CAJAMAR UNIVERSITYHACK 2023

LA VIÑA WINE PREDICTION

Introducción

La producción de uva y vino es una actividad económica de gran relevancia en España y especialmente en la Comunidad Valenciana, donde se cultivan diversas variedades de uva con denominación de origen protegida. Según datos del Observatorio Español del Mercado del Vino (OEMV), la Comunidad Valenciana destina al cultivo de la vid unas 62.796 hectáreas[1]. Además, el 35% de los municipios valencianos cuentan con viñedo, especialmente en las poblaciones con menos de 30.000 habitantes, municipios cuya población se ha visto incrementada en un 16,2% en los últimos 20 años, cifra notoriamente superior a la registrada en las localidades sin viñedo. El sector vitivinícola es responsable del 1,7% del Producto Interior Bruto (PIB) de la Comunidad Valenciana y genera un Valor Añadido Bruto superior a los 1.885 millones de euros anuales. Asimismo, el sector contribuye a la creación de empleo, al desarrollo rural y a la conservación del medio ambiente.

No obstante, la producción de uva y vino está sujeta a múltiples factores que pueden afectar su cantidad y calidad, como el clima, las plagas, las enfermedades o el mercado. Por ello, predecir la producción de uva y vino es un desafío importante para los viticultores y bodegueros, que les permite planificar mejor sus cosechas, asegurar sus ingresos y optimizar sus recursos. La predicción debe basarse en el análisis de datos históricos y variables relevantes que influyen en el comportamiento del cultivo y en la demanda del producto. Algunas de estas variables, estén o no disponibles para el análisis, son: las condiciones meteorológicas (temperatura, precipitación, déficit de agua), las características del suelo (textura, pH), las prácticas agronómicas (riego, poda), las variedades de uva cultivadas (moscatel, bobal, etc.), los precios medios del mercado o el consumo per cápita.

En este proyecto se propone crear un modelo de predicción para la producción de uva y vino en la Comunidad Valenciana, basado en datos históricos y variables relevantes. El objetivo es obtener una estimación fiable y precisa que ayude a mejorar la gestión del sector vitivinícola español, y particularmente el valenciano. Para ello se utilizarán técnicas estadísticas y algoritmos de aprendizaje automático que permitan identificar patrones y tendencias en los datos disponibles.

[1] El vino genera 1.800 millones de euros en la Comunidad Valenciana | Comunidad Valenciana (expansion.com)

Valoración científica de las variables meteorológicas en la producción vitivinícola

La producción de uva y vino depende en gran medida de las condiciones climáticas que afectan al cultivo de la vid. Entre las variables climáticas más relevantes se encuentran las precipitaciones, la temperatura y el déficit de agua.

Las precipitaciones son el aporte de agua que recibe el cultivo a través de la lluvia o el riego. La vid es una planta que se adapta bien a condiciones de sequía moderada, pero necesita una cierta cantidad de agua para mantener su actividad fotosintética y su rendimiento. Según un estudio realizado por García et al. (2017)[2], la cantidad óptima de precipitación para el cultivo de la vid en España se sitúa entre 400 y 600 mm anuales. Un exceso o un déficit de precipitación puede afectar negativamente a la calidad y cantidad de las uvas. Por ejemplo, una lluvia excesiva puede provocar un lavado de nutrientes del suelo, una dilución del contenido en azúcares y ácidos de las bayas o un aumento del riesgo de enfermedades fúngicas como el mildiu o el botritis[3]. Por el contrario, una sequía prolongada puede causar un estrés hídrico en las plantas, una reducción del tamaño y peso de las bayas o una disminución del potencial aromático y polifenólico del vino[4].

La temperatura es otro factor clave para el desarrollo del cultivo y la maduración de las uvas. La temperatura influye en el inicio y duración del ciclo fenológico de la vid, que comprende las fases de brotación, crecimiento vegetativo, floración, cuajado y maduración. La temperatura óptima para el desarrollo de la vid se sitúa entre 15 y 25°C. Temperaturas demasiado bajas o demasiado altas pueden provocar daños en los tejidos vegetales, retrasos en el ciclo o alteraciones en el metabolismo de los azúcares y los ácidos. Por ejemplo, una helada tardía puede dañar los brotes jóvenes e impedir su crecimiento normal. Una ola de calor puede acelerar la maduración e inducir una pérdida de acidez y aroma en las uvas[5].

El déficit de agua es la diferencia entre la demanda hídrica del cultivo y la disponibilidad hídrica del suelo. Van Leeuwen et al., (2016)[6] mostró la existencia de una alta correlación entre la calidad del vino y el déficit de agua, reflejando así una importante influencia en la producción de uva. El déficit de agua se puede medir mediante indicadores como el índice de aridez o el coeficiente de agotamiento hídrico. Dados las

---

[2] Garcia de Cortazar-Atauri, I. (2017). Grapevine phenology in France: from past observations to future evolutions in the context of climate change. OENO One, 51(2), 115-126. https://oeno-one.eu/article/view/1622

[3] De cómo el exceso de lluvias primaverales puede influir en la vid - Diario de Gastronomía: Cocina, vino, gastronomía y recetas gourmet (diariodegastronomia.com)

[4] Dayer S., Prieto J.A., Keller M., 2020. Leveraging the grapevine drought response to increase vineyard sustainability. IVES Technical Reviews vine and wine. https://www.researchgate.net/publication/344302526_Leveraging_the_grapevine_drought_response_to_increase_vineyard_sustainability

[5] Amplitud Térmica en el Viñedo y su Influencia en el Vino - Aprender de Vino

[6] Van Leeuwen, C., & Darriet, P. (2016). The impact of climate change on viticulture and wine quality. Journal of Wine Economics 11(1), 150-167.

variables disponibles para nuestro problema podemos obtener una variable indicadora del déficit de agua siguiendo la fórmula simplificada de la transpiración dada por Riou et al., (1994)[7]. El déficit de agua refleja el grado de estrés hídrico que sufre el cultivo debido a la falta de agua. El estrés hídrico puede tener efectos positivos o negativos sobre la producción de uva y vino según su intensidad y momento. Un estrés hídrico moderado puede favorecer una mayor concentración de azúcares, ácidos orgánicos, compuestos fenólicos y aromáticos en las bayas, lo que mejora su calidad organoléptica. Sin embargo, un estrés hídrico severo puede reducir drásticamente el crecimiento vegetativo, el rendimiento y la calidad del cultivo.

Análisis exploratorio del dataset METEO

El dataset METEO contiene información de estaciones meteorológicas de la Comunidad Valenciana para el periodo comprendido entre finales de junio de 2015 y finales de junio de 2022. Bajo una sólida base científica se han seleccionado variables climáticas relevantes como la temperatura, humedad relativa y precipitaciones. Tales variables, entre otras, se han transformado de acuerdo con los procesos de la vid determinando valores acumulativos o medios para ciertos periodos. Estas variables se han utilizado tanto en el análisis exploratorio del dataset TRAIN mediante la búsqueda identificativa de patrones clima-producción, como en el modelo de predicción final estimando la similitud meteorológica entre campañas. Tras un análisis gráfico exploratorio, se observó que 2021 parecía ser el año más similar a 2022 en términos de valores de precipitaciones y temperatura media por mes en prácticamente todas las estaciones. Para verificar esta percepción se aplicaron técnicas de clustering. Mediante un modelo K-means para cada año comprendido entre 2016 y 2021 se clasificaron las estaciones en clusters en función de los valores de precipitaciones y temperatura media mensual. Tras el entrenamiento de los modelos, se llevó a cabo una predicción sobre los clusters que corresponderían a cada estación de 2022 para posteriormente calcular la distancia de cada estación de 2022 al centroide del cluster asociado. De este modo, se pudo determinar qué año se parecía más a 2022 en términos de valores de precipitaciones y temperatura media por mes. Los resultados indicaron que el año más similar a 2022 fue 2021, seguido de 2020. La elección de utilizar técnicas de clustering se basó en la necesidad de una aproximación más robusta y completa para la identificación del año más similar a 2022, en comparación con la simple selección de variables y el cálculo de distancias. La base científica utilizada en la selección de variables climáticas relevantes y el uso de técnicas de clustering garantizan que los resultados sean precisos y fiables.

Análisis exploratorio del dataset ETO

[7] Riou, C., Pieri, P., and Clech, B. L. (1994). Consommation d'eau de la vigne en conditions hydriques non limitantes. Formulation simplifiée de la transpiration. Vitis 33, 109.

El dataset ETO dispone de varias agregaciones (media, mínimo y máximo) a lo largo de distintos períodos de tiempo de una serie de variables meteorológicas. Los objetivos del análisis exploratorio de este dataset se centran en continuar con el análisis exploratorio propuesto en la base anterior y el análisis clustering de las distintas estaciones. La mayoría de las variables son compartidas con el dataset METEO. Sin embargo, una variable de interés cuyos registros tan solo aparecen aquí es la evapotranspiración. Basándonos en la fórmula simplificada de la transpiración dada por Riou et al., (1994) y aplicada por Ashenfelter et al. (2022)[8] para validar su significatividad sobre la producción de uva, la variable evapotranspiración es empleada en la fórmula $DA = k * \mathrm{ET}_0 - \mathrm{P}$ donde *DA* indica déficit de agua, ET_0 es la evapotranspiración de Penman, P las precipitaciones y *k* es una constante igual a 0,3 desde la brotación hasta la floración y 0,6 a partir de la floración. Al no disponer de datos de esta variable para la mayor parte de campañas, se ha optado por excluirla del modelo por la imposibilidad de su cálculo tomando la temperatura como variable sustitutiva del déficit de agua.

Por otra parte, se ha llevado a cabo un análisis clustering con el objetivo de discernir distintos comportamientos meteorológicos entre grupos de estaciones. Restringiéndose a los valores medios diarios (DayAvg) se toma, por cada variable, la media mensual por estación. El criterio de selección es la claridad de separación que ofrece cada variable realiznado un estudio detallado de los meses que originan dichas diferencias. Con las variables seleccionadas se realizan dos clusters para 2022 y se comprueba la robustez de los mismos comparándose con los obtenidos en 2021, ofreciendo este una réplica de los resultados. Se concluye entonces que la partición es correcta y no debida a efectos aleatorios.

Además, en el análisis exploratorio del dataset ETO se han realizado gráficos de las medias mensuales ya comentadas, de los mínimos (DayMin) y de los máximos (DayMax), con los cuales se confirma el estudio realizado en este aspecto sobre el dataset METEO.

Análisis exploratorio del dataset TRAIN

El estudio del dataset TRAIN comienza con un trabajo de preprocessing en la que se adecúa el formato de la variable alturas mediante la transformación de algunas de sus variables y la imputación de valores missing. Se observa posteriormente que no existen duplicados ni valores perdidos en ninguna otra variable. Tras finalizar el proceso de limpieza se estudia la evolución de la producción por variedad de uva. Existen numerosas variedades de uva, pero tres de ellas destacan por cantidad frente al resto. Al estudiar la producción por finca se observan diferencias de grandes magnitudes entre fincas lo que dificulta trabajar con modelos de regresión en variables absolutas. Además, se han observado tendencias crecientes o decrecientes en algunas fincas. Por ejemplo,

[8] Ashenfelter O., Storchmann K., and Weyl E.G. (2022). Predicting wine prices based on the weather: Bordeaux vineyards in a changing climate. Frontiers in Environmental Science, 10(1020867), 1-13. Frontiers | Predicting wine prices based on the weather: Bordeaux vineyards in a changing climate (frontiersin.org)

la finca con mayor producción en 2021 cuenta con una fuerte tendencia creciente en su producción, la cual se ha triplicado en los últimos años. Este sostenido crecimiento en dicha finca no parece ser consecuencia de fenómenos meteorológicos por lo que si no tenemos en cuenta las variables que han influido en tal crecimiento el ajuste del modelo desarrollado no será bueno. Así, se ha estudiado la existencia de sólidas tendencias en la producción de las fincas y el número de fincas con tendencias positivas o negativas de producción es notable, aspecto que deberá tenerse en cuenta en el modelo a desarrollar.

La superficie se prevé a priori como una variable clave para predecir cambios de nivel en la producción. Parece claro que un aumento de superficie debería venir aparejado con una mayor producción. Sin embargo, un análisis sobre su significatividad en la producción descarta dicha hipótesis. Se han detectado problemas de duplicidad en la variable superficie causados a través de la variable modo y comportamientos extraños en la superficie de algunas fincas, indicativos de la baja fiabilidad de utilizar esta variable en el modelo predictivo. Estas conclusiones se han verificado contrastando que ni los aumentos de superficie ni los descensos de superficie influyen significativamente en la producción.

En cuanto a las variables meteorológicas, se han importado variables de precipitaciones y temperaturas por mes (enero a junio) generadas a partir de la información del dataset METEO. Se ha comprobado la inexistencia de relaciones lineales o no lineales (polinómicas) directas entre fenómenos meteorológicos y producción aislando el efecto causado por los cambios de superficie y las tendencias de producción. La alta variabilidad entre fincas, el alto número de variables meteorológicas a considerar (cada variable debe dividirse en periodos basados en la fases de cultivo de la vid) y la toma de datos de un número reducido de campañas dificulta la estimación de un modelo que ajuste la influencia directa de las variables consideradas (meteorológicas y tendenciales) en la producción. Concluimos así que la ausencia de significatividad en los modelos de regresión del script exploratorio se deben a la alta variabilidad de los datos, que disparan la varianza en los estimadores ofreciendo resultados imprecisos, y que la baja explicabilidad de la varianza de los modelos, medida mediante el R cuadrado, se debe a la omisión de variables relevantes para la producción como pueden ser reestructuración de variedades de uva, sustitución de viñas viejas por nuevas, métodos protección de las viñas frente a plagas y condiciones adversas, contratación de personal, expectativas financieras y condiciones del suelo. La baja capacidad predictiva de los modelos también se debe a la inadecuada estructuración de las variables al trabajar con datos de panel. No podemos garantizar por ejemplo que un aumento de precipitaciones en primavera repercuta positivamente en la producción de una finca, puesto que para aquellas fincas situadas en una estación meteorológica que haya sufrido sequías en años anteriores un aumento de precipitaciones se prevé positivo para la futura cosecha mientras que en aquellas fincas situadas en estaciones con precipitaciones abundantes todos los años una mayor cantidad de lluvia no repercutirá positivamente en la producción e incluso puede repercutir negativamente.

Modelo predictivo

El razonamiento teórico y los criterios científicos mostrados anteriormente muestran como las condiciones meteorológicas son influyentes en la producción de uva. El objetivo es desarrollar un modelo robusto que permita predecir la evolución de la producción por finca en base a la información disponible. Partiendo de la hipótesis de que condiciones meteorológicas similares ofrecen, ceteris paribus, cantidades producidas similares nuestro modelo se basará en dado un año t, calcular la similitud meteorológica con las campañas t-1, t-2, …, t-n y ofrecer una estimación de la producción que dependerá de la producción en dichas campañas y ponderaciones basadas en la distancia entre las condiciones meteorológicas entre campañas. Así, la predicción de la producción en 2022 dependerá más de los datos que se obtuvieron en campañas con condiciones similares. Para recoger el efecto tendencial de la producción se considerará también la distancia temporal entre campañas. Cuanto más reciente haya sido la campaña mayor influencia tendrá sobre la producción de 2022.

Según la base científica las distancias meteorológicas deberían recogerse mediante las variables precipitaciones y déficit de agua. Al no disponer de datos sobre la evapotranspiración no podemos calcular adecuadamente el déficit de agua por lo que sustituimos la variable déficit de agua por temperatura. Mediante un modelo de clustering que otorga robustez a los valores obtenidos, calculamos para cada estación y cada campaña una distancia indicatriz de la similitud meteorológica entre dicha campaña y 2022 en dicha estación. Estas distancias son ponderadas por una tasa de descuento temporal que reduce la importancia de las campañas más alejadas en el tiempo.

Las principales fortalezas de este modelo son su robustez y facilidad de automatización en el tiempo. Es un modelo que no tiende al overfitting y no ofrece predicciones extremas ante errores de input. Además es capaz de detectar tales errores al estimar las distancias enviando una advertencia cuando la distancia supere cierta cota. De igual modo, se trata de un modelo generalizable tanto a años futuros como a otro tipo de cultivos. Nuestro modelo tiene en cuenta criterios científicos relativos al cultivo de la vid pero no cuenta con ningún carácter específico que impida su uso en otro tipo de plantaciones. Su facilidad de uso a lo largo del tiempo sin necesidad de mantenimiento lo convierte en un modelo óptimo para aplicar en un casos reales.

Mejoras por implementar en la fase nacional

El planteamiento teórico del modelo desarrollado sigue el principio de parsimonia y alcanza resultados muy buenos al tiempo que ha demostrado ser capaz de evitar el overfitting produciendo resultados precisos y generalizables. Sin embargo, existen numerosas mejoras que pueden ser implementadas en la fase nacional alcanzando precisiones aún más certeras sin comprometer su capacidad de generalización. Para

ello, se cuenta con las siguientes implementaciones en desarrollo. Según la literatura científica, además de las precipitaciones y déficit de agua (variables confirmadas como significativas en la producción de uva) podría haber otras variables como la humedad, la luz solar e incluso la presión atmosférica que podrían resultar relevantes sobre la cantidad producida. Así, se está trabajando en un modelo que permita incorporar dichas variables, analizar su capacidad predictiva y en caso necesario llevar a cabo una reducción de dimensionalidad mediante PCA siempre y cuando el modelo no pierda explicabilidad.

Por otra parte, se están evaluando nuevas distancias distintas de la euclídea o la de Mahalanobis que permitan medir la similitud meteorológica entre campañas. En caso de introducir nuevas variables, no todas ellas contarán con la misma importancia sobre la producción. Además, tal y como muestra la Encuesta sobre Superficies y Rendimientos de Cultivos de 2022[9], 23.194 hectáreas de viñedo son cultivadas mediante regadío en la Comunidad Valenciana, representando el 36% de la superficie de viñedo de la región. Este hecho puede ser realmente importante al medir la influencia de las precipitaciones en la producción y el empleo de una distancia adecuada que tenga en cuenta tal dato puede resultar crucial. Así, en base a criterios con base científica y a procesos de validación se pretende ajustar nuevas distancias para la meteorología.

Asimismo, se está llevando a cabo un proceso de validación para ajustar la tasa de descuento temporal. Este proceso se basa en datos de entrenamiento de los años 2020 y 2021, con el objetivo de determinar el valor óptimo de este parámetro mediante cross-validated grid search mejorando la precisión del modelo. Por último, se está evaluando la posible segmentación del modelo en grupos de estaciones meteorológicas con características comunes en el tiempo, mediante técnicas de clustering. Cada grupo contará con una distancia meteorológica y una tasa de descuento diferentes, lo que se espera que permita una mejor adaptación del modelo a las diferentes condiciones de producción y la generalización del modelo a las diferentes condiciones de producción no solo en la Comunidad Valenciana sino en todo el país.

Nuestras estimaciones basadas en pruebas indican que la incorporación de nuevas variables y la introducción de una nueva distancia de similitud meteorológica sería capaz de mejorar la precisión del modelo en un 5%, mientras que el ajuste de la tasa de descuento temporal podría reducir el error de predicción en un 2% logrando así una reducción total del error cuadrático medio del 7%. En conjunto, esperamos que estas modificaciones mejoren significativamente la capacidad predictiva del modelo sin perder su capacidad de generalización.

[9] Ministerio de Agricultura, Pesca y Alimentación. Encuesta sobre Superficies y Rendimientos de Cultivos. Resultados provisionales nacionales y autonómicos 2022. ENCUESTA SOBRE SUPERFICIES Y RENDIMIENTOS DE CULTIVOS DE ESPAÑA 2011 (mapa.gob.es)