

DATATHON CAJAMAR 2023

Fase Nacional - Malbecs

Denis Trosman | Vito Stamatti | Sumit Kumar Jethani Jethani



Abril 2023

ÍNDICE

- [Key Points](#)
- [Objetivo](#)
- [Análisis exploratorio e ingeniería de variables](#)
- [Modelos predictivos](#)
- [Resultados](#)
- [Modelización responsable](#)
- [Guía de la entrega](#)



KEY POINTS.

Logramos construir un **modelo predictivo** para la producción de vino que puede ayudar a la bodega a **maximizar la eficiencia de producción, planificación y comercialización** de su producto.

- El modelo seleccionado otorgó un RMSE de **5,280kg** en 2021, y **4,750kg** en 2022. Para contrastar, un modelo baseline de regresión lineal con las variables tal fueron entregadas obtuvo 7,200kg de RMSE para 2021.
- El modelo analiza tanto **datos históricos** de la bodega como **variables climáticas**.

Nuestro modelo es **transparente, explicable y sustentable**

- **Optimizamos la ejecución y transparencia** creando una librería con funciones propias, listas para producción.
- El script corre en local en tan solo **20 segundos**, sin GPU y utilizando los mínimos recursos posibles.
- Escogimos un algoritmo cuya base son árboles, otorgando una **explicabilidad amigable**, a diferencia de otros modelos de caja negra.

El modelo **no discrimina según el identificador de las fincas**, y presenta predicciones justas

- No utilizamos el id de la finca como variable explicativa del modelo, siendo justos con nuevas fincas y evitando sobre-ajustes.
- En cambio, nos enfocamos en variables explicativas que enseñan producciones hipotéticas según características como su superficie, evitando la discriminación entre fincas altamente productivas y otras pequeñas.



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	1.223.660 x 34	Clima por hora
ETO	51.180 x 275	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 continuas que puedan ayudar a mejorar la performance del modelo.
 - Por ejemplo, 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.
 - Así mismo, para no generar dependencia a estos shifts, se han generado producciones hipotéticas según los tamaños de cada finca, teniendo en cuenta otras características como la zona, variedad y el modo.
- 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 no consideramos este año en los modelos finales.
- 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.
- La mejor forma de incorporar el clima a nuestro modelo resultó ser a través de flags que indican **variaciones mayores a un threshold** medido en desvíos estándares, y a través de **análisis de componentes principales**.



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
- Cat Boost, 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 y otras funciones propias.



RESULTADOS.

El mínimo RMSE que se ha obtenido fue de 5.300kg en test, y 4.800kg en train. Este modelo obtuvo un RMSE de 4.750kg 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** u **OrdinalEncoder**.

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.
- El modelo otorgó un resultado de 5,280kg para 2021, y de 5,580kg en promedio de validación cruzada para 2019-2021.

La elección del modelo se basó tanto en resultados como en eficiencia, optimizando sus hiper parámetros.

- Al estar basado en árboles, Random Forest otorga una explicabilidad relativamente simple, a diferencia de, por ejemplo, redes neuronales.
- De la misma manera, este algoritmo no presenta un desafío en términos de consumo computacional/energético (ver [detalle](#)).
- En términos de resultados, no solo el modelo tiene un menor error que el resto, sino que no sobreajusta y es justo en sus predicciones. Así mismo, las predicciones tienen sentido para casos específicos, como tamaño de fincas.



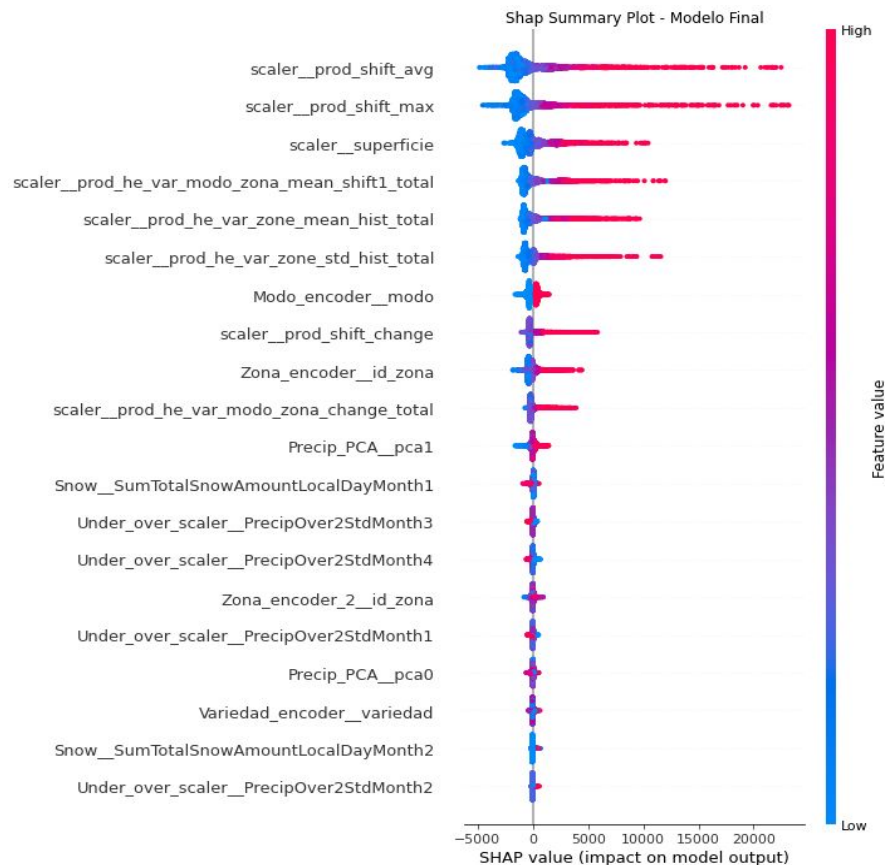
Importancia de variables

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).

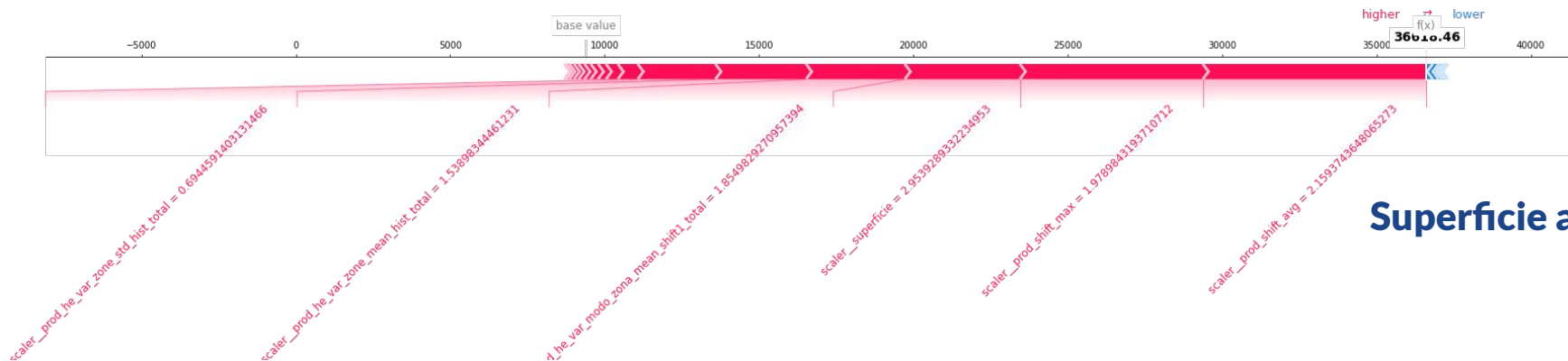
- Le sigue la superficie y aquellas que calculan la producción hipotética según las características de la finca

Sin embargo, cuando vemos las importancias para fincas con superficies bajas, las producciones hipotéticas y el modo aumentan su importancia.

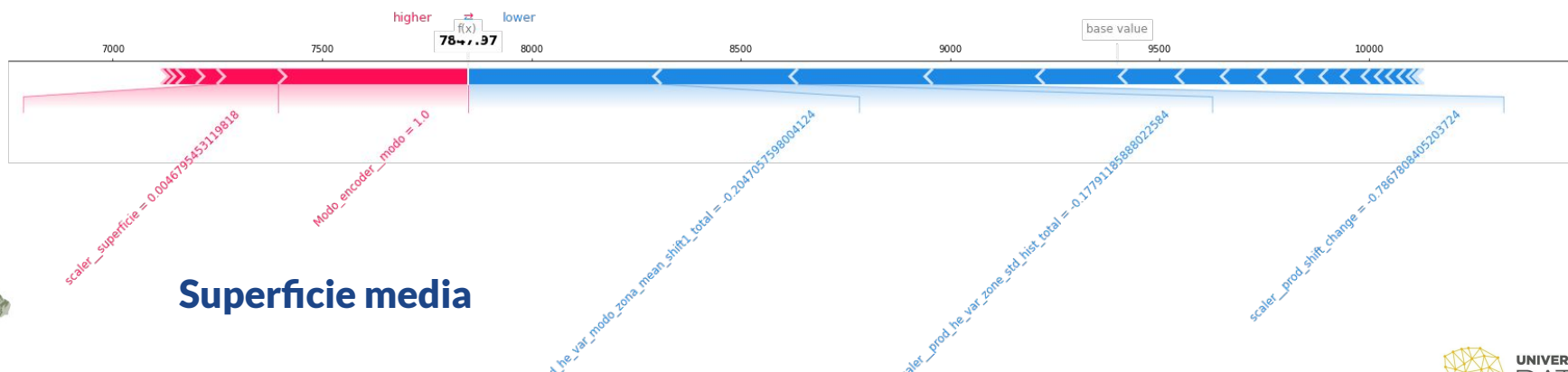
Valores previos de producción **altos** tienen un efecto positivo en la predicción de la producción 2022 (mitad derecha).



Importancia de variables



Superficie alta



Superficie media

MODELIZACIÓN RESPONSABLE.

Construyendo un modelo amigable y eficiente.

Explicabilidad

- A lo largo de la entrega realizamos análisis de los resultados del modelo, tanto de sus predicciones como de sus errores, y de las importancias de las variables explicativas.
- No solo revisamos el valor general, sino también los resultados en casos específicos.
- Para reforzar este punto, también creamos un dashboard en [PowerBI](#) para analizar caso a caso.

Transparencia

- El código presenta comentarios en su totalidad.
- Todas las funciones creadas tienen su documentación propia, en el siguiente [link](#).
- Se detallaron las instrucciones de ejecución en el ReadMe.



MODELIZACIÓN RESPONSABLE.

Construyendo un modelo amigable y eficiente.

Justicia

- Para evaluar el modelo, no tuvimos en cuenta un único set de validación, sino que creamos nuestra validación cruzada para evitar sesgos por campaña, asegurándonos que las distribuciones en estas sean similares.
- Consideramos que nuestro modelo es justo ya que **no discrimina por identificador de la finca**.
- Esto evita sobreajustes y ayuda a mejorar la performance para fincas nuevas en el viñedo.

Sostenibilidad ambiental

- El modelo corre eficientemente, a una velocidad de 22 segundos en local, sin usar GPU ni sobrecargar recursos.
- Esto es contando la lectura de los archivos y sus pre-procesamientos. El modelo en sí, sólo 1 segundo.
- A lo largo de los archivos .py, se dio la opción de visualizar el uso computacional con **memory_profiler**.



Dashboard

Análisis de errores

En esta diapositiva, presentamos la oportunidad de visualizar la comparación de las predicciones del modelo con los valores originales de la producción, también por fincas particulares. A partir de los botones en la parte derecha, pueden cambiar los dos gráficos de esa parte de la diapositiva, logrando ver los errores absolutos y porcentuales según las variables deseadas.

Producción - Predicción vs Original



Seleccionar variable de filtro
(presionar Ctrl y clicar en el elegido)

Seleccionar finca

All

Zona

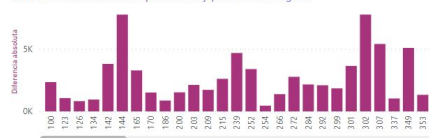
Variedad

Modo

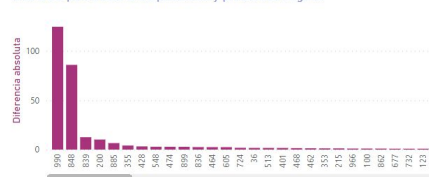
Tipo

Color

Diferencia absoluta entre predicción y producción original



Diferencia porcentual entre predicción y producción original



Análisis de producción

Seleccionar finca

All

Seleccionar zona

All

Seleccionar variedad

All

Seleccionar modo

All

En esta diapositiva presentamos los resultados de producción según campaña. Para un análisis detallado, se puede filtrar por finca, zona, variedad y modo.

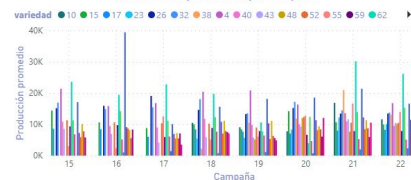
Suma de producción, por campaña



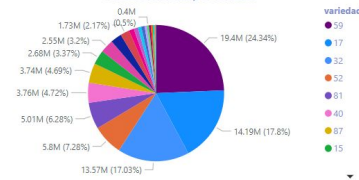
Promedio de producción, por campaña




Promedio de producción, por campaña



Producción total por variedad



UNIVERSITYHACK 2023
DATAATHON



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 archivo **Malbecs.txt adjunto** con las predicciones de 2022.





¡MUCHAS GRACIAS!

Denis Trosman | Vito Stamatti | Sumit Kumar Jethani Jethani