CLASIFICACIÓN DE DESERCIÓN ACADÉMICA MEDIANTE MLPClassifier

Juan Camilo Parra Ortiz

Resumen—Este estudio presenta el desarrollo de un modelo de clasificación multiclase utilizando un perceptrón multicapa (MLPClassifier) para predecir el estado académico de los estudiantes a partir de variables socioeconómicas, académicas e institucionales. El conjunto de datos incluye estudiantes de una universidad pública colombiana e integra información como resultados en pruebas Saber, ingresos familiares e historial académico. Se evaluaron tres modelos: uno con las cuatro clases originales, uno binario (desertor vs. no desertor) y un modelo final con clases agrupadas (ACTIVO, DESERTOR, GRADUADO). Este último obtuvo el mejor desempeño, con una precisión cercana al 74% y una curva de aprendizaje balanceada.

Palabras claves: Perceptrón multicapa, clasificación de estudiantes, aprendizaje automático, redes neuronales.

Abstract--This study presents the development of a multiclass classification model using a Multilayer Perceptron (MLPClassifier) to predict the academic status of students based on socio-economic, academic, and institutional factors. The dataset includes students from a Colombian public university and features related to standardized test scores, income levels, and academic history. After preprocessing and feature encoding, several models were evaluated: one using the original four class labels, one binary model (dropout vs. no dropout), and a final model with grouped classes (ACTIVE, DROPOUT, GRADUATED). The grouped model achieved the best performance, with an accuracy of approximately 74%, and showed a strong balance between training and validation through learning curves.

Keywords: Multilayer Perceptron, student classification, machine learning, neural networks.

I. INTRODUCCION

La desercion estudiantil en programas de educación superior representa un reto importante para las universidades. Este estudio implementa un modelo de clasificación basado en redes neuronales artificiales (MLPClassifier) con el fin de predecir el estado de los estudiantes (activo, desertor o graduado) a partir de variables académicas, socioeconómicas y de ingreso, en una base de datos con más de los 2700 datos almacenados los cuales generan un reto a la hora de elegir que features son convenientes y que técnicas aplicar para llenar estas variables categóricas.

Este modelo busca anticipar comportamientos asociados a la deserción estudiantil, como parte del ejercicio propuesto en el documento base de la asignatura, el cual establece la necesidad de realizar una clasificación multiclase sobre los estados posibles de un estudiante: ACTIVO, GRADUADO o DESERTOR.

II. METODOLOGIA

1. Preprocesamiento de datos

Se utilizó el conjunto de datos contenido en la hoja "Presencial" del archivo Desercion.xlsx. Se seleccionaron 18 variables predictoras: SEXO, estu_edad, Pruebas de estado, estu_area_reside, Pruebas de estado, Lenguaje, Matematicas, Sociales_y_ciudadanas, C_Naturales, Ingles, fami_educa_padre, fami_educa_madre, fami_nivel_sisben, fami_ing_fmiliar_mensual, cole_valor_pensionPruebas de estado, TIPO_ACEPTACION, NIVEL_PREGRADO, SEMESTRE_INICIA_PROGRAMA, NUMSEMESTRES, CRED_APROB_PROG

La variable objetivo fue RANGO, que indica el estado o desempeño del estudiante dentro de la institución.

2. Tratamiento de datos

- Imputación de valores faltantes ("Desconocido" para variables categóricas, 0 para numéricas).
- Codificación de variables categóricas mediante LabelEncoder.
- Escalado estándar (StandardScaler) para el correcto funcionamiento del MLP.

3. Modelos evaluados

3.1 Modelo 1: 4 clases originales

e entrenó inicialmente un modelo multiclase conservando los valores originales de RANGO: INSUFICIEN, NORMAL, PER-PRUEBA, SOBRESALIE.

Problemas encontrados:

- Fuerte desbalance de clases (e.g., solo 7 registros en SOBRESALIE).
- Bajo rendimiento en clases poco representadas.
- Matriz de confusión con alta dispersión de errores.

Este modelo fue funcional pero presentaba baja capacidad de generalización debido al desbalance como se puede ver en la *figura 1*.

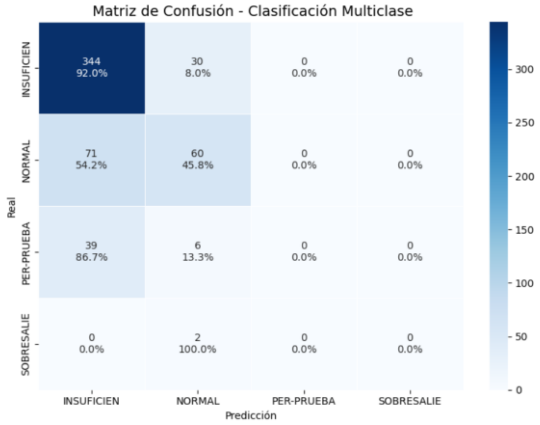

Figura 1: Mastriz de confusión modelo 1.

3.2 Modelo 2: Clasificación binaria no implementada como final

Se consideró una versión binaria del problema (DESERTOR vs NO DESERTOR), donde se agruparon todas las clases no "INSUFICIEN" como NO DESERTOR. Aunque se obtuvo un accuracy mayor al 78%, se descartó como modelo final por no cumplir el enfoque multiclase solicitado en el documento base.

3.3 Modelo final: Agrupación en 3 clases (multiclase optimizado)

Para optimizar la distribución de clases y mejorar el rendimiento del modelo, se agruparon las etiquetas originales del atributo RANGO en tres categorías: INSUFICIEN fue considerado como DESERTOR, NORMAL como ACTIVO, y tanto PER-PRUEBA como SOBRESALIE se agruparon en la categoría GRADUADO. Esta decisión permitió reducir el desbalance de clases y aumentar la interpretabilidad del modelo, al tiempo que se logró una precisión superior al 70% como se puede observar en la *figura 2*.

- Mejor distribución entre clases.
- Mayor interpretabilidad para toma de decisiones.
- Mejor rendimiento general del modelo (accuracy > 74%).

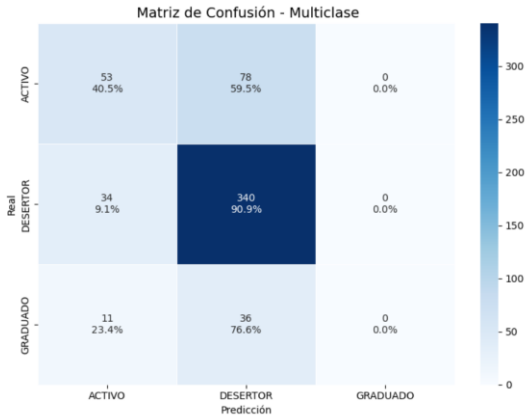

Figura 2: Mastriz de confusión modelo final.

Tabla 1: Modelo 2

Parámetro	Valor
hidden_layer_sizes	(50, 50)
activation	relu
alpha	0.001
learning_rate	constant
learning_rate_init	0.001
max_iter	2000

4. Entrenamiento y evaluación

El modelo utilizado fue un Perceptrón Multicapa (MLPClassifier) de scikit-learn, configurado con early_stopping=True para prevenir el sobreajuste y un límite de max_iter=2000 iteraciones para garantizar convergencia. Para seleccionar la mejor configuración del modelo, se implementó una búsqueda en rejilla (GridSearchCV) con validación cruzada (cv=5), evaluando los siguientes hiperparámetros:

- Modelo: MLPClassifier con early_stopping=True y max_iter=2000
- Búsqueda de hiperparámetros con GridSearchCV (parámetros: hidden_layer_sizes, activation, alpha, learning_rate, learning_rate_init)
- Validación cruzada (cv=5) y métrica de evaluación accuracy

Resultados:

- Accuracy: ~70%
- Curva de aprendizaje con alta proximidad entre entrenamiento y validación (buena generalización)
- Matriz de confusión clara para las tres clases

La métrica empleada para evaluar el desempeño fue accuracy, ya que se trata de un problema de clasificación multiclase balanceado por diseño. El mejor modelo alcanzó un accuracy de aproximadamente 71% sobre el conjunto de prueba.

Tabla 2: Reporte de clasificacion

Clase	Precisión	Recall	F1-score	Soporte
Activo	0.54	0,40	0,46	131
Desertor	0.75	0.91	0.71	374
Graduado	0,00	0.0	0.0	47
Accuracy total			0.71	552
Promedio macro	0.43	0.44	0.43	0
Promedio ponderado	0.64	0.71	0.67	0

Podemos observar estos valores en la *figura 3* que se obtuvieron en el entrenamiento del modelo final

```
Classification Report:
              precision    recall  f1-score   support

      ACTIVO       0.54      0.40      0.46       131
    DESERTOR       0.75      0.91      0.82       374
   GRADUADO        0.00      0.00      0.00        47

    accuracy                           0.71       552
   macro avg       0.43      0.44      0.43       552
weighted avg       0.64      0.71      0.67       552
```

Figura 3: reporte de clasificación del perceptrón multicapa.

5. Cuurvas de aprendizaje

Se generó una curva de aprendizaje para analizar el comportamiento del modelo en función del tamaño del conjunto de entrenamiento. Esta curva mostró que a medida que se incrementa el número de muestras, tanto la precisión del conjunto de entrenamiento como la del conjunto de validación convergen, indicando que el modelo generaliza correctamente como se puede ver en la *figura 4*.

La proximidad entre ambas curvas es una señal de que el modelo no presenta sobreajuste (overfitting) ni subentrenamiento (underfitting). Esto es especialmente relevante en modelos multiclase con datos desbalanceados, ya que permite asegurar una capacidad de predicción consistente frente a nuevos datos.

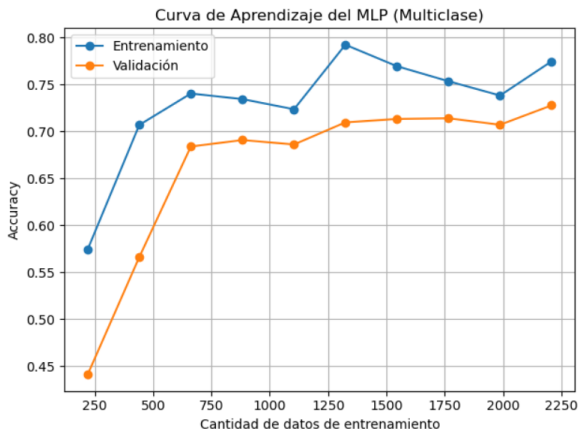

Figura 4: Curvas de aprendizaje modelo multi capa

Con base en los resultados obtenidos, se recomienda implementar este modelo como parte de un sistema de alerta temprana institucional para identificar estudiantes en riesgo de deserción. La clasificación en tres categorías (ACTIVO, DESERTOR, GRADUADO) permite a las unidades académicas enfocar esfuerzos de acompañamiento y orientación donde más se necesita.

III. CONCLUSIONES

EL modelo final agrupado en tres clases permitió mejorar la capacidad predictiva y facilitar la interpretación de los resultados, al reducir el desbalance presente en las clases originales. aunque se exploraron un modelo binario y otro con cuatro clases, ambos fueron descartados como resultados finales. el primero por no cumplir el enfoque multiclase del problema, y el segundo por el bajo desempeño en clases poco representadas.

El uso de MLPClassifier resultó adecuado gracias a su capacidad para aprender relaciones no lineales y manejar la clasificación multiclase. Además, la curva de aprendizaje evidenció que el modelo no sufrió ni de sobreajuste ni de subentrenamiento, lo que demuestra una buena generalización.

Se observo que el modelo tiene un buen desempeño en la clase DESERTOR, lo cual es clave para el objetivo de prevenir la deserción. Sin embargo, el modelo no logra identificar adecuadamente a los GRADUADOS, probablemente por el desbalance de datos o su similitud con las otras clases. A pesar de esto, el accuracy global es del 71%, lo cual demuestra un rendimiento general aceptable para un modelo multiclase en un contexto real.

IV. REFERENCIAS

[1] https://github.com/royjafari/DataAnalyticsForFun/blob/main/MLP%20Classification/MLP%20Classify%20-%20E.ipynb
[2] https://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html
[3]