SOFIS-ticados

Sofia Di Capua, Marc Bosom, Manuel Arnau

# 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.

Dado lo anterior y partiendo de amplios datasets con histórico de producciones de los viñedos que conforman la cooperativa La Viña, así como histórico de la climatología de los mismos, nos han propuesto crear el mejor modelo de predicción de producción de una campaña en base al cual se pueda estimar la cosecha que dispondrá la cooperativa meses antes de la recolección.

# EDA, preprocessing and data mining

A continuación explicaremos cómo hemos tratado y analizado cada uno de los datasets para poder realizar la predicción.

### Dataset TRAIN

* Rellenado los valores NaN que se encontraban en la variable ALTITUD con forward filling y backward filling, y convertido a numérica
* Corregido los valores de superficie 0 de las campañas 20, 21 y 22 con un KNN imputation.
* Estimado los valores de superficie de las campañas anteriores utilizando métodos estadísticos en una primera instancia (Prueba de Levene y Z-test), estudiando la diferencia de producción entre estas campañas y la 20 y la 21. Para aquellos valores que no se han imputado con este método, se ha utilizado un KNN imputation.

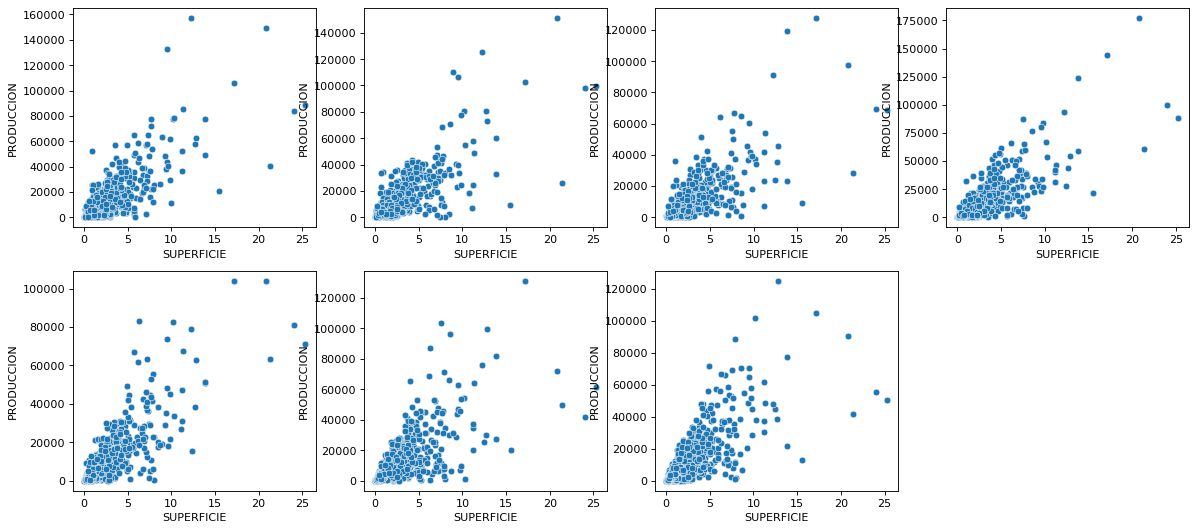


Figura 1: Relación Superfície Producción (de la campaña 14 a la 21)

* Analizado como es la variable objetivo, y sus correlaciones con diferentes atributos como son SUPERFICIE, MODO y TIPO.
* Analizado que la distribución de los atributos en la campaña 22 y en las demás sea parecida
* Detectado outliers de producción por hectárea, algunos probablemente creados en la imputación de superficie, y eliminado en estos valores siguiendo un Z-test.
* Analizado la varianza de la producción por hectárea de cada estación respecto a sus diferentes campañas, lo que ha llevado a ver que los datos climatológicos no serían de tanta ayuda como se pensaba en un principio.
* Calculado la media de producción por variedad

### Dataset METEO Y ETO

* Para ambos datasets se ha dado un mismo formato de fecha, asegurando que sean datetime. Se ha visto también que cubren los mismos días.
* Se ha añadido los atributos Day, Month y Year, que sirven para hacer agrupaciones y poder hacer merge de datasets
* En el caso de METEO, solo era de interés la precipitación y la nieve, ya que los demás datos ya estaban en ETO. Se han rellenado los valores NaN de estas variables con forward filling y backward filling, y luego un groupby por días haciendo agregación, para poder añadir al ETO.
* En el ETO, se ha seleccionado sólo las variables que representan el día completo, y se ha rellenado los valores NaN de Evatransporación, Presión, RadiaciónSolar y Gust usando RandomForests.
* Finalmente se ha unificado ETO y METEO en el primero.

# Data engineering y Data augmentation

### Dataset TRAIN

Con el análisis hecho, nos hemos dado cuenta que realmente superficie no es el valor más indicativo, ya que la densidad de vid en cada finca es diferente (como se ha observado con los outliers). Por lo tanto, hemos calculado una estimación de producción por hectárea por cada entrada del dataset, focalizando sobre todo en la campaña 22.

Para estimar este valor para la campaña 22, hemos tenido en cuenta principalmente las campañas 21 y 20, que son las que tenemos los valores de superficie más fiables, además de tener el mayor porcentaje de fincas repetidas (usando como identificador ID\_FINCA, VARIEDAD, TIPO y en algunos casos MODO y SUPERFÍCIE). Hemos tenido presente que las producciones fueran parecidas y la posibilidad de que el modo de cultivo cambiase de un año a otro. En caso de aparecer solo en la campaña 21, la estimación se ha rellenado con la producción por hectárea de la campaña 21, ya que teniendo en cuenta que la media es el mejor estimador de la esperanza (que es realmente lo que estamos estimando), no debería suponer errores de estimación muy grandes.

Para aquellos valores de diferencia entre la campaña 21 y 22 que sobrepasaban un threshold, añadimos los valores de las demás campañas para calcular la estimación.

Finalmente, para aquellas fincas que sólo aparecen en la 22, hemos calculado la estimación en base a la variedad del cultivo y el modo. Durante todo este proceso también hemos llenado los valores de estimación de las campañas 20 y 21.

Para la estimación de las otras campañas, se ha hecho la media de la producción por hectárea, y aquellas fincas que solo aparecían en una campaña se han eliminado.

Con este valor de estimación de la producción, hemos podido generar una nueva variable (en el dataset se encuentra como FACTOR\_PRODUCCION) para todas las campañas anteriores a la 22 que representa la diferencia entre la producción real y la estimada, y a partir de ella hemos detectado nuevos outliers. También nos ha permitido ver que nuestro método de estimación tiende a estimar a la alza.

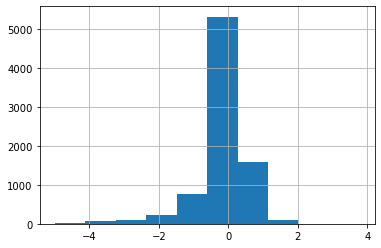


Figura 2: Histograma de diferencias entre estimación y producción

### Series Temporales

Una limitación inicial que presentaba el dataset es que la información climatológica era solo hasta el 30 de junio, pero la campaña no finaliza hasta octubre, donde se da la vendimia.

No obstante, estos datos se pueden tratar como series temporales (una por cada estación), dado que los datos climatológicos tienen estacionalidad, en algunos casos siguen una tendencia y en principio son estacionarios.

Al tener muchas variables, no hemos visto adecuado hacer un análisis de los factores anteriores, ya que es una asunción en principio bastante certera. Tampoco hemos considerado utilizar todas las variables climatológicas, y se ha optado por usar aquellas demostradas en estudios que son relevantes (temperatura, precipitación, …)[[1]](#footnote-0)[[2]](#footnote-1)[[3]](#footnote-2).

Para hacer la predicción de las series temporales hasta noviembre hemos utilizado Prophet. La predicción se ha tenido que hacer por estación y por variable, ya que Prophet no es un modelo que permita encontrar relaciones entre diferentes series temporales como podría ser Temporal Fusion Transformer, pero al tratarse de pocos datos su rendimiento es incluso mejor que modelos neuronales.

Hecha la predicción, se ha juntado con el dataset ETO.

### Datasets Climatológicos

Con toda la información necesaria de todas las campañas, eliminando la información climática de la campaña 15, hemos optado por hacer un dataset por cada variable climatológica, para poder añadir y quitar los datos en la fase en modelaje de una manera más cómoda. Se ha agrupado los datos por meses, siguiendo el calendario de la vid[[4]](#footnote-3). En el caso de la precipitación, la agrupación se ha hecho con una agregación y se ha tenido en cuenta el período de reserva hídrica, para el resto de variables solo se ha hecho la media y se ha tenido en cuenta sólo los meses de brotación/floración, crecimiento y maduración (siguiendo un poco las estaciones del año).

También hemos probado de calcular el índice GDD[[5]](#footnote-4) con los datos presentes, pero en los modelos utilizados no ha resultado ser útil (como alguna de las otras variables climatológicas utilizadas)

# Modelos

Para los diferentes modelos propuestos se ha hecho una forward variable selection de los datos climatológicos, viendo si añadiendo unos datos conseguíamos mejorar el modelo o no. También hemos visto que quitando del set de train aquellas filas con un valor absoluto de FACTO\_PRODUCCION alto (1.6, que representa una diferencia del 160% respecto nuestra estimación) se conseguía una mejora considerable. En todos los casos hemos eliminado las columnas de ID\_FINCA, ID\_ZONA, CAMPAÑA e ID\_ESTACION, ya que supondría un problema para aquellas fincas y zonas que no aparezcan en el set de train.

Primero de todo hicimos un baseline con los datos de train solo (sin nada climatológico) con un RandomForestRegressor, y establecimos la campaña 18 como set de validación ya que hemos observado parecidos climáticos con la campaña 22, y también evaluando la campaña 21.

Los modelos probados más efectivos han sido los siguientes:

* RandomForestRegressor (RFR): Se han añadido los datos de precipitación, viento máximo y radiación solar, además de los datos de train. Dada una buena estimación en el primer modelo, se ha optado por hacer una RandomGridSearch, resultando en los siguientes parámetros finales: {'n\_estimators': 300, 'max\_depth': 25, 'min\_samples\_split': 6}
* XGBoostRegressor: En este caso se han añadido los mismos parámetros climatológicos más las temperaturas medias mensuales. También ha producido una primera buena estimación, por lo que se ha hecho Hyperparameter Search, resultando en los siguientes: {'n\_estimators': 90, 'min\_child\_weight': 5, 'max\_depth': 3, 'subsample': 0.7, 'colsample\_bytree': 0.7, 'reg\_lambda': 0.8}
* KNNRegressor: Para el KNN solo se ha usado los parámetros de superfície y estimación de producción (normalizándolos). Si bien la primera estimación no era mala del todo, sí que era significativamente peor que los anteriores modelos y decidimos no proseguir con este modelo
* Ensemble Modelos: Dado que tenemos dos buenos modelos (RFR y XGBoost) y también una estimación base, se ha buscado la mejor combinación de la predicción del RFR, la del XGBoost y de nuestra estimación base, y los pesos de cada uno han resultado ser: 18% estimación base, 14%XGboost y 68% RFR.

### Tabla Resultados

| MODELO | RMSE CAMPAÑA 18 | RMSE CAMPAÑA 21 |
| --- | --- | --- |
| RFR baseline | 5604.08 | 4657.06 |
| Estimación base | 4867.05 | 3367.82 |
| Primer RFR | 4856.99 | 4021.29 |
| Primer XGBoost | 4774.86 | 3807.31 |
| KNN | 5130.34 | 3985.29 |
| HyperParameter RFR | 4758.16 | — |
| HyperParameter XGBoost | 4870.00 | — |
| Ensemble (0.16,0.14,0.68) | 4735.91 | — |

Para los primeros modelos el set train no contenía los datos de la campaña 21, porque queríamos comprobar cómo era la predicción frente a la estimación base de la campaña 21, y ver que realmente aunque hubiésemos añadido la estimación de esta campaña con información explícita de ella (en otras palabras, haciendo overfitting), se podía obtener un valor bueno.

En los resultados finales (dondé no hay RMSE 21) al train se le añadió la campaña 21, motivo por el cual XGBoost da un resultado peor. Aun así, como la predicción final la hacemos con todos los datos, no nos preocupamos por el empeoramiento de la métrica (además nos muestra que la predicción al añadir más campañas no debería empeorar el resultado, ya que el RFR mejora).

# Entrega

Viendo las métricas obtenidas en cada modelo, y teniendo presente que la predicción final se debería hacer con todas las campañas disponibles, ya que sinó estaríamos desperdiciando información, hemos escogido como modelo final el ensemble de RFR y XGBoost con la estimación base, ya que ha mostrado mejoría al añadir más información y también ha obtenido la mejor métrica para la campaña 18, la cual hemos usado como referencia debido a su parecido en términos climáticos con la campaña 22 y también porque el cálculo de la estimación base se ha hecho de una manera mucho más generalizada (razón por la que el RMSE de la estimación base es mucho mayor que la de la campaña 21).

1. https://repositorio.aemet.es/bitstream/20.500.11765/1969/1/criteriosagro\_cal2001.pdf [↑](#footnote-ref-0)
2. https://www.alice.cnptia.embrapa.br/bitstream/doc/1008614/1/1s2.0S0168192304000115main.pdf [↑](#footnote-ref-1)
3. https://repositorio.unican.es/xmlui/bitstream/handle/10902/17121/AransayGonzalezRaquel.pdf?sequence=1&isAllowed=y [↑](#footnote-ref-2)
4. https://www.vinetur.com/2019111458471/las-7-etapas-de-un-vinedo-el-ciclo-vegetativo-de-la-vid.html [↑](#footnote-ref-3)
5. https://www.evineyardapp.com/blog/2017/03/01/why-the-need-to-calculate-growing-degree-days-in-vineyard/ [↑](#footnote-ref-4)