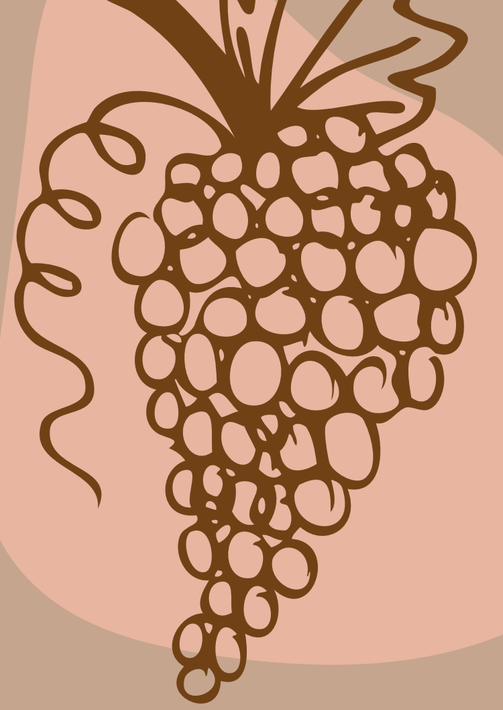


Buscando la predicción de la producción de vino de la campaña del 2022

Pynot

INTRODUCCIÓN

Para degustar una copa de vino se lleva un proceso arduo en donde se deben considerar tanto parámetros propios de la finca como del tipo de uva y por supuesto parámetros meteorológicos que van a afectar definitivamente a la producción del vino.

El crecimiento de la vid tiene varias fases, las cuales comienzan a finales del invierno, entre abril y mayo comienza la siguiente fase, a principios de junio se pueden apreciar los brotes o embriones de flores que van a dar paso a la uva y luego llegando a la fase final donde se da la recolección.

Este proyecto se va a enfocar a predecir la producción de vino en la campaña del 22 para la coorperativa La Viña, en España.

PASOS A SEGUIR

Para comenzar el proyecto se recibieron 3 datasets uno llamado TRAIN, otro METEO y el último ETO. Cada uno con información específica y diferente de los aspectos de la uva.

Se decide seguir el siguiente proceso:

- Investigar sobre el cuidado de la uva y su recolección.
- Exploración de los datasets.
- Depuración de dichos datasets por separado.
- Juntar los datasets.
- Hacer pruebas de modelos y merge entre los datasets.
- Definir el modelo a utilizar.
- Predecir la producción de vino de la campaña 2022.

METODOLOGÍA

En la exploración y depuración se observó que los datos meteorológicos aportados contienen gran cantidad de información. Se eliminaron las variables que contenían gran cantidad de variables nulas, al igual que variables con solo una categoría o incluso con muchos valores nulos que no fueran a influir con el modelo ni aportar grandes cambios. Para poder hacer la predicción se decide unir los datasets después de depurados mediante una agrupación por la suma, la media y el máximo para dejar toda la información útil possible por campaña, asimismo se transpuso la columna de date para poder contar con las variables predictoras por cada mes.

Los missings fueron imputados en cada dataset por separado mediante random, primeramente se imputó por knn pero se determinó que el nivel de procesamiento era mucho mayor y la diferencia no era apreciable. Los outliers fueron otros puntos importantes a la hora de limpiar el dato, para estos se utilizó el método Winsor en los 3 datasets.

Una vez obtenida esta exploración, depuración y unión de la información se decide realizar una matriz de correlación y obtener también el valor de v de Cramer para observar la interacción entre las variables predictoras y la variable objetivo (producción). Al observar detalladamente,la gran correlación entre variables y la gran cantidad de ellas se decide utilizar Análisis de Componentes Principales para reducir la alta dimensionalidad del dato meteorológico. Se buscó recoger al menos un 98% de la variabilidad del dato.

METODOLOGÍA

Se probó primeramente utilizar regresión lineal para generar la predicción de la producción de vino pero al generar los resultados el valor de R2 de estos da un valor bastante bajo, alrededor de 0.4 y un RMSE de 10435.67 lo cual es bastante alto al probar la predicción de los datos entre los años 2015 a 2021. Por lo que se decidió mejorar y cambiar el procedimiento. También se mejoró la depuración y la unión de los datasets.

Se hizo uso de la regresión de random forest sobre únicamente los datos de TRAIN y los resultados mejoraron considerablemente, se obtuvo un R2 de alrededor de 0.6 y un RMSE de aproximadamente 6000 para la predicción de los datos de las campañas 15 y 21 y un valor del alrededor de 8000 para la campaña del 2022. A pesar de que los resultados mejoraron se decide probar con otra modelización que pudiera incluir los datos meteorológicos, aún a riesgo de bajar la performance en test y validación, ya que se consideró que el uso del random forest para la predicción puede generar gran overfitting teniendo en cuenta que realizamos una imputación de la variable superficie con base a los identificadores y dicha variable fue después la que obtuvo casi toda la importancia en la predicción final.

La temperatura, humedad, lluvia, nieve y la presión son variables relevantes a la hora de hablar de producción vinícola y estos estaban siendo ignorados.

MODELO SELECCIONADO

Se decidió continuar con el modelo de Random Forest, pero tratando de dar algo más de importancia a las variables meteorológicas.

Random Forest fue el modelo que mejores resultados dio en validación cruzada, presentando un coste computacional bajo para el data set usado como entrenamiento:

```
Importancia de cada variable para el modelo:
ALTITUD ---> 0.07704137192222321
SUPERFICIE ---> 0.6496942257547647
MODO ---> 0.055023630632002794
PC1 ---> 0.011733456562410137
PC2 ---> 0.010077218507534905
PC3 ---> 0.00886140756956256
PC4 ---> 0.011522463747025487
PC5 ---> 0.009901530869820385
PC6 ---> 0.008682832449928702
PC8 ---> 0.01639636590585099
PC9 ---> 0.013196772117359104
PC10 ---> 0.008385202024109513
PC11 ---> 0.013376965415138666
PC12 ---> 0.014124186285504933
PC14 ---> 0.015968299427490545
PC15 ---> 0.015125752087109588
PC19 ---> 0.009961643263990872
PC22 ---> 0.008597950575883442
PC23 ---> 0.013162326913487378
PC24 ---> 0.0107366920378889
PC25 ---> 0.009126382203534705
PC27 ---> 0.009303323727378517
```

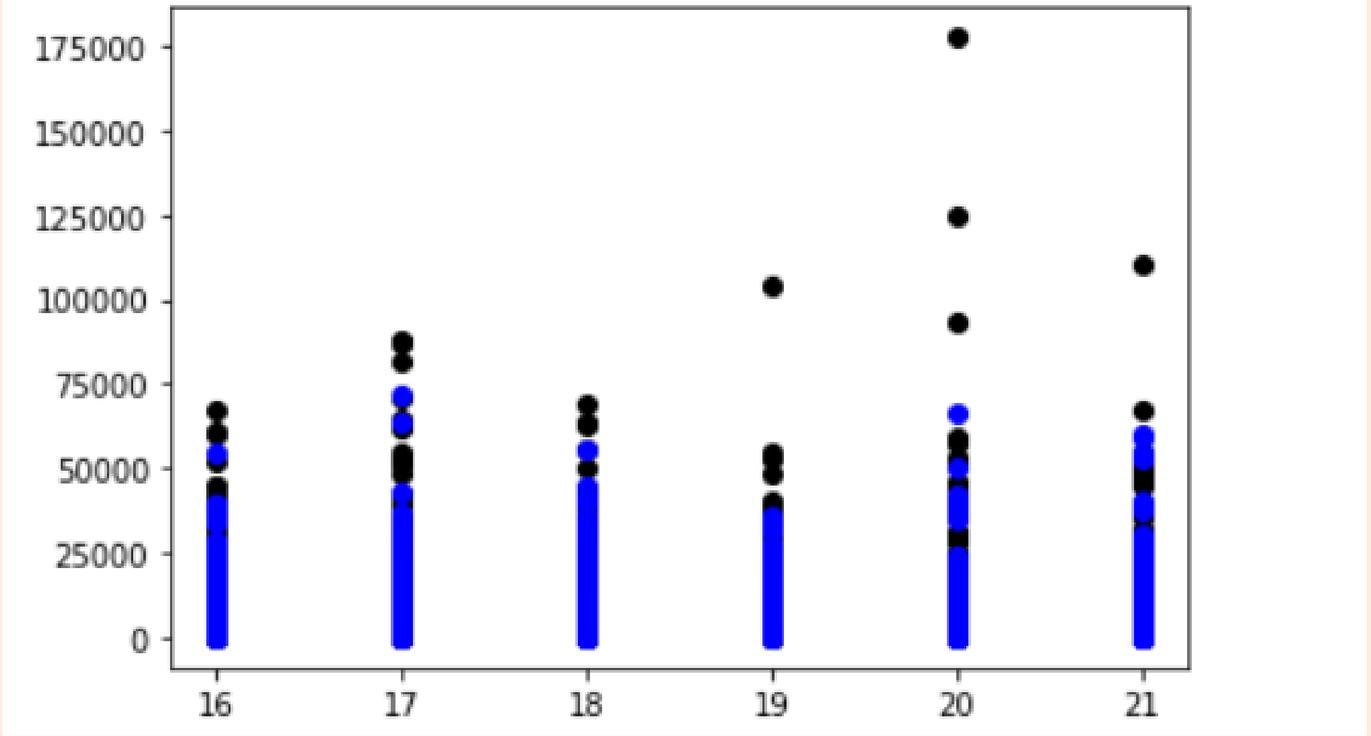
Complejidad = O(6262log(6262)(22)(100))

Tiempo de entrenamiento = 4.078 sec

MODELO SELECCIONADO

Se presenta el gráfico de comparación de la producción contra los años de los valores que se han predicho (valores en azul) y los valores reales para los años del 2015 al 2021 (en color negro). La predicción no se ve distante de lo real lo cual indica gran avance.

```
Root Mean squared error: 10859.20
Accuracy: 0.41
<matplotlib.collections.PathCollection at 0x7f73b0a55070>
```



CONCLUSIÓN

Para finalizar se debe considerar que los valores obtenidos para RMSE y la gráfica mostrada anteriormente son alentadores.

Se propone como siguientes pasos:

1. Revisar y mejorar la unión de los datasets y la depuración de estos especialmente disminuyendo la importancia que los modelos le están dando a superficie específicamente.

2. Probar modelos de regresión múltiple, utilizar series temporales, entre otros.

3. Investigar qué otras variables predictoras pueden eliminarse para mejorar la predicción o cuales mantener.