

DATATHON CAJAMAR 2023

Fase Local - Malbecs

Denis Trosman | Vito Stamatti | Sumit Kumar Jethani Jethani

Marzo 2023

ÍNDICE

- Objetivo
- Análisis exploratorio e ingeniería de variables
- Modelos predictivos
- Resultados
- Guía de la entrega



OBJETIVO.

Analizar y predecir correctamente la producción de los viñedos La Viña.

España es el tercer mayor productor de vino del mundo. Es crucial tener una previsión precisa de la producción agrícola para optimizar la cadena de producción completa, desde la recolección hasta la distribución.

Para lograr este objetivo, **Cajamar** entrega tres datasets, con información pasada de las fincas y del clima en las mismas otorgada por **The Weather Company**.

Cada uno de estos cuenta con sus propios desafíos y características, por lo que analizarlos en profundidad y entender los datos que se están tratando resulta clave para obtener la **mejor predicción posible de la cosecha 2022**.

Datos utilizados		
Dataset	Filas x Columnas	Data
Train	9.601 x 11	Fincas
Meteo	51.180 x 275	Clima por hora
ETO	1.223.660 x 34	Clima agrupado



ANÁLISIS EXPLORATORIO E INGENIERÍA DE VARIABLES.

Script de exploración exhaustivo, formando los primeros pasos para el modelado.

Train

- Se notó que el dataset cuenta con una alta cardinalidad para sus variables ids y con pocas variables continuas.
- Por este motivo, se construyeron nuevas variables buscando ordinalidad y correlación con la producción.
 - Por ejemplo, se han tenido en cuenta los percentiles de las variables IDs con respecto a la producción a lo largo de los años, para que no haya una alta dependencia de los IDs en sí, sino de su potencial cosecha.
 - Así mismo, se generaron *shifts* en las producciones para que los modelos tengan presentes los valores pasados de esta a la hora de generar sus predicciones, entre otros.
- En cuanto a la producción, no se vieron tendencias claras por finca, pero se destacaron los años 2015 y 2018 como aquellos con mayores valores,
 - La producción está altamente correlacionada con la superficie, y ambos tienen un alto rango de valores.



ANÁLISIS EXPLORATORIO E INGENIERÍA DE VARIABLES.

Script de exploración exhaustivo, formando los primeros pasos para el modelado.

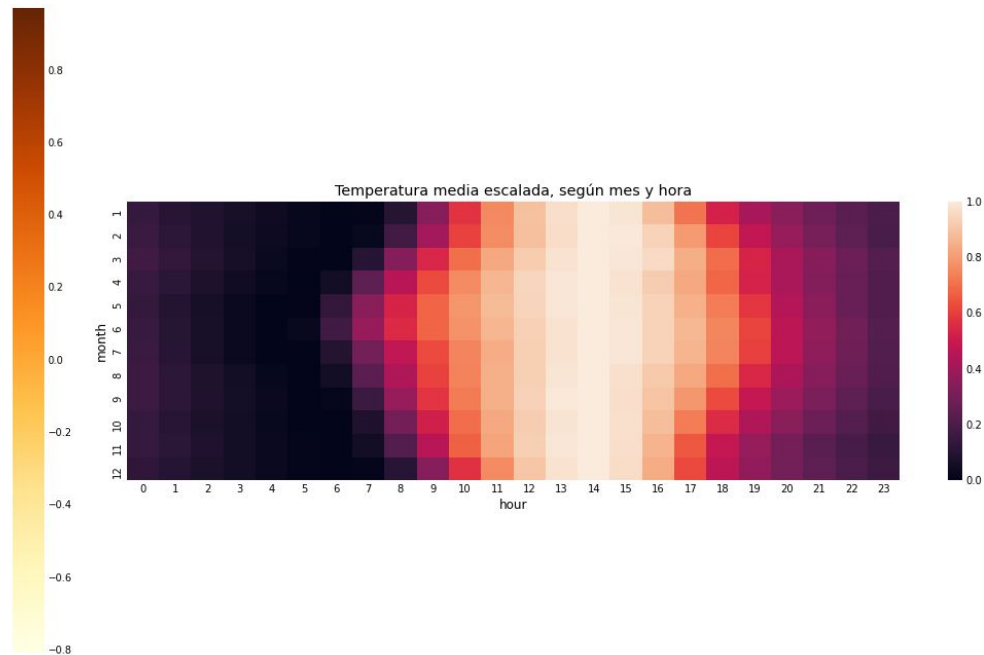
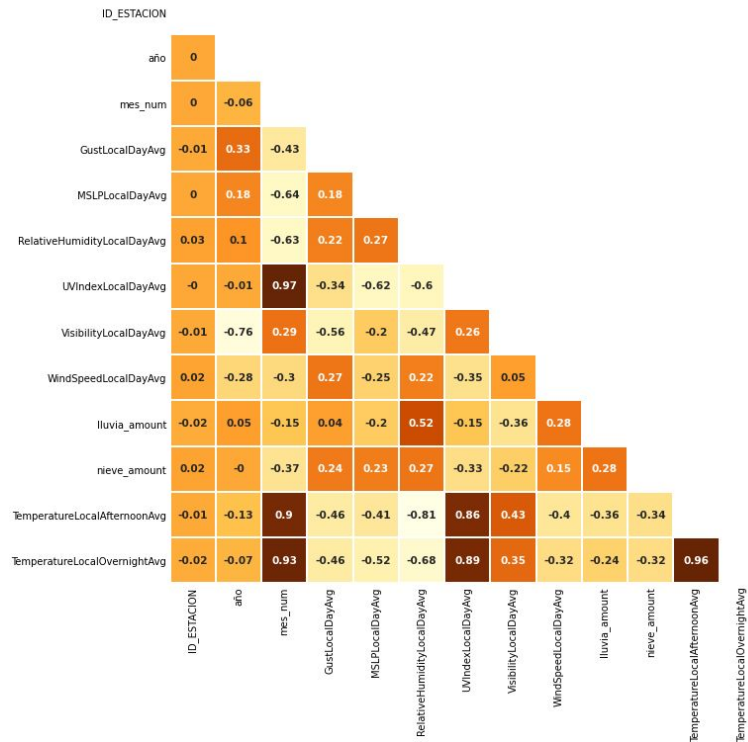
Datasets climáticos

- Con el fin de entender qué variables tienen su mayor importancia, y en qué época del proceso de cosecha, se realizó una investigación de vitivinicultura en España.
 - Para seleccionar variables, nos concentramos en las correlaciones de estas entre sí y las importancias de estas en cada etapa de la cosecha.
 - Se notó la presencia de datos faltantes para 2014, por lo que hubo que realizar las imputaciones correspondientes.
- A lo largo del trabajo se probaron diversos métodos de agregación para no aumentar la dimensionalidad en un alto nivel.
 - Por ejemplo, generar sumas, promedios, máximos y mínimos por mes, estación o año.



Relaciones entre variables climáticas

Correlaciones



MODELOS PREDICTIVOS.

Escoger el mejor modelo predictivo para la cosecha de 2022, teniendo en cuenta el RMSE como medida de scoring.

Metodología: para nuestros intentos, consideramos al set de entrenamiento como aquellas campañas menores a 2021, y separamos a este último como test. Para el cross-validation, el año del test siempre fue mayor al más grande de train.

Modelos: contamos con un problema de regresión, por lo que los modelos probados fueron tales como:

- Regresión lineal
- Random Forest
- Extreme Gradient Boosting
- CatBoost, AdaBoost
- Redes neuronales
- Regresiones de panel (Pooled OLS)
- Stacking

Encoders: en los Pipelines finales, se han probado encoders acordes a los tipos de variables ingresadas al modelo, específicamente aquellos de la librería scikit-learn.



RESULTADOS.

El mínimo RMSE que se ha obtenido fue de 5.400kg en test, y 4.800kg en train. Este modelo obtuvo un RMSE de 4.700kg en la entrega intermedia de 2022.

El mejor modelo predictivo consiste de un pipeline cuyo predictor fue un **Random Forest Regressor**, con sus hiper parámetros optimizados y diversos **Encoders** como, por ejemplo, **StandardScaler** o **KBinsDiscretizer**.

El modelo final fue re-entrenado con la totalidad de Train antes de predecir 2022.

- Para su elección se tuvo en cuenta tanto su resultado en Test como en Train, evitando un sobreajuste.

Notamos que a los modelos de predicción les costaba estimar correctamente los valores altos de producción, por lo que intentamos evitar esto creando las variables que estén relacionadas con la producción en sí.

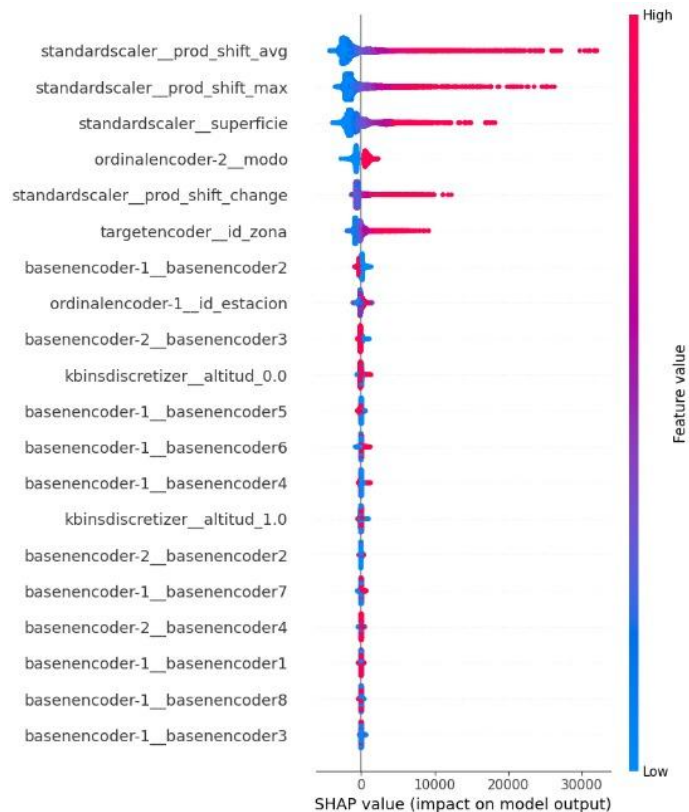
- Sin embargo, hay cambios bruscos de año a año en la producción que no se deben a cambios en la superficie, sino a otras latentes (o variables climáticas cuya relación el modelo no concibe).




Importancia de variables

Las **variables más importantes** para el modelo, como muestra el gráfico a la derecha, son aquellas que enseñan los valores pasados de la producción por finca, variedad y modo (el promedio de sus últimos años o el valor máximo).

Valores previos de producción **altos** tienen un efecto positivo en la predicción de la producción 2022 (mitad derecha).





GUÍA DE ENTREGA.

*Para que el trabajo realizado sea reproducible, se adjuntó un repositorio GitHub con scripts y una librería propia **Malbecs** con funciones de preprocesamiento y transformaciones.*

La entrega cuenta con dos scripts y con dos formatos de ejecución. El script de **exploración**, y el script de **predicción**. Los formatos de ejecución son: **notebooks** y **.py**.

- En el directorio de “./notebooks” se encuentra la entrega en formato .ipynb tanto para la exploración y análisis de datos como para el entrenamiento y predicción del modelo para esta primera fase.
- En el directorio “./scripts” se encuentran los scripts necesarios para la ejecución de las diferentes transformaciones de datos, el entrenamiento del modelo final seleccionado para esta fase, y la generación de predicciones.
 - *run_prep.py* con el preproceso y feature engineering.
 - *run_train.py* con validación y entrenamiento de modelo final.
 - *run_pred.py* con generación de predicciones para 2022
 - *run_all.py* para ejecutar todos los pasos end-to-end.

Se recomienda seguir los pasos en el README para una correcta ejecución.

Así mismo, verán un **txt adjunto** con las predicciones de 2022.



¡MUCHAS GRACIAS!

Denis Trosman | Vito Stamatti | Sumit Kumar Jethani Jethani