Instituto Tecnológico y de Estudios Superiores de Monterrey

Uso de framework o biblioteca de aprendizaje máquina para implementación de una solución.

Israel Sánchez Miranda A01378705

Profesor Jorge Adolfo Ramirez Uresti

31 de agosto de 2022

1. **Sobre el modelo.**

   Para esta actividad se realizó un modelo de *Random Forest* para categorizar la clase de un tipo de vino basado en el dataset wine.csv, el cual se encuentra en el repositorio de Github.

   a. **Descripción del modelo.**

      El modelo lee los datos del archivo wine.csv para posteriormente separarlos en sets de prueba y sets de entrenamiento, la proporción prueba-entrenamiento es solicitada al usuario.

      Una vez separados los datos se crea el modelo de *Random Forest* con n árboles y j nodos hoja como límite. Ambos valores también son proporcionados por el usuario. Cabe destacar que se mantiene una semilla aleatoria constante para reducir la variación de las respuestas y tener un mejor análisis y entendimiento de cómo es que está operando el modelo basado en los hiper-parámetros proporcionados.

      Ya que se creó el modelo este es entrenado con los datos correspondientes para después realizar predicciones con el set de pruebas y así obtener el porcentaje de precisión.

      Posterior a esto, el usuario puede proporcionar valores para cada columna x dentro del data frame para que así el mismo modelo prediga qué clase de vino será el que el usuario ingresó.

      Finalmente, el programa muestra un árbol de decisión aleatorio que conforma el bosque para que el usuario pueda apreciar cómo es que el bosque opera.

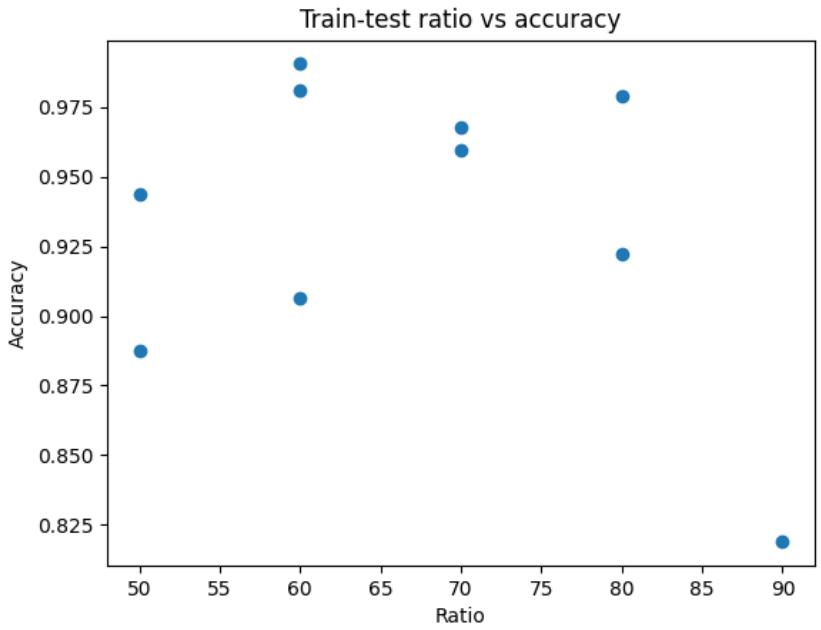
   b. **Hiper-parámetros del modelo.**

      - `n_estimators`: Número de estimadores (árboles de decisiones) que conforman el bosque.
      - `max_leaf_nodes`: Máximo número de nodos hoja que puede tener cada árbol.
      - `RATIO`: Proporción prueba-entrenamiento, indica qué porcentaje del dataset va a ser usado para pruebas, el porcentaje restante será el set de datos de entrenamiento.

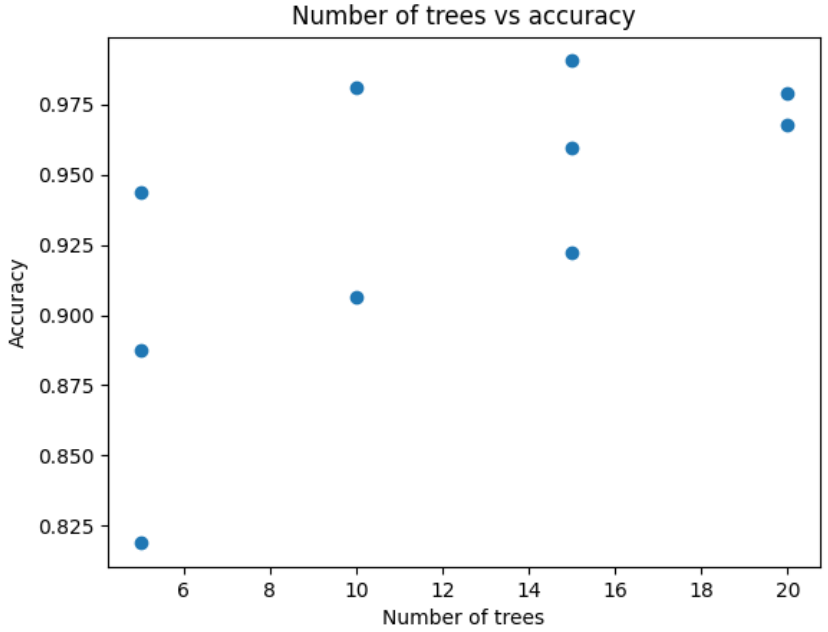
2. **Pruebas del modelo.**

   Se realizaron 10 pruebas en las cuales se variaron los tres hiper-parámetros principales con la finalidad de ver que tan bien se desempeña el modelo.

a. Análisis de pruebas.

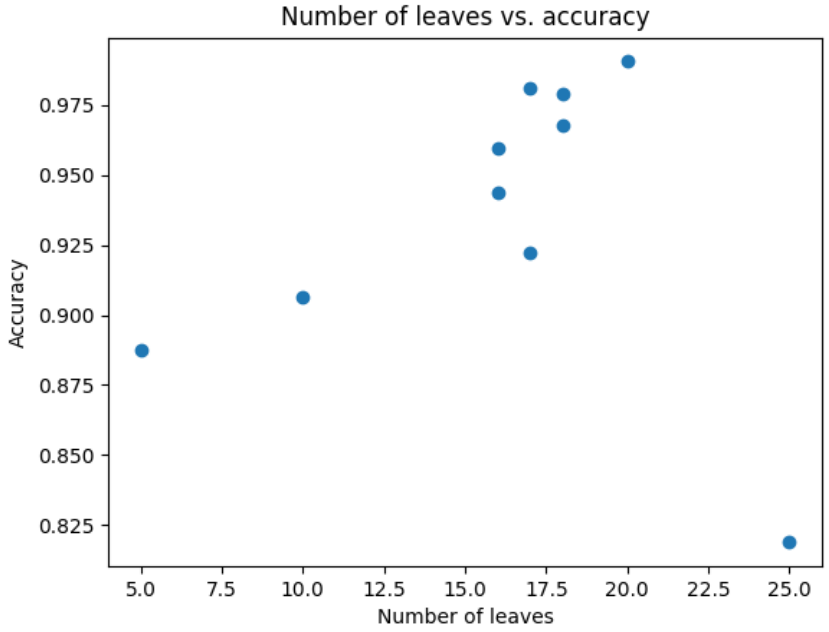


En esta primera gráfica se aprecia la relación entre la proporción entrenamiento-prueba con la precisión del modelo, se puede ver que valores entre el 60% y el 80% de datos de prueba se obtienen modelos con mejor precisión. Sorprendentemente el único registro que usa el 90% de los datos como datos de prueba tuvo la menor precisión, sin embargo, esto igual puede deberse a que los demás hiper-parámetros ocupados no fueron beneficiosos para obtener una precisión más alta.



En este caso, al ver la relación entre número de árboles y precisión sí se ve una mejora considerable en la precisión del modelo con respecto al número de

árboles, los mejores resultados se obtienen con 15 o 20 árboles mientras que 5 árboles dan resultados menos precisos.



De igual manera, el número de hojas influye bastante en la precisión del modelo, dando sus mejores resultados con 16-20 hojas. Sin embargo, hay que asegurarse que al poner muchas hojas no se haga overfitting ya que cada vez el modelo se "especializaría" más con el tipo de datos que se proporcionan.

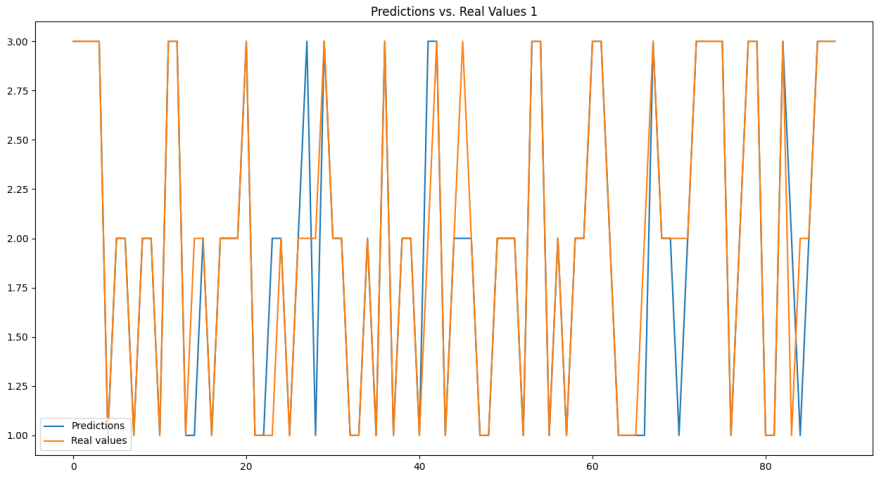
Estas pruebas muestran que, en teoría, las mejores combinaciones del modelo serían una proporción de pruebas entre el 60% y 80%, 15 o 20 árboles con 16-20 hojas.

A pesar de esto, se recomendaría no usar un número de hojas tan elevado para evitar que el modelo tienda al overfitting y así asegurar un modelo preciso y lo más general posible.

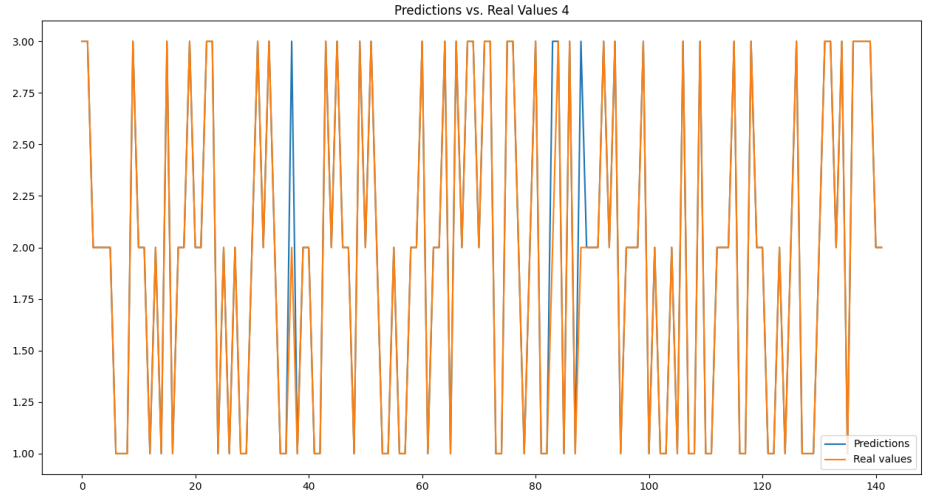
3. **Análisis del modelo.**

Ahora bien, en este apartado se muestran las 3 corridas más interesantes e importantes y se analiza cómo es que los hiper-parámetros afectaron a la precisión del modelo.

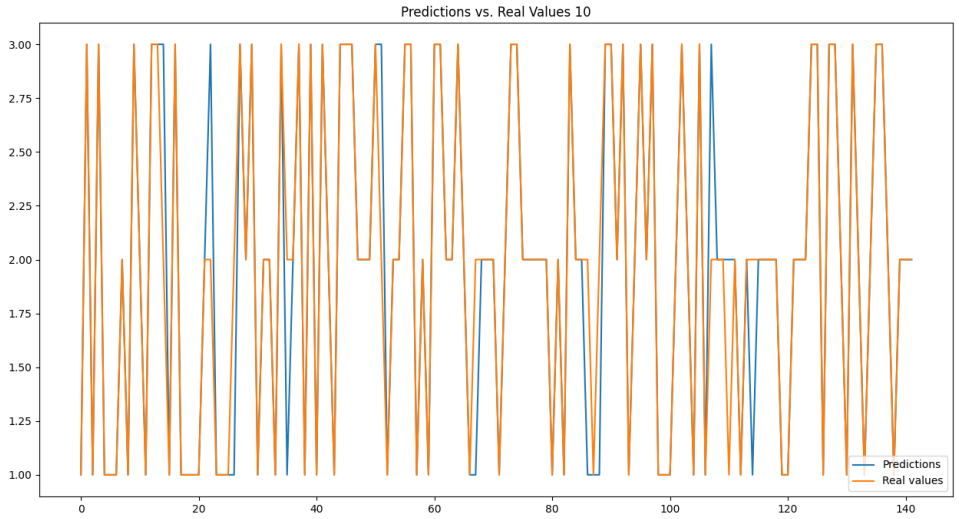
   a. **Precisión del modelo.**



En este caso se utilizaron los valores más bajos para cada uno de los hiper-parámetros, se utilizó una proporción del 50% y 5 árboles con 5 nodos hoja. Esto genera un modelo con una precisión bastante deficiente que, basado en las gráficas anteriores, oscila entre el 81% y 95% lo que al final significa un buen porcentaje pero, a comparación de otras corridas del modelo, puede contar con más registros incorrectos.



Para este otro caso se usaron hiper-parámetros sumamente altos, una proporción del 80% con 20 árboles y 18 hojas. En este caso se obtienen estimaciones sumamente exactas con una precisión entre el 96% y el 97% lo cual es extremadamente alto haciendo que se tengan porcentajes de error muy bajos, sin embargo, esto también puede causar un overfitting del modelo por lo que no es la opción más fiable y viable.



Finalmente, este es el último modelo con los valores que crean una explicación de lo más "general", usando una proporción de 80%, con 15 árboles de 17 hojas cada uno. Esto da precisiones de entre el 91% y el 98%, por lo que puede dar predicciones extremadamente precisas pero igual puede tener un ligero margen de error lo que hace que se evite el overfitting dentro del bosque.

b. Evaluación general del modelo.

En general, el modelo utilizado se apega fielmente para predecir la clase que tendrá un vino basado en las características de entrada, a pesar de que se dan predicciones suficientemente buenas, es necesario tomar en cuenta que los árboles de decisiones y los *Random Forests* pueden tender con facilidad al overfitting por lo que también se recomienda escoger hiper-parámetros que hagan un modelo lo más general posible.