Análisis y Reporte sobre el desempeño del modelo

Diego Emilio Barrera Hernández

A01366802

Introducción:

Se realizó un modelo de regresión logística para analizar los datos recibidos de pasas, y con estos clasificar las pasas según su Clase/Tipo Kecimen o Besni.

- Area
- Perimeter
- MajorAxisLength
- MinorAxisLength
- Eccentricity
- ConvexArea
- Extent

Para más información sobre los datos, se puede consultar directamente la página de referencia dónde se obtuvieron dicho Dataset, en la siguiente URL:

https://archive.ics.uci.edu/dataset/850/raisin

Método empleado:

El modelo de regresión logística empleado para este análisis fue LogisticRegression de la librería Sklearn, usando como hyperparameters:

- Penalty = '11'
- Solver = 'saga'

Al momento de correr nuestro modelo, separamos el dataset en TRAIN y TEST, con la función *train_test_split* de la librería *sklearn.model_selection*, con esta acción se obtuvo un accuracy de ***0.7067*** y un MSE de ***0.2933*** en nuestros datos de train y un accuracy de ***0.6978*** y un MSE de ***0.3022*** en los datos de test.

Se hicieron dos representaciones gráficas de nuestros resultados. En la Fig 1, se pueden ver los resultados obtenidos en la correlación de los datos.

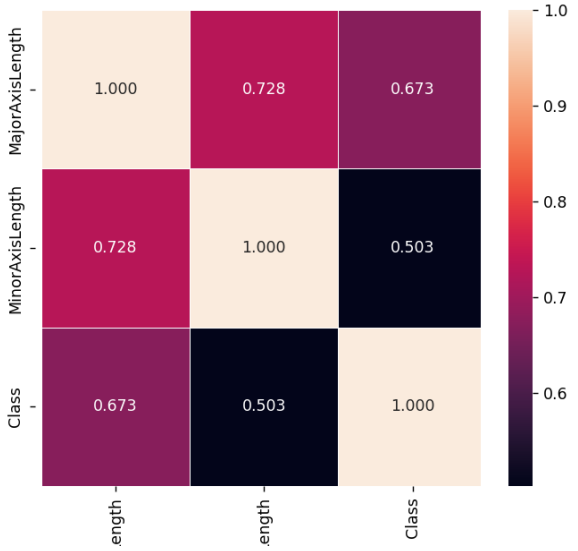

Fig 1: Matriz de correlación de los datos

En esta matriz podemos observar que entre la variable MajorAxisLength y la variable MinorAxisLength existe una correlación de 0.728 lo que significa que existe una correlación positiva altamente moderada entre estas dos variables, pues si una llega a aumentar la otra lo hará también. Por otro lado, tenemos la correlación de MinorAxisLength con Class, esta se encuentra en un 0.503 lo que sugiere que hay una relación positiva moderada, pero no es una relación tan fuerte con la anterior; aún existe una tendencia a que cuando una variable aumenta, la otra también lo hará. Por último, tenemos la correlación de

MajorAxisLength con Class, esta se encuentra en un 0.673, lo que indica correlación positiva, igualmente, moderadamente fuerte, es decir, una relación bastante sólida.

Ahora presentamos la Fig 2, aquí se puede observar los resultados visuales de la predicción de nuestro modelo.

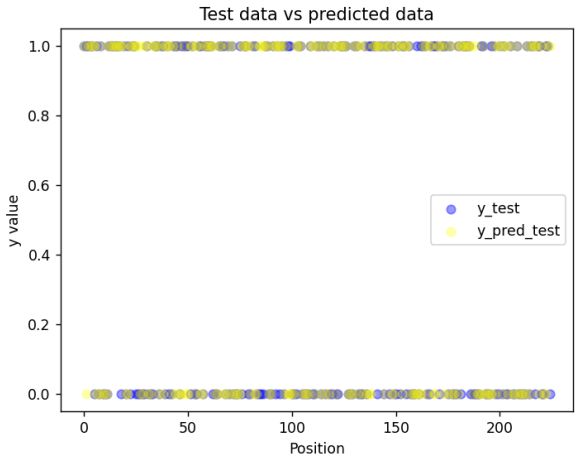

Fig 2: Resultado del modelo 1

En esta gráfica se observar la separación de los datos reales (y_test) en azul y los resultados predichos (y_pred_test) en amarillo. Se puede entender que hay varios puntos amarillos y azules solitarios, esto se debe a que los datos predichos no fueron exactos y no son iguales a los datos reales, por otro lado, se tienen puntos de un color verde, esto nos sugiere que los datos predichos fueron exactos o muy cercanos los datos reales, pues sería el encuentro de ambos datos que nos muestra este resultado de combinación de colores.

Con esta información, podemos concluir que tenemos un **accuracy** de 70% test y 71% en train lo que nos sugiere que nuestro **bias** es ALTO tomando en cuenta nuestro MSE. Por otro lado, la **varianza** es BAJO, debido a que los resultados del modelo se mantienen en el mismo rango de respuesta. Finalmente, podemos decir que nuestro ajuste del modelo es “UNDERFITTING”.

Mejora del modelo:

Al igual que el primer modelo, se utilizó LogisticRegression de la librería Sklearn, solo se cambiaron los hyperparameters default de la librería, es decir, no se especificaron.

Igualmente se separó el dataset en TRAIN y TEST, con la función *train_test_split* de la librería *sklearn.model_selection*, con esta acción se obtuvo un accuracy de ***0.8518*** y un MSE de ***0.1481*** en nuestros datos de train y un accuracy de ***0.8844*** y un MSE de ***0.1155*** en los datos de test.

En la Fig 3, aquí se puede observar los resultados visuales de la predicción de este segundo modelo.

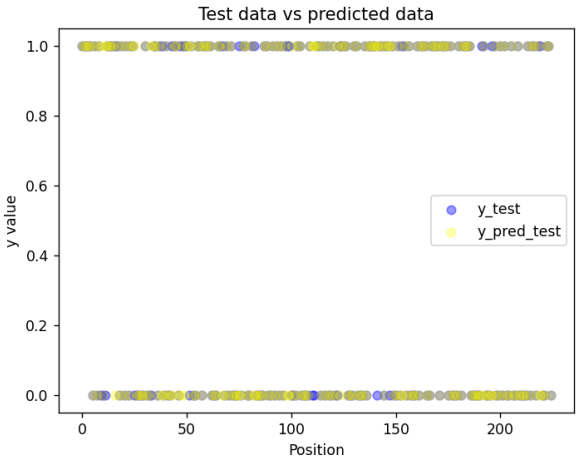

Fig 3: Resultados del modelo

Podemos observar, que, en este caso, existen más puntos de color verdoso que nos confirman mayor exactitud en el modelo.

Con esta información, podemos concluir que tenemos un **accuracy** de 88% test y 85% en train lo que nos sugiere que nuestro **bias** es bajo tomando en cuenta nuestro MSE. Por otro lado, la **varianza** es baja, debido a que los resultados del modelo se mantienen en el mismo rango de respuesta. Finalmente, podemos decir que nuestro ajuste del modelo es “FITT”.

Conclusión

Comparando amos modelos, podemos ver una mejora de **125.71%** en accuracy y una disminución den el error de 0.3022 a 0.1155, por lo que contamos con un mejor modelo, el cual, aunque no llega a tener una precisión de más de 90% ya genera una mejor predicción.