Instituto Tecnológico y de Estudios Superiores de Monterrey

Inteligencia artificial avanzada para la ciencia de datos I

Análisis del modelo

Autores:
Isaac Eduardo Alvarado Pérez - A00828386

Profesor:
Dr. Ivan Mauricio Amaya Contreras

Aprendizaje Máquina (Automático) TC3006C.101

18 de Septiembre de 2022

1. Dataset

El dataset utilizado en este entrega fue el de weatherAUS, el cual cuenta con 23 variables, las cuales sus valores son categóricos completamente, esto para utilizar los modelos vistos en clase, los cuales fueron de clasificación. Es un dataset sacado de kaggle.

La variable objetivo a predecir es la de RainTomorrow", la variable que nos da información acerca de si va a llover el siguiente día o no. Primeramente lo que se hizo en el código fue un filtrado de datos relevantes para la predicción de la variable objetivo, se hizo uso de la matriz de correlación para identificar el nivel de correlación entre las variables, y así, ver el nivel de relación que se tenía con la variable objetivo.

2. Modelo a utilizar

El modelo que se utilizó para las predicciones, fue el de Random Forest Classifier, y se hizo uso de las configuraciones de n estimator, max features y criterion.

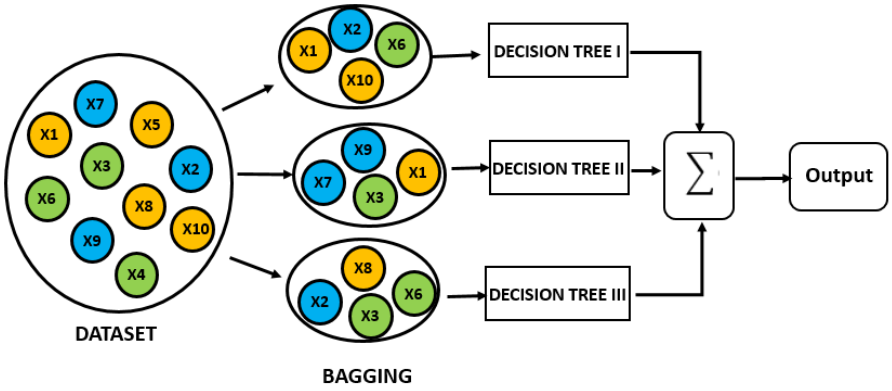

Figura 1: Random Forest Classifier

Una vez hecho el procedimiento para testear los datos, se evalúo la eficacia del modelo hecho.

Se obtuvo una varianza de 0.1264 y un sesgo de 0.0521, en este modeo utilizado (Random-Forest) se hace uso de un método llamado de Ensemble. Dicho método hace uso de Bagging, el cual toma como valor final la media de todas las predicciones hechas, y es aquí donde entra el Random Forest. Este método ayuda a que exista un equilibrio entre la varianza y el sesgo, por lo que podemos observar reflejado en nuestros resultados.

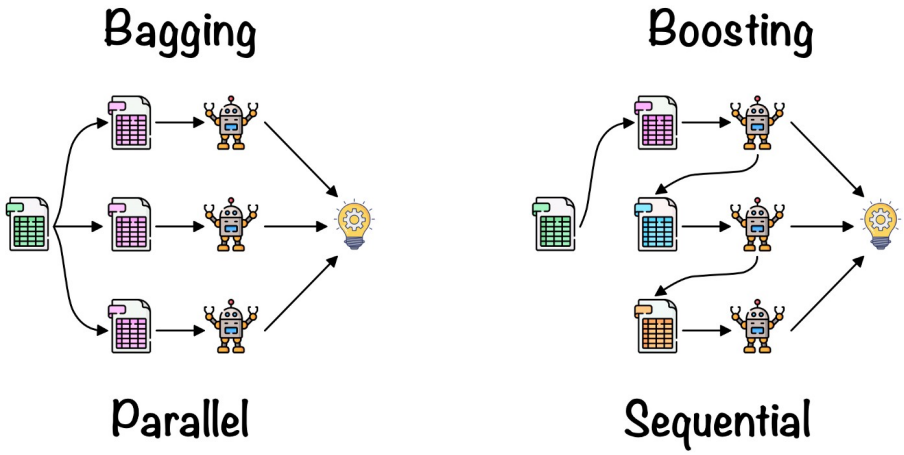

Figura 2: Ensemble

Se analizó la varianza y el sesgo en base a la predicción que se hizo. Y a continuación se mostrará la matriz de confusión con la que también se obtuvo la precisión.

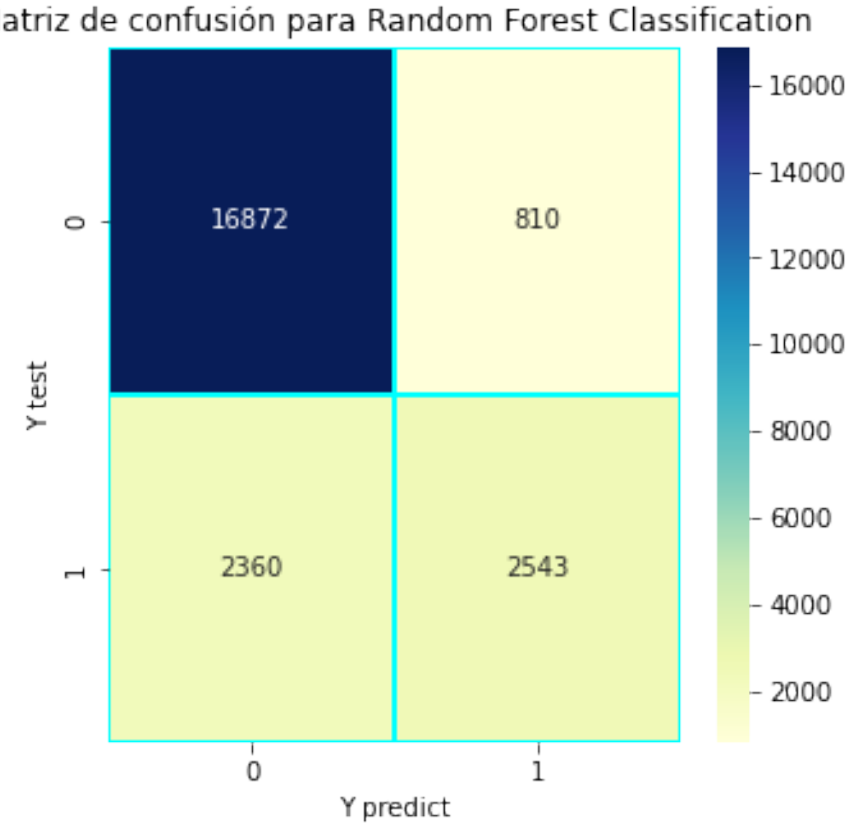

Figura 3: Matriz de confusión

También, para obtener un resultado más contundente se hizo uso de cross-validation con

k-folds, donde se hicieron 5 predicciones para contrastar la precisión obtenida.

3. Tabla resumen

Se obtuvo una ganancia de 0.51 % de precisión, se realizó las comparaciones con ayuda de matrices de confusión. También se hizo uso de Cross-validation para validar los datos.

A continuación se presentará una tabla con los parámetros que se utilizaron desde un principio junto a los mejorados, con su respectivo desempeño.

n_estimators	max_features	criterion	Sesgo	Varianza	score
150	9	gini	0.0521	0.1264	0.859641
50	5	entropy	0.0583	0.1206	0.855302