Momento de Retroalimentación: Módulo 2 Análisis y Reporte sobre el desempeño del modelo.

Adrian Bravo López A01752067

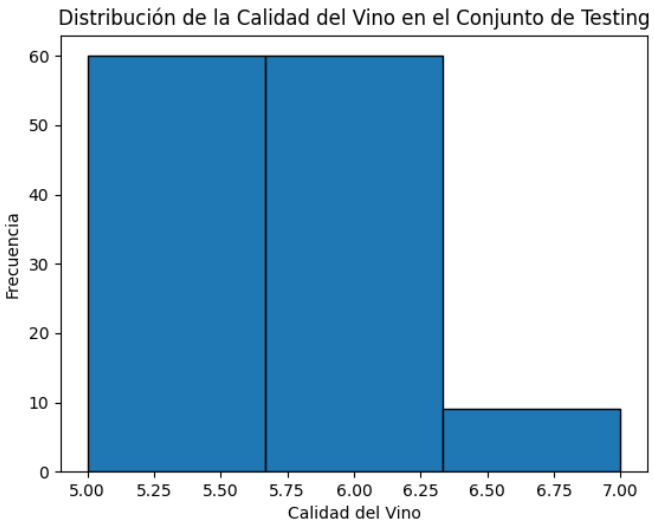
Análisis de implementación de Árbol de Decisión

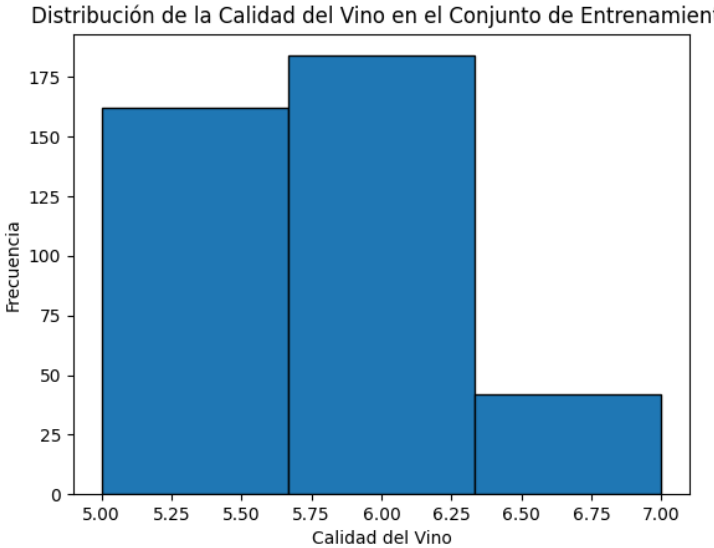
Para realizar este análisis, decidí utilizar la implementación de Arbol de Decision de sklearn. El motivo de este análisis es para evaluar el desempeño de esta particular implementación, tomando en cuenta los siguientes puntos.

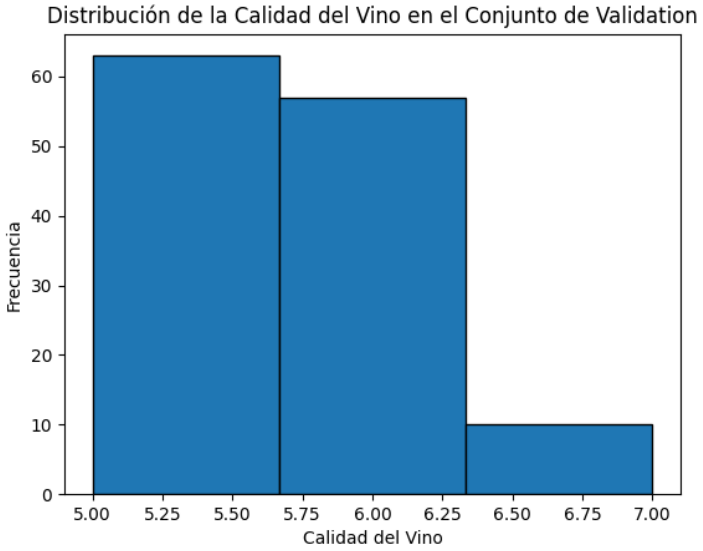
1. Decidí utilizar el dataset de calidad de vino por lo bien que se ajustaba al modelo de predicción que elegí (Árboles). El dataset es estrictamente numérico y no cuenta con valores categóricos, lo cual lo hace ideal para clasificación.

2. Separación y evaluación del modelo con un conjunto de prueba y un conjunto de validación (Train/Test/Validation).

Para el tratado del dataset, se realizaron una serie de pasos antes de dividirlo. Primero se identificaron y eliminaron la mayoría de los outliers, esto con el fin de mejorar el desempeño del modelo ya que los outliers pueden afectar severamente las capacidades de predicción del modelo. Luego se dividió el dataset en uno de Entrenamiento el cual conforma el 80% del dataset y otro de Test/Validación el cual conforma el restante 20%. Se hizo de esta forma debido a que el dataset con el que se cuenta después del recorte de datos es relativamente pequeño y no conforma más de 650 valores en total. El split se realizó usando la librería de sklearn, la cual cuenta con una función que hace la división de manera aleatoria o de manera semi aleatoria si se le proporciona una seed (random_state).

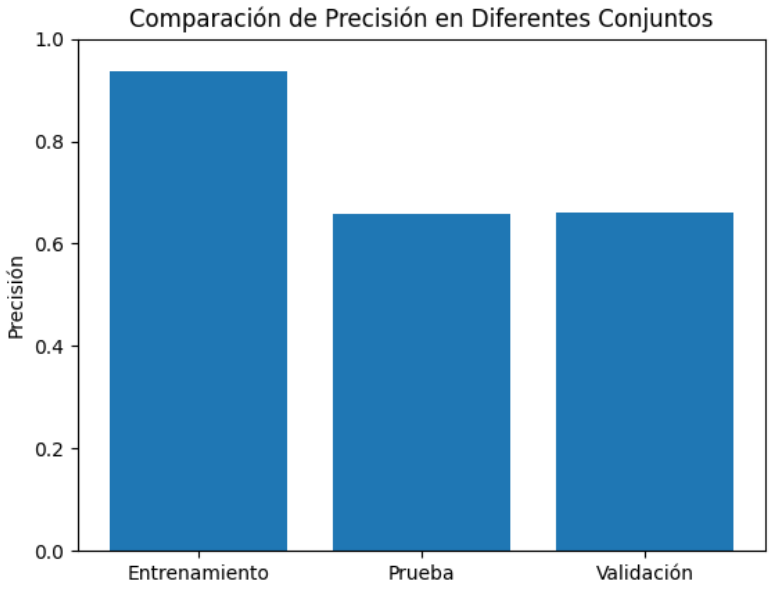






3. Diagnóstico y explicación el grado de bias o sesgo: bajo medio alto

Para revisar el grado de bias se utilizó la siguiente gráfica:

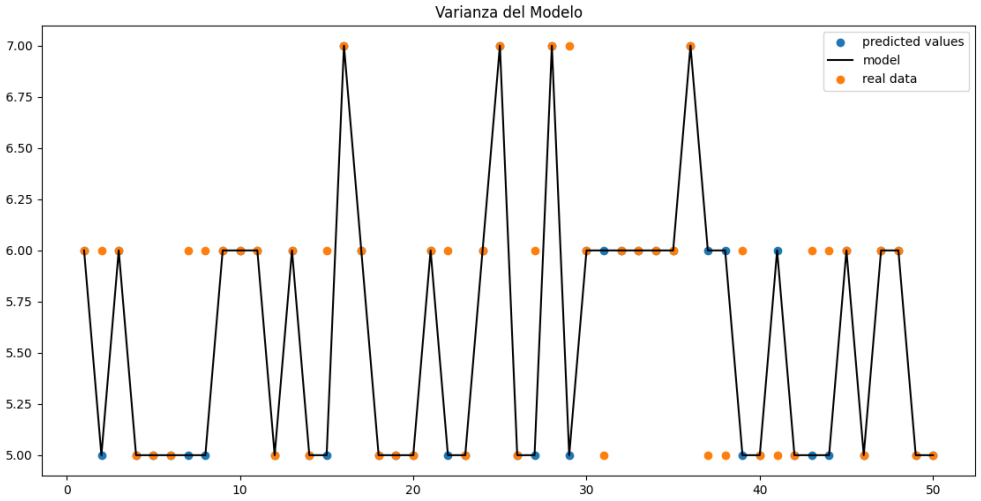


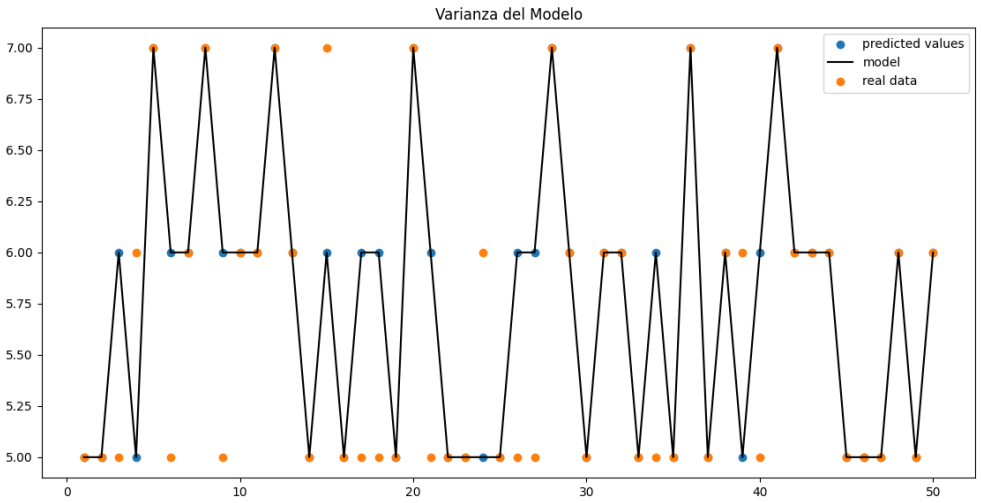
Esta gráfica ilustra la precisión con la que el modelo predice los 3 datasets. Como se puede observar la precisión con la que el modelo predice el set de entrenamiento es mucho mayor que el de los otros dos sets, aproximadamente de 88% para el set de entrenamiento y 61% para el los de prueba y validación. Esta diferencia deja

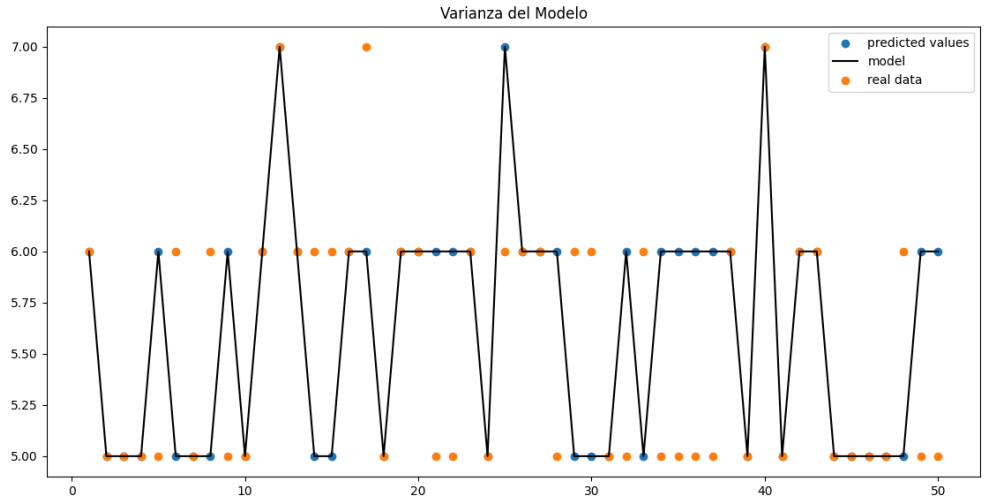
claro que el sesgo del modelo es alto y es una posible señal de que el modelo este siendo overfit.

4. Diagnóstico y explicación del grado de varianza

Para determinar la varianza del modelo se realizó pruebas con 3 diferentes distribuciones de datos de entrenamiento y de prueba.

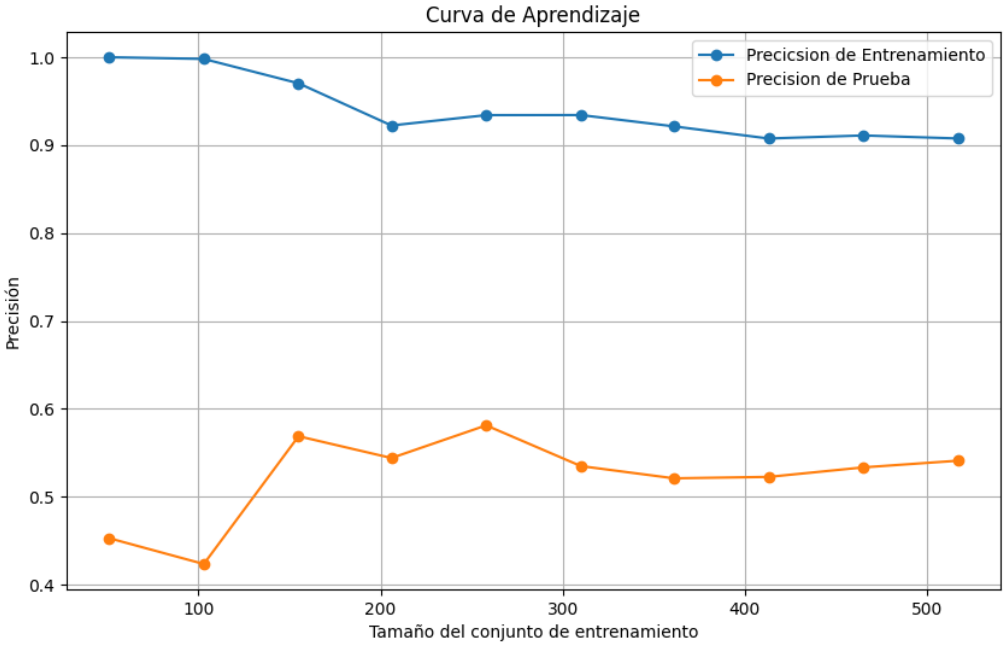






En estas pruebas se tomaron los primeros 50 datos del dataset de prueba y se probaron predecir usando el modelo entrenado. En las 3 gráficas se puede apreciar una comparación entre los valores reales de los datos de prueba contra los predichos por el modelo. Usando estas gráficas y un cálculo de la precisión del modelo, podemos llegar a una conclusión, la cual en este caso puede ser no muy favorable. Dada la diferencia de cantidad de datos predichos correctamente entre las 3 pruebas, en la primera se predijeron el 68% de los datos correctamente y en la segunda se respondieron 76% correctamente lo cual es de por sí una diferencia algo significativa. Finalmente en la última la precisión baja hasta 58%, con lo que podemos concluir que el modelo tiene una alta varianza.

5. Diagnóstico y explicación el nivel de ajuste del modelo

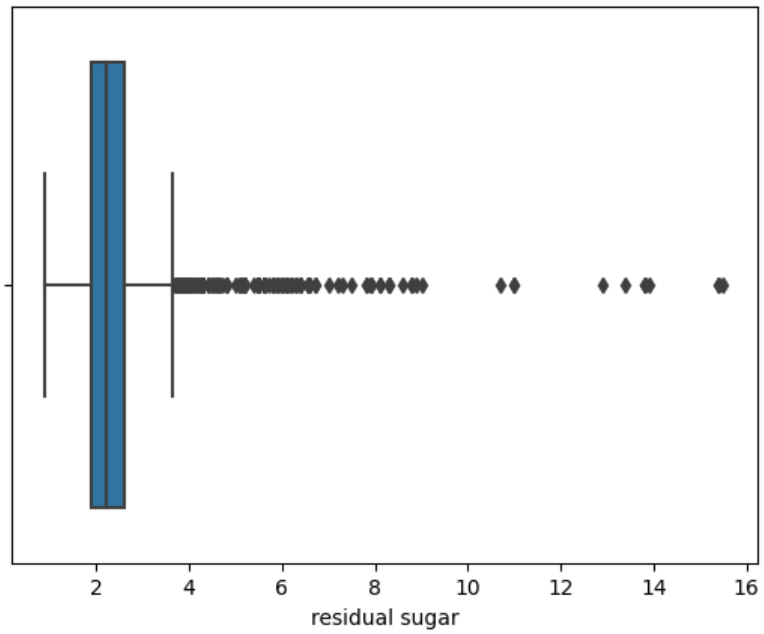


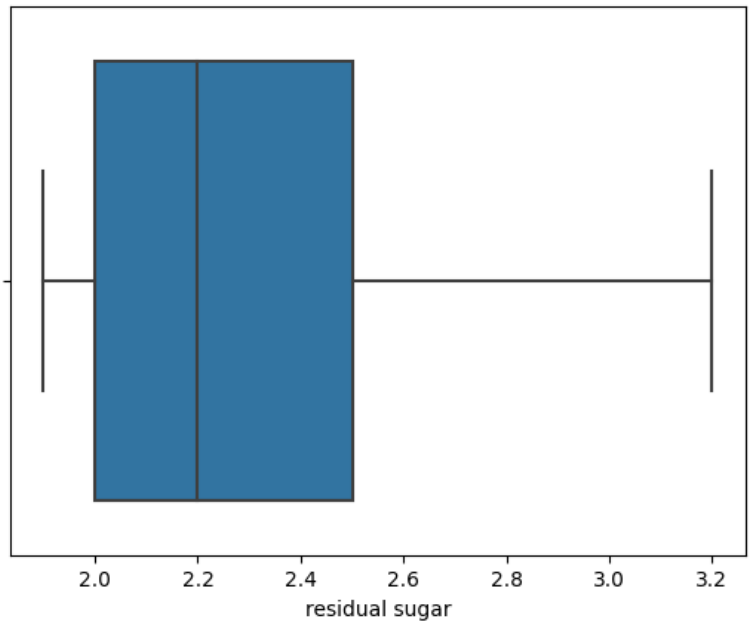
En cuanto a definir si el modelo fue overfit con los datos, esto se podría definir observando los resultados de las pruebas realizadas sobre el modelo anteriormente, de sesgo y varianza, de las cuales la conclusión en ambas es que era alta, lo cual llevaría a la conclusión de que el modelo fue overfit. Para ilustrarlo de otra manera, se puede graficar la curva de aprendizaje, como se ve en la gráfica el modelo es mucho más preciso cuando se trata de predecir los datos de aprendizaje, de hecho llega hasta ser 100% preciso con baja aplicación de datos, mientras que en predecir los datos de prueba con trabajo casi llega a 60% de precisión. Esta gráfica termina solidificando la conclusión de que el modelo fue overfit.

6. Técnicas de Mejoramiento

En cuanto a técnicas de mejoramiento, se aplicaron unas cuantas durante el transcurso del análisis. Lo primero que se hizo fue quitar los datos outliers del dataset, esto con el fin de que la precisión del modelo no fuera afectada por datos ruidosos al momento de hacer predicciones.

Estos son gráficos que ilustran los outliers antes y después de removerlos usando el Z-score.





En retrospectiva, una mejor alternativa hubiera sido cambiar los valores de los outliers por la media o la mediana de los datos con distribución normal para no perder tantos datos.

Después de eso se calculo la mejor profundidad para el árbol de decisión esto se hizo mediante un loop que cambiaba la profundidad hasta encontrar la que mejor precisión tuviera, en retrospectiva hubiera sido una buena idea también cambiar los random_states en cada iteración de la profundidad para mejor clasificar una profundad como la mejor.