Rubén Sánchez Mayén A01378379

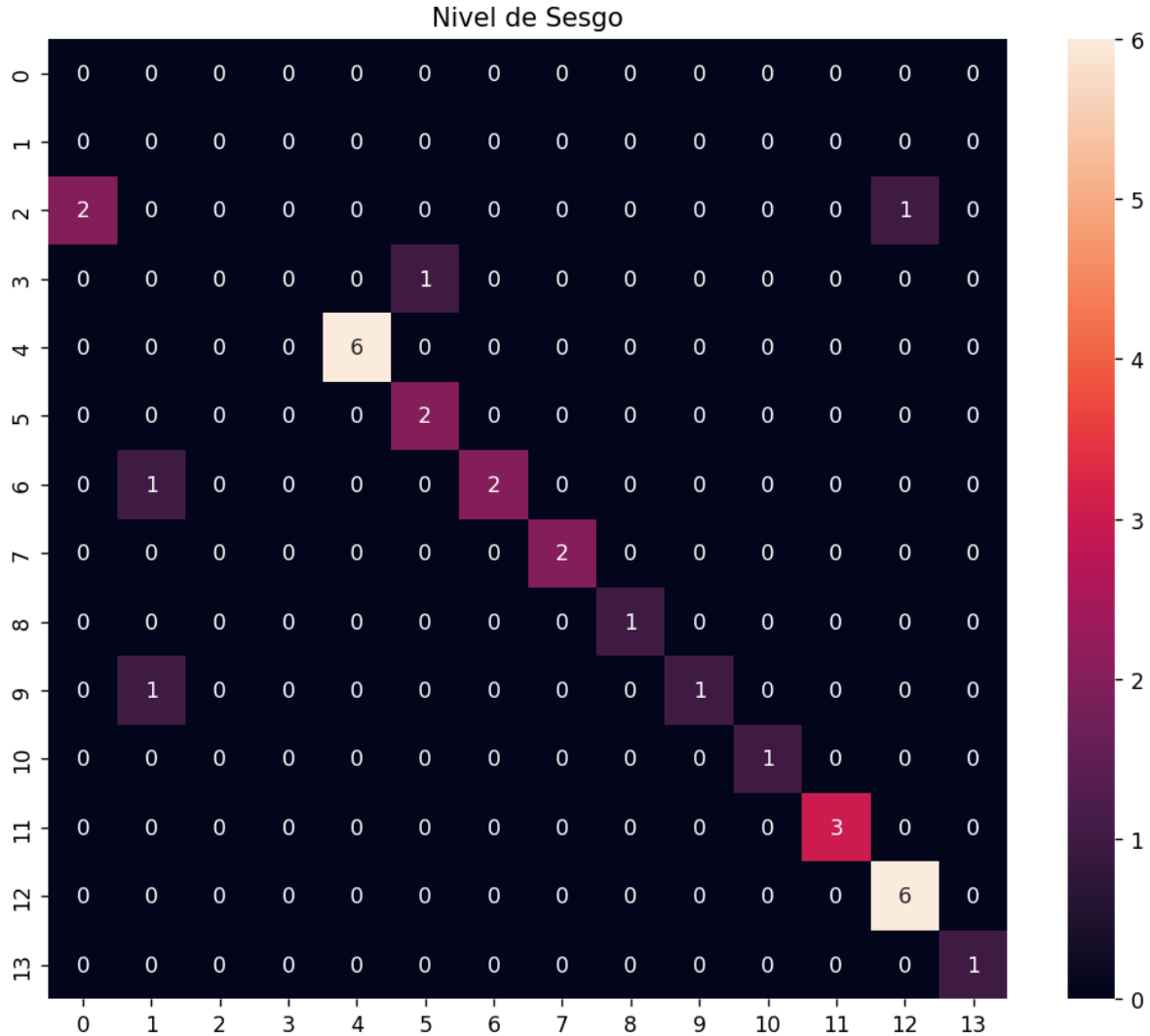
# Reporte sobre modelo de *Random Forests*

El modelo de *Random Forests* clasifica datos utilizando múltiples arboles de decisión. Este modelo introduce aleatoriedad en la creación de cada árbol con la finalidad de generar un conjunto de árboles no correlacionados, los cuáles son promediados.

El *dataset* utilizado para probar este modelo es ***Automobile.csv*** el cuál contiene datos sobre automóviles comerciales de distintas marcas, algunos de estos son precio, tamaño del motor, número de cilindros en el motor, dimensiones, entre otros. Se utilizará el **85% de los datos** para **entrenar** al modelo y el **15%** restante para **probarlo**. El modelo usa **múltiples variables independientes**. La **variable** que se busca **predecir** es la marca del automóvil (columna ***make*** en el *dataset*).

## Sesgo

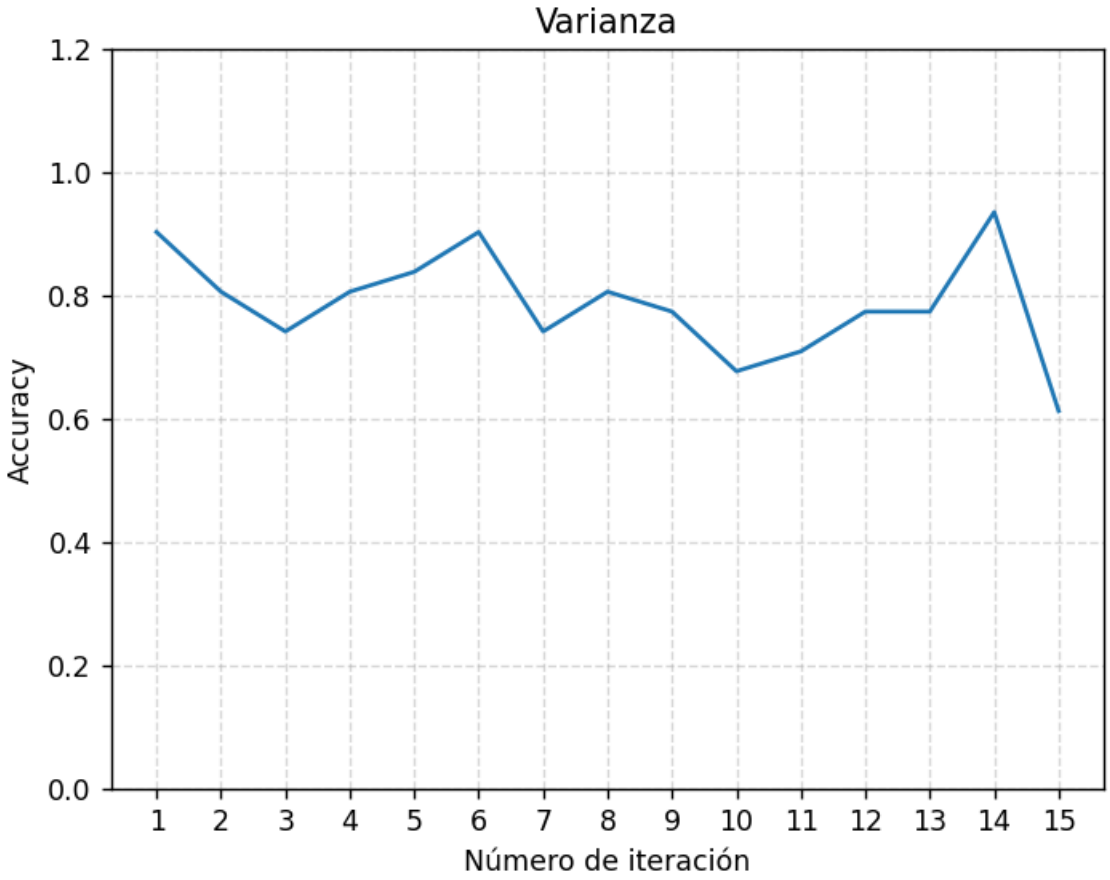
Para revisar el nivel de sesgo, se utilizó una matriz de confusión para revisar el error en las predicciones del modelo. Se utilizaron cómo hiperparámetros **15 estimadores** (árboles), un **máximo de 30 nodos hoja** y un estado de **aleatoriedad de 42**.



Se puede observar que el modelo tiene un **nivel bajo de sesgo**, puesto que predice correctamente la mayoría de marcas de automóviles, no obstante hay clases de automóvil que clasifica cómo erróneas en todas las pruebas (**clase 0 y 1**), lo cual no es deseable.

**Varianza**

Para revisar la varianza se calculó la precisión en distintas iteraciones utilizando los mismos hiperparámetros **15 estimadores** (árboles), un **máximo de 30 nodos hoja** y un estado de **aleatoriedad de 42**. En todas las pruebas se utilizó la **misma relación** entre la **cantidad de muestras** para entrenar al modelo **(85%)** y la cantidad de muestras para probar el modelo **(15%)**.




Se puede observar que el modelo tiene una **varianza media** ya que no siempre muestra el mismo valor de precisión, aunque se utilicen *datasets* de entrenamiento del mismo tamaño y con muestras aleatorias. En ocasiones el modelo puede mostrar una precisión cercana a 9, y en ocasiones una precisión cercana a 6.

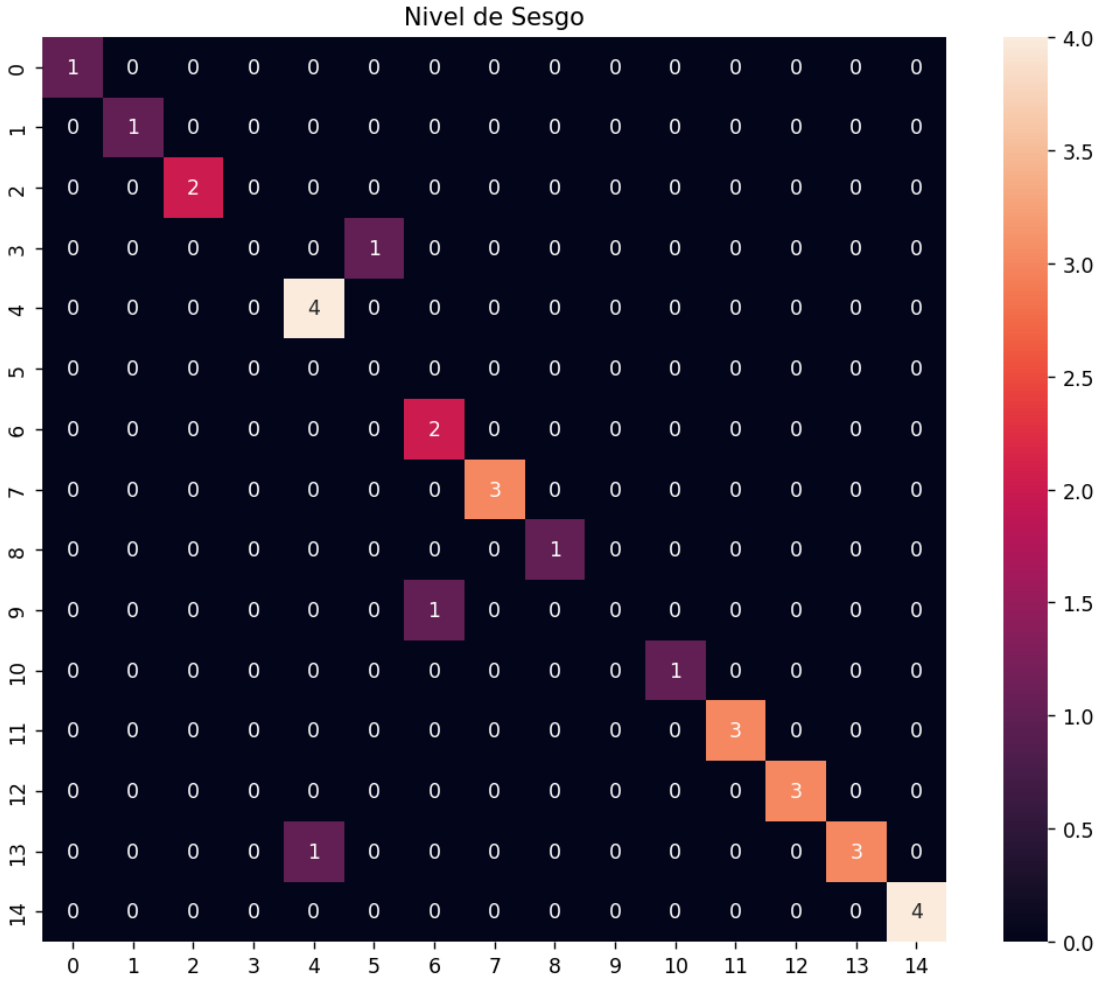
**Overfitting**

Puesto que el modelo tiene un **nivel bajo de sesgo** y una **varianza considerable**, es muy probable que este contenga **Overfitting**. Este fenómeno es común en los árboles de decisión y sus derivados (*bagging, random forests*, etc).

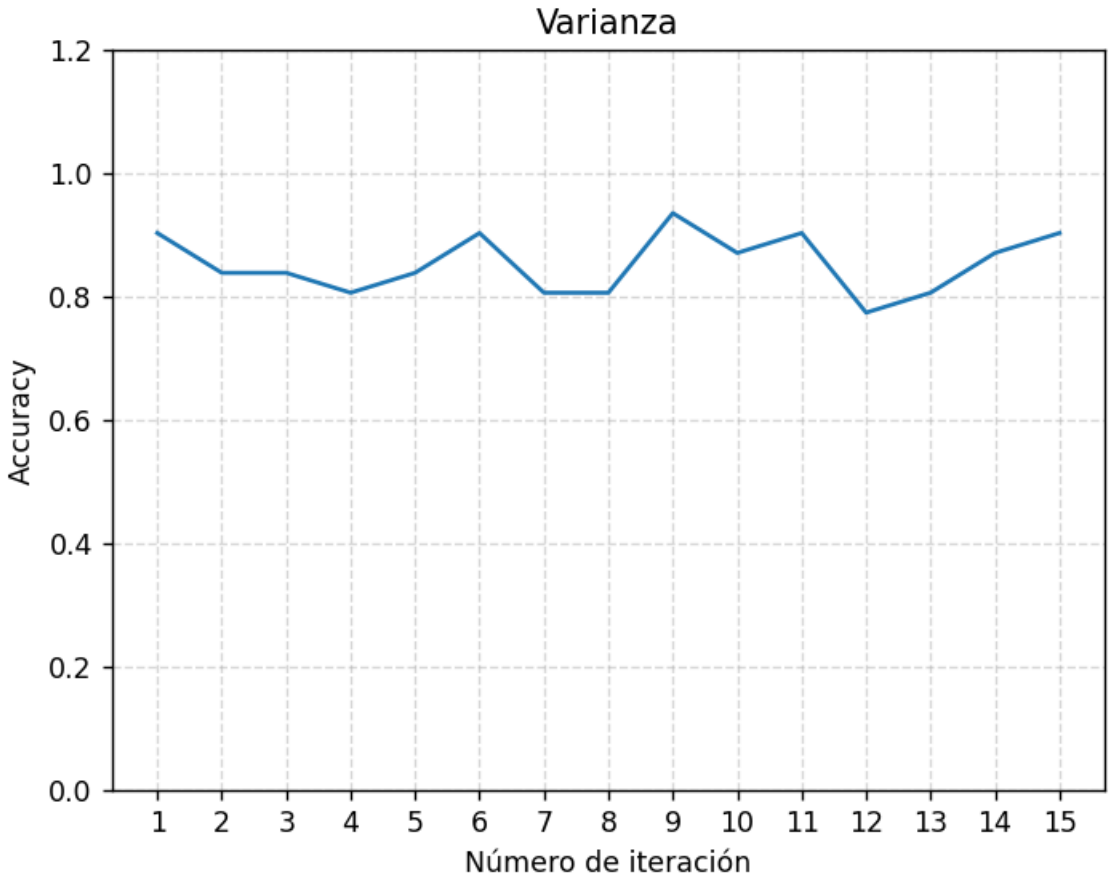
**Mejoras en el desempeño del modelo**

Rubén Sánchez Mayén A01378379

Uno de los objetivos de el algoritmo de ***random forests*** es evitar el ***overfitting*** al promediar múltiples árboles de decisión y modificar sus estructuras según una seríe de hiperparametros comúnmente utilizados las implementaciones de este algoritmo cómo la **cantidad máxima de nodos hoja** o la **profundidad máxima de un árbol**. Es por esto que se modificaron los hiperparametros iniciales, se promediaron **100 árboles en vez de 15** (número de estimadores), se redujo la **profundidad máxima a 15 niveles** y se redujo el **número de nodos hoja máximos** a 25. Los resultados fueron los siguientes:



Rubén Sánchez Mayén A01378379




Se **redujo la varianza del modelo**, mientras que el **nivel de sesgo se mantuvo** en un mismo nivel **(bajo)**, otorgando un modelo con un precisión similar, pero con mayor consistencia.