ENTREGA FINAL DE PROYECTO

Por:

Claudia Yaneth Giraldo Vergara

Introducción a la inteligencia artificial

Profesor:

Raúl Ramos Pollán

UNIVERSIDAD DE ANTIOQUIA

FACULTAD DE INGENIERÍA MEDELLÍN

2023

Contenido

PROYECTO: PREDICCIÓN DE VISAS

1. Introducción

El presente informe ejecutivo resume los avances y resultados obtenidos durante el proyecto de predicción de visas laborales. Se proporciona una descripción detallada de las diferentes etapas del proyecto, desde la exploración descriptiva del dataset hasta las iteraciones de desarrollo de modelos supervisados. Además, se presentan los retos y consideraciones de despliegue identificados durante el proceso y se concluye con las principales observaciones y recomendaciones.

2. Exploración de datos

Vistas de los casos de forma general

En primer lugar, se observó cómo se distribuyen las decisiones finales a las solicitudes que hacen las personas al gobierno de Estados Unidos para obtener una visa laboral.

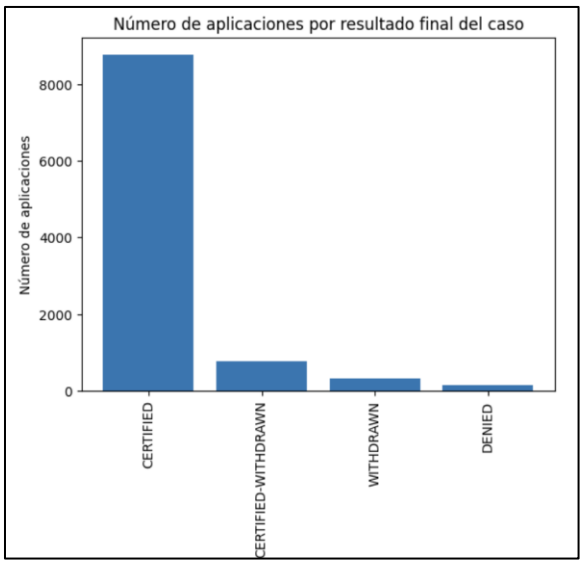

Figura 1. Número de aplicaciones por resultado final de caso

La mayoría de las solicitudes se encuentran en estado CERTIFICADO (con 8776), le siguen CERTIFICADO-RETIRADO (con 767), RETIRADO (con 314) y DENEGADO (con 143).

De los 10 mil datos, solo un 1.43% de los casos totales son denegados y más de un 87% son certificados, eso significa que hay un gran porcentaje de aceptación para este tipo de visa.

Vista de los casos por distribución en el tiempo

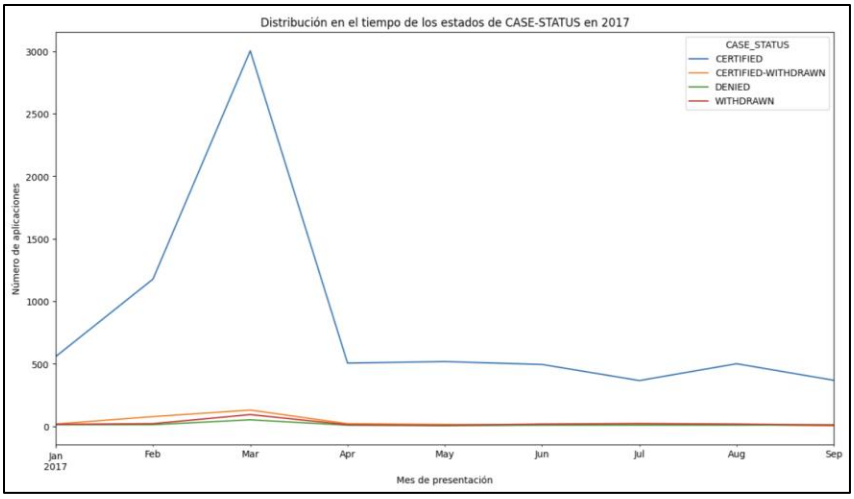

Figura 2. Distribución en el tiempo de los estados de CASE-STATUS

Se observó un aumento considerable de las aplicaciones entre los meses febrero y marzo, la gran mayoría de estas solicitudes obtienen el valor de 'CERTIFICADO'.

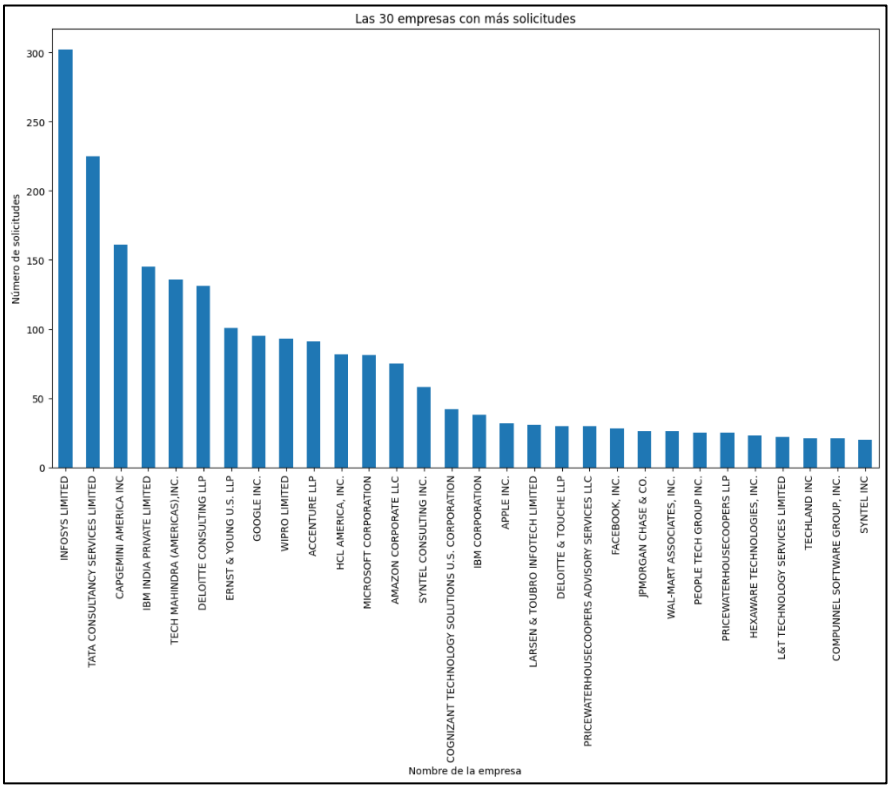

Figura 3. Las 30 empresas con más solicitudes

Observando este gráfico se encuentra algo interesante, y es que 4 de las 5 empresas con más solicitudes de visa laboral son empresas de India (Infosys, Tata Consultancy, IBM India y Tech Mahindra).

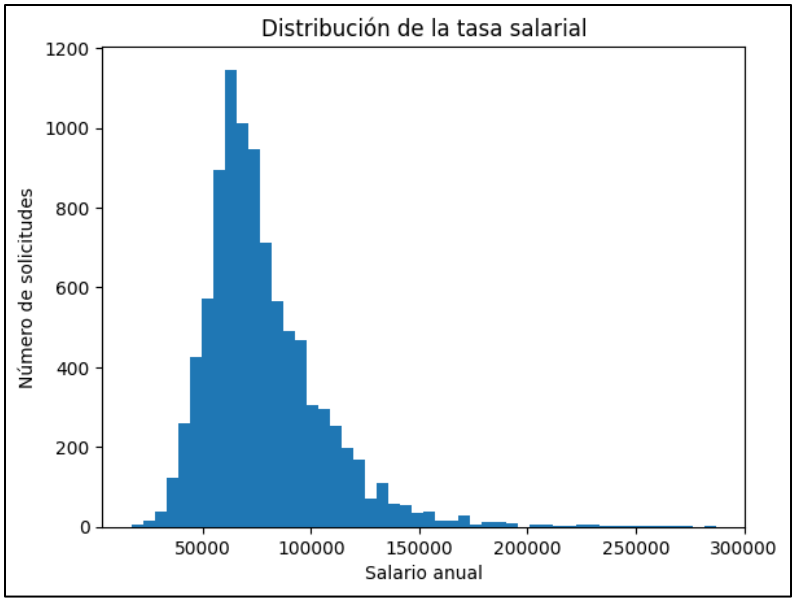

Figura 4. Distribución de la tasa salarial

La distribución de la tasa salarial sigue una distribución normal, con una media ligeramente por encima de los 50 mil dólares anuales. Según la Oficina de Estadísticas Laborales (BLS, por sus siglas en inglés), el salario medio de los trabajadores en los Estados Unidos es de 54 mil dólares anuales por lo que se puede asegurar que este promedio aplica también para los extranjeros.

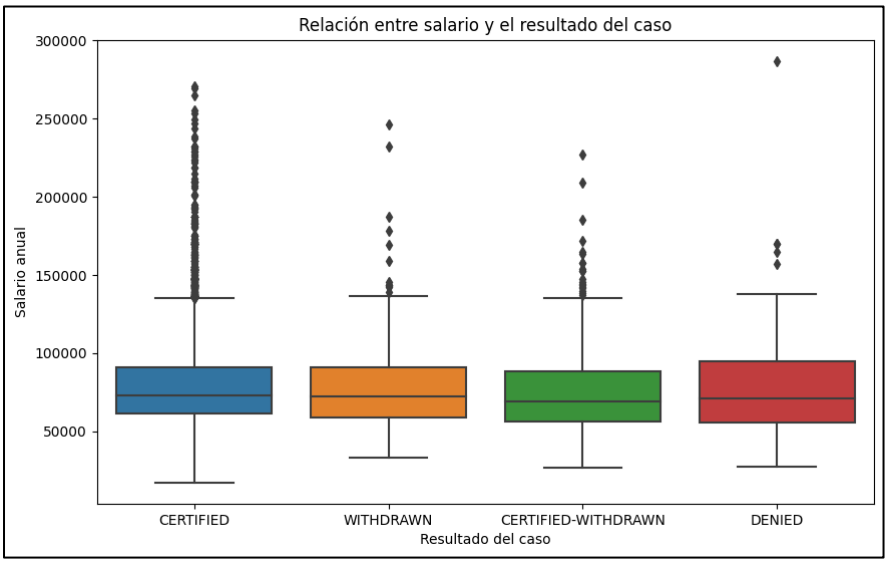

Figura 5. Relación entre salario y el resultado del caso

Mediante este boxplot se puede confirmar que el salario influye en el resultado del caso, se observa que, aunque para los 4 posibles resultados la media ronda entre los 50 mil y 100 mil dólares, los valores atípicos

del resultado 'CERTITIFICADO' van dese 140 mil (aproximadamente) hasta más de 250 mil dólares anuales, es decir, entre más alto sea el salario, más probable es conseguir una visa laboral en Estados Unidos.

3. Iteraciones de desarrollo

En esta sección se presentan las diferentes iteraciones de desarrollo del proyecto, incluyendo el preprocesamiento de datos, los modelos supervisados utilizados, así como los resultados, métricas y curvas de aprendizaje obtenidas en cada iteración.

a. Preprocesado de datos

Se convierten los valores o etiquetas de forma numérica para que sean legibles para el algoritmo. La variable 'CASE_STATUS' se convierte en una variable binaria, se le asignan valores de 1 para indicar que la visa fue aceptada, y 0 para indicar que la visa fue denegada.

	CASE_STATUS	AGENT_REPRESENTING_EMPLOYER	FULL_TIME_POSITION	PW_WAGE_LEVEL	H1B_DEPENDENT	WILLFUL_VIOLATOR	SOC_N	EMP_N
0	0	0	1	1	0	0	11	12
1	0	1	1	1	0	0	11	12
2	0	1	1	3	1	0	11	14
3	0	1	1	3	0	0	11	12
4	0	0	1	2	1	0	11	10

Figura 6. Visualización del dataset después de la categorización de las variables

Desbalance de datos

Se encontró un desbalance de los datos de la variable a predecir ('CASE_STATUS'), como se muestra en el siguiente gráfico:

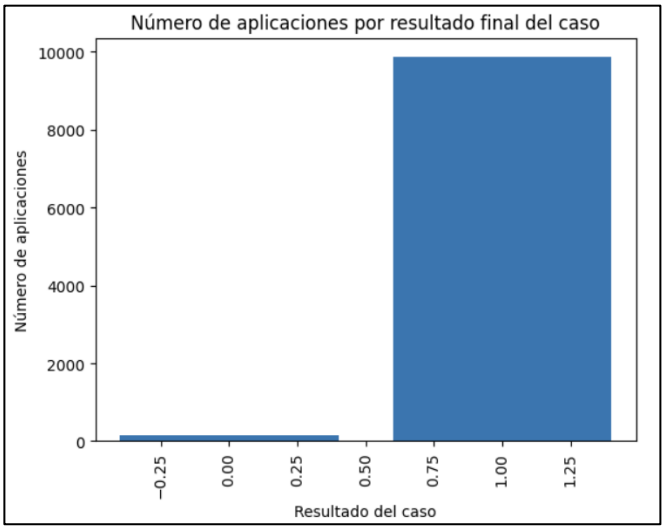

Figura 7. Desbalance del número de aplicaciones por resultado final del caso

Se tienen aproximadamente 10000 datos con resultado ‘1’, es decir, casos donde la solicitud de la visa fue aceptada, en cambio se tienen aproximadamente 200 datos con resultado ‘0’, casos donde la solicitud fue denegada.

Para balancear los datos se opta por realizar una reducción de muestreo de la clase mayoritaria, a pesar de perder un gran volumen de datos, se logra el balance que desea:

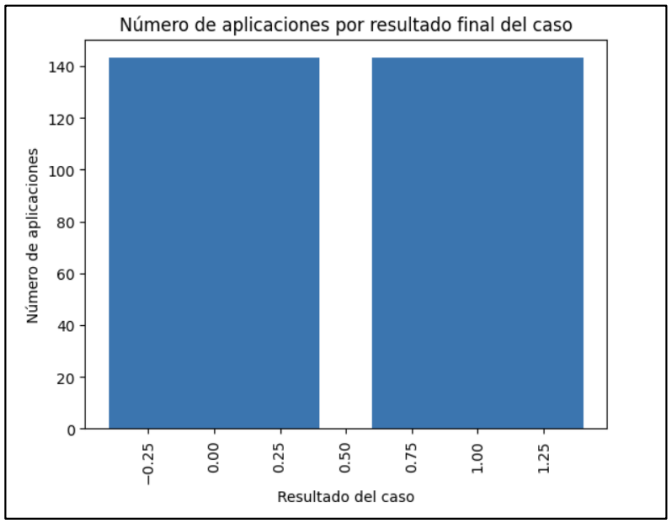

Figura 8. Número de aplicaciones después de la reducción de muestreo

b. Resultados

Modelos supervisados

Se evaluaron los modelos utilizando diversas métricas, como precisión, recall y puntuación F1, y se generaron curvas de aprendizaje para analizar el desempeño de los modelos en función del tamaño del conjunto de datos de entrenamiento.

Métrica

La métrica escogida para medir el desempeño de los modelos es el *‘accuracy’,* obtenido a partir de la matriz de confusión.

A continuación, se presenta los resultados obtenidos con distintos modelos supervisados, se observa que el modelo con Random Forest tiene el mejor accuracy con un 0.69, mientras que el modelo con Gaussian Naive Bayes tiene el peor accuracy con un 0.48.

Regresión logística

	precision	recall	f1-score	support
0	0.60	0.70	0.65	30
1	0.61	0.50	0.55	28
accuracy			0.60	58
macro avg	0.60	0.60	0.60	58
weighted avg	0.60	0.60	0.60	58

Árboles de decisión

	precision	recall	f1-score	support
0	0.63	0.80	0.71	30
1	0.70	0.50	0.58	28
accuracy			0.66	58
macro avg	0.67	0.65	0.64	58
weighted avg	0.66	0.66	0.65	58

Random forest

	precision	recall	f1-score	support
0	0.67	0.80	0.73	30
1	0.73	0.57	0.64	28
accuracy			0.69	58
macro avg	0.70	0.69	0.68	58
weighted avg	0.70	0.69	0.69	58

Gaussian Naive Bayes

	precision	recall	f1-score	support
0	0.50	0.03	0.06	30
1	0.48	0.96	0.64	28
accuracy			0.48	58
macro avg	0.49	0.50	0.35	58
weighted avg	0.49	0.48	0.34	58

Redes neuronales artificiales

	precision	recall	f1-score	support
0	0.64	0.70	0.67	30
1	0.64	0.57	0.60	28
accuracy			0.64	58
macro avg	0.64	0.64	0.64	58
weighted avg	0.64	0.64	0.64	58

Figura 9. Matrices de confusión de modelos supervisados

Curvas de aprendizaje

Regresión logística

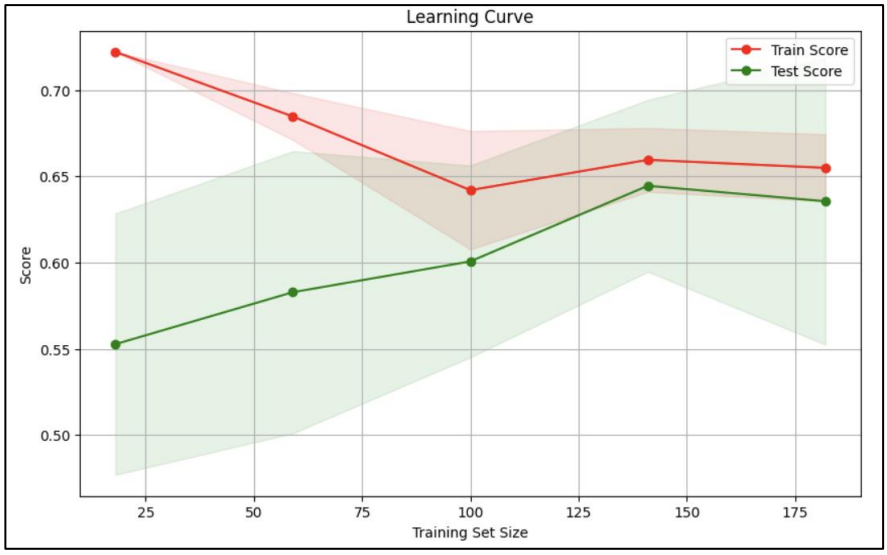

Figura 10. Curva de aprendizaje para regresión logística

Para esta primera curva de aprendizaje se observa un sesgo, esto se puede deber a la simplicidad del modelo.

Árbol de decisión

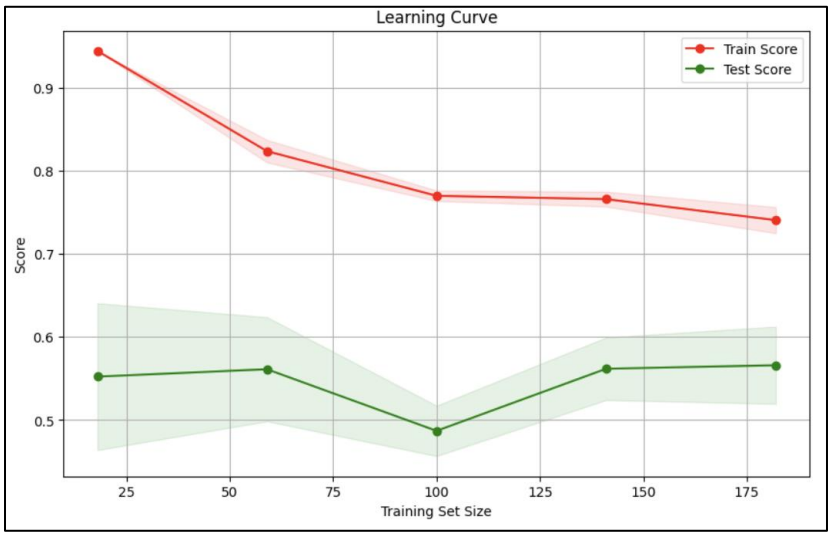

Figura 11. Curva de aprendizaje para árbol de decisión

Se observa un overfitting, el modelo tiene buen rendimiento con los datos de entrenamiento, pero no funciona bien con los datos de prueba.

Random Forest

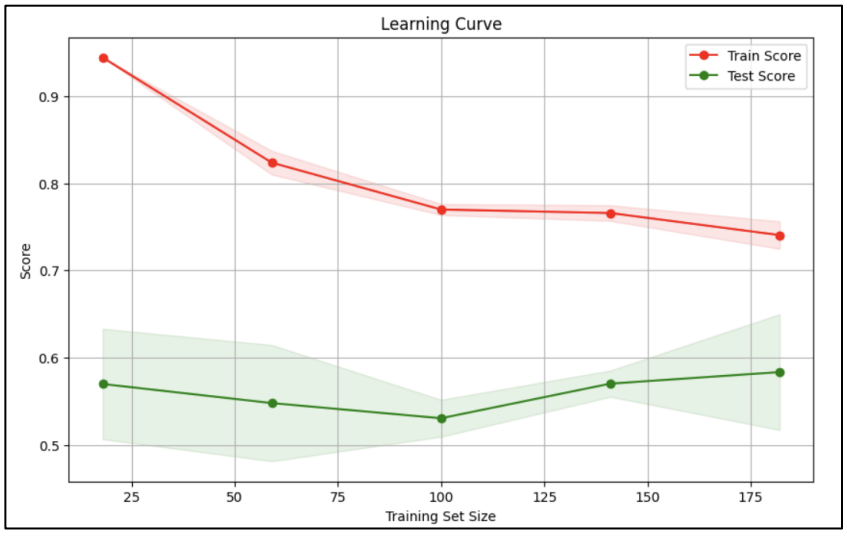

Figura 12. Curva de aprendizaje para Random Forest

El mismo caso que el anterior, se observa un overfitting.

Naives Bayes

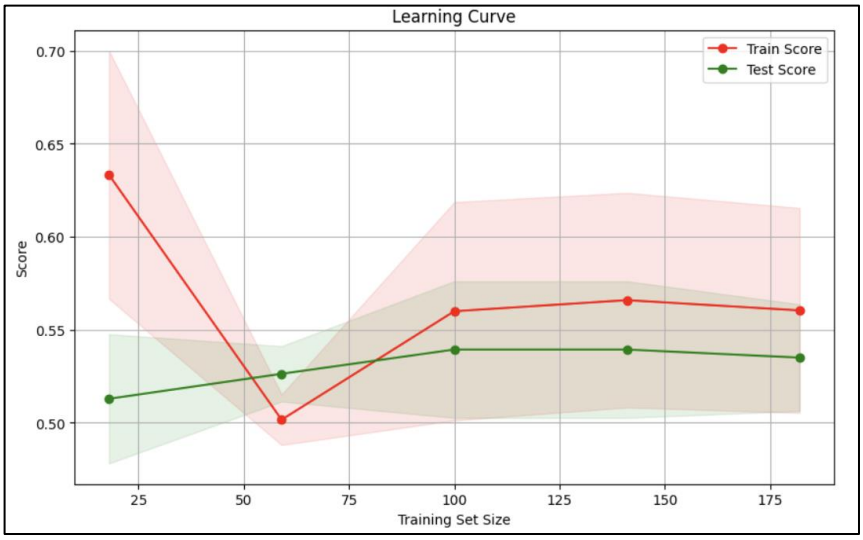

Figura 13. Curva de aprendizaje para Naive Bayes

Esta curva de aprendizaje muestra un buen rendimiento del modelo de Naive Bayes, tanto la línea de train como la de prueba están en un bajo nivel y se estabilizan al tener más de 50 datos.

Redes Neuronales Artificiales

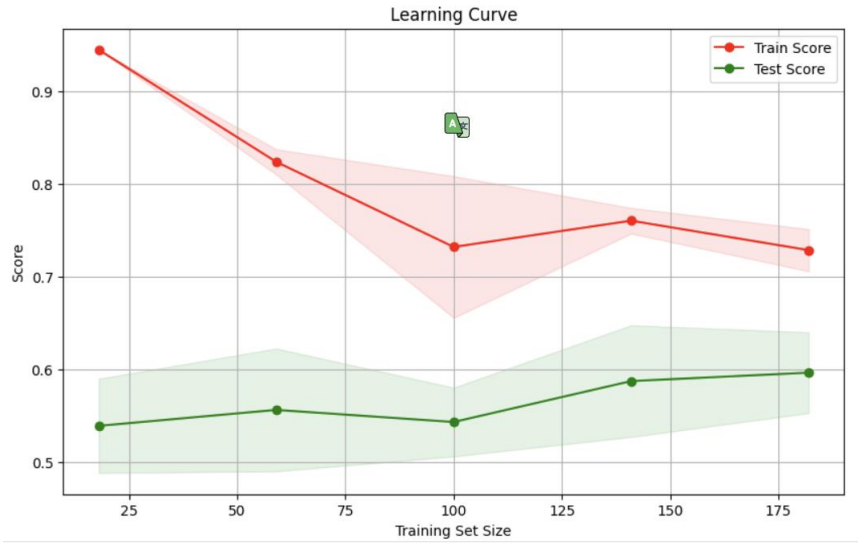

Figura 14. Curva de aprendizaje para Redes Neuronales Artificiales

Mismo caso que las curvas de aprendizaje del árbol de decisión y random forest, se observa un overfitting por lo que no se considera un buen modelo para este caso.

4. Retos y consideraciones de despliegue

Enfrentar el desbalance de datos representó un desafío significativo durante el desarrollo de este proyecto. La disparidad en la distribución de los datos de la variable objetivo ‘CASE_STATUS’ dificultó la generación de modelos de análisis precisos y confiables. Para superar este obstáculo, se implementaron estrategias especializadas, como la reducción de muestreo de la clase mayoritaria. Estas acciones fueron fundamentales para lograr un conjunto de datos más equilibrado y garantizar que el modelo resultante fuera capaz de realizar predicciones más precisas y generalizadas en todas las categorías.

5. Conclusiones

- Se recomienda realizar un análisis exhaustivo de las variables utilizadas en los modelos entrenados, con el objetivo de identificar aquellas que tienen un mayor impacto en los resultados. Este análisis permitirá reducir el error y mejorar la precisión del modelo al enfocar los esfuerzos en las variables más relevantes.

- Es aconsejable explorar técnicas de manejo de desbalance de datos, como el submuestreo o el sobremuestreo, con el fin de abordar este desafío específico. Estas técnicas pueden equilibrar la distribución de clases en el conjunto de datos y mejorar la capacidad predictiva del modelo. Además, se recomienda evaluar el uso de técnicas de generación sintética de datos, como el SMOTE (Synthetic Minority Over-sampling Technique), para aumentar la representación de las clases minoritarias.