Aprendizaje Automático I

Calidad del tinto ‘vinho verde’

facultade de informática da coruña

La presente práctica ha sido realizada por los alumnos *Álvaro Sieira Rama* (alvaro.sieira.rama@udc.es), *Marcos Ramos Gómez* (m.rgomez@udc.es) y *Jorge Crespo Rivas* (j.crespo.rivas@udc.es), estudiantes del Grado en Ciencia e Ingeniería de Datos e integrantes del grupo de prácticas número 3 de la asignatura Aprendizaje Automático I.

Contents.

1. Introducción.

Considerado entre una de las bebidas alcohólicas más consumidas del mundo, y, siendo tras el agua la bebida fermentada más consumida en el globo, el vino se elabora a partir de la uva, sometiendo su mosto o zumo a una larga fermentación alcohólica de entre 1 y 2 semanas. Este constituye un elemento fundamental no solo en la cultura, sino también en la historia de numerosos países europeos, con evidencias de su elaboración que se remontan hasta el año 6000 a.C. en regiones como Georgia, Mesopotamia o el actual Irán. En las últimas décadas, la producción mundial de vino ha superado, de media, los 200 millones de hectolitros anuales, pese a haber experimentado un descenso en los últimos años. En este contexto, varios países europeos desempeñan un papel destacado, dedicando extensas superficies al cultivo de la vid, siendo España uno de los principales productores, con aproximadamente 950.000 hectáreas de viñedos.

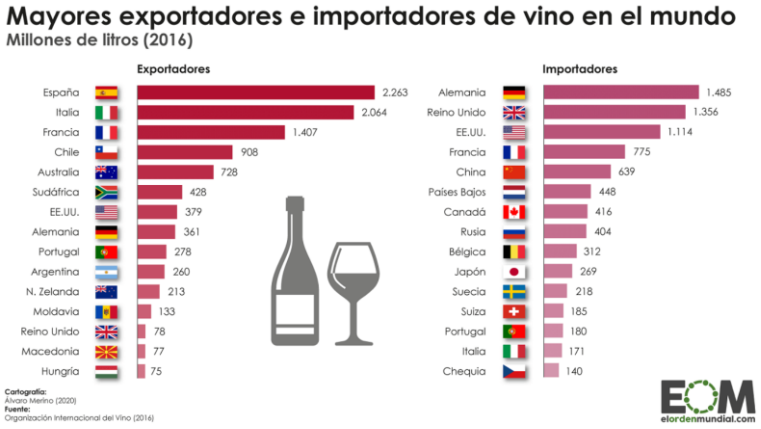

Figure 1: Principales exportadores de vino del mundo. [1]

La variedad de vino analizada, conocida como ‘vinho verde’, se caracteriza por su contenido moderado de alcohol y su sabor afrutado. Originaria de Portugal, esta variedad se produce en la región de Entre Douro e Minho, situada en la Costa Verde. Esta zona es reconocida por exportar vinos de alta calidad tanto a nivel europeo como mundial, y por contar con una notable presencia de pequeños productores.

La calidad sensorial de un vino se determinó tradicionalmente mediante catas a cargo de expertos, un proceso necesariamente subjetivo y costoso en tiempo y recursos. La heterogeneidad de las evaluaciones y la dependencia de la experiencia del catador subrayan la necesidad de métodos complementarios que aporten objetividad y rapidez. Asimismo, parámetros fisicoquímicos como el contenido de azúcares residuales, la acidez fija y volátil, el pH, los niveles de sulfatos y el grado alcohólico ejercen una influencia directa sobre los atributos de aroma, sabor y textura, pero su interpretación conjunta requiere un análisis de carácter más técnico como el abordado a lo largo de esta práctica y sus respectivos resultados.

En este escenario, el aprendizaje automático se presenta como una opción muy interesante para ayudar al enólogo a predecir de forma automática la calidad del vino. Al ser capaz

de manejar grandes volúmenes de datos y captar relaciones complejas entre variables, métodos como las redes neuronales (RR.NN.AA.), las máquinas de soporte vectorial (SVM), los árboles de decisión y el k-vecinos más cercanos (k-NN) han mostrado en estudios previos que pueden distinguir con gran precisión entre muestras de distinta calidad, alcanzando niveles de acierto elevados y ofreciendo un apoyo fiable para la toma de decisiones en la bodega.

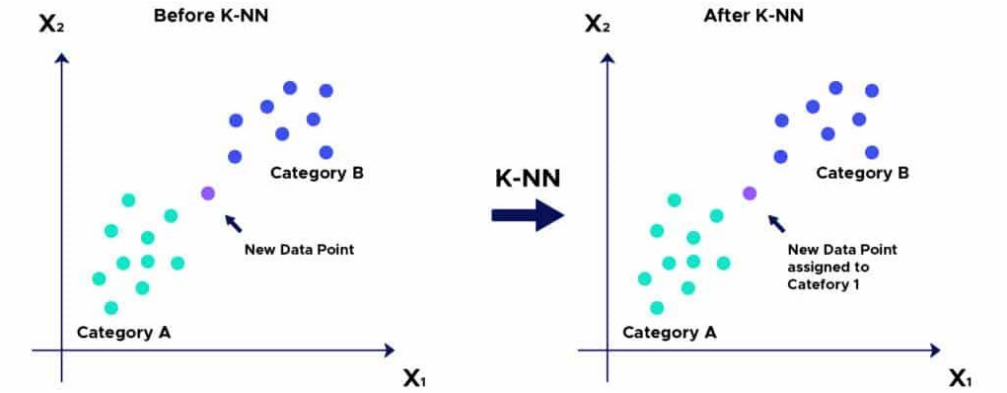

Figure 2: Método k-NN (k-vecinos más cercanos). [2]

2. Descripción del problema.

El problema que se pretende resolver es el de la clasificación de la calidad del vino utilizando métodos de aprendizaje automático. En este contexto, se busca desarrollar un modelo capaz de distinguir entre variantes de 'vinho verde' tinto según su puntuación (del 0 al 10, siendo esta la función objetivo o "target" de la base de datos) basándose en pruebas fisicoquímicas del vino.

Se recopilaron los siguientes datos sobre diversas características de las variantes de 'vinho verde': acidez fija, acidez volátil, ácido cítrico, azúcar residual, cloruros, el dióxido de azufre libre y el total, la densidad, el pH, los sulfatos y el alcohol. Estos datos serán la entrada del modelo de clasificación desarrollado basado en el aprendizaje automático, que servirán para entrenarlo y que, posteriormente, sea capaz de clasificar automáticamente la calidad de una variedad (nueva en la comercialización de la bodega o ya registrada) en base a sus características.

El objetivo final es desarrollar un modelo de clasificación preciso y robusto que pueda ayudar en la identificación y clasificación automatizada de variedades de 'vinho verde', lo cual podría tener aplicaciones en numerosos campos, tales como el control de calidad en bodegas, la optimización de lotes para exportación, el asesoramiento personalizado al consumidor y la investigación enológica para el desarrollo de nuevas variedades y procesos de producción.

2.1. Descripción de la Base de Datos.

La base de datos suministrada para abordar el problema de clasificación de las variedades de 'vinho verde' contiene un total de 4898 instancias. Cada uno de estos registros corresponde a una muestra individual descrita mediante múltiples atributos, junto con su correspondiente valor de nombre que indica la variedad a la que pertenece. En nuestro caso, al haber seleccionado exclusivamente las muestras de vino tinto, todos los datos analizados se refieren a esta variedad, a pesar de que el conjunto original también incluía un dataset con información relativa a variedades de vino blanco.

A continuación, se presenta una tabla que recopila todos los atributos analizados, junto con sus respectivas descripciones. Cabe destacar que ninguno de estos atributos presenta valores nulos y que todos son de naturaleza continua.

Table 1: Descripción de los atributos en la base de datos.

Atributo	**Descripción**
Volatile acidity	Acidez volátil de la variedad de vino en g/dm^3.
Residual sugar	Azúcares no fermentados restantes en g/dm^3.
Chlorides	Concentración de cloruros (sal) en g/dm^3.
Density	Densidad del vino en g/cm^3.
Sulphates	Sulfatos (generalmente sulfato de potasio) en g/dm^3.
Alcohol	Graduación alcohólica expresada en % vol.
Quality	Un número entre 0 y 10 que determina la calidad de la variedad.

2.2. Origen de la Base de Datos.

La base de datos fue obtenida de UC Irvine, Machine Learning Repository.

2.3. Propiedades a destacar de los atributos.

A continuación, se analizarán las características consideradas más relevantes, respaldando su selección mediante técnicas estadísticas. Este enfoque facilitará posteriormente un desarrollo más riguroso de la práctica y una comprensión más amplia del propósito de cada una de las etapas llevadas a cabo en el estudio.

2.3.1. Matriz de correlación.

La siguiente matriz de correlaciones solo contiene las variables que hemos seleccionado como útiles y, por tanto, las variables con las que hemos trabajado. Tras ella se procederá a ahondar en los resultados obtenidos.

	volatile.acidity	residual.sugar	chlorides	density	sulphates	alcohol
volatile.acidity	1.000000000	0.001917882	0.06129777	0.02202623	-0.260986685	-0.20228803
residual.sugar	0.001917882	1.000000000	0.05560954	0.35528337	0.005527121	0.04207544
chlorides	0.061297772	0.055609535	1.00000000	0.20063233	0.371260481	-0.22114054
density	0.022026232	0.355283371	0.20063233	1.00000000	0.148506412	-0.49617977
sulphates	-0.260986685	0.005527121	0.37126048	0.14850641	1.000000000	0.09359475
alcohol	-0.202288027	0.042075437	-0.22114054	-0.49617977	0.093594750	1.00000000

Figure 3: Matriz de correlaciones de atributos seleccionados.

Se han identificado varias correlaciones positivas de magnitud moderada entre dichas variables del tinto. Destaca, entre ellas, la relación entre la densidad y el azúcar residual, donde se observa que un mayor contenido de azúcar residual tiende a estar asociado con un ligero incremento en la densidad del vino. De forma similar, se aprecia una correlación positiva moderada entre la concentración de sulfatos y la de cloruros.

Por otro lado, también se evidencian correlaciones negativas relevantes. En particular, la relación entre la densidad y el contenido alcohólico presenta una correlación negativa moderada, con un coeficiente cercano a -0.5. Este comportamiento resulta coherente desde el punto de vista físico, ya que los vinos con mayor graduación alcohólica tienden a presentar una menor densidad, lo que puede explicarse por una menor proporción de agua u otros factores compositivos. Asimismo, se observan correlaciones negativas de menor intensidad (en el rango de -0.2 a -0.3) entre la acidez volátil y los sulfatos, así como entre la acidez volátil y el alcohol. Estos valores sugieren que, a medida que aumenta la acidez volátil, tienden a disminuir ligeramente tanto la concentración de sulfatos como el contenido alcohólico. De igual manera, se aprecia una correlación negativa entre los cloruros y el alcohol.

El resto de las correlaciones analizadas presentan valores próximos a cero, lo cual indica que, en términos generales, no existe una asociación lineal significativa entre variables como el azúcar residual, los sulfatos, la acidez volátil o el contenido alcohólico, más allá de los casos previamente mencionados.

2.3.2. Rangos intercuartílicos (IQR).

Tras calcular los rangos intercuartílicos, estos son los resultados obtenidos.

volatile.acidity	residual.sugar	chlorides	density	sulphates	alcohol
0.250000	0.700000	0.020000	0.002235	0.180000	1.600000

Figure 4: Rangos intercuartílicos de las variables.

Las variables que podrían estar acotadas son aquellas cuyo rango intercuartílico (IQR) es relativamente pequeño en comparación con su rango total de valores. Esto sugiere que la mayoría de los datos se concentran en un tramo estrecho, lo que indicaría una posible limitación en los valores que pueden tomar.

Basándonos en los IQRs obtenidos:

- volatile.acidity: IQR = 0.25. Muy reducido en relación con el rango global, lo que sugiere que esta variable podría estar acotada.
- chlorides: IQR = 0.02. Extremadamente pequeño, indicando alta concentración de valores y probable acotación.
- density: IQR ≈ 0.0022. Similarmente diminuto, apunta a un rango muy estrecho de densidades.
- residual.sugar: IQR = 0.70. Moderadamente pequeño; aunque hay bastante más variabilidad que en cloruros o densidad, sigue concentrado y podría considerarse acotado.
- sulphates: IQR = 0.10. Ligera dispersión, pero igualmente bastante reducida, lo que también sugiere acotación.
- alcohol: IQR = 1.90. Relativamente grande, denota estilos muy diversos (desde tintos ligeros hasta de alta graduación), por lo que no parece estar acotada.

2.3.3. Explicación de métricas empleadas.

Los estudios preliminares permitirán determinar las métricas más apropiadas para proceder durante el análisis de los datos. En particular, si un atributo presenta valores acotados dentro de un intervalo definido y estos valores se distribuyen de forma aproximadamente uniforme en dicho intervalo, ello sugiere que una estrategia adecuada de normalización sería la basada en la técnica de reescalado Min-Max.

Por ejemplo, si los valores de un atributo están comprendidos entre 0 y 1, y su distribución dentro del intervalo [0, 1] es relativamente uniforme, resulta razonable aplicar la normalización Min-Max. Esta técnica transforma los datos de manera que el valor mínimo se convierte en 0 y el valor máximo en 1, preservando las proporciones relativas entre los valores originales. Este enfoque resulta particularmente útil cuando se requiere comparar atributos que se encuentran en diferentes escalas o unidades, ya que garantiza que todos los valores transformados pertenezcan al mismo rango estándar, lo que favorece la homogeneidad en el análisis posterior.

Sin embargo, es importante resaltar que, aunque la estrategia de reescalado Min-Max ha demostrado ser la opción predeterminada para la mayoría de los métodos evaluados (garantizando que todos los atributos queden acotados en el intervalo [0, 1] y facilitando la comparabilidad entre variables de distinta escala), esta no resulta idónea para el clasificador SVC. En el caso del SVC, la normalización NormalizeZeroMean ha demostrado ser más adecuada, dado que centra cada atributo en torno a una media de cero y, en consecuencia, preserva la eficacia del cálculo de distancias y productos escalares inherentes al kernel. Por tanto, se adopta de forma explícita la normalización Zero-Mean para SVC, mientras que el resto de los métodos emplea el reescalado Min-Max.

3. Análisis bibliográfico.

El presente apartado ofrece un análisis pormenorizado de las investigaciones y manuales fundamentales que sustentan la metodología empleada en este estudio. A continuación,

se describen en detalle cada uno de los trabajos, manteniendo la integridad de las referencias originales y adoptando un tono académico riguroso.

En el manual de Barceló, J. G. (1990) *Técnicas analíticas para vinos* (pp. 31–49). Gab. [1], se examinan en detalle los procedimientos para determinar los principales parámetros de calidad del vino: en primer lugar, la distinción entre acidez total y acidez volátil, explicada a través de protocolos de titulación y criterios de calibración que garantizan mediciones reproducibles; a continuación, los métodos cromatográficos y espectrofotométricos empleados para cuantificar compuestos adicionados como sulfitos y azúcares residuales; y, finalmente, las técnicas de extracción sólido-líquido y la cromatografía de gases acoplada a espectrometría de masas para la detección de pesticidas, acompañadas de los parámetros de validación conforme a normativas internacionales.

Por su parte, Moya Anegón, F., Solana, V. H. y Bote, V. G. (1998) en *La aplicación de Redes Neuronales Artificiales (RNA) a la recuperación de la información* (Bibliodoc: anuari de biblioteconomia, documentació i informació, 147–164) [2], introducen las RNA como herramientas capaces de modelar relaciones no lineales en sistemas de recuperación documental. Tras contextualizar los sistemas algorítmicos tradicionales, describen arquitecturas de perceptrón multicapa, retropropagación y redes recurrentes, detallando funciones de activación y procesos de aprendizaje supervisado, y ejemplifican su aplicabilidad en clasificación temática y filtrado colaborativo para mejorar la eficiencia en la gestión de recursos bibliográficos.

Complementariamente, Valderrama, J. O. y Rojas, R. E. (2009) en *Avances en la predicción de propiedades físicas, fisicoquímicas y de transporte de líquidos iónicos* (Información tecnológica, 20(4), 149–160) [3], aplican RNA a la predicción de la solubilidad de SO_2 en ocho líquidos iónicos, describiendo la obtención de 155 datos experimentales de solubilidad (P-T-x) divididos en 131 para entrenamiento, 16 para validación y 8 para prueba. Analizan distintas configuraciones neuronales variando capas ocultas y neuronas, evaluando su desempeño mediante error cuadrático medio y coeficiente de correlación, y concluyen que las RNA capturan eficazmente las interacciones soluto-solvente, resultando de gran utilidad para diseñar procesos de absorción de gases contaminantes.

En el ámbito médico, Abdulhadi, N. y Al-Mousa, A. (2021, July) presentaron en la conferencia ICIT el trabajo *Diabetes detection using machine learning classification methods* (pp. 350–354). IEEE [4], donde comparan LDA, SVC y árboles de decisión para el diagnóstico de diabetes a partir de variables clínicas como glucemia y presión arterial. Empleando validación cruzada k-fold, evalúan precisión, sensibilidad, especificidad y área bajo la curva ROC, destacando el SVC con núcleo RBF por su superior sensibilidad en la detección de casos positivos, lo que sugiere su integración en sistemas de apoyo diagnóstico.

Asimismo, Lakshmi, K. D., Panigrahi, P. K. y Goli, R. K. (2022) en *Machine learning assessment of IoT managed microgrid protection in existence of SVC using wavelet methodology* (AIMS Electronics and Electrical Engineering, 6(4), 370–384) [5], combinan SVC con análisis wavelet para detectar fallos en microredes IoT. Describen el

muestreo de señales de corriente y tensión, el preprocesamiento mediante transformada wavelet para extraer características temporales y frecuenciales, y la configuración de parámetros de SVC (C y γ), demostrando altas tasas de detección y bajas alarmas falsas en escenarios simulados.

En cuanto a árboles de decisión, Martínez, R. E. B. et al. (2009) en *Árboles de decisión como herramienta en el diagnóstico médico* (Revista médica de la Universidad Veracruzana, 9(2), 19–24) [6], comparan CART y regresión logística en predicciones clínicas (diabetes, hipertensión y cáncer de mama), analizando eficiencia computacional, escalabilidad y capacidad de manejo de variables categóricas y numéricas, y concluyen que los árboles ofrecen un balance óptimo entre interpretabilidad y rendimiento.

El estudio de Santos, F. C. (2009) *Variações do método kNN e suas aplicações na classificação automática de textos* (Instituto de Informática, Universidade Federal de Goiás, Goiânia) [7], aborda las variantes del k-nearest neighbors en clasificación textual, revisando métricas de distancia (euclídea, Manhattan), ponderaciones basadas en frecuencia y técnicas de reducción de dimensionalidad, así como estructuras de índice (k-d tree, LSH) para acelerar la búsqueda de vecinos en espacios de alta dimensión, y evalúa su eficacia en corpus de noticias y literatura científica.

Finalmente, Madariaga Fernández, C. J.; Lao León, Y. O.; Curra Sosa, D. A.; & Lorenzo Martín, R. (2022) en *Empleo de algoritmos KNN en metodología multicriterio para la clasificación de clientes, como sustento de la planeación agregada* (Retos de la Dirección, 16(1), 178–198) [8], proponen una metodología multicriterio que utiliza KNN para segmentar clientes según indicadores socioeconómicos y de comportamiento de compra. Tras normalizar las características, aplican kNN para generar grupos homogéneos y ordenan jerárquicamente dichos grupos conforme a criterios de valor y coste, validando el enfoque en un caso práctico minorista y demostrando mejoras en precisión de previsiones y eficiencia operativa.

4. Desarrollo.

4.1. Descripción del tratamiento de los datos.

Lo primero que se ha llevado a cabo con la base de datos seleccionada ha sido un Análisis de Componentes Principales (ACP).

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11
Standard deviation	1.7604	1.3878	1.2452	1.1015	0.97943	0.81216	0.76406	0.65035	0.58706	0.42583	0.24405
Proportion of Variance	0.2817	0.1751	0.1410	0.1103	0.08721	0.05996	0.05307	0.03845	0.03133	0.01648	0.00541
Cumulative Proportion	0.2817	0.4568	0.5978	0.7081	0.79528	0.85525	0.90832	0.94677	0.97810	0.99459	1.00000

Figure 5: Desviación estándar y proporciones de varianza y acumulada.

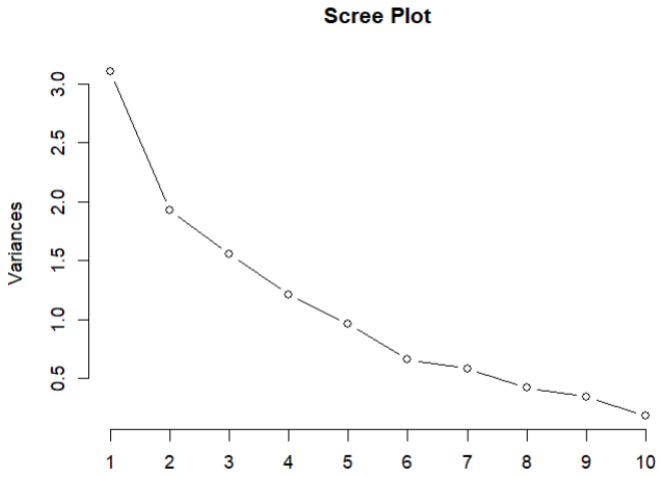

Figure 6: ACP en vino verde usando el método del codo para interpretación.

Como primer paso en el tratamiento de la base de datos seleccionada, se llevó a cabo un Análisis de Componentes Principales (ACP) con el objetivo de reducir la dimensionalidad del conjunto de datos y facilitar su interpretación sin pérdida significativa de información.

En cuanto a los resultados obtenidos del ACP, se analizan tres métricas fundamentales:

- Desviación estándar (Standard Deviation): Indica la magnitud de la variabilidad explicada por cada componente principal. Una desviación estándar más elevada implica una mayor relevancia de la componente asociada.
- Proporción de varianza (Proportion of Variance): Representa la fracción de la varianza total explicada por cada componente individual. Componentes con una mayor proporción de varianza capturan más información del conjunto de datos original.
- Proporción acumulada (Cumulative Proportion): Muestra la cantidad total de variabilidad explicada al considerar un número creciente de componentes. Esta medida resulta útil para determinar cuántos componentes son necesarios para conservar una cantidad representativa de la información original.

A partir de este análisis, se decidió conservar 6 de las 11 variables originales (solamente las representadas en la tabla del apartado 2), al considerar que estas explican un porcentaje suficientemente elevado de la varianza total. Las variables seleccionadas fueron: “Volatile acidity”, “Residual sugar”, “Chlorides”, “Density”, “Sulphates” y “Alcohol”.

Una vez seleccionadas las variables pertinentes, se procedió a su utilización como entradas del modelo. La columna “quality”, ya definida como variable objetivo, se binarizó para simplificar el problema de clasificación. Para ello, se crearon dos categorías: calidad baja (puntuaciones de 1 a 5) y calidad alta (puntuaciones de 6 a 10). Durante este proceso se procuró mantener un reparto equilibrado de instancias entre ambas clases, de modo que el modelo recibiera una muestra representativa y balanceada de cada categoría, favoreciendo así un entrenamiento más estable y fiable.

4.2. Procesamiento de los datos.

El preprocesamiento de los datos se inició con la configuración de una validación cruzada con 10 particiones (folds). Esta técnica permite obtener una estimación más robusta y menos sesgada del rendimiento real de los modelos, al evaluar su desempeño en múltiples subconjuntos del conjunto de datos.

A continuación, se generaron los índices correspondientes a la validación cruzada, los cuales determinan qué muestras se utilizarán como conjunto de entrenamiento y cuáles como conjunto de prueba en cada iteración. Esta segmentación resulta fundamental para garantizar una adecuada evaluación del modelo en todas las particiones.

Posteriormente, se procedió a la normalización de los datos, utilizando diferentes técnicas en función del algoritmo de aprendizaje aplicado:

i. Para el modelo de Redes Neuronales Artificiales (RNA), se aplicó la normalización NormalizeMinMax, dado que estos modelos presentan un mejor rendimiento cuando las variables de entrada se encuentran acotadas en un mismo rango (generalmente entre 0 y 1). Esta práctica ayuda a evitar inestabilidades numéricas durante el entrenamiento, especialmente cuando existen variables con escalas muy dispares.

ii. En el caso del clasificador SVC (Support Vector Classifier), se optó por la normalización NormalizeZeroMean, con el objetivo de centrar las variables en torno a una media de cero. Esta decisión se fundamenta en la sensibilidad del método a las escalas de las variables, ya que el cálculo de distancias y productos escalares —utilizados por el kernel— podría verse afectado negativamente si las variables presentan escalas distintas.

iii. Para el modelo DoME, se evaluaron ambas estrategias de normalización y, finalmente, se seleccionó NormalizeMinMax, debido a que ofreció mejores resultados empíricos y favorece una mayor interpretabilidad de los datos.

iv. En el caso de los árboles de decisión, no se aplicó ninguna técnica de normalización, dado que estos algoritmos no se basan en distancias ni productos escalares, sino en particiones del espacio de entrada mediante umbrales. Por tanto, la escala de las variables no afecta al funcionamiento del modelo.

v. Finalmente, para el algoritmo KNN (k-Nearest Neighbors), se empleó también NormalizeMinMax, ya que este clasificador se fundamenta en la distancia euclídea entre muestras. Normalizar garantiza que todas las variables contribuyan de manera equitativa al cálculo de distancias, evitando que variables con valores más altos dominen el resultado.

Tras la normalización, se procedió nuevamente a generar los índices de validación cruzada con 10 particiones, que se utilizaron para entrenar y evaluar los modelos en todas las fases del experimento.

4.3. Resultados relativos al proceso de entrenamiento.

Para cada una de las técnicas de aprendizaje consideradas, se ha seleccionado la configuración que ofreció los mejores resultados en términos de precisión, sensibilidad y especificidad. Los resultados más destacados han sido agrupados y presentados en una única tabla por cada modelo, seleccionando en cada caso los valores óptimos obtenidos para cada técnica de aprendizaje.

En primer lugar, se presentan los resultados obtenidos con Redes Neuronales Artificiales (RNA). Utilizando la técnica de normalización MinMax, se seleccionaron un total de ocho configuraciones de hiperparámetros diferentes, tras la realización de múltiples pruebas con el objetivo de identificar aquellos que ofrecieran el mejor rendimiento, especialmente en lo referente a la precisión. En todos los casos se trabajó con dos capas ocultas, y los principales parámetros definidos fueron los siguientes:

- learningRate = 0.01, valor comúnmente utilizado por su capacidad para proporcionar estabilidad durante el proceso de entrenamiento.
- maxEpochs = 100, parámetro que establece el número máximo de iteraciones del entrenamiento con el fin de evitar fenómenos de sobreajuste.
- validationRatio = 0.2, lo cual permite reservar una fracción del conjunto de entrenamiento como conjunto de validación interna, habilitando la posibilidad de aplicar early stopping en caso de que no se observe mejora en la pérdida durante 10 épocas consecutivas (maxEpochsVal = 10).
- numExecutions = 5, con el propósito de calcular un promedio de los resultados obtenidos en distintas ejecuciones y así mitigar el efecto de la aleatoriedad inherente al proceso de entrenamiento.

En nuestro caso, la topología [64, 32] fue la que ofreció los mejores resultados para este método y conjunto de datos, alcanzando una precisión cercana al 72%. A continuación, se presenta la tabla con los resultados más destacados obtenidos bajo dicha configuración.

Table 2: Tabla de resultados para RNA.

Métrica	Precisión	Tasa error	Sensibilidad	Especific.	VPP	VPN	F1
Media	0.719	0.281	0.710	0.726	0.710	0.755	0.687
Desv. típ.	0.018	0.018	0.075	0.069	0.051	0.041	0.050

A continuación, se presenta la matriz de confusión obtenida. Para su correcta generación en el caso de las Redes Neuronales Artificiales (RNA), fue necesario modificar el código fuente del archivo soluciones.jl, concretamente en la línea 736 de la función ANNCrossValidation, sustituyendo la función “mean” por “sum”. Este cambio permitió evitar la aparición de valores decimales en la matriz de confusión. Posteriormente, se aplicó una conversión explícita a tipo entero (Int) al momento de imprimir la matriz, con el fin de eliminar la representación en punto flotante (por ejemplo, valores con '.0'), obteniendo así la siguiente salida:

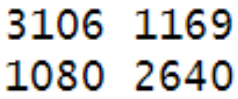

Figure 7: Matriz de confusión para RNA.

Para el algoritmo SVC (Support Vector Classifier), se empleó la normalización ZeroMean, dado que este tipo de modelos requieren un escalado adecuado de los datos con el fin de evitar que atributos con mayor magnitud dominen en el cálculo de distancias y productos escalares.

A continuación, se llevaron a cabo múltiples pruebas experimentales con el objetivo de identificar la configuración de hiperparámetros que ofreciera el mejor rendimiento para nuestro conjunto de datos. En este proceso, se evaluaron los cuatro tipos de kernel disponibles, variando el parámetro C en un rango comprendido entre 0,3 (que confiere al modelo una mayor tolerancia a errores) y 2,0 —que genera un modelo más estricto, es decir, con menor permisividad ante errores de clasificación. Asimismo, se seleccionaron distintos valores del hiperparámetro gamma, encargado de controlar el grado de influencia de cada muestra sobre la frontera de decisión, siendo los valores más elevados los que otorgan mayor relevancia a las observaciones cercanas.

En el caso específico de los kernels polinómico y sigmoide, se estableció coef0 = 0 y se experimentó con diferentes valores del parámetro degree, correspondiente al grado del polinomio.

Tras este proceso de ajuste, se obtuvo el mejor resultado con la configuración de hiperparámetros kernel = rbf, C = 2,0 y gamma = 0,8, alcanzando una precisión de aproximadamente el 77 %. La tabla que se muestra a continuación recoge los resultados más representativos obtenidos con dicha configuración.

Table 3: Tabla de resultados para SVC.

Métrica	Precisión	Tasa error	Sensibilidad	Especific.	VPP	VPN	F1
Media	0.767	0.233	0.758	0.774	0.746	0.787	0.751
Desv. típ.	0.046	0.046	0.059	0.055	0.055	0.046	0.050

Los datos presentados en la tabla evidencian que el kernel RBF, con los valores seleccionados de C y gamma, proporciona un balance óptimo entre sesgo y varianza, así como un alto nivel de precisión en la clasificación de las muestras. Estos resultados sirven como fundamento para comprender el comportamiento del modelo y comparar su desempeño frente a otras configuraciones y algoritmos.

A partir de esta configuración óptima, se procede a evaluar la matriz de confusión correspondiente:

662 180
193 564

Figure 8: Matriz de confusión para SMC.

Para el método DoME, se optó por aplicar la normalización mediante la técnica MinMax, con el fin de evitar que ciertos atributos pudieran dominar durante el proceso de creación de las reglas. Esta estrategia permite garantizar que las variables se encuentren en un rango comparable, lo cual es fundamental para asegurar el buen desempeño del modelo.

A continuación, se realizaron múltiples pruebas para identificar los hiperparámetros más adecuados. Tras una serie de experimentaciones, se concluyó que el rango óptimo de hiperparámetros se encontraba entre los valores de 7 y 10. Sin embargo, también se evaluaron valores ligeramente más altos, alcanzando hasta 25, ya que se observó que valores excesivamente elevados podrían inducir un sobreentrenamiento del modelo, lo cual no es deseable en este contexto.

Con base en estas pruebas, los mejores resultados para este modelo y nuestra base de datos se obtuvieron al utilizar un valor de 7 como hiperparámetro para el número máximo de nodos. Este ajuste resultó en una precisión aproximada del 75 %, la cual representó la mayor precisión alcanzada entre todos los modelos evaluados. A continuación, se presenta la tabla correspondiente a dichos resultados.

Table 4: Tabla de resultados para DoME.

Métrica	Precisión	Tasa error	Sensibilidad	Especific.	VPP	VPN	F1
Media	0.742	0.258	0.758	0.729	0.710	0.776	0.733
Desv. típ.	0.030	0.030	0.032	0.049	0.038	0.026	0.028

El rendimiento obtenido es significativo y subraya la efectividad de este enfoque para nuestro conjunto de datos. A fin de profundizar en la evaluación del modelo, a continuación, se presenta la matriz de confusión correspondiente:

623 180
232 564

Figure 9: Matriz de confusión para DoME.

En el caso de los árboles de decisión, no fue necesario aplicar un proceso de normalización, ya que estos modelos no son sensibles a la escala de los atributos. Esto se debe a que los árboles de decisión se basan en comparaciones entre los valores de las variables y no en cálculos de distancias, lo que los hace independientes de la escala de los datos.

A continuación, se realizaron varias pruebas con diferentes valores de profundidad del árbol con el objetivo de encontrar el valor que ofreciera el mejor rendimiento sin llegar a sobreentrenar el modelo. Es importante señalar que la elección de la profundidad del árbol requiere una consideración cuidadosa. Una mayor profundidad aumenta la capacidad de aprendizaje del modelo, pero también incrementa el riesgo de sobreajuste (overfitting). En contraste, una profundidad más baja tiende a generar un modelo más generalizado, pero puede conducir a un subajuste (underfitting). Tras realizar estas pruebas, se seleccionaron valores de profundidad comprendidos entre 12 y 23.

Los mejores resultados para este modelo y nuestra base de datos se alcanzaron utilizando un valor de 23 para la profundidad del árbol. Este ajuste produjo una precisión cercana al 77 %, el valor más alto alcanzado entre los diferentes modelos evaluados. La tabla que se presenta a continuación refleja los resultados obtenidos con esta configuración.

Table 5: Tabla de resultados para árboles de decisión.

Métrica	Precisión	Tasa error	Sensibilidad	Especific.	VPP	VPN	F1
Media	0.762	0.238	0.742	0.780	0.746	0.777	0.744
Desv. típ.	0.029	0.029	0.034	0.030	0.033	0.027	0.031

Los resultados mostrados en la tabla confirman la efectividad de este enfoque. A fin de obtener una evaluación más detallada del rendimiento del modelo, se presenta a continuación la matriz de confusión correspondiente:

```
667   192
188   552
```

Figure 10: Matriz de confusión para árboles de decisión.

En el caso del método k-Nearest Neighbors (kNN), se aplicó la normalización MinMax, dado que este algoritmo calcula distancias euclídeas entre las muestras y requiere que todos los atributos estén acotados en un mismo rango para evitar que variables de mayor magnitud dominen la medición de similitud.

A continuación, se llevó a cabo un proceso sistemático de ajuste de hiperparámetros, centrado en el número de vecinos (k). Se evaluaron valores comprendidos entre 9 y 19, descartando configuraciones superiores a dicho rango con el fin de prevenir el sobreentrenamiento. El criterio de selección se basó principalmente en la precisión obtenida en las particiones de validación cruzada.

Tras estas pruebas, se determinó que el valor óptimo de k para nuestro conjunto de datos es 11, alcanzando una precisión aproximada del 75 %. Los resultados más representativos de este ajuste se recogen en la siguiente tabla:

Table 6: Tabla de resultados para el método kNN.

Métrica	Precisión	Tasa error	Sensibilidad	Especific.	VPP	VPN	F1
Media	0.742	0.258	0.719	0.761	0.723	0.758	0.721
Desv. típc.	0.048	0.048	0.065	0.042	0.049	0.050	0.055

Los datos expuestos en la tabla demuestran que, con $k = 11$, el modelo alcanza un equilibrio adecuado entre sesgo y varianza, proporcionando un rendimiento estable y satisfactorio. Este hallazgo se alinea con las expectativas teóricas sobre la influencia del número de vecinos en la capacidad de generalización de kNN.

Para completar la evaluación del clasificador, a continuación, se presenta la matriz de confusión correspondiente:

```
651  209
204  535
```

Figure 11: Matriz de confusión para el método kNN.

4.4. Discusión y evaluación.

Tras la aplicación de cinco métodos de clasificación sobre el conjunto de datos reducido por medio de un Análisis de Componentes Principales (ACP) en la fase inicial, se ha podido observar el comportamiento diferencial de cada técnica en términos de precisión a lo largo de los diez pliegues del proceso de validación cruzada. Este esquema de validación se seleccionó para garantizar que cada partición mantuviera la proporcionalidad de las clases y, de este modo, evaluar los modelos con un nivel de fiabilidad elevado.

En cuanto al análisis de resultados, de manera general se consideran satisfactorios. Tras la binarización de la variable objetivo, las precisiones obtenidas oscilan entre el 70 % y el 80 % en prácticamente todos los métodos evaluados. Si bien esta cota no es extraordinariamente alta, sí resulta significativa, pues indica que los modelos son capaces de aprender de los datos y de generalizar de manera adecuada. Además, la ausencia de una desviación estándar elevada entre pliegues revela que los clasificadores no presentan una sensibilidad excesiva a las particiones del conjunto de entrenamiento, lo cual sería indeseable.

Respecto a la influencia de la representación de los datos, la utilización inicial del ACP para excluir determinadas variables contribuyó de forma notable a eliminar la redundancia, a mejorar la visualización y el análisis del comportamiento de los clasificadores, y a mitigar el riesgo de sobreajuste en aquellos algoritmos más sensibles a la dimensionalidad, como, por ejemplo, las redes neuronales artificiales (RNA) y los k-vecinos más cercanos (kNN). La selección de seis componentes principales resultó especialmente acertada, dado que explican una proporción sustancial de la varianza

original y permiten trabajar con la información verdaderamente relevante, eliminando el ruido innecesario.

Finalmente, al comparar los distintos métodos empleados, se identifica que los de mejor rendimiento en este caso son el clasificador de vectores de soporte (SVC) y los árboles de decisión. En particular, el SVC configurado con kernel radial (RBF) alcanzó el mayor nivel de precisión, probablemente gracias a su capacidad para modelar relaciones complejas en un espacio transformado por el ACP.

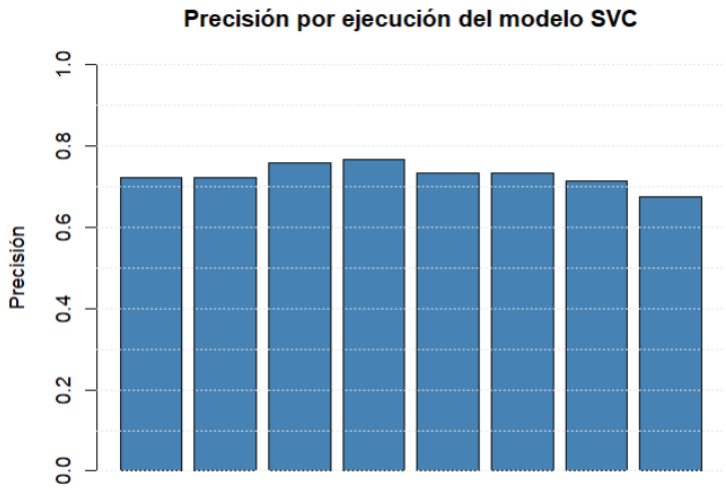

Figure 12: Histograma de precisiones para SVC.

Los árboles de decisión demostraron una notable capacidad de aprendizaje; no obstante, al aumentar excesivamente su profundidad, evidenciaron cierto sobreajuste y, en consecuencia, una merma de rendimiento en algunos pliegues.

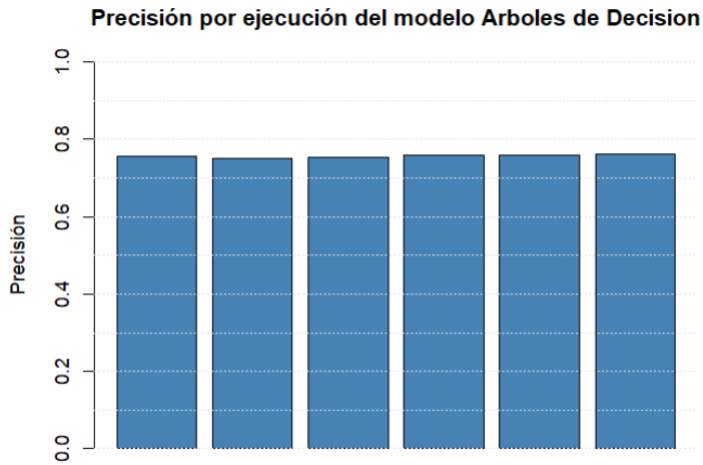

Figure 13: Histograma de precisiones para árboles de decisión.

Por otra parte, los métodos DoME y kNN han exhibido un desempeño intermedio; no obstante, este nivel de eficacia no resulta inferior al de los dos clasificadores mencionados anteriormente, pues las precisiones obtenidas son comparables.

En particular, el método DoME ha mostrado ciertas limitaciones en su rendimiento al contrastarlo con modelos más robustos y parametrizables, como es el SVC. Además, es posible que la reducción de dimensionalidad mediante ACP no le haya favorecido plenamente, dado que este algoritmo podría beneficiarse de un conjunto de características de mayor dimensión.

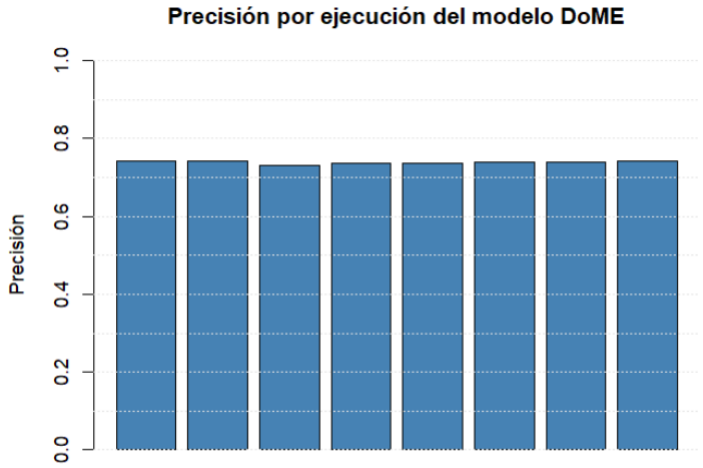

Figure 14: Histograma de precisiones para el modelo DoME.

El algoritmo de los k vecinos más cercanos (kNN) alcanzó un rendimiento satisfactorio, aunque algo inferior al de las demás técnicas anteriores debido a su elevada sensibilidad al ruido en los datos. No obstante, su simplicidad y facilidad de implementación lo convierten en una opción atractiva cuando priman la velocidad de entrenamiento y la interpretabilidad del modelo.

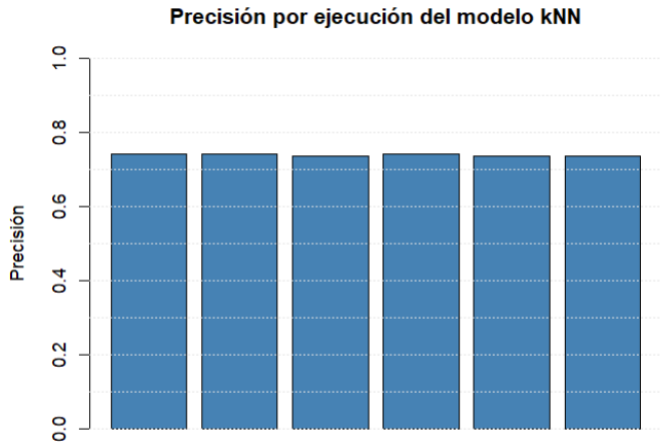

Figure 15: Histograma de precisiones para el método kNN.

Por último, las redes neuronales artificiales (RNA) mostraron un rendimiento notablemente inferior al de los demás métodos, dado que requieren un volumen elevado de datos para aprender patrones con eficacia y el conjunto evaluado carece de la complejidad necesaria para su correcto entrenamiento. En este sentido, la aplicación del ACP pudo no haberles resultado beneficiosa, y un número reducido de épocas de entrenamiento podría haber favorecido su tendencia al sobreajuste.

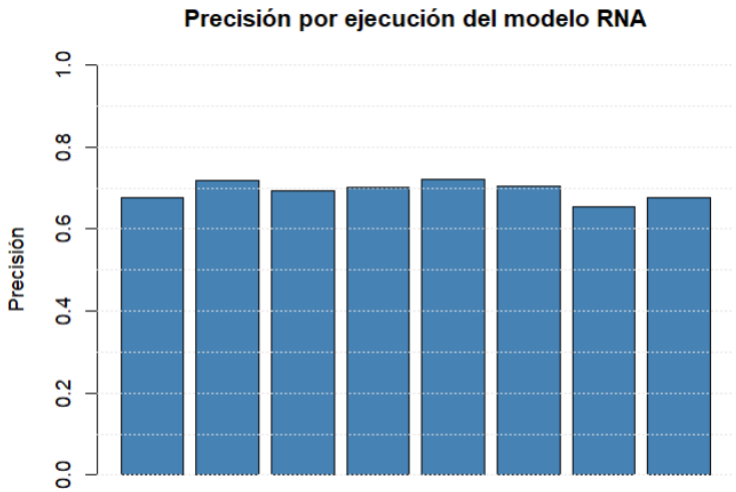

Figure 16: Histograma de precisiones para RNA.

Se ha observado que la aplicación del ACP mejora de forma consistente el rendimiento de los distintos modelos, lo cual pone de manifiesto que trabajar con un número reducido de atributos, pero de mayor relevancia y representatividad, puede resultar más eficaz que emplear un conjunto ampliado que introduzca ruido.

En conclusión, y tras llevar a cabo una adecuada selección de variables, normalización y validación cruzada, los algoritmos que ofrecieron los mejores resultados en este estudio fueron el SVC y los árboles de decisión, mientras que las redes neuronales artificiales registraron el desempeño más bajo.

5. Conclusiones.

Tras el análisis efectuado con los cinco métodos de clasificación aplicados al conjunto de datos de tinto ‘vinho verde’, se han obtenido diversas conclusiones acerca de su rendimiento y su aplicabilidad al problema de predicción de la calidad del vino.

En términos generales, los resultados alcanzados resultan satisfactorios, con precisiones que oscilan entre el 72 % y el 77 % en los distintos métodos empleados. Estos niveles de precisión demuestran que los modelos han sido capaces de identificar patrones relevantes en las características fisicoquímicas del vino y de asociarlos de manera fiable con su calidad, cumpliendo así el objetivo principal del trabajo: desarrollar un sistema

automatizado capaz de predecir la calidad del vino a partir de parámetros objetivos y directamente medibles.

Al examinar en detalle cada método, se aprecia que los mejores desempeños corresponden al clasificador SVC con kernel RBF ($C = 2{,}0$; $\gamma = 0{,}8$) y al árbol de decisión con profundidad máxima igual a 23, obteniéndose precisiones medias de 0,767 y 0,762, respectivamente. No obstante, dado que las desviaciones típicas asociadas (0,046 para SVC y 0,029 para el árbol de decisión) muestran un solapamiento en los intervalos de confianza, no es posible aseverar que uno de ellos supere de manera concluyente al otro, si bien el SVC exhibe una precisión media ligeramente superior.

En un nivel intermedio de rendimiento se encuentran los métodos DoME (maxnodes = 7) y kNN ($k = 11$), ambos con una precisión media de 0,742. Aunque sus resultados son prácticamente idénticos, la desviación típica menor de DoME (0,030 frente a 0,048 de kNN) indica una mayor estabilidad y robustez frente a variaciones en los datos de entrenamiento.

Por último, las redes neuronales artificiales (RNA) presentan el menor rendimiento, con una precisión de 0,719. Esta disminución podría atribuirse a la reducción de la complejidad del conjunto de datos tras la aplicación del Análisis de Componentes Principales (ACP), dado que la transformación a un espacio dimensionalmente reducido podría haber eliminado parte de los patrones no lineales que las RNA emplean para su modelado.

Otro aspecto relevante de este estudio es la eficacia del preprocesamiento mediante ACP. La selección de las seis variables principales demostró ser una decisión acertada, al eliminar redundancias sin comprometer la capacidad predictiva de los modelos. Este hallazgo subraya la importancia de aplicar un adecuado preprocesamiento de datos y de seleccionar cuidadosamente las características más relevantes de la base de datos.

En cuanto a la viabilidad de implementar estas metodologías en un entorno real, los resultados sugieren que los sistemas de aprendizaje automático podrían servir como complemento a las evaluaciones tradicionales de la industria del vino. Aunque una precisión máxima del 77 % podría no bastar para reemplazar por completo el criterio de enólogos y catadores profesionales, estos modelos podrían emplearse para realizar un análisis inicial rápido de grandes volúmenes de muestras y detectar datos atípicos en los parámetros fisicoquímicos que puedan influir en la calidad final del producto.

Durante la ejecución del trabajo se identificaron diversas dificultades, entre las que destacan la optimización de hiperparámetros (especialmente para las RNA y el SVC) y la normalización de los datos. Si bien se emplearon técnicas como MinMax o ZeroMean, no siempre fue posible aplicar cada método de normalización de manera óptima para todos los clasificadores.

En conclusión, este estudio cumple los objetivos inicialmente planteados, demostrando la viabilidad de aplicar técnicas de aprendizaje automático para predecir la calidad del tinto a partir de sus características fisicoquímicas. Los resultados obtenidos sientan las bases para futuras investigaciones que permitan ampliar el alcance y mejorar la precisión de estos modelos, contribuyendo así al objetivo último de optimizar la calidad del vino.

6. Trabajo futuro.

En primer lugar, resultaría de gran interés incorporar nuevas variables que enriquecieran el conjunto de datos empleado. Por ejemplo, la inclusión de información detallada sobre técnicas de cultivo (como tipo de suelo, prácticas de riego o gestión del dosel), así como datos genéticos de las cepas, que permitirían ahondar en los factores que determinan las propiedades sensoriales de cada variedad del tinto 'vinho verde'.

Asimismo, la aplicación de métodos de aprendizaje profundo representaría un camino prometedor, con el objetivo de aumentar tanto la precisión como la capacidad de generalización de los modelos. En este sentido, el empleo de redes neuronales convolucionales podría optimizar el análisis de imágenes de bayas y hojas, mientras que las arquitecturas recurrentes o basadas en transformadores podrían resultar especialmente útiles para procesar series temporales de parámetros agronómicos o enológicos.

Otro aspecto clave sería la validación empírica de los modelos en condiciones reales de producción. Evaluar el rendimiento de los clasificadores en parcelas con microclimas y manejos agronómicos diversos garantizaría su robustez y permitiría confirmar su validez en distintas regiones vitivinícolas. Esta estrategia contribuiría al diseño de herramientas adaptables a contextos geográficos y operativos heterogéneos.

Por otra parte, el desarrollo de soluciones de clasificación en tiempo real podría transformar la toma de decisiones en viñedo. La implementación de sistemas integrados (por ejemplo, aplicaciones móviles o plataformas de escritorio con interfaces gráficas intuitivas) facilitarían mucho la utilización de los modelos por parte de enólogos y técnicos agrónomos, independientemente de su nivel de experiencia en técnicas de inteligencia artificial. Incluso en el contexto de personas no profesionales en los estudios del vino y en situaciones más comunes, la aplicación de un software intuitivo y con facilidades a la hora de su uso facilitaría mucho el rol de un comprador a la hora de elegir qué vino llevarse (como, por ejemplo, a la hora de comprar vino en un supermercado o de pedirse una copa en un restaurante).

Finalmente, promover colaboraciones interdisciplinarias con especialistas en agronomía, biotecnología y ética aplicada ampliaría el alcance y la relevancia de este trabajo. La sinergia entre distintos campos no solo enriquecería el marco científico, sino que también impulsaría el desarrollo de soluciones prácticas y éticamente responsables en la clasificación de variedades de vino.

En conjunto, estas líneas de acción ofrecen un itinerario claro para perfeccionar los resultados obtenidos y abrir nuevas fronteras en la intersección de la inteligencia artificial y la ciencia vitivinícola.

References.

[1] Barceló, J. G. (1990). *Técnicas analíticas para vinos* (pp. 31-49).

(https://shop.gabsystem.com/img/cms/Capitulo%201-6.pdf)

[2] *Bibliodoc: anuari de biblioteconomia, documentació i informació*, 147-164.

[3] Valderrama, J. O., & Rojas, R. E. (2009). Avances en la predicción de propiedades físicas, físico-químicas y de transporte de líquidos iónicos. *Información tecnológica*, *20*(4), 149-160. (https://www.scielo.cl/scielo.php?pid=S0718-07642018000300097&script=sci_arttext)

[4] Abdulhadi, N., & Al-Mousa, A. (2021, July). Diabetes detection using machine learning classification methods. In *2021 international conference on information technology (ICIT)* (pp. 350-354). IEEE. (Dialnet-ServicelearningOAprendizajeservicio-2582784.pdf)

[5] Lakshmi, K. D., Panigrahi, P. K., & kumar Goli, R. (2022). Machine learning assessment of IoT managed microgrid protection in existence of SVC using wavelet methodology. *AIMS Electronics and Electrical Engineering*, *6*(4), 370-384. (https://www.researchgate.net/profile/Dhana-Lakshmi-25/publication/364072267_Machine_learning_assessment_of_IoT_managed_microgrid_protection_in_existence_of_SVC_using_wavelet_methodology/links/635a58d196e83c26eb5c4661/Machine-learning-assessment-of-IoT-managed-microgrid-protection-in-existence-of-SVC-using-wavelet-methodology.pdf?_sg%5B0%5D=started_experiment_milestone&origin=journalDetail&_rtd=e30%3D)

[6] Martínez, R. E. B., Ramírez, N. C., Mesa, H. G. A., Suárez, I. R., Trejo, M. D. C. G., León, P. P., & Morales, S. L. B. (2009). Árboles de decisión como herramienta en el diagnóstico médico. *Revista médica de la Universidad Veracruzana*, *9*(2), 19-24. (http://www.soporte.uv.mx/rm/num_anteriores/revmedica_vol9_num2/articulos/arboles.pdf)

[7] SANTOS, F. C. (2009). Variações do método kNN e suas aplicações na classificação automática de textos. *Instituto de Informática, Universidade Federal de Goiás, Goiânia*. (https://ww2.inf.ufg.br/ppgcc/sites/www.inf.ufg.br.mestrado/files/uploads/Dissertacoes/Fernando%20Chagas.pdf)

[8] Madariaga Fernández, C. J., Lao León, Y. O., Curra Sosa, D. A., & Lorenzo Martín, R. (2022). Empleo de algoritmos KNN en metodología multicriterio para la clasificación de clientes, como sustento de la planeación agregada. *Retos de la Dirección*, *16*(1), 178-198. (http://scielo.sld.cu/scielo.php?pid=S2306-91552022000100178&script=sci_arttext)

Visual references.

[1] https://elordenmundial.com/mapas-y-graficos/paises-exportadores-importadores-vino/

[2] https://datascientest.com/es/que-es-el-algoritmo-knn