UniversityHack 2023 Datathon
Equipo Sarobe - Universidad de Navarra

Diego García Vicente, Diego de Lemos Burgaña, Pedro Sarobe Feijóo

En la realización de este proyecto buscamos proyectar mediante un modelo la producción de una campaña agrícola, con el objetivo de optimizar todos los procesos que se requieren en una cosecha de vino. Para llevar a cabo esto, tomaremos en cuenta los datasets proporcionados de los datos históricos de la cooperativa La Viña. Estos incluyen la producción de los viñedos y los factores meteorológicos que podrían afectar la calidad de una cosecha. Contamos con 3 datasets distintos, Train, Meteo y Eto. Train contiene la información histórica de las fincas dentro de la cooperativa. Meteo incluye datos meteorológicos de varias estaciones climatológicas de la zona. Por último, Eto dispone de información meteorológica detallada y alterada de las mismas estaciones distribuidas en secciones del día.

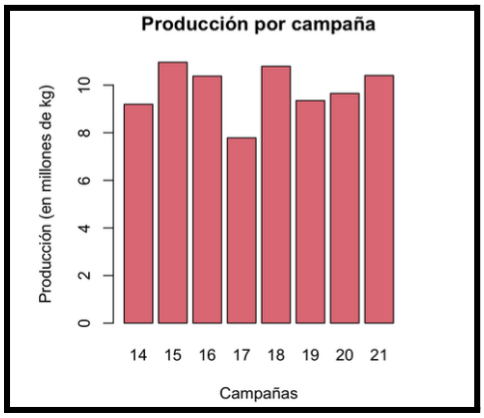
Al insertar los datos en R, exploramos las distintas variables y así entender qué información aporta cada una. Cabe recalcar que los datos meteorológicos entre julio y enero no se pueden tomar en cuenta para estimar nuestro modelo. A su vez, hay datos que no están completos, como es la superficie de las fincas, que solo está visible para los años de cosecha entre 2020 y 2022. El objetivo del modelo será proyectar la producción de la campaña 2022, usando todos los factores que afectan la cosecha.

Dado que disponemos de tres bases de datos distintas, realizamos el análisis exploratorio de los datos por separado. Para empezar tomamos en cuenta los datos de Train, donde vemos que el dataset contiene la siguiente estructura: 9,601 filas y 11 columnas. En este dataset, hay datos de 9 campañas, provenientes de los años 2014 hasta 2022. Las producciones del 2022 son las que hay que predecir, así que no se toman en cuenta los datos de esa campaña. Como mencionamos anteriormente, no hay datos de la producción de cada finca para todas las campañas.

Comenzamos por ver el número de observaciones de cada campaña, la cual es distinta para cada temporada, pero vemos que no hay mucha diferencia entre las campañas.

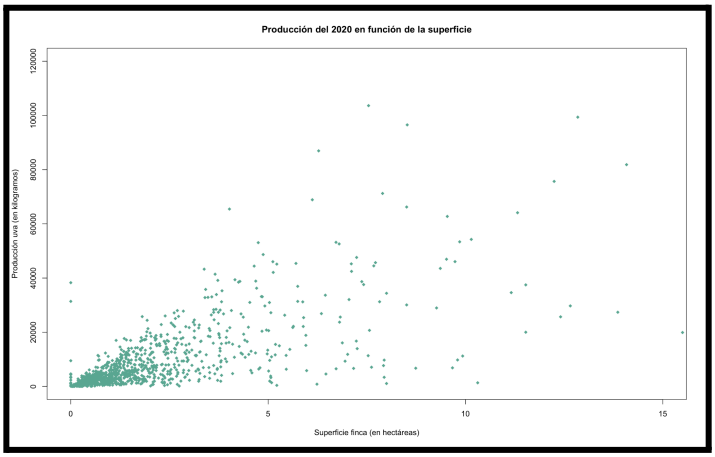
14	15	16	17	18	19	20	21	22
1148	1116	1079	1017	1061	1055	1006	1044	1075

Entendemos que la clave primaria de los datasets debería ser la variable ID Estación, ya que está incluida en las 3 bases de datos. Podemos enumerarlas, viendo que hay 20 estaciones que recogen datos climáticos, de los cuales nos interesa saber la ubicación de cada una de ellas. Con ID Finca, vemos que hay 1,231 fincas distintas en Train, pero para el objetivo final debemos estimar la producción de 1,075 fincas. A su vez, hay 125 zonas con una tipología de suelo común distintas, agrupadas en columna ID Zona. Esto es normal, ya que probablemente los datos de las fincas corresponden a fincas ubicadas por todo territorio español.



Para la producción sacamos un gráfico en el que se ve la producción sacada en millones de kg, seccionadas por campaña. Dentro de la producción, debemos especificar que hay 25 variaciones de uvas, lo cual es un factor para tomar en cuenta a al hora de estimar la producción de una finca. Dentro de las distintas variedades de uvas, solo hay dos métodos de cultivo. Esto se repite en el caso del color, donde contemplamos que solo hay uvas tintas y blancas.

La altitud media de las fincas sobre el nivel del mar es dada dentro de un rango, por lo que hemos cambiado a la media del rango, ya que posteriormente será necesario si queremos hacer un modelo de predicción que incluya esta variable. La variable que menos sentido nos hace es superficie, ya que es imposible que fincas con una superficie de menos de una hectárea tenga una producción de uva tan elevada. Hicimos un gráfico que muestra los kilogramos de uvas producidas en relación con la superficie de la finca de la campaña 2020.



En el dataset de Meteo nos pareció sumamente peculiar que había demasiados N/A en las columnas. Es entendible dado que por ejemplo, las fincas no están ubicadas a la misma altura o en la misma comunidad autónoma, lo que significa que puede estar lloviendo en algunas fincas mientras que en otras hay sequía. Eliminamos variables que no nos parecían relevantes para calcular la producción de una campaña, como lo son la visibilidad, dirección del viento, la velocidad máxima de las rafagas de viento e incluso la variación máxima en la presión atmosférica.

Creamos un nuevo dataset llamado Test, donde incluimos solo los datos de la campaña 2022, los cuales serán proyectados con nuestro modelo. Estos datos ascienden a 1,075 campañas. Una vez separado el rango de altitud y cambiados los formatos de las variables, analizamos aquellas variables que son dummy. Empezamos a evaluar aquellas variables que corresponden a sólo dos metodologías como son modo, tipo o color. Dentro del código del modo de cultivo podemos ver que hay 709 campañas que siguen el modo 2 y 366 del modo 1. En este dataset prácticamente no hay N/A's aparte de la sección de producción, lo que ayuda a visualizar los datos de una forma más efectiva.

En el último dataset proporcionado, Eto, hemos visto que hay unas variables bastante interesantes a analizar. Estas variables son mucho más específicas que las de Meteo, y que afectan de distintas maneras la efectividad de una campaña durante el año. Variables que no conocíamos y se nos complicó interpretar, como lo son Evapotranspiration, y la radiación solar recibida de forma horizontal. Nos pareció curioso que hubiera tantas variables diferentes, ya que todos tienen una métrica de medición muy diferente.

Para la limpieza de los datos y el modelo de proyección vimos más conveniente utilizar Python. Comenzamos el merge verificando que no hubieran duplicados y contabilizando el número de N/A's por tabla. Para el merge usamos left join, agrupando las tablas por el ID Estación, la cual consideramos como la clave primaria de las tablas, y por año/campaña. Todas las variables que tenían mínimo/máximo fueron transformadas a promedio diario, mensual o anual, ya que las diferencias entre el mínimo y el máximo eran leves y entendimos que no era razonable tener dos variables parecidas entre una y otra. Una vez agrupadas las dos primeras tablas de Train y Meteo, hicimos una limpieza en Eto para tener el mismo número de variables y así tener un relación ideal entre las tres tablas.

Para el modelo de proyección realizamos tres algoritmos distintos, para ver cual era el más eficiente o el que más se asemejaba a lo que buscábamos calcular. Empezamos por una regresión lineal, que terminó dándonos un R^2 de 0.725, lo que indica que los números se acercan bastante a la realidad. El error cuadrático medio que sacamos utilizando el número de variables correspondientes, se estimó un valor de 6956.243.

El segundo algoritmo que nos pareció correcto aplicar fue el Gradient Boosting Regressor. Este estimador construye un modelo aditivo en una forma avanzada por etapas; permite la optimización de funciones de pérdida diferenciables arbitrarias. En cada etapa se ajusta un árbol de regresión sobre el gradiente negativo de la función de pérdida dada, lo que nos termina indicando el valor final de producción, el cual era nuestro objetivo. El resultado de este modelo fue un R^2 de 0.8912 con error cuadrático medio de 4376.06.

Por último, aplicamos el modelo de regresión Random Forest, que ajusta una serie de árboles de decisión de clasificación en varias submuestras del conjunto de datos y utiliza el promedio para mejorar la precisión predictiva y controlar el sobreajuste. Esta estimación resultó en un valor de R^2 de 0.7704 con un error cuadrático medio de 6357.990.

Para terminar con nuestro modelo, donde proyectamos la producción de las distintas campañas, comparamos los resultados provenientes de los tres algoritmos de regresión usados anteriormente. Resulta que el método más efectivo para calcular la producción es el Gradient Boosting Regressor, debido a que su R^2 es el más elevado y esto indica que se aproxima más a la producción real de cada campaña.