machine Learning Algorithms. Proceedings of the 2019 8th International Conference on Educational and Information Technology - ICEIT 2019. doi:10.1145/3318396.3318419
20. Liao, S. N., Zingaro, D., Thai, K., Alvarado, C., Griswold, W. G., & Porter, L. (2019). A Robust Machine Learning Technique to Predict Low-performing Students. ACM Transactions on Computing Education, 19(3), 1–19. doi:10.1145/3277569
21. F. Alaieri and A. Vellino, “A Decision Making Model for Ethical (Ro)bots”, IEEE International Symposium on Robotics and Intelligent Sensors (IRIS), 2017, pp. 203-207
22. Y. Bengio, I. Goodfellow and A. Courville, Deep learning. Massachusetts: MIT Press, 2017.
23. K. Lee and H. Shin, "An Application of a Deep Learning Algorithm for Automatic Detection of Unexpected Accidents Under Bad CCTV Monitoring Conditions in Tunnels", 2019 International Conference on Deep Learning and Machine Learning in Emerging Applications (Deep-ML), 2019. Available: https://doi.org/10.1109/Deep-ML.2019.00010. [Accessed 5 May 2021].
24. Stevenson, R., Mikels, J., James, T (2007)Characterization of the Affective Norms for English Words by discrete emocional categories. Behavior Reasearch Methods, Pscychonomic Society, Inc.(4) 1020-1024.
25. Fabg, Xi., Zhan, J (2015) Sentiment analysis using product review data, Journal of Big Data 2, 5.doi: 10.1186/s40537-015-0015-2
26. Korashy, H., Hassan, A., Medhat, W (2014) Semtiment analysis algorithms and applications: A survey. Electrical Engineering. Ain Shams Engineering Journal. <https://doi.org/10.1016/j.asej.2014.04.011>
27. Witten, H., & Frank, E. (2016). Data Mining Practical Machine Learning Tools and Techniques (Second Edition). Morgan Kaufmann Publishers.
28. J. E. Chire Saire and J. F. Oblitas Cruz, "Study of Coronavirus Impact on Parisian Population from April to June using Twitter and Text Mining Approach," 2020 International Computer Symposium (ICS), 2020, pp. 242-246, doi: 10.1109/ICS51289.2020.00056
29. J. Heaton, Introduction to neural networks with Java, 2nd ed. Chesterfield (MO, USA): Heaton Research, 2009, p. 129.