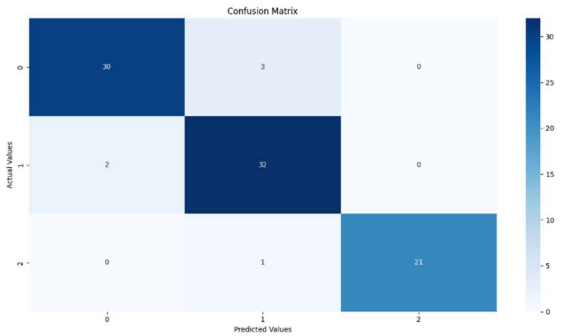
Inteligencia artificial avanzada para la ciencia de datos

Gpo 102

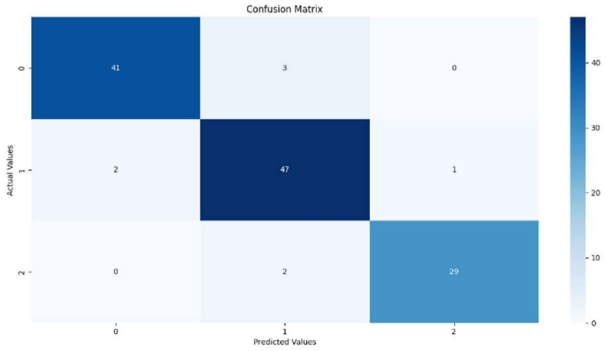
Análisis y Reporte sobre el desempeño del modelo

Héctor Francisco Marin Garrido – A00827714

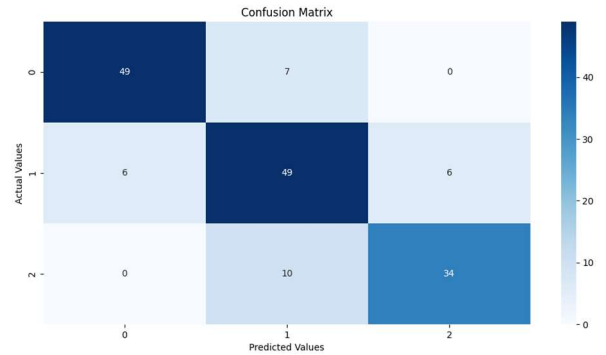
Para este reporte se hizo uso de un modelo de regresión logística para poder hacer predicciones en una base de datos, en este caso se trato de obtener la clase de un vino basándonos en las variables ubicadas dentro de el archivo wine.data que se nos fue proporcionado, en este se encuentran varias variables las cuales podemos utilizar para hacer una predicción acerca de cual de las 3 clases de vino será cada uno en base a estos datos. Primero se tuvo que subir el archivo con los datos, este se subió al repositorio de github para que el código se pueda correr sin necesidad de descargar el archivo ni de modificarlo para poder llegar a los resultados analizados. Para este reporte se analizarán los datos obtenidos mediante el uso de un método con framework basándome en la documentación de scikit-learn. La primera vez que se realizó la simulación se hizo uso de un 50% de datos para entrenar y de el resto para hacer las pruebas. Se obtuvo uno precisión de 98.87% para los datos de entrenamiento y una de 93.25% para los datos de prueba por lo que se puede concluir que hubo un buen fitting aunque también podríamos decir que es posible que se haya presentado un overfitting del modelo, esto se puede deber en parte a que para empezar no hay muchos datos dentro de la base de datos que se analizó. Podemos observar aquí la matriz de los valores reales contra los predichos para e caso en el que se hizo uso un 50% de datos para entrenar.



Podemos observar que hubo muy pocos errores por lo que opte por utilizar aun menos valores para el entrenamiento del modelo haciendo en este caso uso del 70% de los datos para hacer pruebas y solo el 30% para entrenar el modelo. Para este caso se obtuvo una precisión del 100% en los datos de entrenamiento y 93.6% en los datos de prueba por lo que parece ser que la precisión aumento al reducir los datos de entrenamiento por lo que por ahora considero que este ha sido el mejor modelo ya que reducimos la cantidad de datos utilizados para entrenar y tenemos mas espacio para predecir datos.

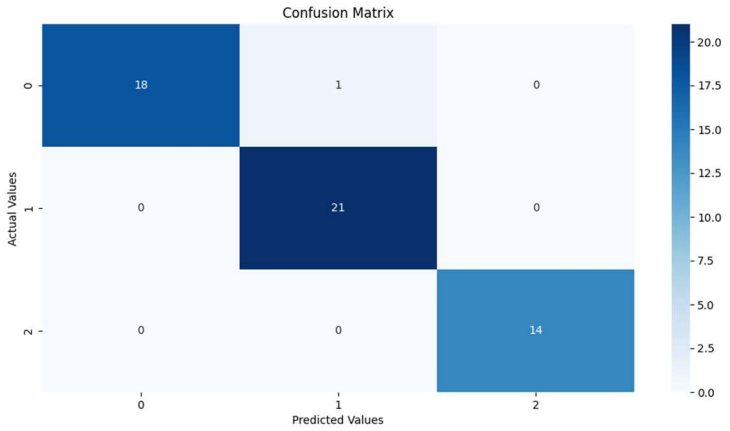


Podemos observar aquí que aumento un poco la precisión en los datos de prueba lo que puede ser atribuido al azar ya que se incremento la cantidad de datos utilizados para hacer la prueba, pero aun así nos indica que este es un buen modelo y llega a hacer las predicciones de una manera rápida y precisa. Después de esto decidí reducir aun mas la cantidad de datos para ver hasta qué punto podría bajar esta cantidad y se empiece a mostrar evidencia de underfitting con las predicciones. Opte por hacer uso del 85% de los datos para hacer pruebas y del 15% para entrenar. Con estos valores se llego a un resultado de 100% de precisión en los datos de entrenamiento y de un 90.13% en los datos de prueba, aquí podemos empezar a ver que aumenta la cantidad de errores obtenidos por lo que pareciera que nos acercamos cada vez mas a un modelo con underfitting pero sigo considerando que los resultados son bastante aceptables. Finalmente, para ver underfitting decidí reducir aun mas la cantidad de datos útilizados para entrenar a un 10% y 90% para probar. Con esto obtuvimos un valor de precisión de 100% para los valores de entrenamiento y una precisión de 81.98% en los datos de prueba por lo que podemos observar que este modelo es peor que los demás y nos muestra un caso de underfitting por lo que este modelo no es el ideal para hacer predicciones a futuro. La matriz muestra los siguientes datos.



Para concluir realice un ultimo modelo tratando de llegar a la mayor cantidad de precisión posible con la menor cantidad de datos posibles para tratar de llegar a un buen punto medio entre under y over fitting,

la mejor distribución a la que llegue fue la de hacer un uso del 70% de los datos para entrenar y el 30% de los datos para hacer pruebas. Con estas proporciones llegamos a un porcentaje de precisión del 95.96% mientras que en los datos de prueba llegamos a una precisión del 98.14% por lo que considero a este el mejor modelo para la predicción de la clase de vino presentado, por lo menos con los datos proporcionados.



Para concluir cabe recalcar que la razón por la cual se obtuvieron estos datos se puede deber a que se tiene acceso a una cantidad pequeña de datos para empezar lo cual puede afectar a nuestros resultados y favorecer y desfavorecer a ciertas proporciones que se utilizaron a lo largo de este análisis con aprendizaje de máquina.