Cajamar La Viña Wine Prediction Datathon 2023

Team Turing

Universidad Complutense de Madrid

1. Introducción

España es el tercer productor mundial de vino. Disponer de una previsión precisa de la producción en una campaña agrícola es cada vez más necesario de cara a optimizar todos los procesos de la cadena: recolección, traslado, procesado, almacenamiento y distribución.

Del mismo modo, datos climáticos como la precipitación y la velocidad del viento en cada estación, así como el tipo de uva y la ubicación geográfica del campo, afectarán el rendimiento de la uva.

El objetivo es predecir los rendimientos futuros a través de datos meteorológicos históricos y datos históricos de siembra.

2. Minería y comprensión de los datos

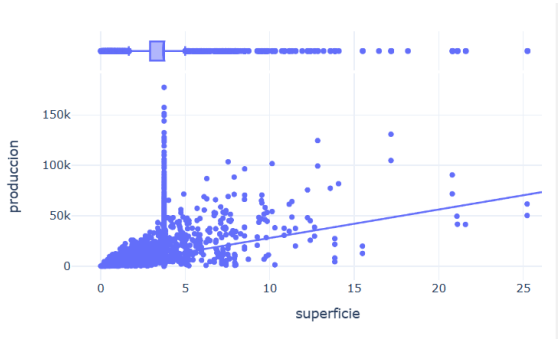
El conjunto de datos consta de tres archivos: TRAIN, METEO y ETO. TRAIN contiene información histórica sobre las fincas que conforman la cooperativa La Viña. METEO y ETO proporcionan información meteorológica detallada de estaciones climatológicas en la zona.

Nos enfocamos en el conjunto de datos METEO ya que creemos que el clima tiene un impacto significativo en las uvas. Al principio, consideramos usar análisis de series temporales para predecir el rendimiento de las uvas.

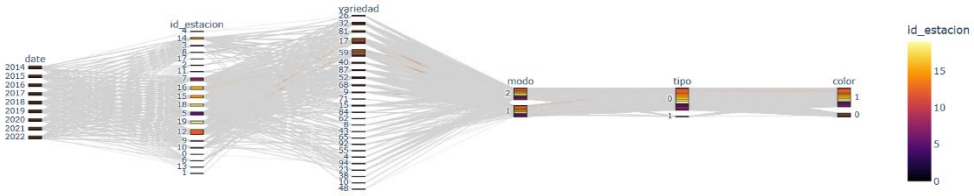
Sin embargo, encontramos algunos problemas con los datos. Por ejemplo, la columna 'altitud' en el conjunto de datos principal tenía un error de codificación y algunos datos estaban marcados como intervalos. Además, había una gran cantidad de valores registrados como 0 en la columna 'superficie', lo cual es muy irrazonable y probablemente fue un error en el registro.

En cuanto a los valores faltantes, los conjuntos de datos meteorológicos contenían una gran cantidad y decidimos eliminar características con demasiados valores faltantes.

También examinamos la distribución del conjunto de datos. En el conjunto principal, la mayoría de las características eran datos discretos. En cuanto a los datos continuos, parecía haber una relación lineal entre la superficie y la variable objetivo.



En otros datos, las estadísticas de datos de tipo y color son desiguales. Las estadísticas también varían mucho entre distintas variedads.



Para estacion de identificacion variable. Originalmente pensamos que se trataba de una variable utilizada para indicar la estación, luego de identificarla pensamos que era la identificación del observatorio meteorológico en la zona donde se encuentra la finca.

3. Preparación de los datos

Durante el preprocesamiento de datos, se realizaron varios cambios para mejorar la calidad de los datos. Se procesó la codificación de fecha en el conjunto de datos principal y se modificó el formato de fecha a fecha y hora. También se cambiaron los nombres de las variables a minúsculas para facilitar la lectura.

Para la variable de altitud en el conjunto de datos principal, que contenía datos con intervalos, se asignó el valor medio y se convirtió en una variable continua. Además, se utilizó el método bfill en fillna para completar otros datos.

En los conjuntos de datos METEO y ETO, se eliminaron entidades con más del 50% de valores faltantes y se utilizó el método bfill para completar el resto.

Finalmente, dado que los datos en el conjunto principal no estaban contados por mes sino por año, se consideró tomar el valor medio de los conjuntos METEO y ETO por año y combinarlos con el conjunto principal según id estacion y fecha. Esto dio como resultado un conjunto completo para análisis.

4. Modelado y discusión

Random Forest es un algoritmo de aprendizaje conjunto que mejora la clasificación o regresión al combinar múltiples árboles de decisión. Introduce aleatoriedad al generar cada árbol mediante

el muestreo aleatorio de muestras de entrenamiento y características.

Al predecir, Random Forest sintetiza los resultados de cada árbol y obtiene el resultado final mediante votación o promedio, reduciendo el sobreajuste y la varianza.

Para la selección de variables, se utilizó el método Embedded y se seleccionó la importancia de las variables en el modelo Random Forest. Basado en los datos totales, se seleccionaron las 10 características principales y se excluyeron las variables con importancia por debajo de 0.01.

Se utilizó una búsqueda en cuadrícula para encontrar los mejores hiperparámetros del modelo y se obtuvieron los mejores hiperparámetros: n_estimators=300, max_depth=None, max_features='log2', min_samples_split=2, min_samples_leaf=1.

Con estos hiperparámetros óptimos, el RMSE del modelo mejoró a 5903 y el R2 mejoró a 0.764. Este fue el modelo final utilizado.

Dado que no había registros estadísticos mensuales en el conjunto de datos original, no se utilizó análisis de series temporales para análisis y predicción.