Predicción de Aprobación o Reprobación de la Calificación del Estudiante con Python

Jhon Marck Leon Cruz[1]
181216@unamba.edu.pe [1]
Jose Luis Goizueta Castañeda[5]
181209@unamba.edu.pe [5]
Alfredo Cervantes Ccasa[2]
172138@unamba.edu.pe [2]
Luis Fernando Leo Huamani[3]
172154@unamba.edu.pe [3]
Yesica Condori Ninai[4]
182200@unamba.edu.pe [4]
Jose Condori Condori[6]
181203@unamba.edu.pe [6]

Student Grade Pass or Fail Prediction with Python

RESUMEN: *Esta propuesta plantea la selección de variables que influyen en la predicción del rendimiento de los estudiantes en la educación secundaria de dos escuelas portuguesas (Gabriel Pereira y Mousinho da Silveira). Se implementaron algoritmos de clasificación a través del lenguaje de programación Python como árbol de decisión para conocer el mejor resultado de predicción. Los atributos de los datos incluyen calificaciones de los estudiantes, características demográficas, sociales y relacionadas con la escuela) y los datos se recopilaron mediante el uso búsqueda en la web, en kagle.com se ofrece una infinidad de dataset's enfocados en todas las áreas para su posterior uso y específicamente el dataset se extrajo de ahí, para luego trabajar con ello, se hizo el análisis de los datos, transformación, limpieza de datos, normalización, definición de entradas, creación de funciones para luego entrenar la red neuronal, finalmente se muestra el valor del error de entrenamiento y el porcentaje de predicción. El algoritmo árbol de decisión arroja mejor exactitud con respecto a los otros algoritmos. Se concluye que las variables que más influyen en el rendimiento académico de los estudiantes de ingeniería son: género, edad, tiempo de estudio, acceso a internet, falta a clases, puntaje del primer ciclo, puntaje del segundo ciclo y la nota final de todo el ciclo, todas estas notas van en un rango de calificación de 0 a 20 para dar una salida binaria de aprobado o desaprobado.*

PALABRA CLAVE: *Algoritmo, Dataset, Entrenamiento, Error De Entrenamiento, Machine Learning, Predicción, Red Neuronal.*

ABSTRACT. *This proposal proposes the selection of variables that influence the prediction of student performance in secondary education in two Portuguese schools (Gabriel Pereira and Mousinho da Silveira). Classification algorithms were implemented through the Python programming language as a decision tree to know the best prediction result. The data attributes include student scores, demographic, social and school-related characteristics) and the data was collected using web search, kagle.com offers a myriad of dataset's focused on all areas to its subsequent use and specifically the dataset was extracted from there, to then work with it, the data analysis, transformation, data cleaning, normalization, definition of inputs, creation of functions was done to then train the neural network, finally it was shows the value of the training error and the prediction percentage. The decision tree algorithm yields better accuracy with respect to the other algorithms. It is concluded that the variables that most influence the academic performance of engineering students are: gender, age, study time, internet access, absence from classes, first cycle score, second cycle score and the final grade of all cycle, all of these grades range from 0 to 20 to give a binary output of pass or fail.*

Keywords: *Algorithm, Dataset, Training, Training Error, Machine Learning, Prediction, Neural Network.*

1 INTRODUCCIÓN

En la actualidad se están presentando diversos cambios en áreas como la medicina, la educación, la economía, y el comercio, entre otros, como producto de la incursión de la analítica de datos en ellos. Y tanto ha sido el apogeo de la analítica que ha permeado el campo de la educación, donde se empieza a procesar un gran volumen datos que contienen la información relacionada con los actores del proceso educativo. Y es aquí donde la ingeniera puede jugar un papel importante al dar su aporte para solucionar múltiples aspectos de índole académico tales como mejorar el aprendizaje, la deserción, el abandono y el rendimiento académico.

Existen diversos factores que permiten medir la eficiencia del proceso educativo, muchos de ellos de carácter multidimensional, como es el caso del fenómeno de retención, definido como la diferencia entre el número de estudiantes que ingresan en primer semestre y los graduados por año (Salcedo, 2010); y del rendimiento académico definido como el principal indicador de éxito o fracaso del estudiante.

2 MARCO TEÓRICO

APRENDIZAJE AUTOMÁTICO

El aprendizaje automático (ML) es el subapartado de la inteligencia artificial (IA) que se centra en desarrollar sistemas que aprenden, o mejoran el rendimiento, en función de los datos que consumen. Inteligencia artificial es un término amplio que se refiere a sistemas o máquinas que imitan la inteligencia humana. El aprendizaje automático y la IA suelen nombrarse juntos, y los términos a veces se usan indistintamente, pero no significan lo mismo. Un aspecto importante que hay que destacar es que, aunque todo aprendizaje automático es IA, no toda la IA es aprendizaje automático.
Hoy en día, el aprendizaje automático está en todas partes. Cuando interactuamos con bancos, realizamos compras online o usamos redes sociales, los algoritmos de aprendizaje automático entran en juego para que nuestra experiencia sea eficiente, fluida y segura. El aprendizaje automático y la tecnología relacionada se desarrollan rápidamente, y apenas estamos empezando a conocer la superficie de sus capacidades.[1]

INTELIGENCIA ARTIFICIAL

Una de las más atractivas soluciones que han ingresado al mercado, son las que se basan o utilizan todo aquello relacionado con Inteligencia artificial y aprendizaje automático. Gracias a que estas tecnologías ayudan a optimizar algunos procesos, reducir gastos y mejorar la efectividad en muchos procedimientos. Esto ha ocasionado un crecimiento exponencial en las inversiones en este sector, algunos medios dan cuenta que la inyección de capital destinado únicamente para la inteligencia artificial (IA) ronda aproximadamente en 1.2 billones de dólares [2].

MACHINE LEARGING

Es una disciplina científica del ámbito de la Inteligencia Artificial que crea sistemas que aprenden automáticamente. Aprender en este contexto quiere decir identificar patrones complejos en millones de datos. La máquina que realmente aprende es un algoritmo que revisa los datos y es capaz de predecir comportamientos futuros. Automáticamente, también en este contexto, implica que estos sistemas se mejoran de forma autónoma con el tiempo, sin intervención humana. Veamos cómo funciona.

Una empresa de telefonía quiere saber qué clientes están en “peligro” de darse de baja de sus servicios para hacer acciones comerciales que eviten que se vayan a la competencia. ¿Cómo puede hacerlo? La empresa tiene muchos datos de los clientes, muchísimos: antigüedad, planes contratados, consumo diario, llamadas mensuales al servicio de atención al cliente, últimos cambios de planes contratados… pero seguramente los usa solo para facturar y para hacer estadísticas. ¿Qué más puede hacer con esos datos? Se pueden usar para predecir cuándo un cliente se va a dar de baja y gestionar la mejor acción que lo evite. En pocas palabras, con Machine Learning se puede pasar de ser reactivos a ser proactivos. Los datos históricos del conjunto de los clientes, debidamente organizados y tratados en bloque, generan una base de datos que se puede explotar para predecir futuros comportamientos, favorecer aquellos que mejoran los objetivos de negocio y evitar aquellos que son perjudiciales. [3].

RED NEURONAL

Las redes neuronales artificiales (también conocidas como sistemas conexionistas) se trata de modelo computacional evolucionado a partir de diversas aportaciones científicas que están registradas en la historia.1 Consiste en un conjunto de unidades, llamadas neuronas artificiales, conectadas entre sí para transmitirse señales. La información de entrada atraviesa la red neuronal (donde se somete a diversas operaciones) produciendo unos valores de salida.
Cada neurona está conectada con otras a través de unos enlaces. En estos enlaces el valor de salida de la neurona anterior es multiplicado por un valor de peso. Estos pesos en los enlaces pueden incrementar o inhibir el estado de activación de las neuronas adyacentes. Del mismo modo, a la salida de la neurona, puede existir una función limitadora o umbral, que modifica el valor resultado o impone un límite que no se debe sobrepasar antes de propagarse a otra neurona. Esta función se conoce como función de activación. [4]

3 METODOLOGÍA

Como se ha mencionado en la introducción, este es un proyecto de desarrollo de software utilizando herramientas de Machine Learning para determinar la probabilidad de deserción de estudiantes, sin embargo, no deja de ser un proyecto de desarrollo de software. Al ser de este tipo lo más recomendable es utilizar una metodología de desarrollo que se encuentre orientada al desarrollo de este tipo de modelos.

En la actualidad existen muchos modelos de desarrollo o marcos de trabajo, de los más conocidos e implementados a nivel empresarial son:

- SCRUM
- KANBAN
- XP

Sin embargo, estos marcos de trabajo se encuentran orientados al desarrollo ágil, pero no están pensados para proyectos de análisis de datos, que en muchos casos requiere múltiples iteraciones en el análisis de los datos para encontrar la información que sea útil para el modelo a implementar.

Teniendo en cuenta lo anterior, se utilizará la metodología de trabajo CRISP-DM, esta metodología está pensada específicamente para proyectos de ciencias de datos y análisis de datos. [5]

4 DESCRIPCIÓN DE LOS DATOS

Estos datos abordan el rendimiento de los estudiantes en la educación secundaria de dos escuelas portuguesas. Los atributos de los datos incluyen calificaciones de los estudiantes, características demográficas, sociales y relacionadas con la escuela) y se recopilaron mediante el uso de informes y cuestionarios escolares. Se proporcionan dos conjuntos de datos sobre el rendimiento en dos materias distintas: Matemáticas (mat) y lengua portuguesa (por). En [Cortez and Silva, 2008], los dos

conjuntos de datos se modelaron bajo tareas de regresión y clasificación binaria/de cinco niveles.[6]

El atributo objetivo G3 tiene una fuerte correlación con los atributos G2 y G1. Esto ocurre porque G3 es la calificación del último año (emitida en el 3er período), mientras que G1 y G2 corresponden a las calificaciones del 1er y 2do período. Es más difícil predecir G3 sin G2 y G1, pero dicha predicción es mucho más útil (consulte la fuente en papel para obtener más detalles). [6]

INFORMACIÓN DE ATRIBUTOS

Nota importante, a diferencia del conjunto de datos original, este convirtió todos los valores binarios en valores booleanos numéricos. Además, no tiene atributos Mjob, Fjob, reason,guardian .

school- escuela del estudiante (binario: 'GP' - Gabriel Pereira o 'MS' - Mousinho da Silveira) (bool: 1 si es GP sino 0)sex- sexo del estudiante (binario: 'F' - femenino o 'M' - masculino) (bool: 1 si es F sino 0)

age- edad del estudiante (numérico: de 15 a 22)

address- tipo de domicilio del estudiante (binario: 'U' - urbana o 'R' - rural) (bool: 1 si es U sino 0)

f**amsize**- tamaño de la familia (binario: 'LE3' - menor o igual a 3 o 'GT3' - mayor a 3) (bool: 1 si es LE3 si no 0).

Pstatus- estado de convivencia de los padres (binario: 'T' - viviendo juntos o 'A ' - aparte) (bool: 1 si es T sino 0).

Medu- estudios de la madre (numérico: 0 - ninguno, 1 - educación primaria (4° grado), 2 – 5° a 9° grado, 3 – educación secundaria o 4 – educación superior)

Fedu- estudios del padre (numérico: 0 - ninguno, 1 - educación primaria (4° grado), 2 – 5° a 9° grado, 3 – educación secundaria o 4 – educación superior).

traveltime- tiempo de viaje de la casa a la escuela (numérico: 1 - 1 hora).

studytime- tiempo de estudio semanal (numérico: 1 - 10 horas).

failures- número de fracasos de clases anteriores (numérico: n si 1<=n<3, si no 4)

schoolsup- apoyo educativo adicional (binario: sí o no) (booleano: 1 si es Sí si no 0)

famsup- apoyo educativo familiar (binario: sí o no) (bool: 1 si es Sí más 0)

paid- clases extra pagadas dentro de la materia del curso (matemáticas o portugués) (binario: sí o no) (bool: 1 si es Sí más 0)).

activities- actividades extraescolares (binario: sí o no) (bool: 1 si es Sí más 0)

nursery- asistencia a guardería (binario: sí o no) (bool: 1 si es Sí más 0)

higher- quiere cursar estudios superiores (binario: sí o no) (bool: 1 si es Sí más 0)

internet- Acceso a Internet en casa (binario: sí o no) (bool: 1 si es Sí más 0)

romantic- con una relación sentimental (binario: sí o no) (bool: 1 si es Sí sino 0)

famrel- calidad de las relaciones familiares (numérico: de 1 - muy mala a 5 - excelente)

freetime- tiempo libre después de la escuela (numérico: de 1 - muy bajo a 5 - muy alto)

goout- salir con amigos (numérico: de 1 - muy bajo a 5 - muy alto)

Dalc- consumo de alcohol en jornada laboral (numérico: de 1 - muy bajo a 5 - muy alto)

Walc- consumo de alcohol en fin de semana (numérico: de 1 - muy bajo a 5 - muy alto)

health- estado de salud actual (numérico: de 1 - muy malo a 5 - muy bueno)

absences- número de ausencias escolares (numérico: de 0 a 93)

G1- nota del primer periodo (numérica: de 0 a 20)

G2- nota del segundo periodo (numérica: de 0 a 20)

G3- nota final (numérica: de 0 a 20)

pass- (bool: 1 si ese alumno aprueba sino 0) objetivo de salida.[6]

5 PREPARACIÓN DE DATOS

Después de haber realizado el análisis descriptivo o descripción de los datos se debe realizar una preparación de la información, esto porque los datos pueden contar con información que no sea relevante para el análisis.

En esta fase se toman los datos, con los que se realizó la etapa anterior y se empieza a identificar qué columnas son necesarias o útiles para la ejecución del modelo. Se debe tener claro que se realizan varias iteraciones donde se completen y estructuren los datos, hasta que se consiga el resultado esperado. Es decir que en cada iteración pueden entrar o salir variables según se crea conveniente.

En la iteración preliminar, se identificaron como útiles las siguientes variables:

- Subir el dataset a Google Colab

```
import pandas as pd
from sklearn.model_selection import train_test_split
```

```
train=pd.read_csv('/content/student-mat-pass-or-fail.csv')
```

```
train.head()
```

Figura 1. importar dataset.

- Verificación de datos

	school	sex	age	address	famsize	Pstatus	Medu	Fedu	traveltime	studytime	...	freetime	goout	Dalc	Walc	health	absences	G1	G2	G3	pass
0	1	1	18	1	0	0	4	4	2	2	...	3	4	1	1	3	6	5	6	6	0
1	1	1	17	1	0	1	1	1	1	2	...	3	3	1	1	3	4	5	5	6	0
2	1	1	15	1	1	1	1	1	1	2	...	3	2	2	3	3	10	7	8	10	1
3	1	1	15	1	0	1	4	2	1	3	...	2	2	1	1	5	2	15	14	15	1
4	1	1	16	1	0	1	3	3	1	2	...	3	2	1	2	5	4	6	10	10	1

5 rows × 30 columns

Figura 2. Comprobación de datos.

- Selección de campos necesarios

```
train=train[['sex','age','studytime','internet','absences','G1','G2','G3','pass']]
train.head()
```

	sex	age	studytime	internet	absences	G1	G2	G3	pass
0	1	18	2	0	6	5	6	6	0
1	1	17	2	1	4	5	5	6	0
2	1	15	2	1	10	7	8	10	1
3	1	15	3	1	2	15	14	15	1
4	1	16	2	0	4	6	10	10	1

Figura 3. selección de datos..

- Limpieza de datos

```
# Este elimina las filas que tengan algun nulo
train=train.dropna()
#Luego eliminados los duplicados
train=train.drop_duplicates()
train.head()
```

	sex	age	studytime	internet	absences	G1	G2	G3	pass
0	1	18	2	0	6	5	6	6	0
1	1	17	2	1	4	5	5	6	0
2	1	15	2	1	10	7	8	10	1
3	1	15	3	1	2	15	14	15	1
4	1	16	2	0	4	6	10	10	1

Figura 4. Limpieza de datos.

- Búsqueda de patrones redes neuronales

```
#Trabajar las ausencias
#Facilita encontrar patrones para RN.
trainDatax['absences']=pd.cut(trainDatax['absences'],[0,15,30,45,60,75])
```

Figura 5.Búsqueda de patrones RN .

- Normalización de datos

```
#Normalizar los datos
from sklearn.preprocessing import StandardScaler
```

```
scaler=StandardScaler()
scaler.fit(trainDatax)
print(scaler.mean_)
```

```
[ 0.52429668 16.69309463  2.03580563  0.83120205 -0.18414322  0.51150895
  0.48081841  0.43478261]
```

Figura 6.Normalización de datos

- Entrenamiento de la red neuronal

```
# Entrenar la RN
import numpy as np
from sklearn.model_selection import train_test_split
```

```
trainDatay=trainDatay.to_numpy()
x_train, x_test, y_train, y_test= train_test_split(trainDatax, trainDatay, test_size=0.3)
```

```
print(x_train[0:5], x_test[0:5], y_train[0:5], y_test[0:5])
```

```
[[-1.04983358  0.24041676 -0.04250198 -2.21906341 -1.23461776 -1.02328902
  -0.85516408 -0.6592119 ]
 [ 0.95253193 -0.54294119  1.14451761  0.45064057 -1.23461776  0.97724101
   0.92339526  0.85697547]
 [ 0.95253193 -1.32629914 -1.22952157 -2.21906341 -1.23461776 -1.02328902
  -0.85516408 -0.6592119 ]
 [ 0.95253193  1.02377472 -0.04250198 -2.21906341  0.27865981  0.97724101
   0.92339526  0.85697547]
 [ 0.95253193  0.24041676 -0.04250198  0.45064057  4.8184925  -1.02328902
  -0.85516408 -0.6592119 ]] [[ 0.95253193 -0.54294119 -0.04250198 -2.21906341  0.27865981 -1.02328902
  -0.85516408 -0.6592119 ]
 [-1.04983358 -0.54294119 -0.04250198  0.45064057  0.27865981  0.97724101
   0.92339526  0.85697547]
 [ 0.95253193  1.80713267 -0.04250198  0.45064057  1.79193737  0.97724101
  -0.85516408  0.85697547]
 [-1.04983358 -1.32629914  2.3315372   0.45064057  0.27865981  0.97724101
   0.92339526 -0.6592119 ]
 [ 0.95253193  1.02377472 -0.04250198  0.45064057  1.79193737 -1.02328902
  -0.85516408 -0.6592119 ]] [[0]
```

Figura 7. entrenamiento de RN

- Error de entrenamiento .

```
import matplotlib.pyplot as plt
plt.plot(errorVec)
plt.title("Error de entrenamiento")
```

```
Text(0.5, 1.0, 'Error de entrenamiento')
```

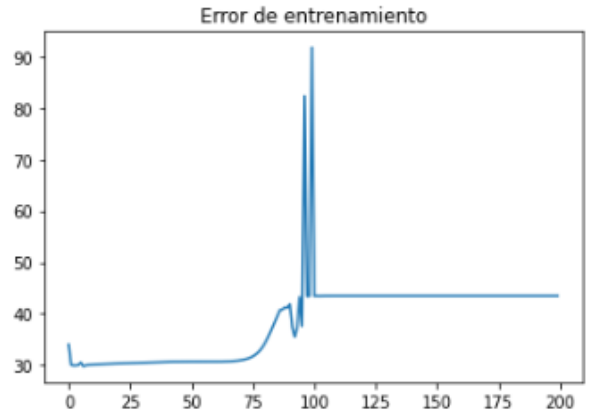

Figura 8.gráfica del Error de entrenamiento.

6 RESULTADOS

En este gráfico se puede visualizar que el modelo ha sido capaz de aprender la relación de los datos de entrada(datos importantes del estudiante) y los resultados(si aprobara el curso

11 Conclusiones

Al desarrollar el presente trabajo, se ha puesto en práctica lo aprendido en computación a la nube y sobre todo la importancia que tiene el Error De Entrenamiento, Machine Learning, Red Neuronal hoy en día. Además, las técnicas que utiliza para describir y predecir errores.

En este estudio se ha demostrado que el uso de red neuronal para la predicción del estado de aprobado/reprobado de los estudiantes en un curso académico, en general, produce la mayor parte resultados exitosos, incluso rindiendo un 86% notable precisiones de clasificación para 10 de 23 modelos, aunque el desempeño de otros clasificadores también ha encontrado dentro de los límites de precisión aceptable.

Redes Neuronales, presenta grandes ventajas con respecto a otros modelos típicos de solución de problemas de Ingeniería, una de ellas es su inspiración en modelos biológicos del funcionamiento del cerebro, lo que facilita su estudio debido a las analogías que pueden introducirse para su análisis.

RECONOCIMIENTO

"Acknowledgment" en inglés americano. Evite las expresiones como "Uno de nosotros (S.B.A.) gustaría agradecer… ." Exponga reconocimientos a patrocinadores y de apoyo financieros. Los Alumnos de la asignatura: _____ del II-2005, desean expresar su agradecimiento a la Universidad de Pamplona por todo el apoyo recibido durante el desarrollo del curso….

12 REFERENCIAS.

[1] Salcedo, A., Desertion in Colombian Universities, http://www.alfaguia.org/alfaguia/files/1319043663_03.pdf, Revista Academia y Virtualidad, 3(1), 50-60 (2010)

[2] Banco interamericano de desarrollo, "Inteligencia Artificial: la región se abre al desarrollo". [Internet]. Disponible en: https://conexionintal.iadb.org/2018/05/30/ideas-2/

[3] Cleverdata. "¿Qué es Machine Learging?" [Internet]. https://cleverdata.io/que-es-machine-learning-bigdata/#:~:text=Machine%20Learning%20es%20una%20disciplina,complejos%20en%20millones%20de%20datos.

[3] González A. Machine Learning, https://www.oracle.com/mx/data-science/machine-learning/what-is-machine-learning/

[4] Salcedo, A.Historia de las redes neurales [url=http://www.itnuevolaredo.edu.mx/takeyas/apuntes/Inteligencia%20Artificial/Apuntes/tareas_alumnos/RNA/Redes%20Neuronales2.pdf

[5] Singular (2016), "CRISP-DM: La metodología para poner orden en los proyectos".https://www.sngular.com/es/data-science-crisp-dm-metodologia/

[6] Kagle,Dinh Anh Predicción de aprobación o reprobación de la calificación del estudiante, https://www.kaggle.com/datasets/dinhanhx/studentgradepassorfailprediction.

[7] J. W. Henry, M. J. Martinko, and M. A. Pierce, "Attributional style as a predictor of success in a first computer science course," Comput. Human Behav., vol. 9, no. 4, pp. 341–352, Dec. 1993.