Tecnológico de Monterrey
Escuela de Ingeniería y Ciencias

Campus Estado de México

Momento de retroalimentación: Módulo 2 Análisis y reporte sobre el desempeño del modelo. (Portafolio Análisis)

Curso:

Inteligencia artificial avanzada para la ciencia de datos I (Gpo 101)

Gabriela Cortés Olvera - A01751655

Profesor:

Jorge Adolfo Ramirez Uresti

Fecha de entrega:

11 de septiembre

Información general del dataset

La fuente de datos proviene de un conjunto de datos proveniente de kaggle (link: https://www.kaggle.com/code/malik9/student-dropout-prediction-with-91-7-accuracy). El conjunto de datos abarca información disponible en el momento de la inscripción del estudiante (trayectoria académica, datos demográficos y factores socioeconómicos) así como el rendimiento académico de los estudiantes al final del primer y segundo semestre.

Este conjunto de datos proporciona el fundamento necesario para desarrollar y entrenar el modelo de análisis discriminante lineal con ayuda de un framwork, aprovechando sus características y atributos para realizar predicciones informadas sobre el abandono escolar. A partir de los resultados obtenidos del framework se comparará de manera general con un modelo previamente realizado sin framework lo cual permititrá comparar dichos modelos.

En las siguientes secciones, se presentarán los resultados obtenidos de la creación de lda para el modelo. Además, se ofrecerán descripciones exhaustivas sobre los aspectos clave del modelo de análisis discriminante lineal, subrayando su funcionamiento y relevancia en el contexto de la predicción de resultados educativos.

El objetivo se centra en la implementación del modelo, para comprender el mecanismo interno y su capacidad para aportar información valiosa en la toma de decisiones relacionadas con la retención estudiantil. De igual manera dicho ejercicio proporcionará información clave de la evaluación del modelo conforme al dataset en cuestión.

El dataset cuenta con 35 variables y 4425 registros, en dicho dataset no se encontraron valores duplicados y faltantes. Por lo anterior, para el tratamiento de los datos no fue necesario tratar datos faltantes, sin embargo, se realizó “hot encoding” para el tratamiento de los datos categóricos. Posteriormente se realizó eliminación de columnas irrelevantes mediante las correlaciones altas, obteniendo como resultado la eliminación de las columnas ‘Nacionality’, ‘Curricular unir 1st sem (enrolled), ‘Curricular unir 2nd sem (credited), ‘Curricular unir 2nd sem (evaluations), ‘Curricular units 1st sem (grade)’, ‘Curricular units 2nd sem (enrolled)’, ‘Curricular units 2nd sem (approved)’, ‘Curricular units 2nd sem (grade)’.

Separación, creación y evaluación del modelo

Posteriomente para la realización del modelo se hizo uso de la librería sklearn para poder ejecutar el split del dataframe para tener 60% de la base de datos para el entrenamiento del modelo, 20% para el test y el 20% para la validación del modelo.

Como una primer corrida se decidió realizar el modelo con el split previamente mencionado y con los hiperparámetros que vienen por default con la función “LinearDiscriminantAnalysis” obtenida de la librería sklearn. Gracias a lo anterior se realizan las predicciones para el set de test, obteniendo como resultado la figura 1 donde se comprueba que se obtiene un accuracy de 0.88.

```
Precisión en el conjunto de test: 0.8801652892561983

Matriz de Confusión (test):
[[425  24]
 [ 63 214]]

Reporte de Clasificación (test):
              precision    recall  f1-score   support

           0       0.87      0.95      0.91       449
           1       0.90      0.77      0.83       277

    accuracy                           0.88       726
   macro avg       0.89      0.86      0.87       726
weighted avg       0.88      0.88      0.88       726
```

Figura 1. Accuracy, matríz de confusión y reporte de clasificación.

Para una segunda prueba se decidió realizar cross validation donde se obtuvieron los resultados que se muestran en la figura 2. Por otra parte para poder visualizar como se comportaron para los diferentes fragmentos del dataframe se realizó el histograma que se muestra en la figura 3 donde se puede ver las diferentes puntuaciones por cada uno de los *fold*. Para terminar de visualizar los resultados se muestra en la figura 4 un box-plot donde nos permite visualizar entre cuales valores se encuentran los scores obtenidos mediante el cross-validation.

```
Mean Score: 0.871723767944354
Standard Deviation Score: 0.015815373648864393
Test Set Score: 0.8801652892561983
```

Figura 2. Resultados de evaluación del modelo con cross-validation.

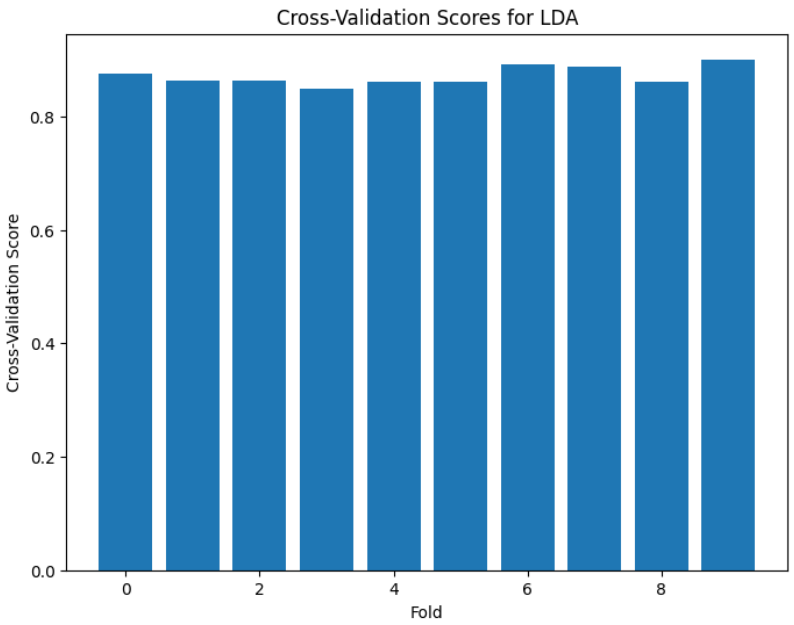

Figura 3. Histograma de scores obtenidos mediante cross-validation.

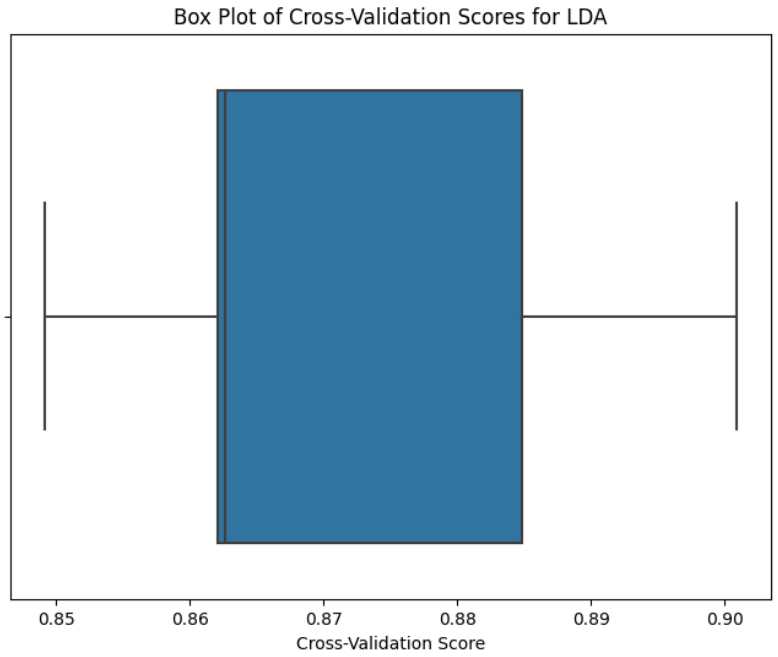

Figura 4. Box-plot scores mediante cross-validation.

Ahora como una tercera opción de evaluación del modelo que involucra al train y test es mediante el uso de cross validation y gridsearch. Para esto se manda el modelo y en el cual se trabajan con los tres diferentes parametros que puede tener

el modelo, con lo anterior es posible evaluar con cuales de los hiperparámetros tiene un mejor resultado en el transcurso de 5 folds para el cross validation. Como resultado de lo anterior, se obtiene que los mejores hiperparámetros para el modelo son los que se muestran en la figura 5 donde a su vez se observa el accuracy para el dataset de train y test.

```
Mejores hiperparámetros: {'n_components': 1, 'solver': 'svd'}
Mejor puntuación en entrenamiento: 0.8712977382276603
Puntuación en prueba: 0.8801652892561983
```

Figura 5. Mejores hipeparámetros junto con sus evaluaciones para train y test.

Ahora que ya tenemos los mejores parámetros para el modelo en cuestión, se decide evaluar el modelo con el dataset de validation con el mejor entrenamiento obtenido el cual se muestra en la imagen previa (figura 5). Para comenzar a hacer las predicciones y evaluaciones se divide aleatoriamente el dataset en 7 diferentes partes con las cuales se obtienen las predicciones para la sección de validación. Posteriormente como se observa en la figura 6 no hay cambios significativos entre las validaciones lo que nos dice que la varianza del modelo es baja. Por útlimo, se grafican los errores cuadrárticos comparativos para observar como funcionan para cada una de las particiones del validation dataset lo cual se muestra en la figura 7.

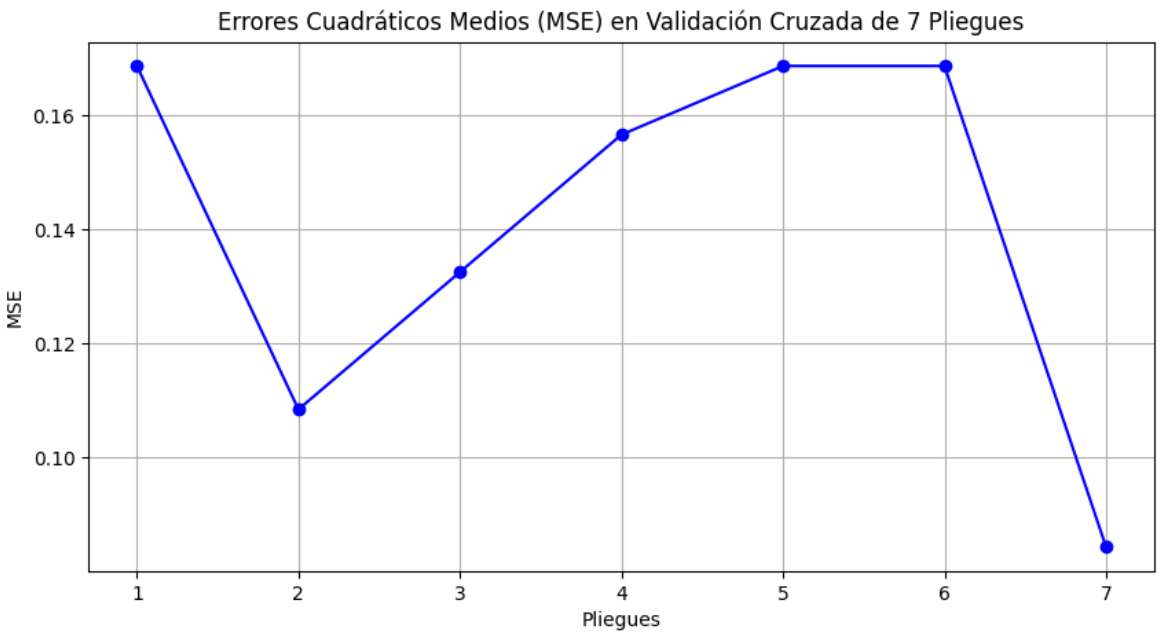

Figura 6. Errores cuadráticos durante la validación cruzada de 7 pliegues.

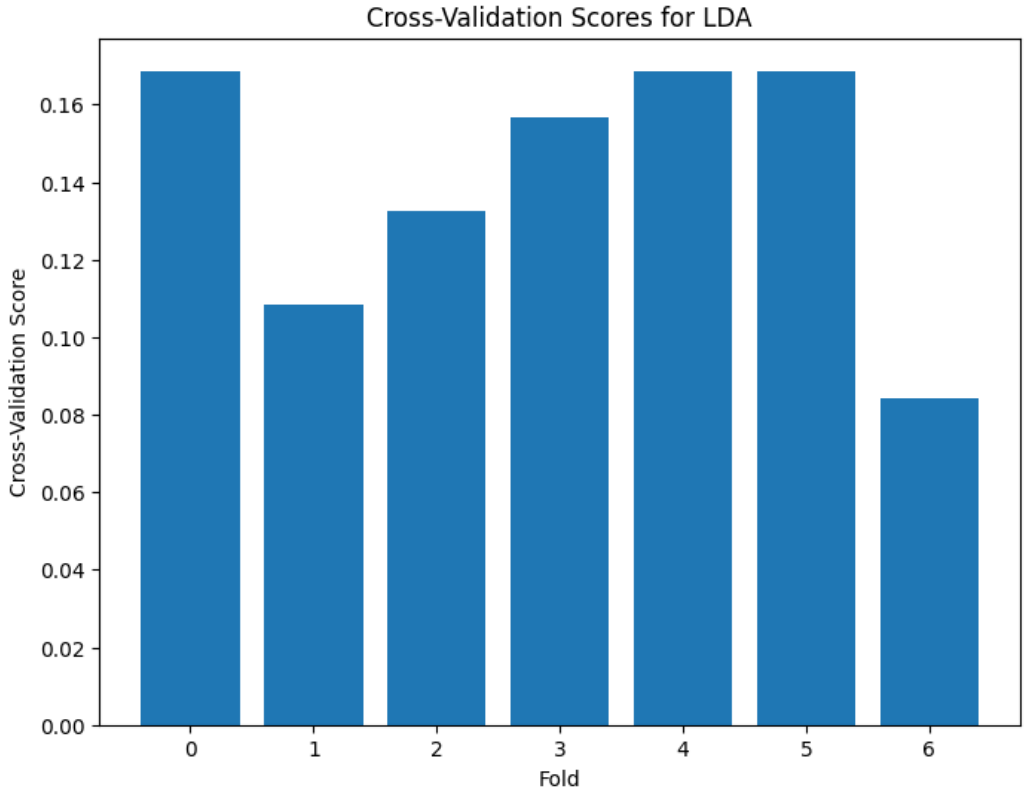

Figura 7. Histogramas para los MSE de cada partición del dataset de validación.

Para terminar con el análisis del modelo se recapitulará las mediciones antes obtenidas del modelo para de esta manera establecer el modelo como bueno o malo.

Para conocer el grado de bias se compara las métricas de evaluación entre en conjunto de test y validación obtienendo como resultado que no hay un cambio significativo en las métricas por lo cual el grado de sesgo es bajo. Posteriormente analizando el grado de varianza se obtiene que es estable por lo que se tiene un bajo nivel de varianza, lo cual nos permite comenzar a concluir que el ajuste del modelo es el adecuado para el dataset en cuestión. Por lo que si englobamos todo el análisis, se puede decir que el nivel de ajuste es adecuado debido a que el rendimiento del modelo es similar en los conjuntos establecidos (entrenamiento y validación) por lo que se tiene un buen ajuste en el modelo.

Anexos:

Link repositorio Github: https://github.com/gabycor/modelo_framework_lda.git

Link colab: https://colab.research.google.com/drive/1pDcLSRxFEYiGeHzkIH3DUTn2L3ubRFPB?usp=sharing