UNIVERSIDAD DE VALLADOLID

E.T.S.I. TELECOMUNICACIÓN

TRABAJO FIN DE GRADO

GRADO EN INGENIERÍA DE TECNOLOGÍAS DE TELECOMUNICACIÓN

Análisis de Técnicas de Aprendizaje Automático en el Sector de la Viticultura

Autor:
Carmen Martín Gallo

Tutor:
Dña. Noemí Merayo Álvarez
Dña. Patricia Fernández del Reguero

TÍTULO: **Análisis de Técnicas de Aprendizaje Automático en el Sector de la Viticultura**
AUTOR: **Carmen Martín Gallo**
TUTOR: **Dña. Noemí Merayo Álvarez**
Dña. Patricia Fernández del Reguero
DEPARTAMENTO: **Teoría de la Señal y Comunicaciones e Ingeniería Telemática (TSCIT)**

TRIBUNAL

PRESIDENTE:	Patricia Fernández del Reguero
SECRETARIO:	Noemí Merayo Álvarez
VOCAL:	Rubén M. Lorenzo Toledo
SUPLENTE:	Ramón J. Durán Barroso
SUPLENTE:	Ignacio de Miguel Jiménez

FECHA:
CALIFICACIÓN:

Resumen de TFG

Este Trabajo Fin de Grado ofrece contribuciones relevantes al estado del arte de la investigación relacionada con la tecnología en el sector de la viticultura.

En primer lugar, se presenta una exhaustiva visión de las técnicas de Inteligencia Artificial empleadas en los últimos años en el ámbito de la vinificación a partir del estudio de artículos que inciden en las técnicas empleadas y cómo estas ayudan a mejorar diversos aspectos, como puede ser la calidad del vino o incluso factores relacionados con la producción o cantidad del vino producido. A partir de estos datos, podemos ofrecer un recorrido documentado sobre las inclinaciones actuales de emplear este gran recurso, la Inteligencia Artificial. Este estudio se centra en las técnicas de Aprendizaje Automático que se pueden integrar en la gestión y procesos de vinificación de viñedos actuales para brindar resultados relevantes y útiles para la industria.

Por otra parte, el segundo componente del trabajo destaca la importancia de las Bases de Datos empleadas, ofreciendo ejemplos y unas breves pinceladas sobre características importantes que influyen a la hora de afrontar un estudio con muestras de vino.

Este documento concluye ofreciendo una interpretación de las nuevas tendencias que se adoptarán en el futuro cercano para mejorar un sector enormemente influyente en nuestro país y a nivel mundial.

Palabras clave

Inteligencia Artificial (IA), vino, enología, *Machine Learning* o Aprendizaje Automático, Bases de Datos (BBDD)

Abstract

This Final Project offers relevant contributions to the state of the art's research related to technology in the viticulture sector.

On the one hand, an exhaustive vision of Artificial Intelligence techniques used in recent years in the field of winemaking is presented. In order to meet that goal, the study of articles that affect the techniques used and how they help to improve various aspects -such as wine quality or factors related to the production or quantity of the wine produced- are used. From these data, we can offer a documented tour of the current inclinations to use this great resource, Artificial Intelligence. This study focuses on Machine Learning techniques that can be integrated into current vineyard management and winemaking processes to deliver industry-relevant and useful results.

On the other hand, the second component of the current work highlights the importance of the databases used, offering examples and a few brief notes on important characteristics that influence when facing a study with wine samples.

This document concludes offering an interpretation of the new trends that will be adopted in the near future to improve a greatly influential sector in our country.

Keywords

Artificial Intelligence (AI), wine, oenology, Machine Learning, Databases (BBDD)

Agradecimientos

A Noemí y Patricia, por ser grandes referentes para futuros ingenieros y brindarme toda su atención y ayuda a lo largo de este proyecto.

A mi familia, tanto de sangre como de corazón, por recordarme que, pase lo que pase, cada día sale el sol.

A todos ellos, gracias.

Índice

Índice de Figuras

Índice de Tablas

Índice de Gráficos

1 Introducción

1.1 Motivación

El sector de la viticultura es uno de los campos más relevantes e influyentes en nuestro país. Desde hace años, la importancia que recae en la producción e importación de vino a nivel económico-social está en auge, de ahí que los estudios para la mejora de la calidad y producción de este brebaje hayan aflorado de forma exponencial.

Por otra parte, los avances tecnológicos llevados a cabo durante los últimos años en términos de Inteligencia Artificial como modo de clasificación, mejora o identificación de patrones, han brindado la posibilidad de emplear estos algoritmos en ámbitos más transversales, como por ejemplo el campo de la medicina y la salud, las comunicaciones, la educación o incluso la agricultura.

Es por esta razón que la incorporación de técnicas de Aprendizaje Automático se lleva a cabo también en el sector de la viticultura de forma sustancial y estas metodologías comienzan a tomar más partido en las bodegas y laboratorios que estudian las fases de maduración y producción enológica.

El Trabajo Fin de Grado presente recorre los últimos estudios realizados para mejorar o clasificar distintos aspectos relevantes en el resultado final del vino, que incluyan técnicas de Inteligencia Artificial para cumplir dicho cometido.

Además, se pretende comparar e identificar patrones y algoritmos clave para una mayor adopción de este recurso por parte de las bodegas en años venideros.

Por último, también se incide y recalca la importancia de utilizar unos datos adecuados que conformen la Base de Datos a utilizar, ya que, dependiendo de este medio, los resultados podrán ser favorables y útiles para los expertos o fracasar en el intento de estudiar y comprender los factores influyentes en el vino.

1.2 Objetivos

A continuación, se definen e identifican los objetivos tanto generales como específicos a tratar en el presente trabajo.

1.2.1 Objetivo General

Los objetivos a cumplir en el escrito aquí expuesto son, principalmente, obtener una **visión global** de las técnicas por ahora adoptadas en el sector del vino, así como visualizar la tendencia que se tomará a futuro para desarrollar **nuevos modelos punteros de Inteligencia Artificial** convergentes con el proceso de vinificación.

- ✓ Estudiar y caracterizar las preferencias actuales nos ayudará a establecer **metas comunes** de implementación de Inteligencia Artificial como técnica de mejora entre expertos enólogos e ingenieros.
- ✓ Comprender las necesidades de un sector ajeno al tecnológico ayudará a focalizar los **puntos de mayor interés** donde aplicar Inteligencia Artificial, así como el fin a conseguir con esta adopción.

1.2.2 Objetivos Específicos

Además de intentar conseguir los objetivos principales nombrados anteriormente, se intentará también lograr los siguientes objetivos específicos:

1. Recoger en un estudio amplio y exhaustivo todas las metodologías relacionadas con la aplicación de la Inteligencia Artificial en la viticultura desarrolladas hasta la fecha. Es decir, realizar una *review* extensa que incluya los artículos relevantes para definir la proyección a futuro de nuevos estudios.
2. Identificar las técnicas de Inteligencia Artificial más utilizadas, así como los índices de satisfacción de los expertos con los resultados obtenidos.
3. Detectar áreas de máxima aplicabilidad de modelos de Aprendizaje Automático, ya sea por una necesitad específica de los enólogos o bien porque supone una mejora sustancial en el desempeño a ejecutar.

4. Descubrir y caracterizar sistemas efectivos ya desarrollados en otros estudios, que servirán como antecedente de futuras investigaciones.
5. Análisis de las Bases de Datos disponibles y empleadas por otros autores, con el fin de averiguar características necesarias para la adquisición de nuevas muestras que conformen nuevos conjuntos de datos.

1.3 Fases y Métodos

En este punto de detallan las fases seguidas en este estudio para cumplir los objetivos mencionados en el punto anterior.

1.3.1 Fase de Análisis

La finalidad de esta primera etapa es crucial en este escrito, ya que se compone de la búsqueda de información, artículos y documentos relevantes para el estudio del *estado del arte* de este Trabajo de Fin de Grado. Los aspectos más relevantes se nombran a continuación:

- ✓ Búsqueda y lectura de informes que mencionen la aplicación de Inteligencia Artificial en el sector del vino.
- ✓ Identificar y agrupar dichos estudios según el impacto o relevancia que los caracterice, para así estructurar este trabajo en distintos bloques con diferentes pautas y conclusiones.

1.3.2 Fase de Implementación

En este escrito la fase de implementación no se lleva a cabo. Esta fase será el objetivo a cumplir por estudios posteriores que tomen este artículo como base de referencia.

1.3.3 *Fase de Pruebas*

Al igual que la Fase de Implementación, este punto no será desarrollado en este estudio y se plantea como objetivo para posibles nuevos modelos que brinden un servicio al sector vitícola.

1.3.4 Fase de Realización de los Informes

Esta fase comprende el proceso de redacción y estructuración en un documento escrito del estudio ya realizado, lo que implica la composición de la Memoria del Trabajo de Fin de Grado del alumno autor.

1.4 Estructura de la Memoria del TFG

La estructura de este artículo consta de 12 puntos completos, incluyendo el presente donde se introduce la temática y se describen las pautas y objetivos que conforman el estudio.

En segundo lugar, compuesto por los Puntos 2 y 3, se distinguen diferentes aspectos en los cuales ya se han aplicado diversos algoritmos de Inteligencia Artificial con los siguientes fines:

Punto 2: Predicción, clasificación y mejora de la **Calidad** del vino.
Punto 3: Empleo de **Espectroscopía** para identificar, en términos generales, diferentes compuestos relevantes en el resultado final del vino.

Después, el Punto 4 incluye una breve conclusión y recopilación de las Bases de Datos detectadas en los artículos incluidos en este estudio.

Finalmente, este Trabajo de Fin de Grado pone su punto final con una conclusión global y predicción de las nuevas metodologías a emplear de cara a futuro y de las líneas de investigación que surgirán en Líneas Futuras en el Punto 5.

Se completa el escrito con un Sumario de Algoritmos -Punto 6- mencionados en el documento, así como con la Bibliografía empleada -Punto 7-, en formato IEEE.

2 Análisis de la Calidad del Vino

2.1 Introducción

En este segundo punto, se estudiarán diversas maneras de predecir, medir e intentar mejorar la calidad del vino.

La calidad del vino producido es de vital importancia para los enólogos y productores de vino. El hecho de poder anticiparse a impedimentos de producción, o incluso el poder controlar ciertas variables fisicoquímicas o ambientales para conseguir lograr un brebaje de la mayor calidad posible es un aspecto muy demandado por las bodegas en la actualidad.

Por ello, dedicaremos un bloque de este estudio al análisis de la calidad del vino empleando técnicas y algoritmos de Inteligencia artificial diversos, que ofrecerán diferentes ventajas y resultados.

2.2 Artículo 1: *A Machine Learning application in wine quality prediction* [1]

El objetivo principal de esta investigación es predecir la calidad del vino mediante la construcción un modelo de Aprendizaje Automático basado en datos sintéticos y datos experimentales recopilados de diferentes regiones de Nueva Zelanda. Utilizamos 18 muestras de vino *Pinot noir* con 54 características diferentes (7 fisicoquímicas y 47 químicas) y calidades comprendidas entre 4,82 y 6,55 medidas de 0 a 10. Se obtienen 1381 muestras artificiales a partir de 12 de las muestras originales utilizando el método SMOTE (*Synthetic Minority OverSampling Technique)*. Las 6 muestras restantes permanecieron inalteradas y se emplearon para realizar la fase de pruebas del modelo. La Figura 1 muestra el diagrama de bloques asociado al estudio.

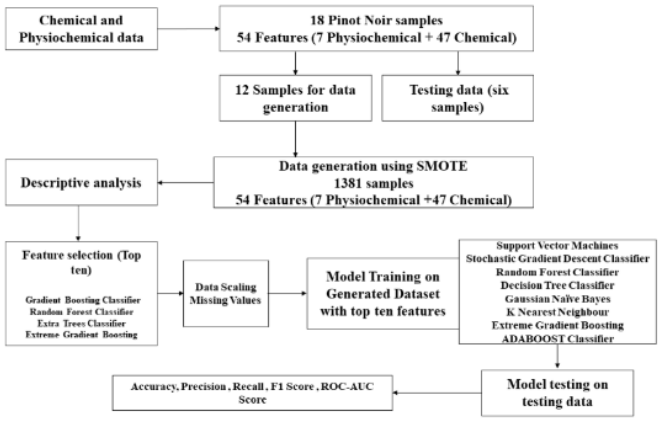

Figura 1. Diagrama de bloques del estudio

Cuatro métodos de clasificación, XGB, *Extra Trees*, *Random Forest* y *Gradient Boosting*, se implementaron con el fin de obtener las 10 características principales de las 54 características globales. Cada algoritmo brinda 10 variables diferentes. Los siguientes pasos de realizan a partir de los datos extraídos de XGB, ya que los modelos de Aprendizaje Automático se comportaron excepcionalmente bien con este conjunto de características.

A continuación, se entrenaron y probaron los siguientes siete algoritmos de Aprendizaje Automático: *Support Vector Machine* (SVM), *Random Forest*, *Decision Tree Classifier* (DTC), *Gaussian Naïve Bayes* (GNB), XGB, K-*closest neighbour* (KNN), *Adaptive Boosting* (AdaBoost), y *Stochastic Gradient Decision Classifier* (SGDC). Los resultados finales se muestran en la Tabla 1.

Classifiers	Precision	Recall	F1	ROC_AUC	MCC	Time (s)
XGB	0.90	0.75	0.78	0.75	0.63	0.127
RF	0.90	0.75	0.78	0.75	0.63	0.52
GNB	0.38	0.38	0.33	0.375	−0.25	0
AdaBoost	1	1	1	1	1	0.31
SGD	0.90	0.75	0.78	0.75	0.63	0.003
SVM	0.90	0.75	0.78	0.75	0.63	0.018
DTC	0.50	0.5	0.49	0.5	0	0.01
KNN	0.90	0.75	0.78	0.75	0.63	0.008

Tabla 1. Resultados de predicción de calidad de vino de los clasificadores sobre las diez variables XGB

También se realizó el mismo experimento tomando las características importantes (llamadas *variables esenciales*) en al menos tres métodos de selección de características. El clasificador AdaBoost mostró una precisión del 100% cuando se entrenó y evaluó sin selección de características, con selección de características XGB y con *variables esenciales*. El rendimiento general de todos los clasificadores excepto KNN mejoró cuando el modelo fue entrenado y probado usando *variables esenciales*, en especial *Random Forest* (RF).

En resumen, se identifican los clasificadores **AdaBoost** y **RF** como los mejores modelos para predecir la calidad del vino en diversas situaciones. Además, también se demuestra la relevancia de selección de características la influencia favorable en el rendimiento de los modelos de variables esenciales seleccionadas de los cuatro métodos de selección de características.

2.3 Artículo 2: *Analysis of white wine using Machine Learning algorithms* [2]

El conjunto de datos utilizado en este trabajo se adquirió del repositorio de Aprendizaje Automático de UCI (UC Irvine)[1] para la variante de vino blanco. Hay un total de 11 parámetros independientes: acidez fija, acidez volátil, ácido cítrico, azúcar residual, cloruros, dióxido de azufre libre, dióxido de azufre total, densidad, pH, sulfatos y alcohol; todos ellos contenidos en las 4898 instancias empleadas para medir la calidad.

Se emplean los siguientes algoritmos de Inteligencia Artificial: *Naïve Bayes*, *Support Vectors Machine* (SVM), *Random Forest*, J48 y *Multi-layer Neural Network* (MLP). Se utiliza la herramienta WEKA para el procesado de datos y generación de resultados en diversos formatos. Los datos de la clasificación se añaden a continuación, en la Tabla 2.

Algorithm	Performance Measures					
	Accuracy	Precision	Recall	F1-Measure	Kappa Statistic	ROC Area
J48	99.895	0.999	0.999	0.999	0.964	0.94
SVM	98.4883	0.985	0.985	0.978	0.0266	0.507
Random Forest	99.895	0.999	0.999	0.999	0.964	0.993
MLP	99.874	0.999	0.999	0.999	0.9565	0.992
Naive Bayes	99.3282	0.995	0.993	0.994	0.8149	0.995

Tabla 2. Resultados de la clasificación de los 11 parámetros

Los resultados demuestran que los algoritmos **J48** y ***Random Forest*** tienen mayor exactitud, con valor igual a 99,895%, precisión 0,999, recuperación 0,999, F1-Measure 0,999 y las estadísticas Kappa 0,964. El área ROC del J48 y el algoritmo *Random Forest* ofrecen índices iguales a 0,94 y 0,993, menor que el algoritmo *Naïve Bayes*.

[1] *Base de Datos UCI disponible en: http://archive.ics.uci.edu/ml/index.php*

Por otra parte, la Tabla 3 ilustra que, en términos de minimizar el error de predicción de la calidad del vino blanco, la mejor opción será el algoritmo de regresión MLP, seguido por J48.

Algorithm	Performance Measures			
	MAE	RMSE	RAE	RRSE
J48	0.0021	**0.0324**	6.8994	**26.3684**
SVM	0.0151	0.1229	49.7216	100.0818
Random Forest	0.0033	0.0327	10.8452	26.6198
MLP	**0.0016**	0.0347	**5.4269**	28.2865
Naive Bayes	0.0077	0.0723	25.2294	58.828

Tabla 3. Resultados de regresión

En cómputo global, la mejor baza para predecir la calidad de vino blanco será el algoritmo J48, ya que obtiene las mejores y más altas tasas, así como consigue minimizar de forma adecuada los posibles errores de predicción ofreciendo índices relacionados con la medición de errores muy bajos.

2.4 Artículo 3: *A hybrid wine classification model for quality prediction* [3]

Este artículo propone un modelo híbrido formado por los dos clasificadores, *Random Forest* y SVM, para predecir la calidad del vino. Para evaluar el desempeño de este modelo híbrido, también se realizan experimentos en los conjuntos de datos de vino para mostrar las ventajas que puede ofrecer. Los conjuntos de datos de entrada son datos de vino tinto y vino blanco, utilizados para entrenar y probar el modelo.

El conjunto de datos de vino de la Base de Datos UCI que consta de 1599 instancias de vino tinto y 4898 muestras de vino blanco. Ambos conjuntos de datos contienen 11 variables fisicoquímicas. La proporción de conjuntos de datos de entrenamiento y prueba es 80/20.

El algoritmo híbrido propuesto, en primer lugar, selecciona *n* modelos del grupo de modelos dado. Luego, se buscan los hiperparámetros mediante el método de búsqueda aleatoria. Los modelos con prestaciones aceptables se fusionan como modelos híbridos. La Figura 2 ilustra el funcionamiento de dicho algoritmo.

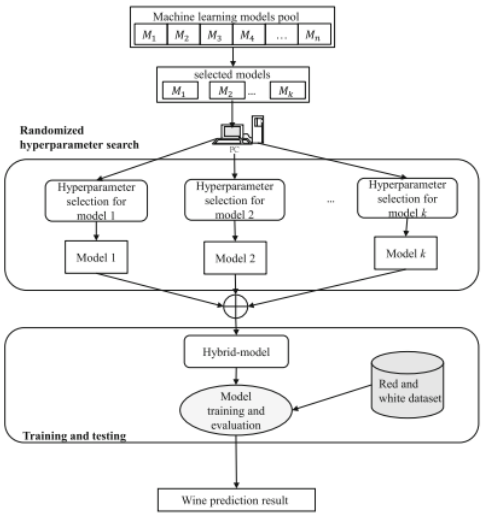

Figura 2. Diagrama de flujo del modelo híbrido propuesto

LOS AUTORES EXAMINAN CUIDADOSAMENTE LOS DATOS PARA ASEGURARSE DE QUE NO EXISTE NINGUNA ANOMALÍA. COMO RESULTADO, PARA EL VINO TINTO DETECTAN 24 INSTANCIAS DUPLICADAS Y PARA VINO BLANCO 937 INSTANCIAS DUPLICADAS. SIN EMBARGO, NO SE CLASIFICAN COMO ANOMALÍA PORQUE TODAS LAS CARACTERÍSTICAS Y LOS VALORES DE LAS ETIQUETAS SON EXACTAMENTE IGUALES.

Models	Accuracy	
	Red wine	White wine
Cortez et al. [6]	0.45	0.51
Apalasamy et al. [17]	0.62	0.65
Proposed model	**0.66**	**0.67**

Testing dataset size percentage	Red wine			White wine		
	10%	*20%*	*30%*	*10%*	*20%*	*30%*
Accuracy	0.69	0.71	0.66	0.70	0.68	0.67
Macro-precision	0.42	0.43	0.34	0.42	0.37	0.47
Macro-recall	0.41	0.38	0.32	0.41	0.34	0.35
Macro-F1 score	0.41	0.39	0.32	0.41	0.34	0.36

Tabla 4. (Izqda. a dcha.) Comparación con artículos [23] y [27] y resultados de este estudio tomando distintos ratios de datos de prueba

Los resultados presentes en la Tabla 4 demuestran que, para el vino tinto, la exactitud fue máxima cuando se probó la proporción del conjunto de datos establecida en 20 %. La precisión y la recuperación tienen valores bajos para todas las proporciones. Este hecho significa que la cantidad de falsos positivos es muy similar a la de falsos negativos. El puntaje F1 para el vino tinto disminuye gradualmente con el aumento de la proporción. El vino blanco muestra un resultado diferente, donde la exactitud siempre es mayor que la recuperación. Cuando la proporción se fijó en 10%, se consiguió la mayor exactitud y puntuación F1. Además, la precisión y recuperación son mucho más cercanas. El bajo valor de puntuación F1 tanto para vino tinto como para vino blanco implica que los datos están muy sesgados en ciertas ocasiones.

2.5 Artículo 4: *Machine Learning approach for attribute identification and quality prediction of red wine* [4]

Este artículo evalúa a través de varios algoritmos de Aprendizaje Automático la predicción de la calidad del vino tinto. El algoritmo *Support Vector Machine* (SVM) produce la máxima precisión del 87,5 %, cuando se utiliza con búsqueda *GridSearchCV*[2] la precisión fue aproximadamente igual al 90%.

Se emplea la Base de Datos de UCI, formada por 1599 instancias de vino tinto, con 11 variables fisicoquímicas en cada muestra.

En la primera etapa del análisis, se realiza un preprocesamiento del conjunto de datos. Después, se obtiene un modelo de datos llamado *modelo de visualización*, que muestra la cantidad de dependencia de la calidad del vino de 11 variables independientes (predictores). A continuación, se evalúan los predictores considerados. Los resultados finales se exponen en la Tabla 5.

Model	Accuracy (in %)	MSE	RMSE	RAS
Logistic regression	86.56	0.134	0.366	0.6218
Decision tree classifier	87.50	0.125	0.3534	0.7506
Random forest classifier	89.5	0.0843	0.2904	0.7744
Stochastic gradient descent classifier	83.12	0.1687	0.4107	0.4959
Naïve bayes classifier	84.68	0.1531	0.3913	0.8221
K nearest neighbor classifier	87.81	0.1218	0.3491	0.6819
Support vector machine	87.50	0.1250	0.3535	0.6185
Support vector machine (Grid search CV)	89.68	0.1031	0.3211	0.6753

Tabla 5. Resultados del análisis de los algoritmos

El algoritmo **SVM** produce la máxima precisión del 87,5 %, proporcionando además una precisión de 89,6875 % cuando se usa con *GridSearch* CV.

[2] *GridSearchCV es una clase que permite evaluar y seleccionar de forma sistemática los parámetros de un modelo, además de evaluar el rendimiento mediante validación cruzada.*

2.6 Artículo 5: *A new red wine prediction framework using Machine Learning* [5]

A partir de la Base de Datos UCI formada por 1599 muestras de vino tinto, se pretende predecir la calidad del brebaje. Para ello, se propone un nuevo marco combinado MF-DCCA (*Multifractal Detrended Cross-Correlation Analysis*) con XGBoost y LightGBM. MF-DCCA investiga la relación dinámica entre las 11 variables fisicoquímicas que forman cada muestra. XGBoost, ofrece una gran mejora respecto a *Gradient Boosting Decision Tree* y, por otra parte, LightGBM es un algoritmo efectivo para resolver tareas de clasificación y regresión que ocupa menos memoria y obtiene una mejor predicción que XGBoost. Se utiliza correlación cruzada para comprobar la correlación entre los indicadores fisicoquímicos del vino tinto y su calidad.

En comparación con otros algoritmos implementados con los mismos datos, los resultados empíricos de este marco tienen una mayor precisión (91,04%), como se aprecia en la Tabla 6. Este hecho se puede atribuir a la optimización provocada por los algoritmos de Aprendizaje Automático.

Test model classifier	*Recall (%)*	*Precision (%)*	*F1 measure (%)*	*Accuracy (%)*
LightGBM	85.63	86.67	86.15	91.04
XGBoost	90.67	90.67	90.67	91.04

Tabla 6. Resultados de los algoritmos tras predecir la calidad de las muestras de vino tinto

LightGMB hizo algunas optimizaciones, usando la resta del histograma para hacer una estrategia de crecimiento de aceleración de hoja-vino, con limitación de profundidad para reducir errores y obtener una mejor precisión, agregando reglas de decisión para ahorrar memoria y reducir gastos computacionales mediante la conversión de características en características multidimensionales únicas. XGBoost ayuda a evitar el sobreajuste de datos, está equipada para detectar y tratar los valores faltantes, siendo un clasificador flexible.

Por la importancia de la correlación y los resultados de clasificación que se obtienen, este enfoque se considera un avance en la clasificación de la calidad del vino tinto. El azúcar residual contribuye más complejamente a la calidad del vino tinto, mientras que las correlaciones cruzadas más débiles son la acidez volátil y los cloruros,

respectivamente. Tanto LightGBM como XGBoost lograron una mayor precisión de clasificación que otros algoritmos de Aprendizaje Automático.

2.7 Artículo 6: *Wine quality analysis using Machine Learning* [6]

Este documento se centra en el estudio comparativo de diferentes algoritmos de clasificación para el análisis de la calidad del vino: SVM, *Random Forest* y *Multilayer Perceptron* (MLP). El conjunto de datos contiene 4898 instancias de vino blanco del repositorio de Aprendizaje Automático de UCI. El algoritmo de Bosque Aleatorio, ***Random Forest***, ofrece el mejor resultado con un porcentaje de precisión de 81,96%, seguido por el algoritmo MLP con un porcentaje de precisión de 78,78% y, por último, SVM, con una precisión de 57,29%. Se adjuntan los resultados obtenidos en la Tabla 7.

Quality	Precision	Recall	F1-score	Support
3	0.00	0.00	0.00	9
4	0.00	0.00	0.00	76
5	0.60	0.65	0.63	669
6	0.56	0.75	0.64	960
7	0.57	0.21	0.31	349
8	0.00	0.00	0.00	82
Average/total	0.53	0.57	0.53	2145

Accuracy score: 0.5729603729603729

Quality	Precision	Recall	F1-score	Support
4	0.98	0.46	0.62	140
5	0.84	0.86	0.85	1469
6	0.77	0.90	0.83	1846
7	0.91	0.67	0.77	730
8	1.00	0.39	0.56	111
9	1.00	0.20	0.33	5
Average/total	0.83	0.82	0.81	4352

Accuracy score: 0.8196231617647058

Quality	Precision	Recall	F1-score	Support
4	0.79	0.56	0.65	140
5	0.85	0.79	0.82	1469
6	0.75	0.87	0.80	1846
7	0.78	0.68	0.72	730
8	0.88	0.40	0.55	111
9	1.00	0.20	0.33	5
Average/total	0.83	0.82	0.81	4352

Accuracy score: 0.787812548529

Tabla 7. Resultados de clasificación de SVM, Random Forest y Multilayer Perceptron, izqda. a dcha.

La razón por la cual el algoritmo *Random Forest* ofrece el mejor resultado, es porque al dividir un nodo durante la formación del árbol, la división elegida será la más efectiva, lo que provoca un aumento de la calidad del modelo.

2.8 Artículo 7: *Red wine quality prediction using Machine Learning techniques* [7]

Este artículo aplica técnicas de minería de datos como *Random Forest*, *Support Vector Machine* y *Naïve Bayes*, utilizando para la obtención de resultados el software RStudio. Los datos se extraen del repositorio de Aprendizaje Automático de UCI formado por 1599 instancias con 11 variables por muestra de vino tinto.

Los datos se separan en un conjunto de prueba que conforma el 30% del total de información y otro conjunto de entrenamiento, formado por el 70% restante.

En el campo del Aprendizaje Automático, una matriz de confusión es una tabla que se utiliza con frecuencia para representar un modelo de agrupación que involucra una gran cantidad de información. Esta investigación utiliza el conjunto de datos de vino tinto y calcula la matriz de confusión, medidas de desempeño relevantes y, finalmente, compara la precisión de diferentes algoritmos de Aprendizaje Automático en función de dichos datos, como ilustra la Figura 3.

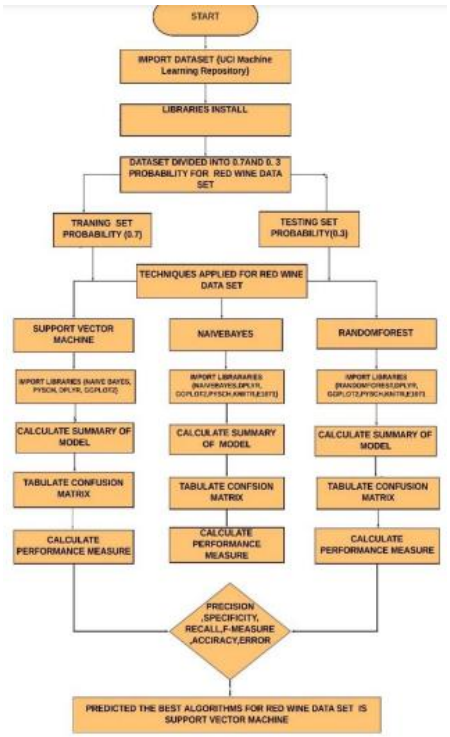

Figura 3. Diagrama de flujo paso a paso desarrollado en este estudio

El primer paso es importar las librerías en el software elegido y desarrollar el modelo para cada uno de los tres algoritmos de Aprendizaje Automático mencionados previamente. Después, se crea la matriz de confusión de 6*6. Dicha matriz variará en función de las observaciones del conjunto de datos y la calidad.

Los resultados del algoritmo *Naïve Bayes* en cuanto a precisión para el conjunto de entrenamiento y el conjunto de prueba son los siguientes: 55,91% y 55,89%; usando el algoritmo **SVM** son 67,25% y 68,64% y usando *Random Forest* son 65,83% y 65,46%.

Los resultados podrían mejorar fusionando loa algoritmos, eligiendo un ajuste adecuado para el hiperplano del algoritmo SVM y seleccionando un árbol balanceado adecuado.

2.9 Artículo 8: Análisis de calidad del vino por medio de técnicas de Inteligencia Artificial [8]

En este artículo, se pretende encontrar relaciones entre variables presentes en el vino que permitan mejorar la calidad de éste. La forma de analizar dicha calidad será mediante el empleo de *Decision Trees*, generados sobre los datos ofrecidos por la plataforma WEKA aplicando el algoritmo de clasificación J48 sobre las variables independientes iniciales.

Los pasos a seguir conforman un modelo donde se comienza construyendo una Base de Datos (BBDD UCI disponible en la web[3] en diferente fuente) formada por 300 registros obtenidos de forma aleatoria entre muestras de vino tinto y blanco indistintamente, para conseguir una primera correlación de variables que centre el estudio de calidad. Con esta información, se construye el archivo WEKA incluyendo las 10 variables independientes frente a la calidad (variable dependiente) y la correlación moderada entre algunas de las variables que se obtuvieron en el paso anterior.

A partir de un proceso de clasificación mediante el Algoritmo J48, se elabora el árbol de decisión que ofrece conclusiones sobre los principales factores químicos influyentes en la calidad del vino. Como último paso, se realiza una optimización del proceso donde se incluye una nueva variable producto de la correlación más baja entre dos de las variables iniciales, alcohol y sulfatos, obteniendo como resultado un árbol de decisión simplificado presente en la Figura 4, que presenta porcentajes de clasificación superiores al 95%.

[3] *Base de Datos empleada en este estudio disponible en: http://www3.dsi.uminho.pt/pcortez/wine/*

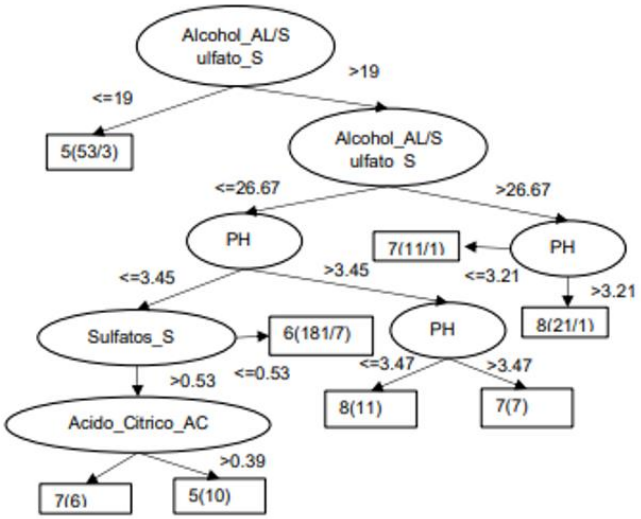

Figura 4. Árbol de Decisión Optimizado

Las variables clave para conseguir mejorar la calidad del vino según este estudio son las siguientes: alcohol, pH, sulfatos y ácido cítrico. Se llega a la conclusión respaldada por el modelo propuesto de que, controlando estas cuatro variables, así como la relación alcohol-sulfatos, se consigue mejorar la calidad del vino.

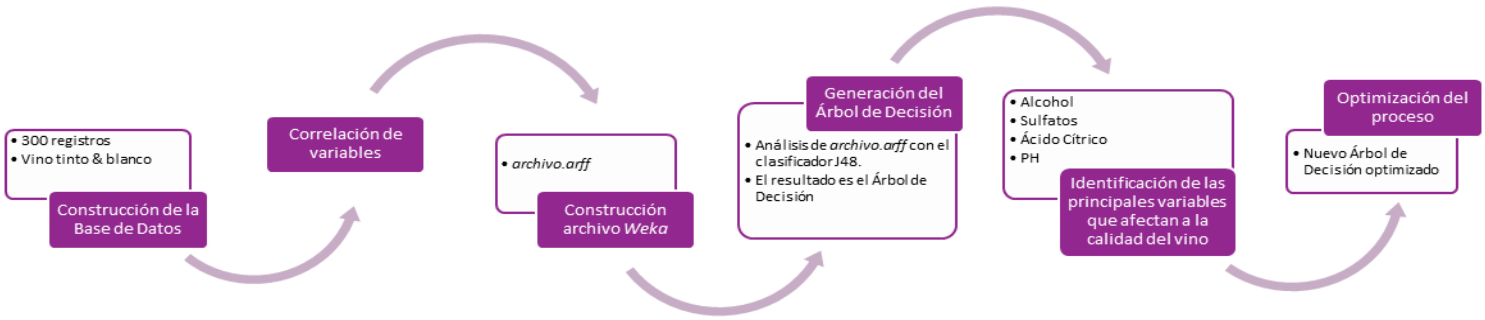

Figura 5. Resumen del proceso de análisis de calidad

En la Figura 5 podemos encontrar un breve esquema de los pasos seguidos en este artículo, así como las variables o características importantes de cada fase o bloque.

2.10 Artículo 9: *Research on red wine quality based on Data Visualization* [9]

En este artículo, se busca mejorar la calidad del vino tinto a partir del estudio de 11 variables fisicoquímicas y sensoriales presentes en el vino; tomando valores desde el 0 (valor mínimo) hasta el 10 (valor máximo). Dichas variables son las siguientes: acidez fija, acidez volátil, ácido cítrico, azúcar residual, cloruros, dióxido de azufre libre, dióxido de azufre total, densidad, pH, sulfatos y alcohol. La salida del modelo a medir será la calidad. Para ello, se pretende aplicar *Data Mining* a partir de *Correlación de Pearson*.

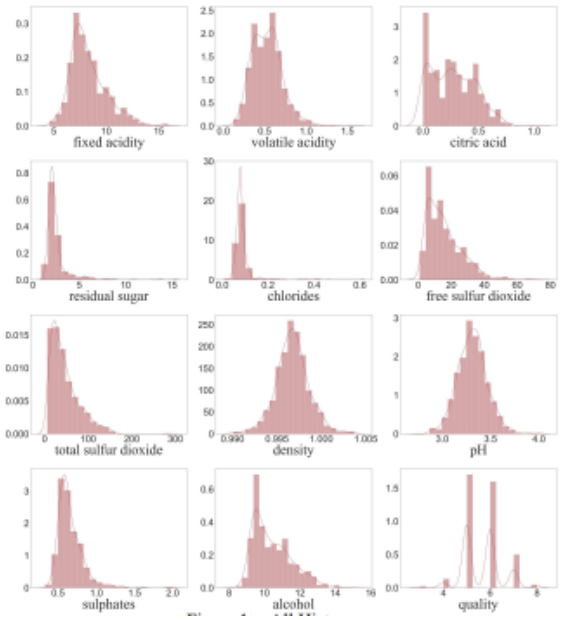

Figura 6. Histograma representando la distribución muestral de cada factor de influencia

Se realiza, en primer lugar, un análisis univariado representado en forma de histograma para observar la distribución muestral de cada factor de influencia. Se emplea la Base de Datos UCI para el vino tinto, formada por 1599 muestras. En la Figura 6 podemos apreciar cómo priman muestras de vino con valores de calidad comprendidos entre 3 y 8, con especial concentración en los índices 5 y 6, es decir, vino de calidad media.

En segundo lugar, se utiliza un Mapa de Calor para analizar la correlación calculando el *Coeficiente de Pearson*.

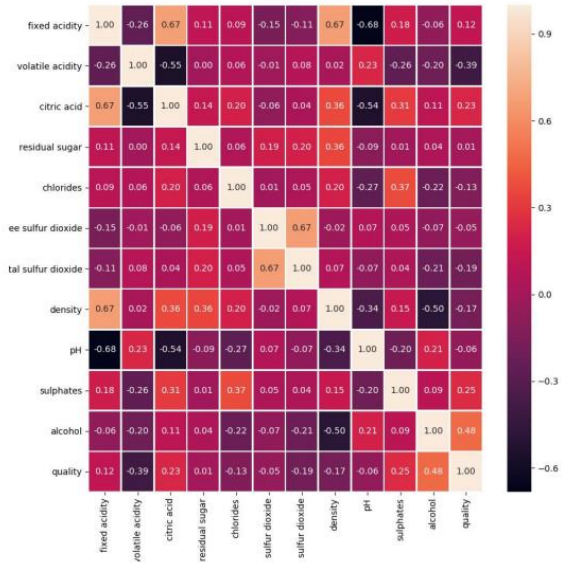

Figura 7. Mapa de Calor de las variables analizadas

Finalmente, partiendo del *Mapa de Calor* representado en la Figura 7, se recalca la tendencia de los factores clave representando los resultados con *box-plot*.

El cómputo de este estudio recae en afirmar que el alcohol, el ácido cítrico, los sulfatos y la acidez volátil son los factores más influyentes que afectan la calidad del vino tinto.

2.11 Artículo 10: *Wine quality prediction using Data Mining* [10]

Este estudio clasificará en tres categorías principales las muestras de vino tinto y vino blanco y comparará y la precisión de los algoritmos *Naive Bayes*, *Simple Logistic*, *KStar*, *JRip* y J48 para predecir la calidad del vino.

El conjunto de datos está formado por 178 muestras de vino con 13 atributos (entradas) y calidad diferenciada en distintas clases “1”, “2” o “3” (la *clase* será la salida). Los atributos son los siguientes: alcohol, ácido málico, cenizas, alcalinidad de cenizas, magnesio, fenoles totales, flavonoides, fenoles no flavonoides, proantocianinas, intensidad de color, tinte, OD280/OD315 de vinos diluidos y prolina. El conjunto de datos se carga en WEKA 3.9. definiendo 80% como conjunto de entrenamiento y el 20% restante como conjunto de prueba.

Naïve Bayes ofrece un 100% de precisión sobre los datos de prueba; mientras que *Simple Logistic*, KStar y J48 ofrecen un 97.22% de precisión. JRip empeora con un porcentaje de 94.44%. El algoritmo J48 brinda un árbol representando las características más relevantes. Este gráfico se ilustra en la Figura 8.

Figura 8. Esquema en árbol del clasificador J48

El clasificador ***Naïve Bayes*** es, por tanto, el algoritmo de clasificación más sencillo y rápido. Maneja valores continuos y discretos para hacer predicciones probabilísticas. Es altamente escalable e insensible a las características irrelevantes. Pero tiene una limitación, si el conjunto de entrenamiento es demasiado grande, otros clasificadores que apliquen validación cruzada de forma recurrente sobre el conjunto de datos ofrecerán mejores resultados.

2.12 Artículo 11: *Wine quality classification with Multilayer Perceptron* [11]

Este artículo trata sobre la clasificación de la calidad del vino con Perceptrón Multicapa utilizando una Red Neuronal Profunda. A diferencia de la mayoría de escritos, en lugar de estudiar varias técnicas se evalúa cómo un modelo de Aprendizaje Profundo predice la calidad utilizando dos funciones de activación diferentes.

La Base de Datos empleada es UCI de vino tinto y blanco, formada por 1599 instancias de vino tinto y 4898 de vino blanco, con 11 atributos por muestra.

Se emplea el lenguaje R con *Keras* y *Tensorflow*. La Red Neuronal Profunda se construye utilizando un modelo de Perceptrón Multicapa con múltiples capas ocultas. Se

intenta ajustar el modelo con diferentes partes temporales, tamaños de lote y tamaño de grupo de validación.

Las distintas funciones de activación serán la función de activación de la unidad lineal rectificada, *Rectified Linear Unit* (RELU) y *tanh*, para vino tinto y blanco. Cuando se emplea RELU en el conjunto de datos de vino tinto, se obtiene un 53% de precisión de validación. Al usar la función *tanh* se obtiene una precisión de validación del 52% con 200 *epochs* y 66 tamaño de lote con división de validación en 0,15. Para el conjunto de datos de vino blanco, RELU obtiene precisión de validación del 48%, mientras que *tanh*, obtiene una precisión de validación del 53 % con 200 *epochs* y 66 tamaño de lote con división de validación en 0,15.

Los resultados de los experimentos variando valores de *epochs* y tamaño de lote sugiere que los datos se sobreajustan o se subajustan aumentando *epochs* y tamaños de los lotes. Si bien el uso de una función de activación diferente no afecta mucho.

2.13 Artículo 12: *Prediction of quality for different type of wine based on different feature sets using Supervised Machine Learning techniques* [12]

Este artículo tiene como objetivo evaluar la calidad del vino en base a los atributos del vino que influyen de forma directa en la calidad de este.

Para este estudio, la Base de Datos empleada es la desarrollada por UCI, estudiando el vino tinto y el vino blanco de forma independiente. Partiendo del número total de muestras, se realiza una división arbitraria de las muestras para formar el grupo de entrenamiento y el grupo de pruebas.

Se emplean diferentes técnicas de selección de características: Algoritmo Genético (*Genetic Algorithm*, GA) basado en selección de características y, por otro lado, Recocido Simulado (*Simulated Annealing*, SA) basado en selección de características. Los clasificadores empleados en ambos algoritmos de selección de características son RPART, C4.5, PART, *Bagging* CART, *Random Forest*, *Boosted* C5.0, SVM, LDA y NB.

Los resultados afirman que el clasificador **SVM** destaca respecto a los demás algoritmos evaluados. El resultado muestra que el clasificador SVM funciona mejor para ambos tipos de algoritmos de selección de características, GA y SA; aunque los resultados sobre conjuntos de características basadas en SA ofrecen mejores resultados. SA obtiene los índices más altos en términos de precisión, sensibilidad y especificidad, como ilustra la Figura 9.

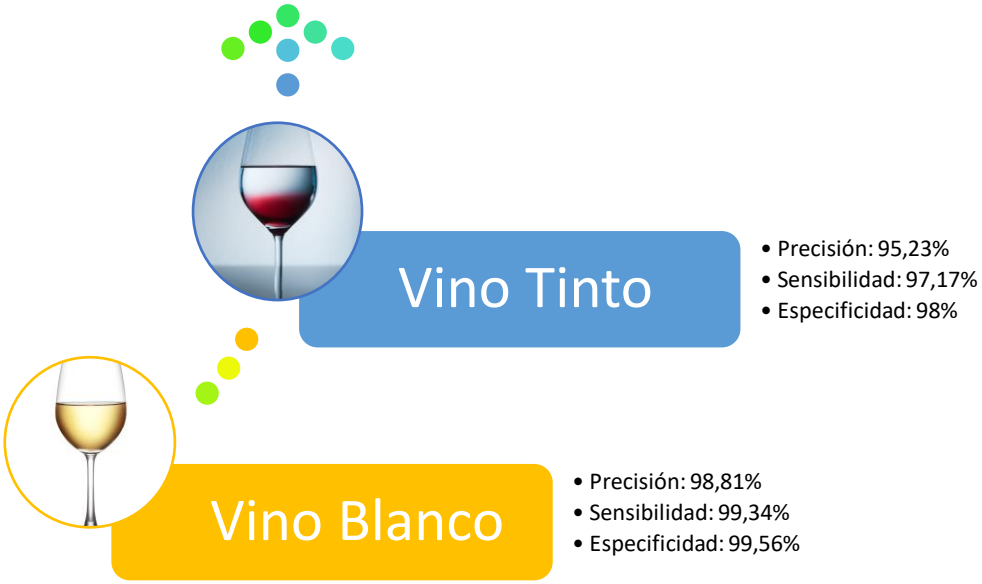

Figura 9. Resultados obtenidos por SVM, los cuales le posicionan como el mejor clasificador evaluado

2.14 Artículo 13: *A classification approach with different feature sets to predict the quality of different types of wine using Machine Learning techniques* [13]

Este estudio pretende predecir la calidad de distintos tipos de vino empleando, para ello, los algoritmos de clasificación RPART, C4.5, PART, *Bagging* CART, *Boosting* C5.0 y Árboles de Decisión no lineales. El fin de evaluar estos clasificadores reside en validar métodos de Inteligencia Artificial para poder concretar la calidad del vino a partir de variables fisicoquímicas de su composición.

Para realizar el análisis de los resultados, se compararán los índices obtenidos por cada clasificador en cuanto a precisión, sensibilidad, especificidad, valor predictivo positivo (PPV) y valor predictivo negativo (NPV).

Además, previo a la obtención de resultados, se somete a las muestras (provenientes de la BBDD UCI de vino tinto y vino blanco, que cuenta con 11 variables fisicoquímicas y como salida la calidad del vino) a un proceso de selección de características y reducción de variables mediante los algoritmos: Análisis de Componentes Principales *(Principal Component Analysis,* PCA) y Eliminación de Características Recursivas (*Recursive Feature Elimination,* RFE).

Los resultados finales afirman que el clasificador ***Random Forest*** es el mejor algoritmo a la hora de predecir la calidad del vino, tanto incluyendo PCA como RFE, a la hora de realizar una selección de variables. Los valores obtenidos para **RF + RFE** en cuanto a la capacidad de predecir la calidad del vino son de 94,51% para vino tinto, y 97,79% prediciendo la calidad del vino blanco. Los resultados se recogen en la Tabla 8.

RFE+RF	*Precisión*	*Sensibilidad*	*Especificidad*	*PPV*	*NPV*
Vino Blanco	97,79%	98,82%	99,12%	99,56%	0,9976%
Vino Tinto	94,51%	95,75%	97,56%	98,89%	98,91%

Tabla 8. Mejores resultados del estudio, obtenidos por RF + RFE tanto para vino blanco como tinto

Los valores de precisión empleando RF + PCA son algo inferiores, pero superiores al 92% para ambos tipos de vino.

2.15 Artículo 14: *Wine quality detection through Machine Learning algorithms* [14]

Este artículo compara los resultados obtenidos por dos técnicas de *Data Mining*[4] distintas: *Random Forest* y *Logistic Regression* (LR), para concretar qué técnica ofrece mejores resultados en cuanto a detección de la calidad del vino.

[4] *Data Mining*: Tarea de encontrar patrones ocultos en un gran conjunto de datos que es imposible analizar a partir de datos resumidos. Una de las técnicas de *Data Mining* es la clasificación, dentro de la cual se encuentran estos dos algoritmos empleados en dicho documento: *Logistic Regression* y *Random Forest.*

La calidad, atributo incluido en la información ofrecida por la Base de Datos UCI para vino tinto, se define como 'buena' si adquiere una puntuación por encima de 5, 'mala' si el valor es inferior.

En primer lugar, se aplica una fórmula de estandarización para trabajar con **datos normalizados**. El preprocesamiento de datos incluye la definición de variables y función de salida, manejo de datos faltantes, escalado de características y división de datos en datos de tren y datos de prueba, codificación de datos, etc., todo esto mejora los datos estándar. Del conjunto de datos con 1599 muestras, 438 se eliminan en la detección de valores atípicos.

El conjunto de datos se divide aleatoriamente, conformando el conjunto de entrenamiento un 80% del total de datos y el 20% restante empleado como datos de prueba y test. Los datos de entrenamiento se utilizan para ajustar mejor el modelo y los datos de prueba se utilizan para examinar el rendimiento del modelo.

A continuación, se aplican los algoritmos de clasificación *Logistic Regression* y *Random Forest* con 50 árboles de decisión, obteniendo en cada caso una Matriz de Confusión diferente. Después, se procede a comparar los resultados obtenidos por ambas. La precisión del modelo para RF es de un 84%, mientras que para LR su tasa de precisión es del 76%, datos que podemos encontrar en la Tabla 9.

	Logistic Regression	Random Forest
Accuracy	76%	84%
Precision	0.75	0.82
Recall/ Sensitivity	0.81	0.90
Specificity	0.71	0.79
F1_score	0.77	0.85
Out-of-bag rate	0.23	0.15
Error at $\alpha = 0.05$	0.28	0.20
K-fold Accuracy mean	0.72	0.77
K-fold standard deviation	0.04	0.05

Tabla 9. Comparación de resultados entre Logistic Regression y Random Forest haciendo validación cruzada k-fold con 10 validaciones

Podemos observar en la Tabla 9 el parámetro *F1_score*. Este parámetro ayuda a la hora de comparar objetivamente, dado que se trabaja con un conjunto de datos **no** equilibrado que cuenta con un número diferente de instancias en cada categoría.

Se observa, finalmente y respaldándose en los datos volcados en la Tabla 9, que ***Random Forest*** basado en árboles de decisión generó mejores predicciones que *Logistic Regression*.

2.16 Artículo 15: *Fuzzy logic tool for wine quality classification* [15]

El objetivo del presente estudio es el desarrollo de un modelo de toma de decisiones multicriterio de Lógica Difusa para clasificar objetivamente la calidad del vino en función de los atributos de la uva seleccionada. El modelo desarrollado se comparará con los datos brindados por 11 panelistas profesionales con un mínimo de 10 años de experiencia en vinificación de vino tinto Agiorgitiko en la región de Nemea.

Se recolectaron muestras de **trece viñedos** comerciales en Nemea, Grecia plantados con *Vitis vinifera cv. Agiorgitiko* y se vinificaron durante dos años consecutivos (2012 y 2013). Se tomó una submuestra de 300 bayas por viñedo para determinar los valores medios de bayas por viñedo. Las **8 variables** medidas en la cosecha y utilizadas como entradas son: sólidos solubles totales, pH, volumen de bayas, infección por botritis, coloración de la semilla de uva, extractabilidad de antocianina, densidad óptica (OD 520) y fenoles de la piel (Dpell).

FMCDM, *Fuzzy Logic Multi Criteria Decision Making*, se basa en la extracción de la experiencia y el conocimiento de expertos enólogos y literatura para la selección de las variables, así como para la evaluación de la importancia de cada parámetro. Se construyeron variables lingüísticas y reglas para presentar la relación entre la uva y la calidad del vino.

Utilizando el sistema FMCDM obtenemos el valor de salida que va de 0 a 1 y se interpreta a través de los *membership degrees* o grados de calidad de los diferentes conjuntos difusos. Además, la puntuación de salida del sistema FMCDM se normalizó en el mismo rango que la evaluación sensorial.

Los resultados de FMCDM y la evaluación de expertos fue de concordancia 83,46% y 76,92% para la cosecha 2012 y 2013, respectivamente. Por ello, se concluye anotando la eficacia del sistema FMCDM, ya que clasificó los vinos de manera similar a los expertos.

2.17 Artículo 16: *Selection of important features and predicting wine quality using Machine Learning techniques* [16]

Este escrito emplea *Linear Regression*, Redes Neuronales y S*upport Vector Machine* para predecir la calidad del vino tinto y blanco de dos formas distintas: en primer lugar, determinar la dependencia de la variable objetivo dentro de las variables independientes y, en segundo, predecir el valor que tomará dicha variable. Para ello, se seleccionarán las variables importantes.

La Base de Datos empleada es la de UCI, siendo 1599 las muestras de vino tinto, y 4898 asociadas al vino blanco. Dichas muestras contienen 11 variables fisicoquímicas. Se utiliza transformación lineal para **normalizar** todos los valores entre 0 y 1 y evitar más influencia de algunas muestras. Se logra dividiendo todos los valores de entrada por el valor de variable máximo.

Después, se aplica *Linear Regression* o Regresión Lineal para determinar la dependencia de calidad sobre las 11 variables independientes (predictores). Se utiliza para predecir el valor de una variable en función del valor de dos o más variables. Los resultados de los experimentos realizados se adjuntan en la Tabla 10.

R= .60045958	R^2= .36055170		Adjusted R^2= .35611948			p < 0.0000
	Standard regression coefficient (b*)	Standard error of b*	Raw regression coefficient (b)	Standard error of b	t value	p-value
fixed acidity	0.053879	0.055944	0.0250	0.02595	0.96308	0.335653
volatile acidity	**0.240261**	**0.026851**	**1.0836**	**0.12110**	**8.94780**	**0.000000**
citric acid	-0.044038	0.035502	-0.1826	0.14718	-1.24044	0.214994
residual sugar	0.028513	0.026192	0.0163	0.01500	1.08860	0.276496
chlorides	**0.109230**	**0.024436**	**1.8742**	**0.41928**	**4.47007**	**0.000008**
free sulfur dioxide	**0.056491**	**0.028124**	**0.0044**	**0.00217**	**2.00864**	**0.044745**
total sulfur dioxide	**0.132979**	**0.029684**	**0.0033**	**0.00073**	**4.47983**	**0.000008**
density	-0.041789	0.050558	-17.8812	21.63310	-0.82657	0.408608
pH	**0.079080**	**0.036628**	**0.4137**	**0.19160**	**2.15897**	**0.031002**
sulphates	**0.192336**	**0.023999**	**0.9163**	**0.11434**	**8.01430**	**0.000000**
alcohol	**0.364470**	**0.034948**	**0.2762**	**0.02648**	**10.42901**	**0.000000**

R= .53091465	R^2= .30675369		Adjusted R^2= .28025362			p < 0.0000
	Standard regression coefficient (b*)	Standard error of b*	Raw regression coefficient (b)	Standard error of b	t value	p-value
fixed acidity	**0.062430**	**0.019889**	**0.066**	**0.02087**	**3.1389**	**0.001706**
volatile acidity	**0.212048**	**0.012951**	**1.863**	**0.11379**	**16.3733**	**0.000000**
citric acid	0.003019	0.013087	0.022	0.09577	0.2307	0.817589
residual sugar	**0.466653**	**0.043109**	**0.081**	**0.00753**	**10.8249**	**0.000000**
chlorides	-0.006100	0.013483	-0.247	0.54654	-0.4524	0.650973
free sulfur dioxide	**0.071681**	**0.016210**	**0.004**	**0.00084**	**4.4219**	**0.000010**
total sulfur dioxide	-0.013712	0.018142	0.000	0.00038	-0.7558	0.449791
density	**0.507528**	**0.064417**	**150.284**	**19.07451**	**7.8788**	**0.000000**
pH	**0.117021**	**0.017967**	**0.686**	**0.10538**	**6.5131**	**0.000000**
sulphates	**0.081374**	**0.012936**	**0.631**	**0.10039**	**6.2905**	**0.000000**
alcohol	**0.268840**	**0.033656**	**0.193**	**0.02422**	**7.9878**	**0.000000**

Tabla 10. Valores de las características obtenidas para vino tinto (izquierda) y para vino blanco (derecha)

Para el vino tinto, el valor de R^2 ajustado es 0,3561, lo que muestra una dependencia de la calidad del 35,61 % en todos los predictores en su conjunto. Si este valor es inferior al 50 % significa que hay uno o más predictores; dichos predictores no son buenos para predecir el valor de la calidad. Al mismo tiempo, el valor *P* es mucho menor que 0,05 como se muestra en la Tabla 10.1, lo que indica que R^2 ajustado es significativamente diferente de cero y niega la hipótesis nula. Los predictores como acidez volátil, cloruros, dióxido de azufre libre, dióxido de azufre total, pH, sulfatos y alcohol son predictores importantes de la calidad, ya que el valor *p* para estos predictores es inferior a 0,05 (intervalo de confianza del 95%). El coeficiente de regresión *b* para la acidez volátil es 1,0836, lo que indica que si todos los demás predictores están controlados (constantes), entonces el incremento de una unidad de acidez volátil aumenta la calidad en ese mismo valor, 1,0836.

La Tabla 10.2 indica que la calidad del vino blanco depende en un 28,02 % de todos los predictores en su conjunto. También muestra que el valor *p* para predictores individuales como acidez fija, acidez volátil, azúcar residual, dióxido de azufre libre, densidad, pH, sulfatos y alcohol es mucho menor que 0,05, lo que significa que estos predictores predicen la calidad de manera más significativa en comparación con otros. El coeficiente de regresión no estandarizado *b* es 0,06 para acidez fija, lo que indica que si todos los demás predictores están controlados (constantes), el incremento de una unidad en la acidez fija aumenta la calidad en 0,06. La misma declaración se puede hacer para otros predictores.

Por último, la calidad del vino se predice con ayuda de Redes Neuronales y *Support Vector Machine* (SVM) considerando todos los predictores, así como los predictores seleccionados.

Network name	Training error	Test error	Validation error
(All features) MLP 11-5-1	0.187312	0.195660	0.169588
(Selected features) MLP 8-5-1	0.145736	0.146024	0.140383

Network name	Training error	Test error	Validation error
(All features) MLP 11-5-1	0.234133	0.241568	0.243491
(Selected features) MLP 8-5-1	0.190578	0.207456	0.199758

Figura 10. Resultados de aplicar Redes Neuronales para vino tinto (izquierda) y vino blanco (derecha)

Document No.	Original Quality value	Quality Output (11-5-1) for all features	Quality Output (7-5-1) for selected features
1	5.000000	4.825105	5.023361
82	5.000000	5.118246	4.989526
228	5.000000	5.260592	5.116120

Figura 11. Ejemplo de resultados obtenidos con SVM para vino tinto

Se puede resumir en base a los resultados obtenidos en las Figuras 10 y 11, que SVM es mejor técnica de Aprendizaje Automático para las predicciones de calidad del vino, consiguiendo índices de clasificación de calidad muy próximos a los reales. Al mismo tiempo, se demuestra que tanto SVM como Redes Neuronales pueden hacer predicciones más precisas utilizando predictores específicos en lugar de todos los atributos.

2.18 Artículo 17: *An analytical toast to wine: Using Stacked Generalization to predict wine preference* [17]

Este trabajo presenta un enfoque multi-método para predecir el nivel de calidad de una muestra de vino dadas sus propiedades fisicoquímicas. Los resultados se contrastan con los obtenidos en [27] -punto 2.28-, explorando las posibilidades que ofrecen las siguientes técnicas *Machine Learning*: Regresión lineal (LR), Redes Neuronales (NNs) y *Suport Vector Machines* (SVMs).

La generalización apilada/acumulada se considera un tipo de meta-aprendizaje debido a la transferencia de información de nivel base al meta-nivel. La capacidad de aprender de los aprendices base (*base-learners*) ayuda a reforzar las fortalezas y debilidades de cada aprendiz base para una mayor precisión en el metanivel. Por lo tanto, la generalización apilada funcionará al menos tan bien, si no mejor, que el mejor estudiante base para un problema dado. El proceso puede ser organizado de la siguiente manera:

1. Construir un conjunto de datos que consista en predicciones de espera de un conjunto de estudiantes base usando un conjunto de entrenamiento y otro de prueba. Dicho *dataset* serán los metadatos.
2. Ejecutar un meta-aprendiz que utilice las predicciones realizadas en el nivel anterior como entradas, es decir, entrenar el meta-aprendiz en los metadatos.

El marco de generalización apilado propuesto se ejecuta en el conjunto de datos de vino blanco formado por 4898 muestras provenientes de la Base de Datos del repositorio de *Machine Learning* UCI (*UCI Machine Learning Repository database*).

Se implementa únicamente el estudio sobre el vino blanco ya que la generalización apilada funciona mejor para volúmenes grandes de datos, excluyendo en este estudio, por tanto, el análisis de vino tinto. El análisis se desarrolla en R (versión 3.3.2) con el paquete 'CARET', *Classification And REgression Training*.

En este documento, se trabaja con un conjunto de 20 aprendices base (*base-learners)* diversos y se combina con una *formulación bayesiana rápida* para maximizar el poder predictivo. Además, una '*variable importance scheme'* se deriva para ayudar a la generalización apilada. Con una metodología experimental rigurosa, los resultados predictivos muestran que esta estrategia de conjunto puede proporcionar un rendimiento significativamente mejor que el propuesto en [27].

La variable predictora más influyente dada por generalización apilada es el porcentaje de alcohol por unidad de volumen. Esto es consistente con la teoría enológica, los aumentos en el nivel de alcohol presente típicamente culminan en vinos de mayor calidad. Los porcentajes obtenidos por cada variable de calidad superan con una diferencia significativa los resultados obtenidos por algoritmos de Aprendizaje Automático probados en [27].

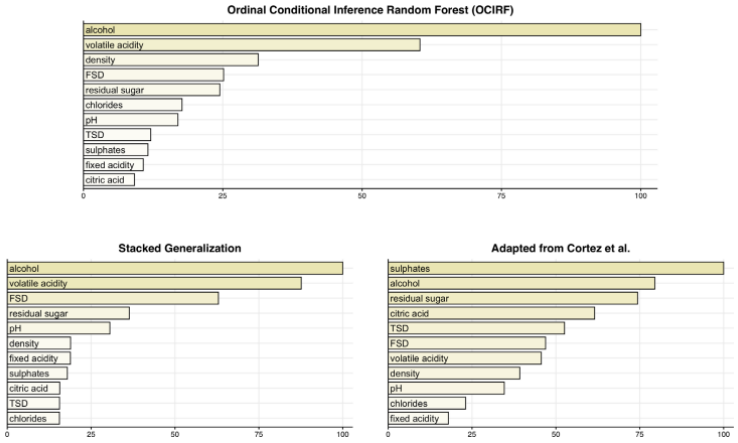

Figura 12. Importancia de cada variable en el estudio actual (generalización apilada) comparado con los resultados obtenidos en [27] y OCIRF

En la Figura 12 se ilustra el porcentaje (%) de importancia que representa cada variable tanto en el estudio pertinente, como contrapuesto con los resultados obtenidos en [27]. Además, se incluye una tercera gráfica, *Random Forest* de Inferencia Condicional Ordinal, *Ordinal Conditional Inference Random Forest* (OCIRF), que permite establecer una línea base imparcial para evaluar la importancia de cada variable en el conjunto de datos de vino blanco.

2.19 Artículo 18: *Classification of Wine Quality with Imbalanced Data* [18]

En este documento se intenta conseguir una predicción de la calidad del vino a partir del uso de los algoritmos *Decision Tree, Adaptive Boosting y Random Forest* con distinto número de árboles de decisión, sobre los datos iniciales y sobre los obtenidos tras aplicar previamente un proceso de balanceamiento de los datos utilizando el algoritmo SMOTE (*Synthetic Minority OverSampling Technique)*.

La *técnica de sobremuestreo de minorías sintéticas* (SMOTE) es un método de sobremuestreo para abordar el problema del desequilibrio de datos. La idea básica es volver a muestrear el espacio de datos para crear más puntos sintéticos de la clase menos dominante. En este código, el *wine_data* (conjunto de datos iniciales) se alimenta al algoritmo SMOTE con el parámetro de sobremuestreo 600 para generar casos de la clase minoritaria y con parámetro de submuestreo 100 para seleccionar casos de la clase mayoritaria. Los casos mayoritarios y minoritarios originales (provenientes de la BBDD UCI, con calidades mayoritarias entre 5 y 6) fueron 4535 y 363; que tras aplicar el algoritmo SMOTE pasaron a tomar valores 2178 y 2541, respectivamente.

Este artículo implementa *System R* para la preparación y análisis de datos. Se aplicaron tres métodos de clasificación, *Decision Tree,* AdaBoost*, Random Forest,* en los datos antes y después de aplicar el algoritmo SMOTE para equilibrar los datos. Para cada modelo, se utilizó el 75% de las instancias de datos para el entrenamiento, el 15% para la validación y un 15% para pruebas. Tras comparar diversos parámetros de rendimiento, se llega a la conclusión de que **RF** aporta los mejores resultados en cuanto a tasas de error, así como valores ROC, como ilustra la Tabla 11.

(a) Before applying SMOTE

Models	Error Rate	Specificity $(TN\ /\ N)$	ROC
Decision tree	7.2%	0/53	0.50
AdaBoost	6.8%	4/53	0.83
Random Forest	5.4%	17/53	0.88

(b) After applying SMOTE

Models	Error Rate	Specificity $(TN\ /\ N)$	ROC
Decision tree	28.7%	276/388	0.73
AdaBoost	12.4%	350/388	0.93
Random Forest	4.7%	373/388	0.99

Tabla 11. Resultados que ofrecen los algoritmos de clasificación

TODOS LOS ALGORITMOS OFRECEN MEJORES RESULTADOS TRABAJANDO CON LOS DATOS TRAS APLICAR SMOTE QUE SIN REALIZAR EL PREBALANCEO.

Este artículo también utiliza el estudio realizado para identificar las variables determinantes en la calidad del vino. Los cálculos se realizan sobre las 11 variables iniciales obteniendo la *Precisión de disminución media* y la *Disminución media de Gini.* De este modo, identificamos aquellas variables que están más involucradas en la determinación de la calidad del vino. Los resultados indican que la acidez volátil, el dióxido de azufre libre y el alcohol tienen los valores más altos, por lo que son los parámetros más relevantes; mientras que el dióxido de azufre total juega el papel menos importante en la determinación de la calidad de un vino.

2.20 Artículo 19: *The classification of white wine and red wine according to their physicochemical qualities* [19]

En este estudio se pretende predecir el tipo de vino y su calidad tanto para vinos tintos como blancos. Para ello, se emplearon conjuntos de datos tomados del repositorio de Aprendizaje Automático de UC Irvine, formado por 1599 instancias para vino tinto y 4898 instancias para vino blanco; con 11 características de datos fisicoquímicos. La Base de Datos agrupa las muestras de vino tinto en seis grupos distintos de calidad, mientras que agrupa en siete grupos las instancias de vino blanco. Este escrito compara los resultados obtenidos de los siguientes algoritmos: *k-nearest-neighbourhood*, *Random Forest* y *Support Vector Machines.*

En primer lugar, se desarrolló un proceso de clasificación de las instancias empleando el método de validación cruzada con k = 10 (valor de k más favorable[5]), y, de forma simultánea, también mediante *Percentage Split* (división porcentual). La Base de Datos se divide aleatoriamente en dos grupos, siendo el 80% de entrenamiento y el resto de prueba. El mejor resultado de clasificación del tipo de vino se obtuvo mediante el

[5] *El k-fold es una prueba experimental donde la Base de Datos se divide aleatoriamente en k bloques de objetos disjuntos. El algoritmo de minería de datos se entrena usando k-1 bloques y el bloque restante se utiliza para probar el rendimiento del algoritmo, este proceso se repite k veces.*

algoritmo ***Random Forest*** tanto para las muestras de vino tinto como para las de vino blanco. La precisión con este algoritmo es 99,5229% y 99,4611% respectivamente.

Después, se usaron los tres algoritmos de minería de datos para clasificar la calidad tanto del vino tinto y del vino blanco. ***Random Forest*** obtuvo los mejores resultados prediciendo la calidad del vino. La precisión de validación cruzada y división porcentual con este algoritmo para vino blanco es de 70,3757% y 68,6735% respectivamente. De igual forma, *Random Forest* también obtuvo los mejores resultados para vino tinto, 69,606% y 71,875% respectivamente.

Por último, se realiza una última prueba haciendo reducción de variables: Análisis de Componentes Principales (PCA). Se llega a la conclusión de que realizar una reducción de variables en la selección de características aumenta la tasa de éxito de clasificación en el algoritmo *Random Forest* para el vino tinto (aumentó de 69,606% a 71,232% para validación cruzada y de 71,875% a 73,4375% para el modo de división porcentual), mientras que redujo los valores obtenidos para medir la calidad de las muestras de vino blanco (disminuyó de 70,3757% a 69,9061% para validación cruzada, y de 68,6735% a 67,449% para el modo de división porcentual).

2.21 Artículo 20: *A new mathematical modelling approach for viticulture and winemaking using Fuzzy Cognitive Maps* [20]

Este artículo pretende modelar y predecir la calidad y la cantidad del vino empleando *Fuzzy Cognitive Maps* (FCM) entrenados con el algoritmo *Non Linear Hebbian Learning* para la cosecha de un vino concreto. Para crear el *Fuzzy Cognitive Map*, se realiza una labor de estudio con expertos que ayudan a definir las variables influyentes (entradas del modelo) así como las relaciones entre estas.

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14
C1	-	W	VS	M	M	M	M	-	M	-	-	-	VS	S
C2	S	-	S	VS	S	S	W	M	M	-	-	-	VS	S
C3	W	-	-	W	S	M	M	W	S	-	-	-	VS	S
C4	W	-	M	-	M	S	M	-	W	-	-	-	VS	S
C5	-	-	-	-	-	W	M	-	W	-	-	-	S	S
C6	-	-	-	W	S	-	VS	W	M	-	-	-	S	VS
C7	-	-	-	-	S	S	-	M	S	-	-	-	S	-
C8	-	-	-	-	W	M	-	-	S	-	-	-	VS	-
C9	-	-	-	-	-	-	-	-	-	-	-	-	VS	-
C10	W	-	-	-	-	M	-	-	-	-	-	-	VS	VS
C11	W	-	-	-	-	M	-	-	M	-	-	-	VS	VS
C12	-	-	-	-	-	-	-	-	-	-	-	-	VS	VS
C13	-	-	-	-	-	-	-	-	-	-	-	-	-	-H
C14	-	-	-	-	-	-	-	-	-	-	-	-	-H	-

Tabla 12. Variables lingüísticas que describen las relaciones entre conceptos

Se consideran como entradas al modelo elaborado de *Fuzzy Cognitive Map*, por tanto, las variables numeradas desde C1 hasta C12 representadas en la Tabla 12: características especiales del suelo, condiciones climáticas en el viñedo de la región, variedad de uva, factor humano (labores de cultivo y cuidados agrícolas en el viñedo), fermentación alcohólica, enfermedades, alteraciones y deterioro de la calidad del vino, sustancias enológicas adicionales al vino, almacenamiento de vino en barricas, maduración /crianza del vino, lluvia antes de la cosecha, tiempo de cosecha y poda. Las salidas de dicho esquema a estudiar mostrado en la Figura 13 serán la calidad y la cantidad de vino.

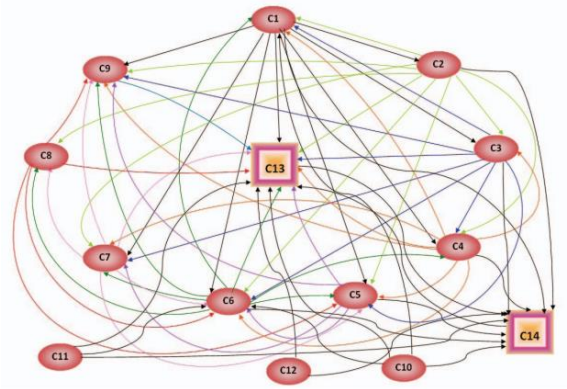

Figura 13. Mapa Cognitivo Difuso

El primer caso consta de la variante de uva *Roditis*, uva de color rosa. La época de la cosecha se realiza a finales de septiembre y la poda se realizó en marzo. Los resultados satisfactorios de dicha predicción fueron validados por las estadísticas de *Achaia Clauss*, una bodega griega ubicada en Patras, en Peloponeso. En este caso, con estas condiciones casi perfectas, la calidad y cantidad del vino (C13, C14) son los mejores obtenidos. Ambos valores están por encima de 0,9. Eso significa que, intentando replicar estas condiciones, los resultados de la producción de vino serán lo mejores posible.

En el segundo ejemplo, se analiza un caso especial: producción de 2014 de variante *Bombino Bianco*, Puglia; donde la producción se vio reducida en un 30% debido a causas naturales, por lo que la cosecha de ese año comenzó con retraso y se extendió hasta finales de octubre. En este caso, ambas variables de salida ofrecen peores resultados. La cantidad es la variable que se ve más afectada, ya que su valor es de tan solo 0.3320 (C14= 0,3320), el más bajo en el punto de equilibrio. En cuanto al estudio de la calidad, ésta no se ve tan afectada (C13=0,68), por lo que la producción de ese año se considera satisfactoria.

Los resultados de ambos esquemas reflejan condiciones de calidad y cantidad coherentes que aportan una primera evaluación y toma de contacto a los laboratorios sobre el vino que van a analizar.

2.22 Artículo 21: *Assessing wine quality using a Decision Tree* [21]

Este artículo persigue encontrar una predicción de la calidad del vino empleando el algoritmo C4.5 para crear el árbol de decisión a estudiar y parte del conjunto de datos de calidad del vino presente en la Base de Datos UCI (1599 muestras de vino tinto y 4898 muestras de vino blanco; 11 atributos). Encontramos los valores medios que deberán estimar los algoritmos en la Tabla 13.

Attribute	Red wine Mean (Range)	White wine Mean (Range)
Fixed acidity	8.3 (4.6 - 15.9)	6.9 (3.8 - 14.2)
Volatile acidity	0.5 (0.1 - 1.6)	0.3 (0.1 - 1.1)
Citric acid	0.3 (0.0 - 1.0)	0.3 (0.0 - 1.7)
Residual sugar	2.5 (0.9 - 15.5)	6.4 (0.6 - 65.8)
Chlorides	0.08 (0.01 - 0.61)	0.05 (0.01 - 0.35)
Free sulfur dioxide	14 (1 - 72)	35 (2 - 289)
Total sulfur dioxide	46 (6 - 289)	138 (9 - 440)
Density	0.996 (0.990 - 1.004)	0.994 (0.987 - 1.039)
pH	3.3 (2.7 - 4.0)	3.1 (2.7 - 3.8)
Sulphates	0.7 (0.3 - 2.0)	0.5 (0.2 - 1.1)
Alcohol	10.4 (8.4 - 14.9)	10.4 (8.0 - 14.2)

Tabla 13. Valores medios y rangos de los datos fisicoquímicos en el conjunto de datos de calidad del vino de la BBDD UCI

Compara los resultados obtenidos aplicando el Árbol de Decisión de C4.5 con los datos ofrecidos por otros algoritmos *Machine Learning* que proporciona la plataforma WEKA, obteniendo grandes resultados, como apreciamos en la Tabla 14.

Prediction model	Precision (%)	Recall (%)	F-Measure (%)	Accuracy (%)
LibSVM	55.70	58.00	55.70	58.03
BayesNet	56.40	58.30	56.90	58.28
MultiPerceptron	57.70	59.70	58.30	59.66
Proposed	60.10	60.70	60.30	60.66

Tabla 14. Comparativa de los algoritmos Machine Learning empleados en el artículo

La metodología propuesta (algoritmo **C4.5**) consigue los valores más altos de exactitud (60,10%), *F-Measure* (60,70%), recuperación o *recall* (60,30%) y precisión media global del algoritmo (60,66%) comparados con los algoritmos LibSVM, BayesNet y *MultiPerceptron* ofrecidos por WEKA.

Además, también se realiza un estudio sobre las preferencias gustativas tanto para vino tinto como blanco, consiguiendo porcentajes de precisión de 60,7% para el vino blanco y 58,7% para el vino tinto, como vemos en la Tabla 15.

Quality	White wine			Red wine		
	Precision (%)	Recall (%)	F-Me (%)	Precision (%)	Recall (%)	F-Me (%)
3	7.7	10.0	8.70	–	–	–
4	24.4	20.8	22.4	27.3	23.9	25.5
5	48.2	72.2	70.2	60.0	61.9	60.9
6	57.9	57.7	57.8	63.9	64.6	64.3
7	55.7	48.7	52.0	52.8	51.9	52.4
8	10.0	60.7	60.3	36.7	31.4	33.8
Avg	61.1	60.7	60.3	58.2	58.7	58.5
Ac (%)	60.7			58.7		

Avg: Average, Ac: Accuracy, F-Me: F-Measure

Tabla 15. Precisión de los datos obtenidos en cuanto a preferencias gustativas

2.23 Artículo 22: *Modeling wine preferences from physicochemical properties using Fuzzy techniques* [22]

Este artículo compara un modelo híbrido de Lógica Difusa (*fuzzy logic techniques*) GFS-GPG-R, MOGUL-IRLHC-R, MOGUL-TSK-R con los resultados obtenidos en el artículo [27] --punto 2.28 de este documento- *Multiple Regression* o MR, NN y SVM. Utiliza la herramienta *Visual-FIR* (sobre MATLAB) para trabajar con el algoritmo FIR[6], mientras que emplea KEEL (*opensource java software tool*) para GFS[7].

La Base de Datos empleada en este estudio es la Base de Datos UCI de vino blanco (4898 muestras, 11 atributos) con datos datados entre May2004 y Feb2007.

Se han realizado 100 experimentos provenientes de 20 simulaciones con validación cruzada (*5-fold cross-validation*), que ofrecen la siguiente tabla de resultados, Tabla 16, donde **FIR** destaca en la mayor parte de las métricas: error MAD (*Mean Absolute Deviation)* más bajo, así como la mayor precisión para tolerancias T=0,25 y T=1. Además, FIR destaca sobre la segunda mejor opción, SVM, en cuanto a tiempo computacional.

	MR	NN	SVM	GFS-GPG-R	MOGUL-IRLHC-R	MOGUL-TSK-R	FIR
MAD	0.59	0.58	0.45	0.63	0.58	0.56	**0.44**
$Accuracy_{T=0.25}$	25.6%	26.5%	50.3%	31.3%	30.6%	25.1%	**51.2%**
$Accuracy_{T=0.50}$	51.7%	52.6%	**64.6%**	46.3%	50.4%	53.0%	63.3%
$Accuracy_{T=1.00}$	84.3%	84.7%	86.8%	79.4%	83.8%	86.0%	**88.7%**

Tabla 16. Resultados de este experimento comparados con los obtenidos en [27]: MR, NN y SVM

FIR emplea el algoritmo EFP, *equal frequency partition*, para la discretización de las variables de entrada. A continuación, realiza la función-selección durante el proceso

[6] FIR, *Fuzzy Inductive Reasoning* es un modelo cualitativo, no paramétrico y superficial basado en lógica difusa. Dos procesos principales: *selección de características* (desarrolla un modelo) y *predicción o simulación*, (utiliza el modelo para inferir el comportamiento futuro del sistema).

[7] GFS, *Genetic-Fuzzy Systems* es un sistema difuso aumentado por un proceso de aprendizaje basado en computación evolutiva, que incluye algoritmos genéticos, programación genética y estrategia evolutiva. Tres tipos en este documento: MOGUL-TSK-R, MOGUL- IRLHC-R y GFS-GPG-R.

de modelado, ofreciendo las variables más relevantes en la calidad del vino**:** alcohol, acidez fija, dióxido de azufre libre, azúcar residual y acidez volátil.

2.24 Artículo 23: *Classification-based Data Mining approach for quality control in wine production* [23]

El objetivo de este artículo es identificar anomalías en los conjuntos de datos de vino para detectar adulteraciones en la producción de vino. Se emplea para el estudio la Base de Datos de UCI para vino tinto, formada por 1599 instancias, así como la de vino blanco con 4898 muestras; 11 atributos. Se emplea 10-*fold-cross-validation* para separar los datos de prueba y entrenamiento (los datos se dividen de forma aleatoria en 10 grupos de igual tamaño y se ejecutan los algoritmos 10 veces).

Los algoritmos a probar son *Decision Tree* (ID3) y *Naïve Bayes*. ID3, según el artículo, es el algoritmo más preciso del grupo de árboles de decisión; por lo que construye el árbol completo de arriba hacia abajo sin retroceso, haciendo de este método un algoritmo rápido. *Naïve Bayes*, por el contrario, ofrecerá información sobre la correlación de variables siendo un algoritmo escalable y veloz.

Se realiza un proceso de normalización de todos los datos presentes utilizando la herramienta WEKA. También se realiza una selección de atributos, es decir, reducción de variables, empleando la selección automática de WEKA, método *Info Gain Attribute Eval*. Las variables seleccionadas se muestran en la Tabla 17.

Red wine dataset	White wine dataset
Volatile Acidity (2)	Volatile Acidity (2)
Total Sulfur Dioxide (7)	Citric Acid (3)
Sulphate (10)	Chlorides (5)
Alcohol (11)	Free Sulfur Dioxide (6)
	Density (8)
	Alcohol (11)

Tabla 17. Atributos más relevantes tras aplicar reducción de variables

Los resultados muestran que *Decision tree* (**ID3**) predomina sobre *Naïve Bayes* en términos de clasificación del vino, así como en tiempo de computación. ID3 ofrece precisiones de 60,0% y 52,3% para vino tinto y blanco, respectivamente. *Naïve Bayes* consigue un 58,8% y 50,2% de precisión sobre vinos tintos y vinos blancos.

No obstante, ambas tasas son bajas y no se consideran algoritmos óptimos para clasificar el vino. SVM presente en [27] ofrece tasas de precisiones mayores.

Los atributos más relevantes en la calidad del vino son el alcohol y la acidez volátil. Además, se observa que el vino blanco es más sensible a cambios fisicoquímicos ya que son más variables las predominantes en la calidad del brebaje, por lo que controlar los índices de los atributos del vino es crucial para controlar la calidad.

2.25 Artículo 24: *Data Mining techniques for modelling seasonal climate effects on grapevine yield and wine quality* [24]

En este estudio, se estudian datos climáticos y su influencia en la calidad del vino implementando técnicas de Minería de Datos.

Los datos sobre las condiciones climáticas locales han sido proporcionados por NIWA[8] a través de su portal web, donde se analizan datos de rendimiento de un viñedo en el norte de Nueva Zelanda. Estos datos corresponden al intervalo temporal comprendido entre junio de 1996 y diciembre de 2009, recopilados por una estación meteorológica cercana. Por el contrario, los datos de rendimiento de vid facilitados por la bodega corresponden a la cosecha 1998-2006. Este rendimiento de 12 años se clasificó como alto, bajo y moderado por enólogos y son **4** los **atributos** seleccionados: rendimiento por hectárea, Brix, ácidos y pH.

Los datos se analizan colectivamente usando una Neurona Artificial no supervisada basada en Minería de Datos (método de Mapa Autoorganizado de *Kohonen*), Árboles de Decisión y, finalmente, Análisis Discriminante para comprender mejor las asociaciones entre los conjuntos de variables.

[8] *National Institute of Water and Atmosphere*. [Online] http://cliflo.niwa.co.nz

Analizando los datos obtenidos con el SOM, *self-organizing map*, y la matriz de coeficientes de correlación presente en la Tabla 18, el pH muestra ser el atributo con correlación entre las propiedades analizadas más alto.

	Yieldt/ha	*Brix*	*Acid*	*pH*	*rateNum*
Yieldt/ha	1				
Brix	0.050256	1			
Acid	-0.45827	-0.3148	1		
pH	0.571357	0.44143	-0.59227	1	
rateNum	0.698016	0.012479	-0.03556	0.473365	1

Tabla 18. Matriz de coeficientes de correlación de los atributos

El Árbol de Decisión creado con el software ISEE (www.rulequest.com/see5-info.html) muestra que las temperaturas mensuales máximas de marzo y mínimas de noviembre se asocian a las clases anuales de rendimiento. Se describen a continuación y en la Figura 14 las reglas derivadas de este análisis del Árbol de Decisión:

Regla 1: temperatura máxima de marzo > 24,5ºC y mínima en noviembre > 10,5ºC

Regla 2: temperatura máxima marzo> 24.5ºC y mínima noviembre <= 10.5ºC

Regla 3: temperatura máxima de marzo <=24ºC

```
Rules:

Rule 0/1: (3, lift 2.4)
        MarmaxT > 24.2
        NovminT > 10.5
        ->  class moderate  [0.800]
Rule 0/2: (3, lift 2.4)
        MarmaxT > 24.2
        NovminT <= 10.5
        ->  class low  [0.800]
Rule 0/3: (3, lift 2.4)
        MarmaxT <= 24.2
      ->  class high [0.800]
```

Figura 14. Reglas del árbol de decisión

Las pruebas realizadas utilizando las reglas del Árbol de Decisión enumeradas en la Figura 14, brindaron resultados de precisión 100% en la clasificación de las clases de rendimiento de este viñedo.

2.26 Artículo 25: *Data Mining application for upgrading quality of wine production* [25]

Este artículo implementa *Data Mining* con el fin de realizar una clasificación del vino, para después analizar la calidad de éste. Se pretende predecir la calidad de la producción de vino comparando tres algoritmos diferentes: Análisis de Regresión, (*Linear/Multiple Regression*) empleando el algoritmo *Back Propagation Algorithm;* en segundo lugar, *Decision Tree* y, por último, Redes Neuronales, *Neural Network (NN).* Las variables a estudiar son las siguientes: acidez fija, acidez volátil, ácido cítrico, azúcar residual, cloruros, dióxido de azufre libre, dióxido de azufre total, densidad, pH, sulfatos y alcohol.

La salida a estudiar es la *calidad,* que, para simplificar el estudio, se divide en dos grados, 1 y 0. La Base de Datos empleada incluye 3655 muestras brindadas por *SAS EM Insight node*[9], y a la cual NO tenemos acceso ni conocemos la variedad.

El marco de investigación compuesto por preparación de los datos, análisis de éstos y obtención de resultados *y* propuesta de mejoras, concluye afirmando que el percentil de respuesta de las **NN** es más alto que el obtenido empleando Árboles de Decisión y Regresión Lineal. Partiendo de esta base, se evalúa ahora el porcentaje de precisión de las Redes Neuronales. Se crea una Matriz de Predicción, *Forecast Matrix,* representada en la Figura 15, empleando 3615 muestras.

	0(forecast)	1(forecast)
0(actual)	1002	249
1(actual)	306	2058

Figura 15. Matriz de Predicción del modelo

[9] *SAS Enterprise Miner para el análisis de datos, es un software/herramienta interactiva para exploración de información y análisis.*
https://www.sas.com/es_es/curiosity.html?utm_source=google&utm_medium=cpc&utm_campaign=brand-global&utm_content=GMS-158646-gbc-cf&dclid=&gclid=Cj0KCQjwt-6LBhDIARIsAIPRQcL-LOG8wGutMEZLg2jNWlHxfU_rRouxJQ6QY-NDdN-IFv13G0LY6KoaAqs8EALw_wcB

El grado de precisión final obtenido es 84% de para el algoritmo dominante, es decir, Redes Neuronales.

2.27 Artículo 26: *Wine vinification prediction using Data Mining tools* [26]

Este artículo estudia Árboles de Decisión (DT), Redes Neuronales Artificiales (ANN) y Regresión Lineal (LR) como técnicas de Minería de Datos. Se lograron muy buenos resultados con precisiones entre el 86% y el 99% obtenidas para todos los modelos.

Este trabajo adoptó parte de los datos recopilados durante la fase de producción de vinos "tintos verdes" en la finca vinícola en la región de Minho (norte de Portugal) durante cuatro años. Los datos son preprocesados, descartando los registros con valores en blanco, quedando un total de 362 muestras.

El conjunto de datos tiene los siguientes 14 atributos: fermentación (SFTM), tipo de clarificación (polivinilpolipirrolidona, albúmina, gelatina, caseína más el "testigo", con letras *p, a, g, c* y *t* respectivamente), tipo de vinificación (Vinificación por Maceración Pelicular Fermentativa (C), Vinificación por Maceración Carbónica (CM) y Vinificación por Cubo Rotativo (CR)), pH, A420, A520, A620, *Chemical Age* CA, IFC, Ant, sabor, color, aroma, espuma.

Se pretende analizar la variación de los parámetros químicos y el valor predictivo de los parámetros subjetivos en términos de dos clases "A" y "B" correspondientes a la categoría "media" o "buena" asociada al atributo. Los datos fueron divididos en conjuntos de entrenamiento equilibrados y no equilibrados. En cada simulación, los datos disponibles se dividieron de forma aleatoria en particiones mutuamente excluyentes: el conjunto de entrenamiento, con 2/3 de los datos disponibles, y el conjunto de prueba con 1/3 de las instancias restantes.

Se prueban dos enfoques diferentes: el primer enfoque se basa en la clasificación utilizando DT y ANN, y el segundo utilizando LR. Estos dos enfoques se compararán en

términos de exactitud. Se realizaron 10 ejecuciones aplicadas en todas las pruebas, obteniendo la precisión de las estimaciones mediante el método *Holdout*.

Ambos enfoques obtienen valores de precisión alrededor del 90%. Analizando los resultados, podemos comprobar que **no hay mejoras reseñables cuando se utilizan conjuntos de entrenamiento balanceados frente a no balanceados**. Los resultados revelan que el Modelo 1 (CART, Árbol de Decisión de Microsoft) es más preciso a la hora de obtener los valores de los atributos de las muestras, con 85% a 98% de precisión, que el Modelo 2 (J48 + WEKA *Decision Tree*), con 83% a 92% de precisión.

Por otro lado, ambos modelos coinciden en que el atributo SFTM es el más relevante para clasificar diversos atributos subjetivos. El segundo atributo más importante es el tipo de clarificación. Para la clasificación del atributo "aroma" los atributos más relevantes son "tipo de vinificación" para la herramienta WEKA y "tipo de vinificación y edad química" para la Herramienta de Microsoft, a pesar de que la precisión es similar para ambas herramientas (91,18% y 92%).

En cuanto a términos de clasificación empleando la Red Neuronal Artificial (10 neuronas, una para cada variable de entrada, una capa oculta con 10 neuronas y una capa de salida que consta de 2 neuronas (una para cada clase "A" – "Media" y "B"- "Bueno"), se consiguen índices de clasificación del orden de 97%.

El análisis de los resultados muestra que hay diferencias entre la utilización de la DT y ANN. En mismas condiciones de entrada, los **ANN's** presentaron mejores índices.

2.28 Artículo 27: *Modeling wine preferences by Data Mining from physicochemical properties* [27]

En este artículo, uno de los más relevantes y base de muchos estudios realizados hasta la fecha, se pretende modelar preferencias del gusto basadas en datos analíticos empleando el entorno de desarrollo R usando *RMiner*.

La Base de Datos contiene muestras tomadas entre May2004 – Feb2007[10] y contiene 1599 instancias de vino tinto y 4898 de vino blanco, ambos de la variante *vinho verde*, Portugal (BBDD UCI).

En cuanto a las preferencias, cada muestra fue evaluada por un mínimo de tres evaluadores sensoriales (usando catas a ciegas), que clasificaron el vino en una escala de 0 (muy malo) a 10 (excelente). Por otra parte, se tienen en cuenta las siguientes variables fisicoquímicas, siendo un total de 11 atributos: acidez fija, acidez volátil, ácido cítrico, azúcar residual, cloruros, dióxido de azufre libre, dióxido de azufre total, densidad, pH, sulfatos y alcohol.

Este artículo presenta un método que realiza mediciones simultaneas de variables y selección de modelos empleando L*ineal/Multiple Regression* (MR), *Neural Networks* (NN) y *Support Vector Machine* (SVM). La **selección de variables** se basa en el análisis de sensibilidad; método computacionalmente eficiente que mide la relevancia de entrada y guía el proceso de selección de variables. Además, se propone un método de búsqueda para seleccionar el mejor parámetro del *kernel* SVM con un bajo esfuerzo computacional.

Se lograron grandes resultados, destacando el modelo SVM sobre las técnicas NN MR, aun requiriendo un mayor coste computacional (*5-fold cross-validation testing* tarda máx. 26 min). Además, se logra una mejora de rendimiento configurando la tolerancia para aceptar respuestas que son correctas de entre las dos más cercanas (T= 1,0), obteniendo una precisión global de 89,0% (tinto) y 86,8% (blanco), frente a valores de 88,6% (tinto) y 84,3% para MR y 88,8% (tinto) y 84,7% (blanco) considerando NN. Observamos la comparativa con los resultados ya comentados en la Tabla 19.

[10] http://www3.dsi.uminho.pt/pcortez/wine/

	Red wine			White wine		
	MR	NN	SVM	MR	NN	SVM
MAD	0.50 ± 0.00	0.51 ± 0.00	**0.46** $\pm 0.00^{a}$	0.59 ± 0.00	0.58 ± 0.00	**0.45** $\pm 0.00^{a}$
$\text{Accuracy}_{r-0.25}$ (%)	31.2 ± 0.2	31.1 ± 0.7	**43.2** $\pm 0.6^{a}$	25.6 ± 0.1	26.5 ± 0.3	**50.3** $\pm 1.1^{a}$
$\text{Accuracy}_{r-0.50}$ (%)	59.1 ± 0.1	59.1 ± 0.3	**62.4** $\pm 0.4^{a}$	51.7 ± 0.1	52.6 ± 0.3	**64.6** $\pm 0.4^{a}$
$\text{Accuracy}_{r-1.00}$ (%)	88.6 ± 0.1	88.8 ± 0.2	**89.0** $\pm 0.2^{b}$	84.3 ± 0.1	84.7 ± 0.1	**86.8** $\pm 0.2^{a}$
$\text{Kappa}_{r-0.5}$ (%)	32.2 ± 0.3	32.5 ± 0.6	**38.7** $\pm 0.7^{a}$	20.9 ± 0.1	23.5 ± 0.6	**43.9** $\pm 0.4^{a}$
Inputs ($\overline{I}$)	9.2	9.3	9.8	9.6	9.3	10.1
Model	–	$\overline{H} = 1$	$\overline{\gamma} = 2^{0.19}$	–	$\overline{H} = 2.1$	$\overline{\gamma} = 2^{1.55}$
Time (s)	**518**	847	5589	**551**	1339	30674

[a] Statistically significant under a pairwise comparison with MR and NN.
[b] Statistically significant under a pairwise comparison with MR.

Tabla 19. Resultados del modelado de vino; mejores valores obtenidos aparecen en negrita

Todos los métodos presentan precisiones específicas por encima del 90%. La superioridad de **SVM** sobre NN se deba probablemente a las diferencias en la fase de entrenamiento. El algoritmo SVM garantiza un ajuste óptimo, mientras que el entrenamiento NN puede caer en un mínimo local.

2.29 Comparativa de todos los Artículos

2.29.1 Tabla Global. Resumen de los artículos de calidad

A continuación, el siguiente punto expone una recopilación en forma de tabla, Tabla 20, de todos los artículos incluidos en el bloque asociado al estudio y predicción de la calidad del vino empleando técnicas de Inteligencia Artificial. Después, se adjuntan algunos gráficos con observaciones y comparaciones interesantes realizadas sobre las instancias de la Tabla 20.

Título del Artículo	Año	Objetivo	Base de Datos		Técnica IA		Resultados	
			Nº Muestras	Nº Variables %Muestras/Fase	Reducción de Variables	Clasificador	Info. Variables	Info. Algoritmos
2.2 *A machine learning application in wine quality prediction*	2022	Predicción de calidad de vino Pinot noir	18 muestras *Pinot noir*. 12 muestras generan 1381 muestras artificiales y las 6 restantes para *test*. (Vino tinto)	54 características (7 fisicoquímicas y 47 químicas) Se reduce a 10 variables	XGB, *Extra Trees*, *Random Forest* y *Gradient Boosting*	*Support vector machine* (SVM), *Random Forest*, *Decision Tree Classifier* (DTC), *Gaussian Naive Bayes* (GNB), XGB, *K-closest neighbour* (KNN), *Adaptive Boosting* (AdaBoost) y *Stochastic Gradient Decision Classifier* (SGDC)	Se utiliza el método de reducción XGB para el estudio (ofrece los mejores resultados) Los mejores clasificadores con XGB son AdaBoost, con precisión, recuperación, F1, ROC, MCC del 100% y tiempo de 0,31 sg; seguido por RF	
2.3 *Analysis of white wine using machine learning algorithms*	2021	Predecir la calidad del vino blanco	UCI vino blanco: 4898 instancias	11 parámetros independientes[11]	*Naïve Bayes*, *Support Vectors Machine* (SVM), *Random Forest*, J48 y *Multi-layer Neural Network* (MLP)		Algoritmo J48 ofrece los mejores resultados en la mayoría de campos: Exactitud 99.895%, precisión 0.999, recuperación 0.999, F1-Measure 0.999, estadísticas Kappa 0.964, área ROC 0.94; Términos de error: MAE 0.002, RMSE 0.032, RAE 5.4269 y RRSE 26.3684	
2.4 *A Hybrid Wine Classification Model for Quality Prediction*	2021	Predecir la calidad del vino tinto y blanco	UCI vino tinto y blanco 1599; 4898 instancias	11 variables 80% entrenamiento 20% prueba + ratios variables	Modelo híbrido de *Random Forest* y SVM		Precisión de 66% para vino tinto y 67% para vino blanco. La exactitud (*accuracy*) fue máxima empleando un 20% de las muestras para la fase de prueba en vino tinto y un 10% para vino blanco.	
2.5 *Machine learning approach for attribute identification and quality prediction of red wine*	2021	Predecir la calidad del vino tinto	UCI vino tinto 1599 instancias	11 variables por muestra	LR, DT, RF, SGDC, *Naïve Bayes*, KNN, SVM, SVM + *GridSearchCV*		El algoritmo SVM produce la máxima precisión del 87.5 %, cuando se utiliza con búsqueda GridSearchCV la precisión fue de un 90%.	
2.6 *A new red wine prediction framework using machine learning*	2020	Predecir la calidad del vino tinto y ofrecer resultados sobre las variables estudiadas	UCI vino tinto 1599 instancias	11 variables fisicoquímicas	MF-DCCA con XGBoost y LightGBM. MF-DCCA investiga la relación dinámica entre las 11 variables		MF-DCCA: El **azúcar residual** contribuye favorablemente; y las correlaciones cruzadas más débiles son la acidez volátil y los cloruros.	LightGBM ofrece resultados de: Recuperación 85.63%, precisión 86.67%, F1 measure 86.15% y exactitud 91.04 % XGBoost ofrece los siguientes resultados: Recuperación: 90.67%, precisión: 90.67%, F1 measure: 90.67% y exactitud: 91.04 %

2.7 *Wine Quality Analysis Using Machine Learning*	2020	Predecir la calidad del vino tinto	UCI vino blanco 4898 instancias	11 variables	SVM, *Random Forest* y *Multilayer Perceptron*	*Random Forest*, ofrece el mejor resultado, con precisión de 81.96%, seguido por *Multilayer Perceptron* con un porcentaje de precisión de 78.78% y, por último, SVM, con precisión de 57.29%.	
2.8 *Red Wine Quality Prediction Using Machine Learning Techniques*	2020	Predecir la calidad del vino tinto	UCI vino tinto 1599 instancias	11 variables 70% entrenamiento 30% prueba	*Random Forest*, SVM y *Naïve Bayes*	Precisión para el conjunto de entrenamiento y el conjunto de prueba: SVM: 67.25% y 68.64%, *Random Forest*: 65.83% y 65.46% y *Naïve Bayes* : 55.91% y 55.89%	
2.9 *Análisis de calidad del vino por medio de técnicas de Inteligencia Artificial*	2020	Encontrar relaciones entre variables presentes en el vino que permitan mejorar la calidad de éste	UCI vino tinto y blanco 300 registros aleatorios entre muestras de vino tinto y blanco	10 variables	Algoritmo J48 (*Decision Tree)*	Alcohol, PH, sulfatos, ácido cítrico y relación alcohol-sulfatos	Con el Árbol de Decisión optimizado se obtienen porcentajes de clasificación superiores al 95%
2.10 *Research on Red Wine Quality Based on Data Visualization*	2020	Mejorar la calidad del vino tinto	UCI vino tinto 1599 instancias	11 variables	*Data Mining* a partir de *Correlación de Pearson*	Alcohol, ácido cítrico, sulfato y acidez volátil	-
2.11 *Wine Quality Prediction Using Data Mining*	2019	Predecir la calidad y clasificar el tipo de vino	Vino tinto y blanco BBDD indefinida. 178 muestras	13 atributos[12] 80% entrenamiento 20% prueba	*Nave Bayes*, *Simple Logistic*, KStar, JRip y J48	*Naïve Bayes* ofrece un 100% de precisión sobre los datos de prueba; mientras que *Simple Logistic*, KStar y J48 ofrecen un 97.22% de precisión. JRip empeora con un porcentaje de 94.44%. Si el conjunto de entrenamiento es demasiado grande, clasificadores que apliquen validación cruzada de forma recurrente ofrecerán mejores resultados que *Naïve Bayes*.	

[11] Parámetros usuales: acidez fija, acidez volátil, ácido cítrico, azúcar residual, cloruros, dióxido de azufre libre, dióxido de azufre total, densidad, pH, sulfatos y alcohol.

[12] Los atributos son los siguientes: alcohol, ácido málico, cenizas, alcalinidad de cenizas, magnesio, fenoles totales, flavonoides, fenoles no flavonoides, proantocianinas, intensidad de color, tinte, OD280/OD315 de vinos diluidos y prolina.

2.12 *Wine Quality Classification with Multilayer Perceptron*	2018	Predecir la calidad del vino	UCI vino tinto y blanco 1599; 4898 instancias	11 variables	Lenguaje R con *Keras* y *Tensorflow* + Red Neuronal con funciones de activación: RELU y *tanh*		RELU + vino tinto: 53% de precisión de validación. *tanh* + vino tinto: precisión de validación del 52% RELU + vino blanco: 48% de precisión de validación. *tanh* + vino blanco: precisión de validación del 53% Los resultados variando valores de *epochs* y tamaño de lote sugiere que los datos se sobreajustan o se subajustan aumentando *epochs* y tamaños de los lotes. Si bien el uso de una función de activación diferente no afecta mucho.
2.13 *Prediction of Quality for Different Type of Wine based on Different Feature Sets Using Supervised Machine Learning Techniques*	2018	Predecir la calidad del vino tinto y blanco	UCI vino tinto y blanco 1599; 4898 instancias	11 variables 50% entrenamiento 50% fase de prueba	*Genetic Algorithm*, GA y *Simulated Annealing*, SA	RPART, C4.5, PART, *Bagging* CART, *Random Forest*, *Boosted* C5.0, SVM, LDA y NB	SVM destaca respecto a los demás algoritmos evaluados tanto con GA como con SA. Vino tinto: Precisión: 95.23%, Sensibilidad: 97.17%, Especificidad: 98% Vino Blanco: Precisión: 98.81%, Sensibilidad: 99.34%, Especificidad: 99.56%
2.14 *A Classification Approach with Different Feature Sets to Predict the Quality of Different Types of Wine using Machine Learning Techniques*	2018	Predecir la calidad de distintos tipos de vino	UCI vino tinto y blanco 1599; 4898 instancias	11 variables	*Principal Component Analysis*, PCA y *Recursive Feature Elimination*, RFE	RPART, C4.5, PART, *Bagging* CART, *Boosting* C5.0 y Árboles de Decisión no lineales	*Random Forest* es el mejor algoritmo a la hora de predecir la calidad del vino, tanto incluyendo PCA como RFE. Precisión RF + RFE en vino tinto: 94,51%; vino blanco: 97,79% Precisión RF + PCA en ambos tipos de vino > 92%
2.15 *Wine Quality Detection through Machine Learning Algorithms*	2018	Predecir la calidad del vino tinto	UCI vino tinto 1599 instancias	11 variables 80% entrenamiento 20% validación y test	*Fórmula de estandarización para trabajar con datos normalizados: se eliminan 438 muestras (valores atípicos)	*Random Forest* con 50 árboles de decisión y *Logistic Regression*	La precisión del modelo para RF es de un 84%, mientras que para LR es del 76%

2.16 *Fuzzy logic tool for wine quality classification*	2017	Clasificar objetivamente la calidad del vino en función de los atributos de las bayas de uva	Vino tinto 300 bayas por viñedo. 13 viñedos	8 variables[13]	Decisiones Multicriterio de Lógica Difusa (FMCDM)		Los resultados de la FMCDM coincidieron en un 83.46% y 76.92% en comparación con los datos catalogados por enólogos para la cosecha 2012 y 2013, respectivamente.	
2.17 *Selection of important features and predicting wine quality using machine learning techniques*	2017	Predecir la calidad del vino tinto y blanco y selección de características	UCI vino tinto y blanco 1599; 4898 instancias	11 variables	*Transformación lineal* para normalizar todos los valores		**Vino tinto:** acidez volátil, cloruros, dióxido de azufre libre, dióxido de azufre total, pH, sulfatos y alcohol. **Vino blanco:** acidez fija, acidez volátil, azúcar residual, dióxido de azufre libre, densidad, pH, sulfatos y alcohol	SVM es mejor técnica de Aprendizaje Automático, con gran capacidad de aproximación a los valores de calidad reales
					Linear Regression para determinar la dependencia de calidad sobre las 11 variables	Redes Neuronales y *Support Vector Machine* (SVM) para predecir la calidad		
2.18 *An analytical toast to wine: Using stacked generalization to predict wine preference*	2017	Encontrar variables influyentes en la calidad del vino blanco	UCI vino blanco: 4898 instancias	11 variables	Generalización Apilada + *Linear Regression* (LR), Redes Neuronales (NNs) y *Suport Vector Machines* (SVMs) + R (análisis de datos. Paquete CARET)		La variable predictora más influyente dada por generalización apilada es el porcentaje de alcohol por unidad de volumen	-
2.19 *Classification of Wine Quality with Imbalanced Data*	2016	Predecir la calidad del vino blanco y encontrar variables influyentes en la calidad de este	UCI vino blanco: 4898 instancias (4535 y 363, clase mayoritaria y menoritaria)	11 variables 75% entrenamiento, 15% validación y 15% para pruebas	Precisión de disminución media y la Disminución media de Gini para estudiar variables influyentes.	*SMOTE* (balanceo de datos) + *Decision Tree*, AdaBoost y *Random Forest* + *System R* (preparación y análisis de datos)	Parámetros más relevantes: acidez volátil, el dióxido de azufre libre y el alcohol. Parámetro menos relevante: dióxido de azufre total	RF aporta los mejores resultados de tasas de error y valores ROC. Todos los algoritmos ofrecen mejores resultados trabajando con los datos tras aplicar SMOTE

[13] Las variables medidas son: sólidos solubles totales, pH, volumen de bayas, infección por botritis, coloración de la semilla de uva, extractabilidad de antocianina, densidad óptica (OD 520) y fenoles de la piel (Dpell).

2.20 *The Classification of White Wine and Red Wine According to Their Physicochemical Qualities*	2016	Predecir el tipo de vino y su calidad tanto para vinos tintos como blancos	UCI vino tinto y blanco 1599; 4898 instancias	11 variables 80% entrenamiento 20% fase de prueba	Método de validación cruzada o *Percentage Split* + *K-nearest-neighbourhood*, *Random Forest* o *Support Vector Machines*	*Random Forest* es el mejor clasificador distinguiendo el tipo de vino con precisión: 99.5229% y 99.4611% *Random Forest* para la clasificación de la calidad el vino tinto aumentó de 69.606% a 71.232% para validación cruzada y de 71.875% a 73.4375% para el modo de división porcentual; mientras que para clasificar la calidad del vino blanco disminuyó de 70.3757% a 69.9061% y de 68.6735% a 67.449%
2.21 *A New Mathematical Modelling Approach For Viticulture And Winemaking Using Fuzzy Cognitive Maps*	2016	Predecir y modelar la calidad y la cantidad del vino en base a cosecha real	Datos cosecha real: Variante *Roditis* de bodega griega en Patras, Peloponeso; Variante *Bombino Bianco* 2014, Puglia (Vino blanco)	Entradas: 12 variables[14] + Salidas: calidad y cantidad	*Fuzzy Cognitive Maps* entrenados con el algoritmo *Non Linear Hebbian Learning*	Caso *Roditis:* Cosecha a finales de septiembre y poda en marzo. Ambos valores (calidad y cantidad) están por encima de 0.9. Caso ideal. Caso *Bombino Bianco:* 30% menos de producción por causas naturales. La cosecha se extendió hasta finales de octubre. Peores resultados: Cantidad muy baja, C14= 0.3320 y calidad media C13=0.68
2.22 *Assessing wine quality using a Decision Tree*	2015	Predecir la calidad del vino tinto y blanco en cuanto a las preferencias gustativas	UCI vino tinto y blanco 1599; 4898 instancias	11 variables	*Decision Tree* (Algoritmo C4.5 para crear el árbol de decisión) LibSVM, BayesNet y *MultiPerceptron* ofrecidos por WEKA	*Decision Tree* es el mejor algoritmo de clasificación de valores medios de los 11 atributos. Obtiene valores de exactitud 60.10%, *F-Measure* de 60.70%, *recall* 60.30% y precisión media global 60.66%. Clasificando la calidad en cuanto a preferencias gustativas, *Decision Tree* consigue porcentajes de precisión de 60.7% para el vino blanco y 58.7% para el vino tinto

[14] Variables estudiadas: Características especiales del suelo, condiciones climáticas en el viñedo de la región, variedad de uva, factor humano (labores de cultivo y cuidados agrícolas en el viñedo), fermentación alcohólica, enfermedades, alteraciones y deterioro de la calidad del vino, sustancias enológicas adicionales al vino, almacenamiento de vino en barricas, maduración /crianza del vino, lluvia antes de la cosecha, tiempo de cosecha y poda

2.23 *Modeling Wine Preferences from Physicochemical Properties using Fuzzy Techniques*	2015	Predecir la calidad del vino blanco y encontrar variables influyentes en la calidad de este	UCI vino blanco: 4898 instancias	11 variables	FIR emplea el algoritmo EFP, *equal frequency partition*, para la discretización de las variables de entrada y selección de variables en el proceso de modelado	Modelos híbridos de lógica difusa GFS-GPG-R, MOGUL-IRLHC-R, MOGUL-TSK-R y FIR. Los compara con MR, NN y SVM [27]. Se realizan 100 experimentos provenientes de 20 simulaciones con *5-fold cross-validation*	Variables más relevantes en la calidad del vino blanco: Alcohol, acidez fija, dióxido de azufre libre, azúcar residual y acidez volátil	FIR ofrece los mejores índices menos en tolerancia T=0.5, superada en 1.1% por SVM (64.6%). Los valores FIR obtenidos son: error MAD más bajo: 0.44, así como mayor precisión para tolerancias T=0'25 y T=1, de valor 51,2% y 88,7%. FIR destaca también en términos de tiempo computacional
2.24 *Classifcation-based Data Mining Approach for Quality Control in Wine Production*	2012	Control de la calidad y detección de anomalías	UCI vino tinto y blanco 1599; 4898 instancias	11 variables 10-*fold-cross-validation* para separar datos de prueba y entrenamiento	Reducción de variables empleando WEKA, método *Info Gain Attribute Eval*.	*Decision Tree* (ID3) y *Naïve Bayes*	**Vino tinto**: acidez volátil, dióxido de azufre total, sulfatos y alcohol **Vino blanco**: acidez volátil, ácido cítrico, cloruros, dióxido de azufre libre, densidad y alcohol	*Decision Tree* (ID3) predomina sobre *Naïve Bayes*. ID3 ofrece precisiones de 60.0% y 52.3% para vino tinto y blanco, respectivamente. *Naïve Bayes* consigue un 58.8% y 50.2%
2.25 *Data Mining Techniques for Modelling Seasonal Climate Effects on Grapevine Yield and Wine Quality*	2010	Influencia en la calidad del vino con datos climáticos	Datos climáticos NIWA y datos bodega de Nueva Zelanda	4 atributos: Rendimiento por hectárea, Brix, ácidos y pH	Neurona artificial no supervisada (método de mapa autoorganizado de Kohonen, SOM)	*Decision Tree* (creado con el software ISEE) + Análisis discriminante	pH	*Decision Tree* obtuvo resultados de precisión 100% en la clasificación de las clases de rendimiento
2.26 *Data mining application for upgrading quality of wine production*	2010	Clasificar y predecir la calidad del vino	Base de Datos *SAS EM Insight node*; 3655 instancias	11 variables[10]	Análisis de Regresión, (*Linear/Multiple Regression*) empleando el algoritmo *Back Propagation Algorithm*; en segundo lugar, *Decision Tree* y, por último, *Neural Network* (NN)		El percentil de respuesta de las NN es superior al obtenido empleando DT y LR. Partiendo de esta base, se evalúa el porcentaje de precisión de las NN. Se crea una *Forecast Matrix* y se obtiene un grado de precisión de 84%	

2.27 *Wine Vinification prediction using Data Mining tools*	2009	Predicción de vinificación	Vino tinto 362 muestras	14 atributos[15]	Árboles de Decisión (DT) con J48 y CART, Redes Neuronales Artificiales (ANN) y Regresión Lineal (LR)	SFTM, tipo de clarificación	Se lograron precisiones entre el 86% y el 99% obtenidas para todos los modelos. En mismas condiciones de entrada, los ANN's presentaron mejores índices que DT
2.28 *Modeling wine preferences by data mining from physicochemical properties*	2008	Predecir la calidad del vino tinto y blanco y ofrecer resultados químicos sobre las variables para mejorar la calidad	UCI vino tinto y blanco 1599; 4898 instancias	11 variables	*Data Mining para Neural Networks* (NN) y *Support Vector Machine* (SVM) + *5-fold cross-validation*	SVM destaca aun requiriendo un mayor coste computacional. Se logra una mejora de rendimiento configurando la tolerancia T= 1.0, obteniendo una precisión global de 89% (tinto) y 86.8% (blanco), frente a valores de 88.6% (tinto) y 84.3% para MR y 88,8% (tinto) y 84.7% (blanco) considerando NN. Todos los métodos presentan precisiones específicas por encima del 90%	

Tabla 20. Tabla global con los artículos que tratan el estudio de la calidad del vino

[15] El conjunto de datos tiene los siguientes 14 atributos: fermentación (SFTM), tipo de clarificación (polivinilpolipirrolidona, albúmina, gelatina, caseína más el "testigo", con letras p, a, g, c y t respectivamente), tipo de vinificación (Vinificación por Maceración Pelicular Fermentativa (C), Vinificación por Maceración Carbónica (CM) y Vinificación por Cubo Rotativo (CR)), pH, A420, A520, A620, Chemical Age CA, IFC, Ant, sabor, color, aroma, espuma.

2.29.2 Gráfico con los algoritmos que ofrecen mejores resultados

A continuación, se adjuntan unos gráficos resumen de los algoritmos que han obtenido mejores resultados en los artículos del apartado dedicado a la *calidad del vino*.

Para la elaboración de este primer gráfico, Gráfico 1, se tienen únicamente en cuenta los algoritmos calificados como "mejores" en cada artículo. Dado que los estudios se han realizado sobre variedades de vino tinto, blanco, ambas variedades y/o no especifica el tipo de vino, encontraremos 5 columnas para cada caso. Cada vez que un algoritmo obtiene los mejores resultados en uno de los estudios recogidos en este documento, se le asociará un punto. El sumatorio de todas las veces que dicho algoritmo ha obtenido las mejores predicciones frente a otros algoritmos o bien ha sido el único algoritmo contemplado será la puntuación final mostrada en el gráfico en %. Por ejemplo, para el caso de *Random Forest*, según cinco de los artículos que miden la calidad del vino ha demostrado ser la mejor técnica para predecir la calidad del brebaje; uno de ellos referido a vino tinto, dos de ellas para vino blanco y otros dos estudios que miden ambas variedades. Puntuará, por tanto, 1 punto en la columna asociada a vino tinto, 2 puntos en la asociada a vino blanco y 2 puntos en la columna de ambas variedades.

La última columna mostrará de forma global sin distinguir entre el tipo de vino cuántas veces el algoritmo elegido ha obtenido predicciones destacables. Esta columna es de gran interés para elegir unas metodologías u otras de cara a nuevos estudios.

El Gráfico 1 expresa la aportación de cada algoritmo en tanto por ciento (%) en función de los artículos que componen cada columna, lo que ayuda al tratamiento y comprensión de los datos. Para el mismo ejemplo de antes, *Random Forest*, según los datos recogidos, ha mostrado ser el mejor algoritmo a la hora de predecir la calidad del vino blanco con un porcentaje de excelencia de 50% entre los artículos del documento. Eso implica que la mitad de los artículos que han realizado un estudio centrado en variedades de vino blanco coinciden en que los resultados más favorables para mejorar y/o predecir la calidad del vino sucederán empleando este algoritmo. Si, por el contrario, realizamos un análisis de vino tanto blanco como tinto, los algoritmos que han ofrecido mejores tasas en estudios previos son dos, SVM y RF, empatados con un porcentaje de 20% sobre el total.

DE FORMA VISUAL PODEMOS IDENTIFICAR LOS ALGORITMOS MÁS POPULARES SEGÚN LA VARIEDAD DE VINO ELEGIDA

RF Y SVM PREDOMINAN EN CÓMPUTO GLOBAL

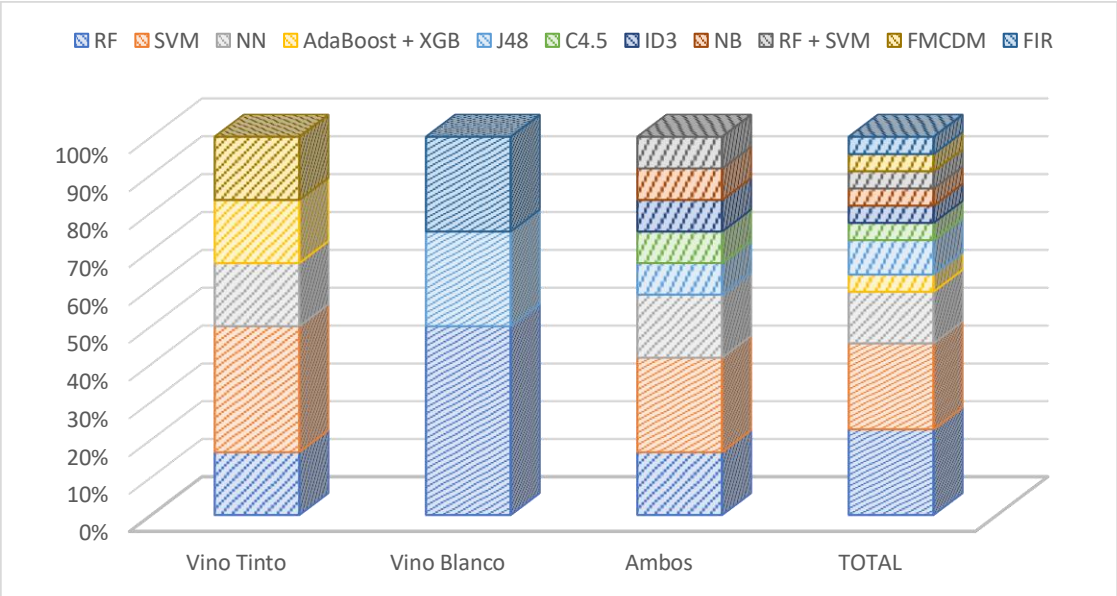

Gráfico 1. Algoritmos más populares para predecir la Calidad del vino

Si deseamos realizar un estudio sobre ambas variedades de vino –tinto y blanco- optar por *Random Forest* será una elección más acertada, ya que ninguno de los artículos centrados en el estudio de la calidad del vino blanco nombra SVM como mejor opción para conseguir el objetivo. En cambio, a la hora de predecir la calidad del vino tinto, ambos algoritmos -SVM y RF- han resultado ser el mejor modo de predicción en al menos uno de los estudios de esta variedad de vino. Por ello, un resumen del Gráfico 1 se muestra en el Gráfico 2.

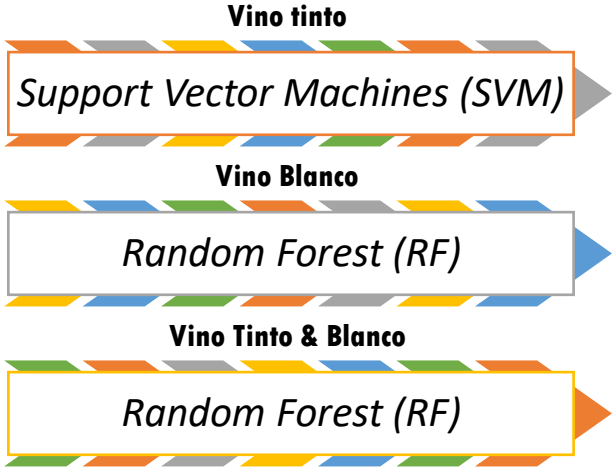

Gráfico 2. Mejor algoritmo para predecir la calidad del vino según la variedad

2.29.3 Gráfico con atributos relevantes en la calidad del vino

Uno de los principales objetivos de emplear técnicas de Inteligencia Artificial es identificar atributos influyentes en la calidad del vino. Esto ayudará a los enólogos a establecer pautas que favorezcan el control de las variables más relevantes para mejorar la calidad del vino. Por ello, el Gráfico 3 ofrece una mejor visualización de los resultados obtenidos por los estudios realizados durante los últimos años, donde apreciamos claramente que, para ambos tipos de vino, la variable más influyente es el alcohol (más dominante para las variedades de vino blanco). El Gráfico 4 incluye las variables clave según el tipo de vino.

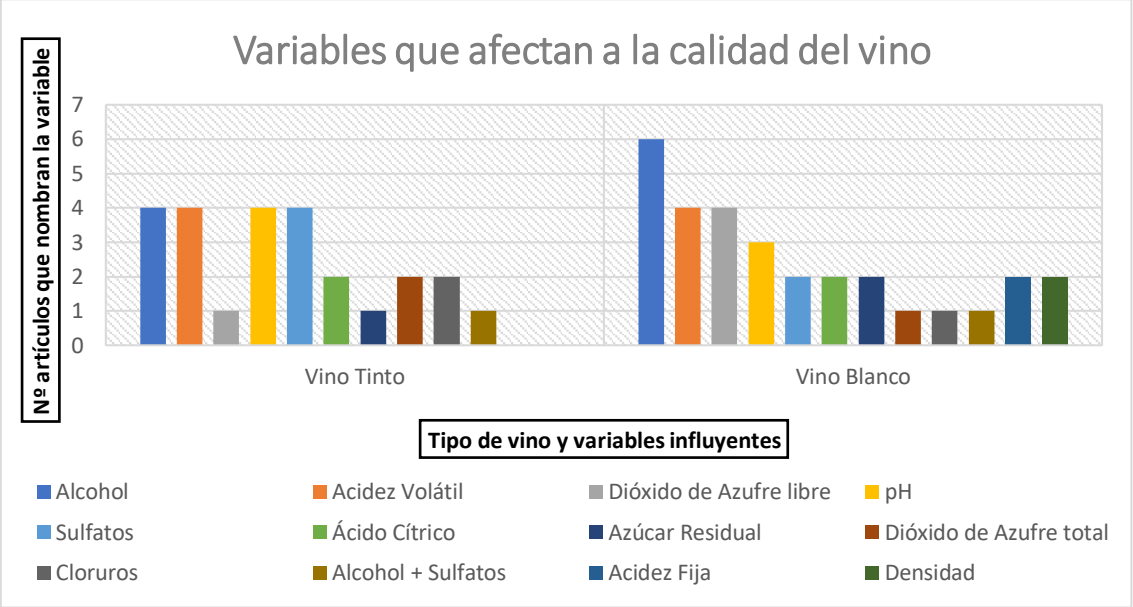

Gráfico 3. Variables que afectan a la calidad del vino según su variedad

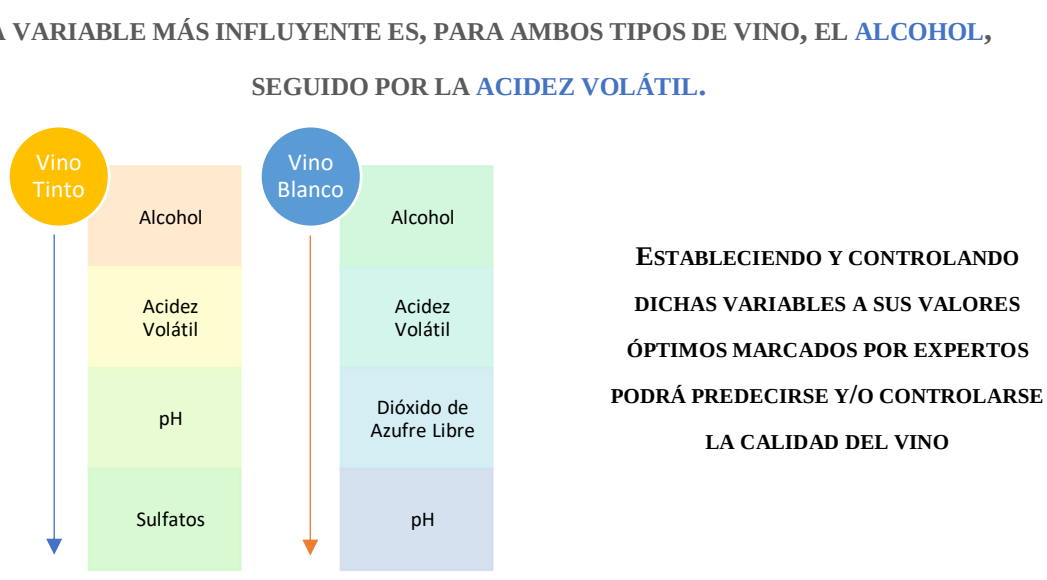

Gráfico 4. Variables más relevantes según la variedad de vino

2.30 Conclusiones

La predicción de la calidad del vino es un aspecto importante para los enólogos, ya que el sector de la viticultura tiene un impacto social de gran magnitud que busca, en muchos casos, la excelencia.

Debido al gran abanico de variedades y posibles calidades de vino que ofrece el mercado, puede resultar de gran interés para los enólogos medir la calidad de los brebajes obtenidos, así como conocer variables relevantes a tener en cuenta a la hora de la preparación de estos. De esta forma, podrán ofertar a los clientes finales distintas variedades según los índices de los atributos relacionados con la calidad, creando varias gamas dentro de una misma bodega.

El empleo de técnicas de Inteligencia Artificial ayuda a estos cometidos. Los algoritmos que han ofrecido mejores índices de predicción sobre datos ya categorizados han sido *Random Forest* (RF)y *Support Vector Machine* (SVM), siendo *Random Forest* más favorable a la hora de trabajar con datos pertenecientes a vino tinto y *Support Vector Machine* para vino blanco. En estudios donde se tenían en cuenta ambas variedades, SVM ha sido el algoritmo que ofreció un mejor resultado global en más ocasiones, seguido, de nuevo, por RF.

En términos de variables, mejorar la calidad del vino dependerá en gran parte, según los estudios realizados, de mantener los niveles de alcohol y acidez volátil en niveles óptimos establecidos por los expertos. Controlando dichos valores, podrá mejorarse la calidad del vino final.

En cuanto a Bases de Datos, se observa que, cuanto mayor número de muestras compongan en el *dataset*, mejor entrenado será el algoritmo al contar con un conjunto de entrenamiento más amplio y, por tanto, podrán esperarse resultados más precisos. Esto servirá de gran importancia a la hora de realizar clasificaciones según la variedad, por ejemplo.

3 Espectroscopía y Mediciones del nivel de Azúcar, pH y concentración de Antocianinas

3.1 Introducción

Este tercer punto está compuesto por el estudio de los artículos que hablan (en su mayoría) de las mediciones del nivel de azúcar, pH y/o concentración de antocianinas realizados mediante espectroscopía.

El nivel de azúcar es un factor relevante que afecta a la calidad, por lo que realizar un seguimiento de dicho valor resulta de especial interés para los enólogos. Para ello, se mide el grado Brix del vino. Los grados Brix (°Bx) miden la concentración total de sacarosa disuelta en un líquido, es decir, miden el dulzor de los alimentos. Este dulzor más o menos intenso caracteriza, en muchas ocasiones, los matices y el sabor del vino, haciendo de este un brebaje único y particular. Los índices de pH y antocianinas presentes en el vino resultan también de interés para las bodegas, siendo los encargados de dotar a la bebida de distinta acidez y pigmentación -relacionado de forma directa con el color-, respectivamente.

Emplear técnicas hiperespectrales ayudará, desde un marco teórico, a analizar y localizar zonas espectrales de mayor influencia o variación de las variables a medir, las cuales, recordemos, afectarán de forma directa al resultado final brindado en la botella de vino. También se hará uso de Inteligencia Artificial para predecir los niveles de azúcar, pH y/o antocianina del vino empleando diferentes algoritmos que actuarán sobre las imágenes hiperespectrales tomadas en cada estudio.

El siguiente punto está dedicado a hablar sobre la importancia y adquisición de imágenes hiperespectrales, debido a la gran influencia de esta metodología para evaluar distintas variables que afectan a calidad del vino.

3.1.1 Imagénes hiperespectrales

Como aventurábamos previamente, las imágenes hiperespectrales han supuesto un gran avance a la hora de obtener información sobre muestras frescas de uva. Esto implica de forma directa una mejora sustancial para el futuro análisis y tratamiento de datos, donde la Inteligencia Artificial juega un papel esencial.

Las imágenes hiperespectrales, al igual que otras imágenes espectrales, recopilan y procesan información de todo el espectro electromagnético.

COMBINA EL PODER DE LA IMAGEN DIGITAL Y LA ESPECTROSCOPIA.

3.1.1.1 Captación de imagénes hiperespectrales

Para obtener imágenes hiperespectrales se utiliza una cámara hiperespectral. Estas cámaras especializadas están formadas por diversos sensores que captan distintas medidas según el modo seleccionado:

- **REFLECTANCIA** (R): Relación de la luz reflejada de la muestra dividida por la luz total disponible, en general, esto se define a partir de un material sólido.
- **TRANSMITANCIA** (T): Relación de la energía que atraviesa una muestra, comparada con la luz total disponible: en general, aplicado a líquidos.
- **ABSORDANCIA** o Absorbencia: Medición indirecta de la absorción, que utiliza Log[1/T] o Log[1/R].

Estos dos primeros modos espectroscópicos (reflectancia y transmitancia) difieren principalmente en la posición de la muestra con respecto a la fuente de luz y el detector. En el modo de reflectancia, la fuente de luz y el detector (cámara hiperespectral) se colocan sobre la muestra, mientras que, en el modo de transmitancia, la fuente de luz y el detector se colocan en lados opuestos de la muestra y la luz debe atravesar la muestra. El modo de transmitancia tiene la desventaja de requerir una mayor proximidad a la muestra porque la intensidad de la luz se atenúa significativamente al atravesar las muestras. Por esta razón, generalmente se prefiere el modo de reflectancia [31].

Para cada píxel de una imagen, una cámara hiperespectral adquiere la intensidad de la luz para un gran número de bandas espectrales contiguas. Cada píxel de la imagen contiene así un espectro continuo (en reflectancia) y se puede utilizar para caracterizar los objetos de la escena con gran precisión y detalle.

La imagen hiperespectral en modo reflectancia ha demostrado ser una alternativa a técnicas clásicas de determinación de parámetros enológicos importantes para la evaluación de la madurez y la fecha óptima de cosecha.

La Inteligencia Artificial ayuda a esta caracterización, ofreciendo un manejo y análisis de la información recogida en las imágenes hiperespectrales.

3.1.1.2 Ejemplo de modelo experimental

Para entender mejor el modo de operación de esta técnica aplicada sobre el vino combinando imágenes hiperespectrales e Inteligencia Artificial, tomaremos como ejemplo el modelo empleado en [31], es decir, el punto 3.2.4 de este documento, aunque cabe destacar que es la metodología por la que optan la mayor parte de los autores que emplean imágenes hiperespectrales como obtención de información de muestras de uva.

En la Figura 16 observamos un gráfico que muestra de forma esquemática e ilustrativa el procedimiento de los artículos que emplean imágenes hiperespectrales para formar la Base de Datos que tratarán posteriormente en la fase de predicciones los algoritmos de Inteligencia Artificial.

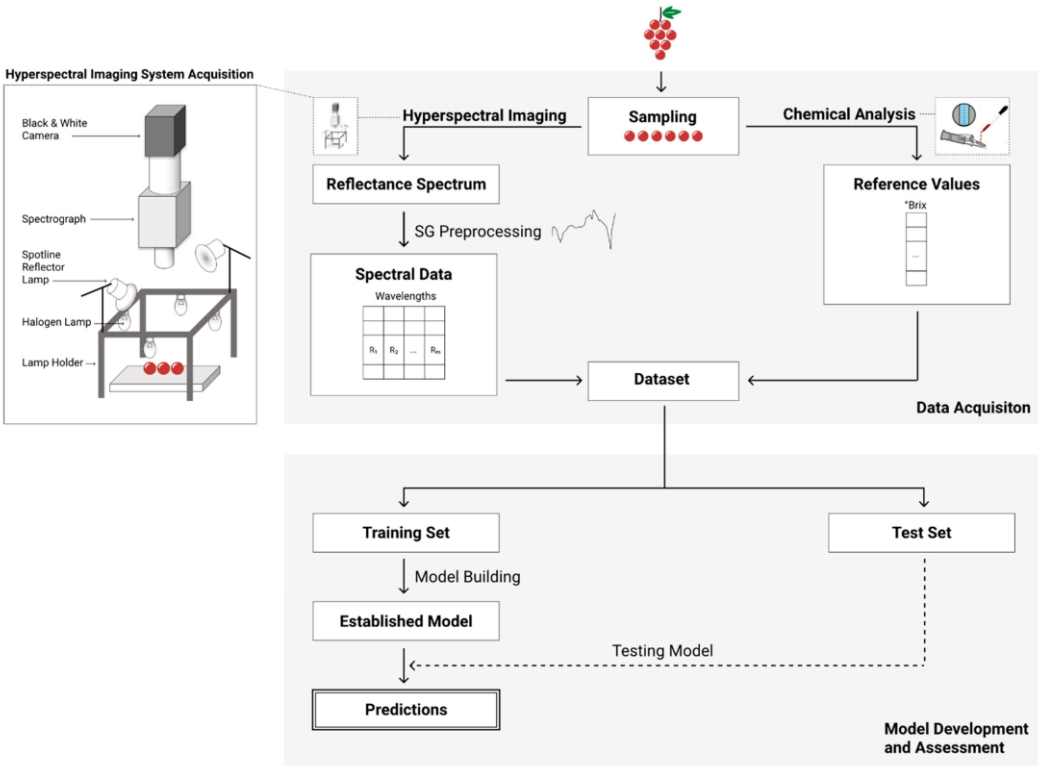

Figura 16. Representación gráfica del procedimiento experimental adoptado en [31]

Dichos procedimientos se aplican en las fases presentes en la Figura 17:

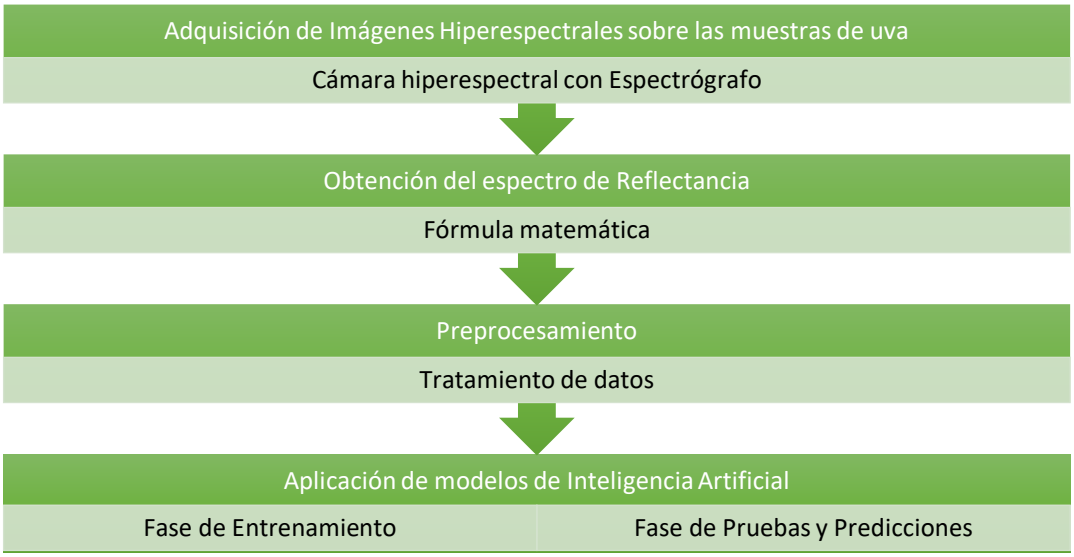

Figura 17. Fases del modelo experimental

En primer lugar, a partir de las muestras de uva recogidas, se emplea la cámara hiperespectral para obtener las imágenes hiperespectrales. Las mediciones hiperespectrales para cada muestra se realizan a lo largo del "ecuador" de la baya (tomando el pedicelo -soporte delgado y alargado que sostiene un solo fruto- como refeencia) y para tres posiciones diferentes de la baya correspondientes a rotaciones de aproximadamente 120°. Posteriormente, tras hallar los valores de reflectancia y aplicar preprocesamientos, los datos obtenidos conformarán la Base de Datos con información química y analítica sobre los espectros asociados a las uvas. A partir de ahí, podrán aplicarse técnicas AI para buscar patrones y obtener resultados sobre el vino.

Los valores de **reflectancia** se calculan con el objetivo de corregir las variaciones de la señal causadas por la iluminación y la cámara hiperespectral tras la toma de las instantáneas. La reflectancia se define como la relación entre la intensidad total de la luz reflejada por una muestra y la intensidad total de la luz que incide sobre la muestra. Por lo tanto, para una longitud de onda dada, λ y, debido al uso de imágenes hiperespectrales, también para una determinada posición, x, la reflectancia, R, se calcula como:

$$R(x,\lambda) = \frac{GI(x,\lambda) - DI(x,\lambda)}{SI(x,\lambda) - DI(x,\lambda)}$$

donde *GI* es la intensidad de la luz reflejada por las uvas, *SI* la intensidad de la luz proveniente del blanco de referencia y *DI* la señal de corriente oscura (ruido electrónico) asociada a la salida de la cámara hiperespectral manteniendo el obturador de la cámara cerrado.

Los artículos que adoptan esta metodología para la adquisición de espectros e información enológica, toman 32 imágenes hiperespectrales para SI(x, λ), DI(x, λ) y para cada conjunto de posiciones en GI(x, λ) con el objetivo de minimizar el ruido de medición. Las imágenes hiperespectrales finales se obtienen promediando las 32 imágenes y, tras la identificación de las bayas de uva, se calculan los valores de reflectancia a través de la ecuación R(x, λ) presente en el párrafo anterior.

Para crear un espectro de reflectancia único para cada muestra, los espectros de reflectancia medidos para todos los puntos de las bayas se promedian sobre la dimensión espacial y las posiciones.

A la hora de aplicar algoritmos de Inteligencia Artificial, el proceso dependerá del algoritmo seleccionado. En el caso de [31], se aplica una arquitectura de Red Neuronal Convolucional Unidimensional, 1D CNN, como la mostrada en la Figura 18.

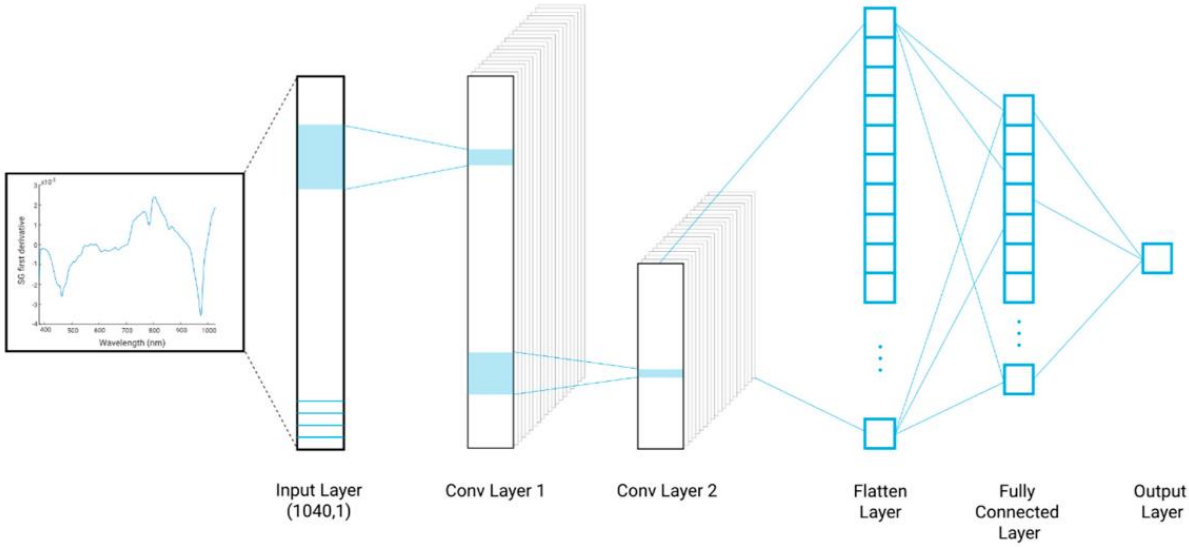

Figura 18. Arquitectura 1D CNN empleada en [31]

Esta Red Neuronal 1D CNN trabaja sobre el espectro de reflectancia final tras aplicar preprocesados. En este caso, se aplicó la 1ª derivada de *Savitzky-Golay* como preprocesado espectral.

3.1.1.3 Importancia de la Inteligencia Artificial combinada con imágenes hiperespectrales en el vino

Junto con la ayuda de métodos efectivos de Aprendizaje Automático, las imágenes hiperespectrales permiten **estimar múltiples parámetros enológicos** a partir de los datos espectrales, ya que la luz reflejada depende de la composición química de las uvas. De hecho, debido a la compleja estructura de longitud de onda espacial con picos superpuestos, es imperativo utilizar AI para convertir los datos espectrales en información enológica. Estos métodos tienen la capacidad de, a partir de muestras de entrenamiento provenientes de espectros medidos en las imágenes hiperespectrales y la información enológica medida de las muestras (AI), generar modelos de aprendizaje favorables. Por lo tanto, una vez que se establece el modelo, pueden predecir la información química de interés para nuevos conjuntos de muestras [29].

La principal desventaja de emplear imágenes hiperespectrales junto con algoritmos de Inteligencia Artificial es la dimensionalidad de los espectros de reflectancia. En el caso de las Redes Neuronales, las entradas deben ser lo más pequeñas posible para proporcionar buenos resultados de generalización. Además, el tiempo de entrenamiento de etapa aumenta con el tamaño de la Red Neuronal. Por lo tanto, es importante que se utilicen procedimientos eficientes de reducción de dimensionalidad. Una de las técnicas más utilizadas es el Análisis de Componentes Principales, PCA

Una gran parte de los artículos de este tercer punto hace uso de la combinación de imágenes hiperespectrales e Inteligencia Artificial para evaluar distintas variables que afectan a la madurez, sabor o color del vino.

3.2 Predicción de variables enológicas empleando técnicas de espectroscopía

En este apartado se incluyen los artículos cuyo objetivo es predecir y/o medir los valores enológicos de algunas variables presentes en el vino, como son el contenido de azúcar, el pH o las antocianinas. La práctica de espectroscopía más elegida por los autores es el empleo de imágenes hiperespectrales como modo de obtención de los espectros de las muestras de uva tomadas. A partir de esos espectros, se aplican técnicas IA.

3.2.1 Artículo 1*: Towards robust Machine Learning models for grape ripeness assessment* [28]

El presente trabajo aborda la aplicabilidad del modelo de Redes Neuronales profundas (DNN) para la evaluación de la calidad de las uvas de vino a través de 1D-CNN, utilizando mapas de activación de regresión (RAM) para mostrar la contribución de cada longitud de onda para la **predicción del contenido de azúcar**. De esta manera, se identifican las regiones espectrales relevantes relacionadas con el parámetro enológico.

Se aplica como preprocesamiento *Savitzky-Golay* (SG), formando el modelo híbrido SG + 1D CNN. Valores de reflectancia fueron calculados para corregir posibles variaciones de las imágenes hiperespectrales.

Con el fin de desarrollar y probar el modelo SG + 1D CNN se reunieron un total de 1748, 454 y 463 muestras de uva de los años 2012, 2013, 2014, 2016, 2017 y 2018 para *Touriga Franca*, y de los años 2013, 2014, 2016 y 2017 para *Touriga Nacional* y *Tinta Barroca*, respectivamente.

Las imágenes hiperespectrales adquiridas a partir de las muestras de uva tenían formato 1040 x 1392: las 1040 longitudes de onda oscilando entre 380 y 1028 nm, con aproximadamente 0,6 nm de separación entre longitudes de onda; los 1392 píxeles de dimensión espacial sobre las muestras que fue de aproximadamente 110 mm de ancho.

La estructura de la Red Neuronal Convolucional unidimensional consiste en una capa de entrada, dos capas convolucionales unidimensionales, capas de sondeo, una capa plana, una capa completamente conectada y una capa de salida, además de una capa 1D de agrupación promedio entre la capa de salida y el último bloque convolucional.

Finalmente, se generó un mapa de calor del espectro. Cinco espectros aleatorios de *Touriga Franca* fueron seleccionados. La Figura 19 ilustra el puntaje de contribución de cada longitud de onda y en cada espectro seleccionado aleatoriamente (1 a 5) a través de un mapa de calor, utilizando los mapas de activación de regresión obtenidos anteriormente.

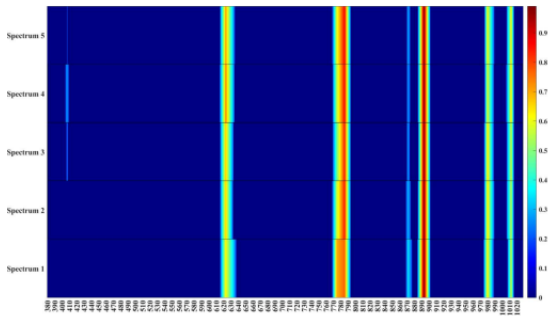

Figura 19. Mapa de calor ilustrando el puntaje de contribución de cada longitud de onda

El grado de oscuridad del color corresponde con el valor de los mapas de activación de regresión, dando valores normalizados entre 0 y 1. El color rojo más oscuro representa el valor de activación más alto.

La Figura 20 muestra una ampliación de las bandas de longitud de onda más relevantes de la Figura 19 para poder identificar mejor las regiones espectrales más importantes.

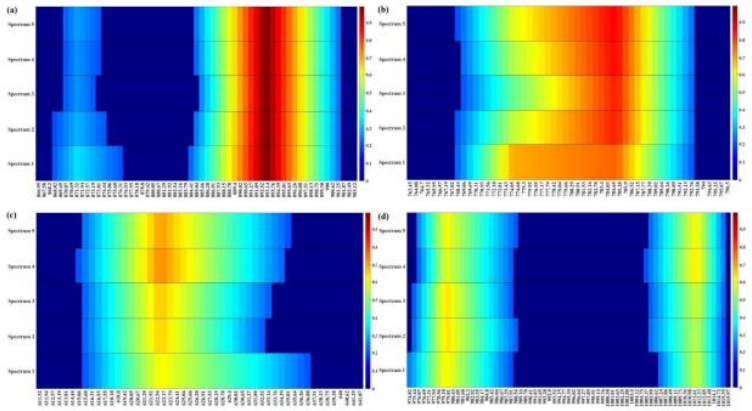

Figura 20. Ampliación de los puntos de interés identificados en la Figura 19

Se observa que hay 3 regiones que predominan en importancia y otras dos que aportan de forma más secundaria. Las principales regiones son: 888-896 nm, 770-790 nm y 615-630nm. Las regiones espectrales 975-900 y 1008-1015 nm también muestran algunas contribuciones importantes. Además, los cinco espectros seleccionados presentan similares bandas espectrales destacables.

De la Figura 20A también es posible identificar las bandas en torno a 893 nm como las más importantes para la estimación del contenido de azúcar. Los resultados aquí obtenidos coinciden con lo estipulado de forma analítica para predecir el contenido de azúcar en las uvas.

3.2.2 Artículo 2*: Application of hyperspectral imaging and Deep Learning for robust prediction of sugar and ph levels in wine grape berries* [29]

Este trabajo presenta un modelo de arquitectura de Red Neuronal Convolucional unidimensional para la predicción del contenido de azúcar y pH, usando datos de diferentes añadas provenientes de imágenes hiperespectrales en modo de reflectancia para evaluar la capacidad de generalización del modelo SG + 1D CNN.

Para la obtención de las imágenes hiperespectrales se utilizó una cámara hiperespectral -cámara en blanco y negro JAI Pulnix- y un espectrógrafo -Specim Imspector V10E (Specim, Oulu, Filand)-. No especifica el rango de operación.

> *Las imágenes hiperespectrales permiten recopilar información sobre la intensidad de la luz reflejada por las uvas en función de su longitud de onda. Además, las imágenes hiperespectrales permiten la adquisición de un gran número de muestras para evaluar la madurez de la uva localmente en el viñedo, siendo un importante valor añadido para la industria.*

Al igual que en el artículo anterior, con el fin de desarrollar y probar el modelo SG + 1D CNN, se reunieron un total de 1748, 454 y 463 muestras de uva de los años 2012, 2013, 2014, 2016, 2017 y 2018 para *Touriga Franca*, y de los años 2013, 2014, 2016 y 2017 para *Touriga Nacional* y *Tinta Barroca*. La adquisición de imágenes hiperespectrales se realizó utilizando muestras de dichas uvas recogidas. Cada muestra comprendía seis o 12 bayas de uva, recolectadas al azar de un solo racimo.

La estructura del modelo CNN 1D propuesta, desarrollada mediante la herramienta KERAS *package* 2.2.4 en Python, consistía en una capa de entrada, dos capas convolucionales unidimensionales, capas de sondeo, una capa plana, una capa completamente conectada y una capa de salida, como se ilustra en la Figura 21.

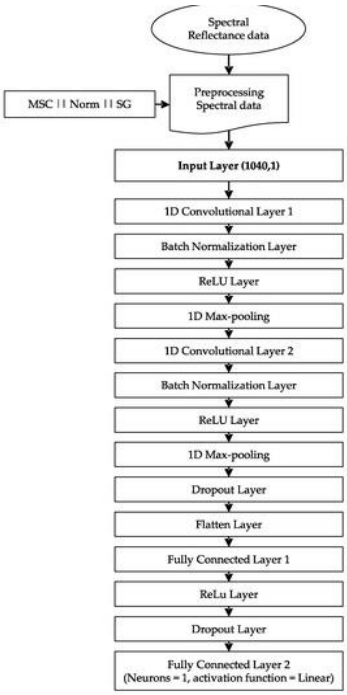

Figura 21. Estructura del modelo CNN 1D propuesto

También se introdujeron fases de normalización de lotes y capas de abandono. Diversos parámetros como el número de filtros, el tamaño del *kernel* o tasas de aprendizaje, entre otros, fueron optimizados a través de la optimización Bayesiana por medio de un proceso Gaussiano (BOGP) y el error cuadrático medio (MSE). Una breve descripción de la estructura se presenta a continuación.

Los resultados muestran que *Savitzky–Golay* + arquitectura 1D CNN es la mejor técnica para la predicción del contenido de azúcar y pH, obteniendo el error cuadrático medio más bajo y el mejor rendimiento general. Se mostró un gran desempeño de generalización con valores RMSEP (*root-mean-square error of predictions*) de 1,118 ºBrix y 1,085 ºBrix para el contenido de azúcar y 0,199 y 0,183 para pH, para conjuntos de prueba de diferentes variedades y diferentes cosechas, respectivamente. Se observa una correlación entre el aumento RMSEP y diferentes distribuciones de contenido de azúcar y pH.

También se observa que el modelo desempeñó mejor para el contenido de azúcar que para pH, con un menor aumento RMSEP. El modelo maneja mejor los valores de azúcar comprendidos entre 17 y 21 ºBrix (percentiles 25 y 50 en el conjunto TB y TN), que pertenecen al rango de valores delimitado por los percentiles 25 y 75 del conjunto de entrenamiento TF.

3.2.3 Artículo 3: *A review of different dimensionality reduction methods for the prediction of sugar content from hyperspectral images of wine grape berries* [30]

Se aplicaron varias técnicas de reducción de dimensionalidad (preprocesamientos) a imágenes hiperespectrales de vino en modo de reflectancia para elaborar un estudio de la eficiencia de los modelos de Aprendizaje Automático en la predicción del contenido de azúcar. Los resultados obtenidos demostraron que el modelo predijo favorablemente los valores del parámetro enológico de un año de cosecha a otro. Se emplea el modo de reflectancia y una cámara hiperespectral operando en las bandas 380 – 1028 nm en el proceso de espectroscopía.

Las instancias seleccionadas pertenecen a la variedad *Touriga Franca*, tomando un total de 608 muestras entre los años 2012 y 2015, cada muestra compuesta por la media de 6 bayas. La Figura 22 muestra un ejemplo de las mediciones de reflectancia obtenidas para el conjunto de muestra *Touriga Franca* 2013.

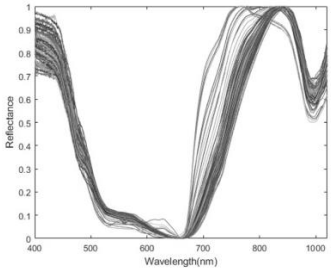

Figura 22. Valores de reflectancia obtenidos para todas las muestras Touriga Franca 2013

Se comparan técnicas de reducción de dimensionalidad tanto lineales (PCA) como no lineales globales (técnicas globales no lineales, Kernel PCA y M*ultilayer Autoencoders*) y no lineales locales (*Local Linear Embedding* (LLE) y *Laplacian Eigenmaps*). Se emplea 10-*fold cross validation* para seleccionar el número de muestras de entrenamiento y validación, aplicado para comparar los resultados aplicando *Neural Networks* (NNs) y *Support Vector Regression* (SVR).

Los resultados muestran que, para las técnicas de reducción de dimensionalidad estudiadas, PCA supera a las técnicas no lineales para el caso de datos hiperespectrales del mundo real. Además, las técnicas locales superan a las técnicas globales, con LLE obteniendo resultados superiores a los otros métodos, pero sin superar los índices ofrecidos aplicando PCA.

En cuanto a los resultados con diferentes regresores, ambos obtienen buenos resultados, lo que podría indicar que el desempeño de los métodos de reducción de dimensionalidad es independiente del regresor.

Los mejores resultados para cada añada son los ilustrados en la Figura 23:

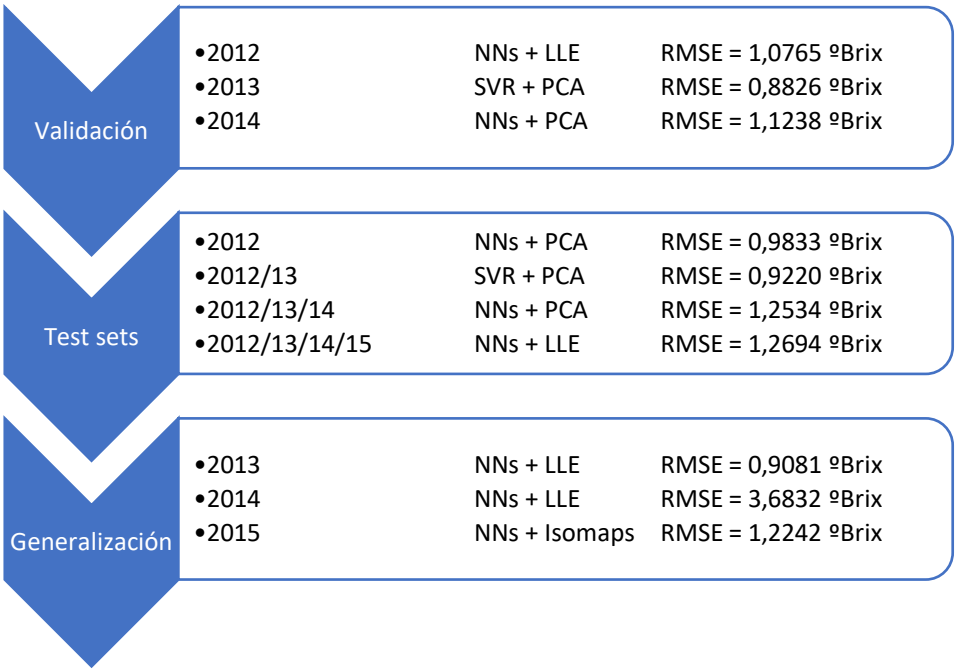

Figura 23. Gráfico ilustrativo de los resultados RMSE obtenidos por fase y añada

3.2.4 Artículo 4: *Prediction of sugar content in port wine vintage grapes using Machine Learning and hyperspectral imaging* [31]

Este artículo, tomado como base de explicación de los modelos experimentales empleando imágenes hiperespectrales presente en el punto 3.1.1.2 de este documento, estudia el contenido de azúcar de las bayas de vino centrándose en la fase de cosecha haciendo uso de imágenes hiperespectrales preprocesadas mediante *Savitzky−Golay* (SG) para extraer información de las bayas que servirá como entrada de los algoritmos de Aprendizaje Automático que producen estimaciones del nivel de azúcar.

Para la obtención de las imágenes hiperespectrales se utilizó una cámara hiperespectral -cámara en blanco y negro JAI Pulnix- y un espectrógrafo -Specim Imspector V10E (Specim, Oulu, Filand)- operando en las bandas 380 – 1028 nm. Se

reunieron un total de 1748, 454 y 463 muestras de uva de añadas diferentes de los vinos *Touriga Franca* (TF), *Touriga Nacional* (TN) y *Tinta Barroca* (TB).

Se comparan cuatro métodos diferentes de aprendizaje, todos ellos probados sobre la herramienta MATLAB R2019b. Dichos métodos son: *Ridge Regression*, o Regresión de Cresta, *Partial Least Squares* o Mínimos Cuadrados Parciales, Redes Neuronales y Redes Neuronales Convolucionales, es decir, RR, PLS, NN y 1D CNN.

Los valores de TF van desde 7,87 hasta 30,26 ºBrix, mientras que en TN y TB tienen mínimas de 6,95 y 5,48 ºBrix y las máximas de 29,66 y 29,95 ºBrix. Los resultados muestran que los modelos estimados pueden predecir con éxito el contenido de azúcar a partir de datos hiperespectrales, siendo 1D CNN el método que ofrece mejores prestaciones.

3.2.5 Artículo 5: *Determination of sugar, ph, and anthocyanin contents in port wine grape berries through hyperspectral imaging: an extensive comparison of linear and non-linear predictive methods* [32]

Este artículo compara 16 métodos de regresión lineal y no lineal diferentes para predecir el contenido de azúcar, pH y antocianina de las uvas a través de imágenes hiperespectrales en modo de reflectancia. Todos los cálculos se realizaron en el entorno MATLAB R2019b.

Las bayas de uva de vino consideradas en el presente trabajo son de la variedad portuguesa, *Touriga Franca* (*Vitis vinifera L.*), contando con un total de 240 racimos que conforman -a partir de 6 uvas por muestra- los 240 espectros.

La configuración utilizada para adquirir los datos espectrales (imágenes hiperespectrales de exploración lineal) consistía en un espectrofotómetro UV/Vis (Shimadzu) empleado para medir la concentración de antocianinas totales de forma fotométrica mediante el método *SO2 bleaching*, y una cámara hiperespectral de 1040 × 1392 píxeles, en los que los 1040 píxeles son los canales de longitud de onda medidos con un ancho de aproximadamente 0,6 nm (rango de 380 a 1028 nm), y los 1392 píxeles

indicaban la dimensión espacial (una línea sobre la muestra) con un ancho de aproximadamente 110 nm. Un ejemplo de estas imágenes obtenidas por la cámara hiperespectral se muestra en la Figura 24.

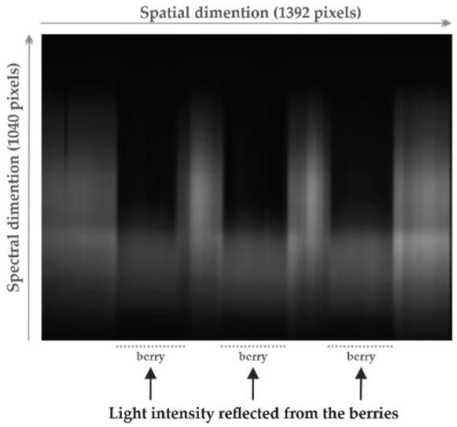

Figura 24. Imagen hiperespectral de barrido lineal adquirida antes de la segmentación, considerando tres uvas fotografiadas simultáneamente

Después, se realizan diversos procesamientos y promediados para obtener un único espectro de reflectancia para cada muestra. En la Figura 25 observamos el espectro de reflectancia de las 240 muestras de la variedad *Touriga Franca* tomadas a partir de los datos hiperespectrales recopilados.

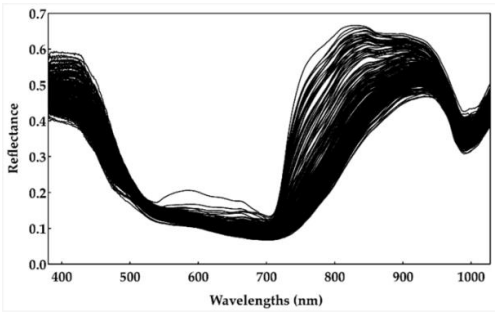

Figura 25. Espectro de reflectancia de las 240 muestras

Se desarrolla un marco de análisis predictivo lineal y no lineal (L&NL-PAC), totalmente integrado con cinco técnicas de preprocesamiento -*Multiplicative Scatter Correction* (MSC), *Standard Normal Variate* (SNV), Normalización, *Savitzky−Golay* (SG) + 1ª derivada y SG + 2ª derivada- y cinco clases diferentes de métodos de regresión con 16 métodos en total, como muestra la Figura 26, comparado a través un robusto esquema de división de datos estratificados de doble validación cruzada de Monte Carlo.

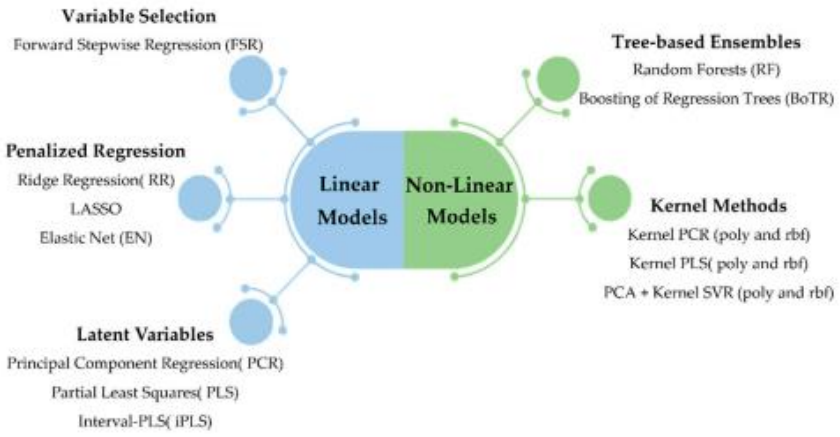

Figura 26. Modelos de regresión empleados en el estudio

L&NLPAC permitió la identificación de los mejores métodos de regresión, así como longitudes de onda más contribuyentes a la variabilidad de cada parámetro enológico.

- En cuanto al preprocesamiento, SNV, MSC y la primera derivada de *Savitzky-Golay* (SG) ofrecen un desempeño similar, mientras que la segunda derivada de SG proporcionó malos resultados. Por tanto, la primera derivada de SG fue el enfoque de preprocesamiento seleccionado para realizar las comparaciones entre los métodos de regresión. De los 16 métodos de regresión probados, los mejores resultados se obtuvieron con la regresión de cresta (RR). Se muestran los valores obtenidos por los mejores métodos distinguiendo entre el parámetro enológico en la Tabla 21.

Enological Parameter	Ranking Three Best Methods (1st, 2nd and 3rd Positions)	RMSEP (Mean ± Sd Values)
Sugar content	RR	0.998 ± 0.102
	PLS	1.055 ± 0.127
	KPLS rbf	1.060 ± 0.128
pH	RR	0.168 ± 0.014
	KPLS rbf	0.172 ± 0.015
	PLS/ KPCR polynomial	0.172 ± 0.015/ 0.175 ± 0.016
Anthocyanin concentration	EN	19.773 ± 2.019
	PCR	19.864 ± 2.231
	RR	19.961 ± 2.061

Tabla 21. Resultados de niveles de azúcar, pH y antocianinas según el modelo empleado

DESTACAN AQUELLAS REGIONES OSCILANTES ENTRE 700 Y 960 NM PARA LOS TRES PARÁMETROS ENOLÓGICOS, CON ALGUNOS OTROS PICOS IMPORTANTES PARA LA MEDICIÓN DE ANTOCIANINAS ENTRE 400 Y 520 NM.

LAS REGIONES ESPECTRALES ENTRE 400 Y 500 NM TAMBIÉN FUERON SUTILMENTE RELEVANTES PARA EL CONTENIDO DE AZÚCAR Y EL PH, PERO TENDÍAN A SER MÁS RUIDOSAS.

3.2.6 Artículo 6: *Development of predictive models for quality and maturation stage attributes of wine grapes using Vis-NIR reflectance spectroscopy* [33]

El objetivo principal del siguiente trabajo fue desarrollar modelos predictivos de calidad y atributos de la etapa de maduración de las uvas de vino utilizando espectroscopía de reflectancia visible/infrarrojo cercano (Vis-NIR) mediante un espectro-radiómetro FieldSpec® 3 operando en longitudes de onda comprendidas entre 350 - 2500 nm, aunque se trabaja únicamente en 450–1800 nm por excesivo ruido.

Un total de 432 *Syrah* y 576 bayas de *Cabernet Sauvignon* fueron recolectados y sus espectros de reflectancia fueron adquiridos y preprocesados. En un paso posterior se determinaron como patrones de referencia los sólidos solubles totales, las antocianinas totales y los flavonoides amarillos.

Regresión de Componentes Principales (PCR) y Regresión de Mínimos Cuadrados parciales (PLSR) se utilizaron como modelos predictivos utilizando tanto el espectro de datos completo (450–1800 nm) como un conjunto más pequeño de muestras espectrales seleccionadas por el método *Jack-Knife*. Los resultados fueron más favorables empleando este último método. También se desarrollaron otros modelos predictivos utilizando Regresión Lineal Múltiple (MLR) con atributos de calidad.

Los estados de maduración se discriminaron mediante Análisis de Componentes Principales + Análisis Lineal Discriminante (PCA-LDA), Análisis de Componentes Principales + Análisis Discriminante Cuadrático (PCA-QDA), Análisis de componentes principales + LDA utilizando la distancia de Mahalanobis (PCA-LDA+Mahalanobis), y técnicas de clasificación de Análisis Discriminante de Mínimos Cuadrados Parciales (PLS-DA).

Los modelos de regresión PCR, PLSR y MLR han proporcionado predicciones favorables para los sólidos solubles totales y antocianinas ($R^2 \geq 0{,}90$), así como precisión igual o superior al 70% para el contenido de flavonoides. Además, fue posible diferenciar las distintas etapas de maduración de las vides con un 93,15 % de precisión utilizando PLS-DA.

Se identifican firmas espectrales de los 3 atributos de calidad medidos para diferentes longitudes de onda en las regiones espectrales visible e infrarroja, lo que permite desarrollar instrumentos optoelectrónicos enfocados a las necesidades del sector vitivinícola:

- La región **visible** (**380–780 NM**) está relacionadas con la presencia de pigmentos verdes (clorofila) y rojos (**ANTOCIANINAS** y flavonoides)
- La región espectral **infrarroja** (780–2500 nm) está relacionada con las bandas de absorción por: **AZÚCARES** (alrededor de **980 NM**); agua (alrededor de 973, 1324 y 1581 nm) y **ANTOCIANINAS** (alrededor de **1154 NM**)

3.2.7 Artículo 7*: Using Support Vector Regression and hyperspectral imaging for the prediction of oenological parameters on different vintages and varieties of wine grape berries [34]*

El principal objetivo de este trabajo es predecir la concentración de antocianinas, pH y el contenido de azúcar en bayas de vino medidas mediante espectroscopía basada en imágenes hiperespectrales en modo de reflectancia de 380–1028 nm. Se utiliza para ello SVR con distintos tipos de *kernel* -lineal, sigmoideo, polinómico y de base radial gaussiana- destacando el *kernel* de base radial gaussiana.

Se midieron las siguientes variedades de vino: *Touriga Franca* (TF) 2012-2015 - empleada para entrenar, validar y probar la metodología SVR-, *Touriga Nacional* (TN) 2013 y *Tinta Barroca* (TB) 2013 - probar la capacidad de generalización del enfoque SVR-. Se utilizaron 552 muestras de TF, 60 de TN y 84 de TB. Cada muestra fue recolectada durante el proceso de maduración utilizando imágenes hiperespectrales en el rango de 380–1028 nm, como apreciamos en la Figura 27.

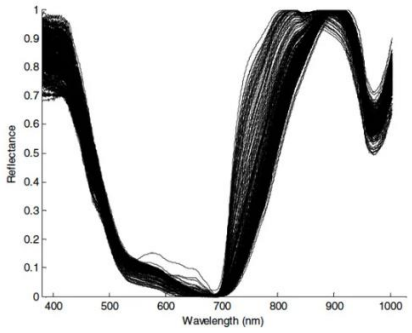

Figura 27. Mediciones de reflectancia para las muestras de la variedad TF 2012

A continuación, se llevó a cabo un Análisis de Componentes Principales (PCA) presente en la Figura 28.

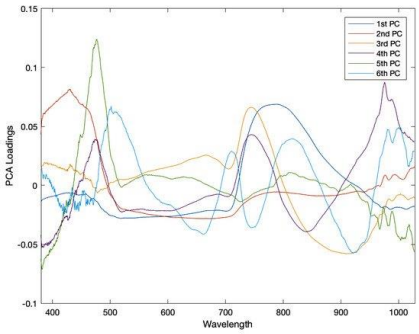

Figura 28. Gráfico de Componentes Principales TF 2012

Observando la Figura 28 cabe destacar la gran cantidad de picos a lo largo del espectro, con énfasis entre **400-700 NM**, que suelen estar relacionados con compuestos químicos como la clorofila, los carotenoides y las **ANTOCIANINAS**, mientras que los picos entre **700-800 NM** son comúnmente asociado con **AZÚCARES.**

Por último, se procede a realizar un análisis de algunos componentes cruciales en el vino. Los mejores resultados de R^2 obtenidos por el modelo fueron 0,89, 0,81 y 0,90, con valores RMSE de 35,6 $mg \cdot L^{-1}$, 0,25 y 3,19 ◦Brix, para concentración de antocianinas, índice de pH y contenido de azúcar, respectivamente.

3.2.8 *Artículo 8: Comparison of different approaches for the prediction of sugar content in new vintages of whole Port wine grape berries using hyperspectral imaging* [35]

Este artículo comparar los resultados de PLS y NN para monitorizar la calidad de las uvas utilizando predicciones de contenido de azúcar basadas en imágenes hiperespectrales tomadas con cámara hiperespectral con espectrómetro integrado (380 – 1028 nm).

La variedad de vino empleada fue *Touriga Franca*, contando con 324 muestra. De las muestras recolectadas en 2012, había 210 muestras disponibles para entrenamiento y validación y 30 muestras para probar los modelos. Todas las muestras recolectadas en 2013 fueron dispuestas para la fase de prueba.

Los resultados para la regresión PLS y NN en cuanto al conjunto de prueba (muestras de 2012) fueron 0,94 °Brix y 0,96 °Brix RMSE, y 0,93 y 0,92 para los coeficientes de correlación al cuadrado (R^2), respectivamente. Los resultados obtenidos empleando el conjunto de prueba (muestras de 2013), ofreció valores RMSE de 1,34 °Brix y 1,35 °Brix, y valores R^2 0,95 y 0,92. PLSR y NNs muestran un rendimiento similar, siendo ambas técnicas capaces de predecir el contenido de azúcar.

3.2.9 Artículo 9: *Characterization of Neural Network generalization in the determination of pH and anthocyanin content of wine grape in new vintages and varieties* [36]

Este trabajo mide la capacidad de generalización de imágenes hiperespectrales (380 - 1028 nm) combinada con Redes Neuronales (NN) en la estimación de pH y contenido de antocianinas durante la maduración de la uva.

Se entrena el algoritmo NN con muestras de uva de la variedad *Touriga Franca* (TF) cosechadas en 2012 para, a continuación, probar los resultados sobre la añada de 2013 y dos nuevas variedades, *Touriga Nacional* (TN) y *Tinta Barroc*a (TB) de 2013. Cada muestra contenía una pequeña cantidad de bayas enteras. Se contó con 210 muestras disponibles TF 2012 para entrenamiento y validación y 30 muestras para la prueba. Para TF, TN y TB 2013 hubo 81, 60 y 84 muestras disponibles, respectivamente, todas empleadas en la fase de prueba.

Para obtener las imágenes hiperespectrales y los espectros asociados a partir de dichas muestras, se tomaron instantáneas de seis bayas de uva en para tres posiciones diferentes rotando las bayas aproximadamente 120°. Para crear un espectro de reflectancia único, todos los puntos de las bayas se promediaron sobre la dimensión espacial y la rotación. Por último, el espectro resultante se normalizó.

La mejor estructura de Redes Neuronales contaba con una capa oculta con dos neuronas y 18 componentes principales (PC) como entrada para la estimación del pH y 14 PC de entrada para la determinación de antocianinas. El número óptimo de PC se determinó minimizando el error cuadrático medio de la validación cruzada *n-fold*. Esta Red Neuronal se entrenó mediante *Levenberg-Marquardt*, que incluye *backpropagation.*

Los resultados estimando el pH son muy favorables (RMSE_pH de TF, TN y TB: 0,191, 0,170 y 0,176), en cambio, para medir las antocianinas, los resultados son favorables para TF y TN 2013 pero se obtiene una mala generalización para la variedad TB de 2013 (RMSE_Anth de TF, TN y TB: 22,1, 23,2 y 51,3 $mg \cdot L^{-1}$).

3.2.10 Artículo 10: *Visible-Near Infrared reflectance spectroscopy for nondestructive analysis of red wine grapes* [37]

Este artículo busca analizar distintos componentes presentes en las uvas de vino tinto mediante espectroscopía en modo de reflectancia Vis-NIR.

Las muestras se toman de tres variedades de vino distintas en dos añadas consecutivas: 43 y 36 de Cabernet Sauvignon, 83 y 80 de *Cabernet Franc* y 38 y 36 de *Syrah*, respectivamente para los años 2009 y 2010. Los espectros de reflectancia sobre las muestras recogidas fueron tomados con un espectrómetro de matriz de diodos en un rango de longitud de onda de 350 a 850 nm.

Se usó PLS sujeto a selección de variables por eliminación recursiva de características. Se realizaron análisis químicos para el contenido de sólidos solubles, Brix, pH, acidez titulable (TA), fenoles totales y antocianinas totales para todas las muestras. Se utilizaron diversas técnicas de preprocesamiento: normalización, SG y SNV.

Los modelos de mejor rendimiento para Brix, pH, TA, fenoles y antocianinas en 2009 tenían errores cuadráticos medios (RMSEP) de 0,65, 0,05, 0,59 g/L, 31,2 mg/L y 75 mg/L, respectivamente, con el correspondiente R^2 de 0,84, 0,58, 0,56, 0,27 y 0,65. Los mejores modelos de 2010 tenían RMSEP de 0,65, 0,05, 0,86 g/L, 27,9 mg/L y 111 mg/L, respectivamente, con valores R^2 de 0,89, 0,81, 0,58, 0,25 y 0,17.

Las calibraciones de 2009 se usaron para estimar el Brix y el pH a partir de los datos espectrales de las muestras recolectadas en la siguiente temporada de crecimiento. En dicho proceso se obtuvo un rendimiento RMSEP de 0,87 y 0,05 y valores R^2 de 0,71 y 0,56, respectivamente. La descomposición del análisis de componentes principales

(PCA) de los datos de reflectancia de 2009 y 2010 mostró similitudes en las cargas resultantes, lo que indica una estructura de datos subyacente similar.

3.2.11 *Artículo 11: Predicting the anthocyanin content of wine grapes by NIR hyperspectral imaging* [38]

Este artículo tiene como objetivo demostrar la viabilidad del uso de imágenes hiperespectrales en modo de reflectancia para predecir cambios en el contenido de antocianinas en uvas de vino durante la maduración. En este caso, se empleó un espectrógrafo hiperespectral para la obtención de imágenes hiperespectrales (ImSpector N17E, Spectral Imaging Ltd., Finlandia) con un rango espectral de 900 a 1700 nm.

Se recolectaron 120 muestras (50 uvas/muestra) de uva de la variedad *Cabernet Sauvignon* en 6 tomas diferentes y se midió el contenido de antocianinas empleando el sistema de imagen hiperespectral. Sobre 72 de esas muestras se formó el conjunto de espectros de calibración y las 48 muestras restantes dieron lugar a los espectros del conjunto de validación. Los espectros asociados a dichas 6 recolectas diferentes se muestra en la Figura 29.

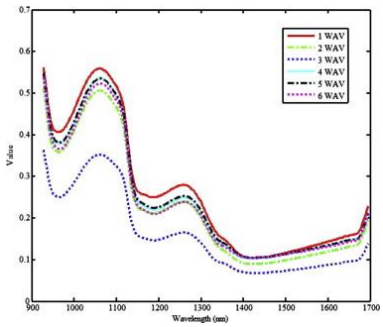

Figura 29. Valores espectrales medios de las muestras de uva Cabernet Sauvignon en las 6 recolectas

Después, se realizó un proceso de preprocesamiento incluyeron suavizado Savitzky-Golay (SG), Variación Normal Estándar (SNV), Corrección de Dispersión Multiplicativa (MSC), primera derivada y segunda derivada.

Se desarrolló un modelo cuantitativo empleando tanto Mínimos Cuadrados Parciales de Regresión (PLSR) como Regresión de Vector de Soporte (SVR) para calcular el contenido de antocianinas.

Los mejores resultados se obtuvieron empleando PLSR + SVM, ofreciendo un coeficiente de validación (P-R^2) de 0,9414 y RMSEP de 0,0046, superior al modelo PLSR, que tenía un P-R^2 de 0,8407 y un RMSEP de 0,0129. Por lo tanto, la imagen hiperespectral puede ser un método rápido y no destructivo para predecir el contenido de antocianinas de las uvas de vino durante la maduración.

3.2.12 Artículo 12: *Brix, pH and anthocyanin content determination in whole Port wine grape berries by hyperspectral imaging and Neural Networks* [39]

Este artículo pretende mejorar las mediciones no destructivas y simultáneas de varios parámetros enológicos utilizando imágenes hiperespectrales en modo de reflectancia de 380 a 1028 nm combinadas con Redes Neuronales (NN) aplicando PCA y *7-fold-cross-validation* sobre los 240 espectros de *Touriga Franca*.

El espectro de cada muestra se adquirió mediante imágenes hiperespectrales sobre 6 bayas enteras de uva. Los espectros fueron convertidos a parámetros enológicos, es decir, las variables a medir como resultado final, por Perceptrones Multicapa.

La Red Neuronal tenía únicamente una capa oculta con dos o tres neuronas y una sola neurona de salida. La función de activación de las neuronas ocultas y de salida fue la tangente hiperbólica y la identidad, respectivamente. La función de activación de las neuronas ocultas fue no lineal para poder representar cualquier función de salida. El entrenamiento se realizó utilizando *Levenberg-Marquardt*: tipo de retropropagación con tasa de aprendizaje variable, repetido para 100 pesos iniciales aleatorios diferentes.

El conjunto de prueba con 30 muestras reveló valores R^2 de 0,73, 0,92 y 0,95 y RMSE de 0,18, 0,95 °Brix y 14 mg/l para pH, azúcares y contenido de antocianinas, respectivamente.

3.2.13 Artículo 13: *Determination of technological maturity of grapes and total phenolic compounds of grape skins in red and white cultivars during ripening by near infrared hyperspectral image: A preliminary approach* [40]

Este estudio obtiene imágenes hiperespectrales de uvas intactas durante la maduración usando un sensor hiperespectral infrarrojo cercano (900–1700 nm).

Se recolectaron 213 muestras espectrales (99 de uva roja y 114 muestras espectrales de uva blanca) de variedades *Vitis vinífera L. cv. Zalema*, *Tempranillo* y *Syrah* de cuatro viñedos ubicados en la Denominación de Origen Condado de Huelva D.O. (Andalucía, España). Observamos los espectros provenientes del sensor asociados a las muestras en la Figura 30.

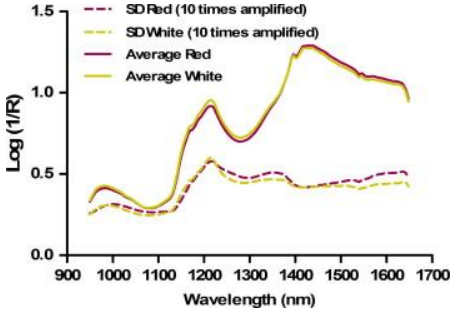

Figura 30. Espectros promediados y desviación estándar de las muestras de uva roja y blanca en zona NIR

A diferencia de muchos otros estudios, el sensor empleado por los autores opera en un rango espectral superior, por lo que las curvas de los espectros ofrecen resultados sobre variables enológicas distintas. No se especifica qué regiones espectrales corresponden a cada parámetro enológico.

Los datos espectrales se han correlacionado con distintos parámetros enológicos -concentración de fenoles totales de la piel de la uva, concentración de azúcar, acidez titulable y pH- mediante Regresión de Mínimos Cuadrados Parciales Modificada (MPLS). Se desarrollaron modelos de calibración distintos para uva roja y blanca.

Los resultados obtenidos (coeficiente de determinación RSQ y error de la validación cruzada SEP, respectivamente) para el modelo global de muestras de uva tinta y blanca fueron: 0,89 y 1,23 mg g^{-1} de piel de uva para concentración de fenoles totales, 0,99 y 1,37 ºBrix para concentración de azúcar, 0,98 y 3,88 g L^{-1} para acidez titulable y para pH 0,94 y 0,12.

3.2.14 Artículo 14: *Determination of sugar content in whole Port Wine grape berries combining hyperspectral imaging with Neural Networks methodologies* [41]

Este estudio se centra en el desarrollo de modelos de calibración avanzados haciendo uso de imágenes hiperespectrales en modo de reflectancia entre 308 y 1028 nm combinadas con Redes Neurales para medir el contenido de azúcar en racimos de uva.

El contenido de azúcar se estimó a partir de los espectros utilizando *feedforward Multiplayer Perceptrons* **en dos Redes Neuronales diferentes** entrenadas cada una con un conjunto de datos de un año diferente (2012 y 2013) correspondiente a la variedad *Touriga Franca*. La validación para ambas Redes Neuronales se realizó mediante validación cruzada *n-fold* y el conjunto de prueba utilizado fue de 2013.

En el presente trabajo, las Redes Neuronales fueron entrenadas usando el algoritmo de *Levenberg-Marquardt*[16], así como un proceso de reducción de variables mediante PCA. La fase de entrenamiento se repitió para 100 pesos iniciales diferentes generados de forma aleatoria y el entrenamiento se detuvo en el número de etapas que producían el error cuadrático medio más bajo para patrones de validación. La función de activación utilizada para calcular las capas neuronales ocultas fue la *tangente hiperbólica* y la de salida fue la función *identidad*.

Experimento 1: Datos del 2012 para el entrenamiento de la Red Neuronal

El número de componentes principales óptimos, PC, fue determinado minimizando el error cuadrático medio de la raíz de la red creada. Del total de 240 muestras recogidas en 2012, se utilizaron 210 para entrenar la red y determinar la mejor estructura de Red Neuronal. La mejor Red Neuronal tenía una capa oculta con dos neuronas y 18 PC para entradas. La red fue entrenada con *7-fold-cross-validation*.

[16] *Levenberg-Marquardt* es un enfoque de retropropagación con una tasa de aprendizaje variable que ofrece resultados más eficientes que el enfoque de algoritmo convencional.

Los resultados del conjunto de validación presentaron un R^2 y RMSE de 0,892 y 1,09 ºBrix, respectivamente. Los conjuntos de prueba presentaron un R^2 y RMSE de 0,924 y 0,955 ºBrix para muestras de 2012 y un R^2 y RMSE de 0,906 y 1,165 ºBrix para muestras de 2013. El percentil 90 para el error absoluto (AE) fue de 1,78 ºBrix y en el caso de prueba fue de 1,58 ºBrix en 2012 y 2.04 ºBrix en 2013. Estos valores corresponden al 11,4%, 10,1% y 13,1% de la variación total en los valores de contenido de azúcar en los datos recopilados, que fueron de 15,6 ºBrix. En términos de porcentaje absoluto de error, el percentil 90 fue del 10,1% en la validación, 9,30% y 14,5% en 2012 y 2013 conjunto de prueba, respectivamente.

Experimento 2: Datos del 2013 para el entrenamiento de la Red Neuronal

Del total de 84 muestras recolectadas, 60 fueron utilizadas para la fase de entrenamiento. Se implementó un procedimiento de validación cruzada similar al experimento 1. Los resultados se obtuvieron para la mejor estructura de Red Neuronal -una capa oculta con dos neuronas y 11 componentes principales.

Los valores R^2 fueron 0,884 para el conjunto de validación y 0,959 para el conjunto de prueba. En términos de RMSE, los valores fueron 1,185 ºBrix y 1.026 ºBrix, respectivamente. El percentil 90 para el error absoluto fue de 1,81 ºBrix para datos de validación y 1,72 ºBrix para datos de prueba, que corresponden al 11,6% y 11,3% de la variación total, respectivamente. En cuanto al error porcentual absoluto, el percentil 90 representa 10,1% para validación y 11,2% para conjunto de prueba, respectivamente.

Ambas redes entrenadas presentan desempeños similares. Los resultados R^2 presentados en este estudio superan los resultados existentes previamente en el modo de reflectancia y funcionan de manera similar en el modo de transmitancia. Por otro lado, para valores RMSE, los resultados aquí presentes con las muestras de 2013 son ligeramente más altos. Sin embargo, los errores de predicción obtenidos para ambas Redes Neuronales están en un rango aceptable y las actuaciones logradas fueron realmente muy satisfactorias, lo que lleva a creer en la robustez de la metodología.

3.2.15 *Artículo 15: Optimization of NIR spectral data management for quality control of grape bunches during on-vine ripening* [42]

Este artículo emplea espectroscopia NIR -medida a partir de un espectrofotómetro de matriz de diodos (380–1700 nm)- como técnica no destructiva para la evaluación de los cambios químicos en los principales atributos de calidad de las uvas para vino *Vitis vinifera L*. durante la maduración en la vid y en la cosecha.

Se utilizaron un total de 363 muestras de 25 variedades de uva blanca y roja de los años 2006, 2007 y 2008 para construir modelos de predicción de calidad, dividiendo el conjunto de calibración con 251 muestras (73% del total) y el conjunto de validación formado por las 93 muestras restantes (27%).

Se probaron dos enfoques de regresión (MPLS y LOCAL) para la cuantificación de cambios del contenido de sólidos en soluble (SSC, medido en ºBrix), contenido de azúcares reductores, valor de pH, acidez titulable, ácido tartárico, ácido málico y contenido de potasio.

Los resultados de la validación cruzada indicaron que la tecnología NIRS proporcionó una precisión excelente para los parámetros relacionados con el azúcar (R^2 = 0,94 para los parámetros SSC y contenido de azúcar reductor) y una buena precisión para los parámetros relacionados con la acidez (R^2 entre 0,73 y 0,87) para el modo de análisis de racimos ensayado utilizando la regresión MPLS. No obstante, LOCAL ofreció valores más precisos.

3.2.16 Artículo 16: *Determination of anthocyanin concentration in whole grape skins using hyperspectral imaging and Adaptive Boosting Neural Networks* [43]

Este artículo utiliza AdaBoost + PCA para la estimación de la concentración de antocianina de la uva utilizando imágenes hiperespectrales. Las entradas de la Red Neuronal fueron los componentes principales de los espectros de las uvas.

Los datos hiperespectrales se recogieron en modo de reflectancia para 46 uvas enteras individuales de la variedad *Cabernet Sauvignon* durante 2009, utilizando una cámara hiperespectral (400–1000 nm). Se realizó generalización de la calibración mediante la combinación de parada temprana y *leave-one-out-cross-validation* (LOOCV).

Se adjunta en la Figura 31 el espectro promedio asociado a las 46 uvas recolectadas. Observamos los picos de las variables enológicas en las longitudes de onda cercanas a los 400 nm y en torno a los 850-900 nm.

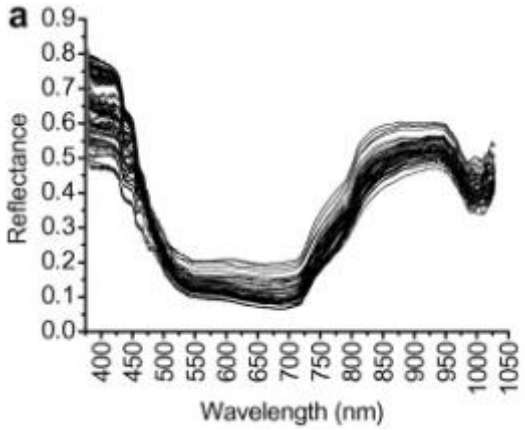

Figura 31. Espectro de reflectancia de uvas Cabernet Sauvignon

Los valores medios de contenido de azúcar de las muestras variaban entre 14,6 y 20,2 ºBrix y, empleando AdaBoost compuesto por cuatro Redes Neuronales, el coeficiente de correlación al cuadrado de concentración de antocianina obtenidos por las calibraciones tuvo un valor de 0,65.

Al intentar aplicar PLS a los datos expuestos, el coeficiente de correlación al cuadrado fue de 0,25. Esto sugiere que el uso de las Redes Neuronales y AdaBoost ofrece grandes resultados frente a otras líneas de investigación.

La cancelación de errores en las salidas de la Red Neuronal se realizó promediando estas salidas, por lo que los resultados de AdaBoost pueden mejorarse incluso si la Red Neuronal recién agregada tiene un coeficiente de correlación al cuadrado, error absoluto medio o error porcentual absoluto medio peor que las Redes Neuronales de AdaBoost. Por ello los resultados de AdaBoost no siempre mejoran cuando se agrega una nueva Red Neuronal.

3.2.17 Artículo 17: *Soluble solids content and pH prediction and varieties discrimination of grapes based on visible–near infrared spectroscopy* [44]

Este estudio hace uso de espectroscopia de reflectancia visible e infrarroja cercana para la predicción del contenido de sólidos solubles (SSC) y el pH de las uvas. Se emplean algoritmos genéticos (GA) para resaltar longitudes de onda relevantes. Los modelos establecidos en base a longitudes de onda efectivas seleccionadas superaron a los que utilizan datos espectrales completos.

Espectros de 439 muestras de uva (293 muestras de uvas para la calibración y 146 muestras para la predicción) de tres variedades fueron analizadas por algoritmo genético (GA). Después, con los espectros obtenidos mediante un espectror radiómetro FielSpec Handheld Pro FR (325–1075 nm), se ejecutó la discriminación de variedades y la predicción del contenido de los componentes en función de la Máquina de Vectores de Soporte de Mínimos Cuadrados (LS-SVM).

Los resultados de GA indicaron las longitudes de onda más efectivas para la discriminación de variedades y la predicción de contenido: 636, 649, 693 y 732 nm. Las longitudes de onda efectivas de las muestras de uva indicadas por GA fueron 446, 489, 504, 561 nm para pH, y 418, 525, 556, 633, 643 nm para SSC.

Se logró una tasa de predicción correcta del 96,58% para la discriminación de variedades. Los valores RMSE fueron 0,9579 para GA-LS-SVM y 0,9252 para PLS midiendo SSC, mientras que para pH los valores RMSE fueron 0,1257 y 0,1487, respectivamente.

3.2.18 Artículo 18: *Soluble solids content and pH prediction and varieties discrimination of grapes based on visible–near infrared spectroscopy* [45]

Este estudio mide el potencial de la espectroscopia de fluorescencia frontal para la discriminación de las fechas de maduración de la uva *Cabernet Franc* de tres parcelas situadas en el Valle del Loira y durante seis fechas de maduración.

Los 18 lotes fueron analizados por espectroscopía de fluorescencia frontal y espectroscopia visible mediante un espectrofluorímetro *FluoroMax-2* (Spex-Jobin Yvon, Longjumeau, Francia) con espectros de excitación entre 250–310 nm.

A continuación, las bayas se dividieron en 4 lotes: el primero estaba compuesto por 20 bayas para espectroscopia de fluorescencia; el segundo estaba compuesto por 50 bayas para espectroscopia visible; el tercero estaba compuesto por 50 bayas para el análisis de compuestos fenólicos localizados en las pieles, y el último estaba compuesto por 50 bayas para análisis tradicionales (TSS y acidez total). Los espectros de uva se clasificaron por Análisis Factorial Discriminante (FDA).

Las fechas de maduración fueron predichas de forma correcta por ambos espectros: las longitudes de onda del espectro visible correspondientes a la absorción de **ANTOCIANINAS** coinciden con las longitudes de onda provenientes de la espectroscopía de fluorescencia frontal, alrededor de los puntos inicial y final (**263 Y 292NM**).

Después, se investigaron modelos de regresión PLS para predecir los sólidos solubles totales (TSS), acidez total, malvidina-3G, antocianinas totales y contenido de fenoles totales de visible y espectros de fluorescencia.

- Para predecir indicadores tecnológicos (TSS y acidez total), el modelo PLS con espectros visibles (RMSECV = 0.82◦Brix o 0.96 $g \cdot L^{-1} \cdot H_2SO_4$) fue mejor que aquellos con fluorescencia (RMSECV = 1.39◦Brix o 2.06 g L−1 H2SO4).
- Para malvidina-3G y antocianinas totales, todas R^2 fueron superiores a 0,90 y los valores de RMSECV fueron bajos.
- Para las espectroscopias visible y de fluorescencia se logró predecir el contenido de antocianinas de forma favorable.
- En cuanto al fenólico total, la mejor predicción la proporcionó la espectroscopia de fluorescencia.

3.2.19 Artículo 19: *The prediction of total anthocyanin concentration in red-grape homogenates using visible-near-infrared spectroscopy and Artificial Neural Networks* [46]

Este estudio compara el rendimiento del análisis de Regresión de Mínimos Cuadrados Parciales (PLS) y las Redes Neuronales Artificiales (ANN) para la predicción

de concentración total de antocianinas en uva roja a partir de sus espectros visible-infrarrojo cercano usando un espectrómetro FOSS NIRSystems 6500 Vis-NIR (Foss NIRSystems, Silver Springs, MD, EE. UU) operando entre 400 y 2500 nm. Se compara también la combinación de ambos métodos, ya que combina las ventajas de reducción de datos PLS con las capacidades de modelado no lineal de ANN.

Se obtuvieron 3134 muestras de uva de las añadas 1999-2003 como conjunto de calibración y 250 de la cosecha de 2004 como conjunto de prueba. Las muestras pertenecen a 9 variedades y 11 regiones del sur de Australia. Se detectaron picos fuertes en el rango espectral visible de **400 A 700 NM** asociados a compuestos de **ANTOCIANINA** y picos cerca de 990, 1200, 1450 y 1935 nm debidos principalmente a agua.

Se demostró con este método híbrido PLS-ANN que el modelo es más rápido y fácil de entrenar que el uso de datos de espectro completo sin procesar. Obtuvo índices de precisión comprendidos entre 81% y 90% según el número de componentes principales empleados. También fue más lineal y preciso que los modelos PLS globales y locales, reduciendo la necesidad de actualizar la calibración con muestras de nuevas añadas, requiriendo también menos entradas que utilizando PCA (error de predicción más alto es de 0,23 $mg \cdot g^{-1}$ para la regresión LOCAL, comparado con un valor de 0,16 $mg \cdot g^{-1}$ para PLS global o 0,18 $mg \cdot g^{-1}$ empleando el modelo híbrido).

3.2.20 *Artículo 20: Maturity, Variety and Origin Determination in White Grapes (Vitis Vinifera L.) Using near Infrared Reflectance Technology* [47]

El objetivo de este artículo es estudiar si la espectroscopia de reflectancia en el infrarrojo cercano (NIR 800-500 nm) medido por un refractómetro podría usarse para determinar el contenido de sólidos solubles (parámetro directamente relacionado con la madurez) e identificar diferentes variedades y orígenes de uvas.

Se realizó una Regresión de Mínimos Cuadrados Parciales (PLS) para calibrar los espectros NIR. Los resultados demuestran que la tecnología NIR es adecuada para determinar el contenido de sólidos solubles, aunque requiere un modelo de calibración para cada variedad.

Se recolectaron muestras de uva *Viura* y *Chardonnay* en dos localidades (Cadreita y Villamayor de Monjardin) con distintas condiciones ambientales.

Los modelos de regresión para uva *Chardonnay* fueron más robustos que para uva Viura, presentando coeficientes de determinación para calibración y validación de 0,75 y 0,70 y un error estándar de validación cruzada de 1,27. El modelo combinado de variedades mostró una desviación predictiva residual de 1,33, mientras que los modelos *Viura* y *Chardonnay* mostraron 1,54 y 1,88, respectivamente. Los resultados del análisis discriminante utilizando variables del espectro NIR mostraron un porcentaje de uvas bien clasificadas según la variedad a la que pertenecían del 97,2% y un 79,2% de uvas *Chardonnay* bien clasificadas según su origen.

3.3 Espectroscopía aplicada a objetivos de índole variada

Los siguientes artículos, aunque incluyen técnicas de espectroscopía de igual forma que los comentados en el punto anterior, no obedecen a un patrón claro de estudio, por lo que su tratamiento y análisis se realiza de forma independiente.

3.3.1 Artículo 1*: Grapevine variety identification using "Big Data" collected with miniaturized spectrometer combined with Support Vector Machines and Convolutional Neural Networks* [48]

El objetivo de este trabajo es identificar y diferenciar variedades de vino a partir de los datos brindados por un espectrómetro (60 - 1028 nm) y un análisis de Inteligencia Artificial. Un total de 35.833 espectros de hojas de 626 plantas de 64 variedades diferentes, todas ellas del año 2017, fueron recolectadas para este estudio.

Estos datos obtenidos en modo de reflectancia fueron preprocesados mediante SG, logaritmo, MSC, *Standard Normal Variate* (SNV), 1ª derivada y 2ª derivada, como muestra la Figura 32.

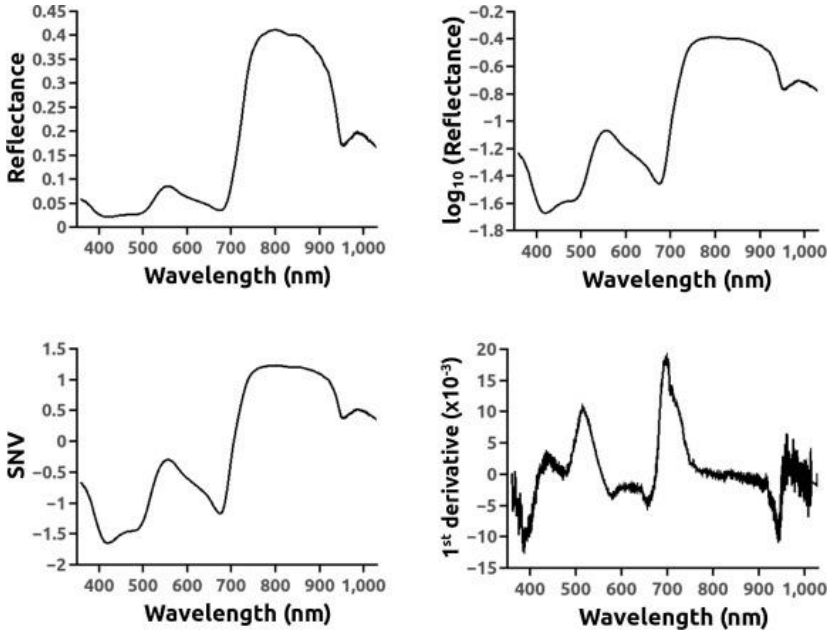

Figura 32. Ejemplo de preprocesamiento de espectro

Los datos preprocesados válidos para este estudio (dependiendo de la variedad emplea unos preprocesados u otros) se usaron para construir máquinas de vectores de soporte (SVM con *software* LIBSVM) y Redes Neuronales convolucionales (CNN entrenada empleando KERAS) para diferenciar *Touriga Nacional* (TN) de otras 63 variedades o *Touriga Franca* (TF) de 63 variedades; y analizar las eficiencias de clasificación obtenidas.

Se demuestra que es posible separar espectros de hojas de las variedades de vid TN o TF de espectros de otras 63 variedades al utilizar más de 35.000 espectros, aunque la eficiencia puede variar en función de la variedad. Se consiguen índices de precisión de clasificación de TN de 63,02% al emplear SVM y 81,90% al identificar el resto de espectros. Para TF se obtienen mejores resultados usando CNN, consiguiendo porcentajes de clasificación correcta del 91,63% al identificar otros espectros y 93,82% al diferenciar TF. El trabajo también ha demostrado que es posible recopilar grandes cantidades de datos en una cantidad de tiempo relativamente pequeña, es decir, 4 días.

3.3.2 Artículo 2: *Assessment of grapevine variety discrimination using stem hyperspectral data and AdaBoost of Random Weight Neural Networks* [49]

El siguiente estudio tiene su origen en la necesidad de diferenciar correctamente los tipos de plantas presentes en los cultivos de vid. Los datos espectroscópicos en modo de reflectancia se procesaron con una combinación de AdaBoost y Redes Neuronales de peso aleatorio (RWNN).

Se utilizan diez variedades de vid, 5 de tintos y 5 de blancos: Cabernet Sauvignon (CS) y otras nueve variedades portuguesas, incluyendo *Touriga Franca* y *Touriga Nacional* (TN). Se recogieron 30 muestras por día y por variedad, es decir, 300 muestras por día y 1200 en total, todas ellas pertenecientes al año 2016. A partir de dichas muestras, se adquirieron los espectros por medio de un espectrómetro portátil Flame-S de OceanOptics operando entre 350 y 1028 nm. Cada espectro corresponde a un promedio de 100 escaneos, lo que significó un tiempo de medición total por muestra de 4 sg.

Sobre dichos espectros obtenidos se formaron dos conjuntos diferentes de entrenamiento, uno con los 1200 datos originales, y un segundo con 37.200 instancias provenientes de crear 31 variantes de reflectancia a partir de cada espectro medido.

El porcentaje de correcta clasificación al separar las diez variedades varió entre 41,7 y 70,8%, dependiendo de la variedad. La tasa de falsos positivos (FPR) varió entre 2,5 y 7,3%. TN y CS se clasificaron correctamente con porcentajes del 70,8 y 70% únicamente con FPR de 5,3 y 3,3%, respectivamente. El mejor clasificador creado parece ser más apropiado para identificar una o dos variedades de entre las diez en vez de identificar todas las variedades simultáneamente debido al número de clasificaciones incorrectas.

Los resultados se compararon con *Random Forest* y SVM, pero la metodología propuesta en este artículo consiguió los mejores índices: **AdaBoost + RWNN**. La combinación permitió mejorar los porcentajes de correcta clasificación entre 5,3 y 10 puntos porcentuales en relación con el uso de RWNN. Emplear el segundo grupo de entrenamiento con muestras provenientes del espectro de reflectancia también mejoró los porcentajes de clasificación entre 1,6 y 6,7 puntos de porcentaje.

3.3.3 Artículo 3: *Identification of grapevine varieties using leaf spectroscopy and Partial Least Squares* [50]

Este artículo pretende clasificar variedades de vid a partir de espectroscopia foliar. El método consiste en un clasificador basado en Mínimos Cuadrados Parciales (PLS) que discrimina entre variedades de vid utilizando imágenes hiperespectrales (380 - 1028 nm) de una hoja medida en modo de reflectancia.

El clasificador fue creado con información proveniente de los espectros asociados a 300 hojas, 100 de cada una de las variedades *Vitis vinifera L*. provenientes de varias regiones de España: *Tempranillo*, *Garnacha* y *Cabernet Sauvignon*. En la Figura 33 aparece dibujado el espectro promedio proveniente de todas las muestras de hoja.

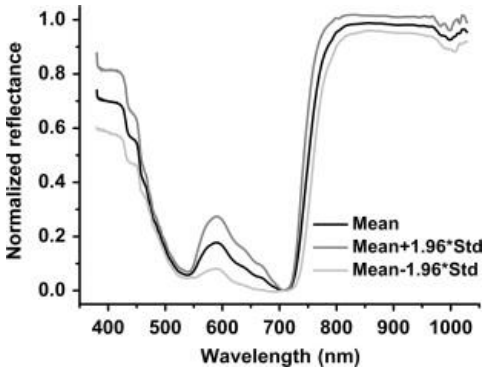

Figura 33. Espectro de reflectancia medio para todas las hojas utilizadas

La validación cruzada de Monte-Carlo (PLS+MCCV) confirmó el desempeño del clasificador para las 3 variedades, que superó el 92% en todos los casos. Se obtuvieron los mejores resultados para los conjuntos de validación con 13 componentes; 92,0%, 94,2% y 94,6% para *Tempranillo, Garnacha y Cabernet Sauvignon*, respectivamente.

3.3.4 Artículo 4: *Discrimination of varieties of red wines based on independent component analysis and BP Neural Network* [51]

En este estudio, se pretende diferenciar 5 variedades de vino tinto -*Dynasty, Louis Lanza, Gran Dragón, Shangeli·La* (cebada desnuda) y *Shangeli·La* (*Cabernet Sauvignon*)- empleando la técnica de espectroscopía *Visible/Near Infrared Spectroscopy* (Vis/NIR)

El número de muestras para cada variedad fue 35, formando un total de 175 muestras. Se seleccionaron aleatoriamente 150 muestras de vino tinto (30 muestras de cada variedad) para el conjunto de calibración, mientras que las 25 muestras restantes (5 de cada variedad) se separaron para el conjunto de validación.

Para cada muestra, 3 espectros en modo de absordancia fueron escaneados por un dispositivo espectral FieldSpec Pro-FR (325–1075 nm).

20 componentes independientes (IC) extraídos por Análisis de Componentes Independientes (ICA) se emplearon como entradas de la Red Neuronal de Retropropagación (BPNN) empleada en la calibración. Se compuso un modelo de 3 capas modelo con la función de transferencia sigmoidea. Por tanto, la matriz del conjunto de calibración se compuso de 150 muestras y 20 variables.

Se consiguieron tasas de reconocimiento del 100% y se seleccionaron dos bandas características para vinos tintos: 400-430 y 512-532 nm, como ilustra la Figura 34.

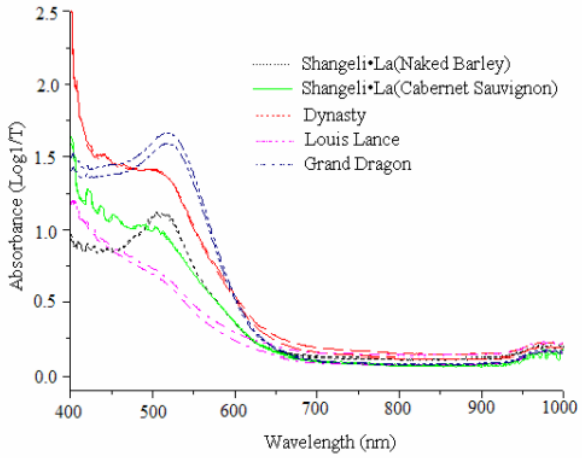

Figura 34. Espectros Vis/NIR de vinos tintos

Se puede concluir que la espectroscopia Vis/NIR se puede utilizar para identificar las variedades de vino tinto de forma rápida y precisa. ICA combinado con la Red Neuronal BP es una buena optimización de métodos tradicionales de reconocimiento de patrones, que puede mejorar la precisión del reconocimiento.

3.3.5 Artículo 5: *Discrimination of rice wine age using visible and near infrared spectroscopy combined with BP Neural Network* [52]

Se estudia clasificar vinos de arroz de diferentes añadas mediante espectroscopía de infrarrojo cercano y visible (Vis/NIR) combinado con métodos quimiométricos.

Se obtuvieron espectros de 240 muestras de vino *Kuaijishan* (80 para cada año, de 1, 3 y 5 años) en la región Vis/NIR (325-1075nm) en el espectro-radiómetro en modo de transmitancia. Se seleccionaron aleatoriamente 60 muestras de cada añada haciendo una suma de 180 muestras para el conjunto de entrenamiento, mientras que las 60 muestras restantes se utilizaron para el conjunto de validación.

Se aplicó el análisis de Mínimos Cuadrados Parciales (PLS) para extraer los componentes principales (PC) como nuevos vectores propios para representar la información de los espectros sin procesar. Luego, los 5 primeros PC se utilizaron como entradas de la Red Neuronal BP. Finalmente, se desarrolló un modelo de Redes Neuronales de BP de 4 capas. El modelo PLS + BPNN logró índices de discriminación del 96,67%.

En cuanto a regiones espectrales relevantes, presentes en la Figura 35, para el primer componente principal, PC1, fue alrededor de 379 nm, y PC4 alrededor de 375 nm. Esta región espectral podría estar determinada principalmente por pigmentos de color.

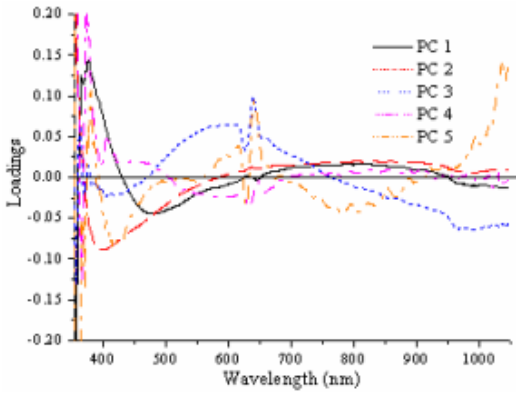

Figura 35. Información espectral de los cinco primeros componentes principales

También observamos en la Figura 35 cómo PC2 muestra una correlación a los 400 nm. PC3 y PC5 tuvieron una correlación positiva alta alrededor de 640 nm, y PC5 también una correlación positiva alta con la región alrededor de 1036 nm.

3.3.6 Artículo 6: *Detection of six kinds of acid in red wine with infrared spectroscopy based on FastICA and Neural Network* [53]

Para la detección rápida de 6 tipos de ácidos en el vino tinto, se analizaron los espectros infrarrojos (IR) de 44 muestras de vino *Cabernet Sauvignon* empleando un escáner de vino que usa un interferómetro Michelson para generar los espectros FT-IR. Se propone un método basado en regresión de Redes Neuronales Artificiales de Retropropagación (BP-ANN) y Análisis Rápido de Componentes Independientes (FastICA): ICA-NNR.

La Red Neuronal empleada, BP-ANN, consta de tres capas (capa de entrada, capa oculta y capa de salida) utilizando la función *tansig* como función de transferencia entre la capa de entrada y la capa oculta, y la función *purelin* entre la capa oculta y la capa de salida. La cantidad de neuronas en la capa de entrada es la cantidad de filas de la matriz, que está determinada por la cantidad de ICs, es decir, 6: acidez total, acidez volátil, ácido málico, ácido láctico, ácido cítrico y ácido tartárico. Los coeficientes de correlación (R) entre los valores de referencia y los valores predichos por el modelo son 0,9833, 0,9759, 0,9585, 0,9989, 0,9643 y 0,9884, respectivamente.

Comparado con métodos comunes, el método ICA-NNR tiene ventajas tanto en el coeficiente de correlación (resultados incluidos en el anterior párrafo) como en el error estándar de calibración, que en el peor de los casos es del orden de 0,081 como muestra la Tabla 22.

Items	PCR			PLS			ICA-NNR		
	PCs	*SEC*	*R*	Fs	*SEC*	*R*	ICs	*SEC*	*R*
Total acid	4	0.201	0.932	7	0.114	0.973	5	0.081	0.983
Volatile acid	6	0.156	0.926	9	0.138	0.955	5	0.028	0.976
Malic acid	6	0.175	0.915	10	0.035	0.962	6	0.055	0.959
Lactic acid	8	0.102	0.965	9	0.032	0.989	10	0.012	0.999
Citric acid	8	0.133	0.929	7	0.072	0.953	8	0.033	0.964
Tartaric acid	9	0.085	0.957	11	0.076	0.966	7	0.053	0.988

PCs—Number of principal components in PCR; Fs—Number of main factors in PLS.

Tabla 22. Comparación de los resultados de los modelos de calibración

3.3.7 Artículo 7: *Application of Least Squares Support Vector Machines for discrimination of red wine using visible and near infrared spectroscopy* [54]

Se utilizan métodos de espectroscopia y quimiometría -modo de transmitancia visible e infrarroja cercana (Vis/NIR) FieldSpec Pro-FR (325–1075 nm)- para discriminar variedades de vino tinto. Las 175 muestras de 5 variedades de vino tinto -*Shangri-La*, *Shangri-La Cabernet Sauvignonse*, *Dynasty*, *Louis Lance* y *Grand Dragon Cabernet Sauvignon*-se separaron aleatoriamente en un conjunto de calibración con 125 muestras (25 instancias/ variedad) y conjunto de validación con 50 muestras (10 instancias/variedad).

Los componentes principales (PC) del conjunto de calibración, seleccionados por PLS del espectro original (la banda espectral (400-604 nm) se utilizó para realizar el análisis PLS), se utilizaron como entradas para entrenar las Máquinas de Vectores de Soporte de Mínimos Cuadrados (LS- SVM).

Después, se logró en el conjunto de validación una tasa de reconocimiento del 94% al predecir las variedades en función de sus PC con error predictivo ±0,1; con umbral de error predictivo ±0,2 el reconocimiento fue del 100%. Los valores RMSEP y R^2 fueron 0,0531 y 0,9986 respectivamente; valores que muestran la eficiencia del modelo.

3.4 Comparativa de todos los artículos

3.4.1 Tabla Global. Resumen de los artículos de mediciones del nivel de azúcar, pH y concentración de antocianinas

A continuación, se expone una recopilación de todos los artículos incluidos en el tercer bloque asociado al estudio y medición de niveles de azúcar, pH y concentración de antocianina, variables que afectan a la calidad del vino. En primer lugar, aparece la Tabla 23, con los artículos que se centran en la predicción de variables enológicas. Después, la Tabla 24 recoge artículos con objetivos de clasificación, detección y discriminación de diversos factores. Todos los artículos presentes en las Tablas 23 y 24 emplean tanto técnicas de espectroscopía como de Inteligencia Artificial.

	OBJETIVO	TÉCNICA IA ESPECTROSCOPÍA	PRE PROCESADO	RMSE *root-mean-square error of predictions* AZÚCAR EN ◦BRIX	R/T/A	MUESTRAS TOTALES	RESULTADOS
3.2.2 ***2021***	Predicción del contenido de azúcar y pH	SG + 1D CNN Imágenes Hiperespectrales	SG	Azúcar Variedades: 1,118 Cosechas: 1,085 pH 0,199 y 0,183	R	*Touriga Franca: 1748* *Touriga Nacional:* 454 *Tinta Barroca:* 463	SG + 1D CNN ofrece muy buenos resultados para la predicción del contenido de azúcar, mejor que para pH, con un menor aumento RMSEP
3.2.3 ***2021***	Predicción del contenido de azúcar aplicando distintos métodos de reducción de dimensionalidad	*NNs y SVR + 10-fold cross validation* Imágenes Hiperespectrales (380 - 1028 nm)	-	Azúcar Entre 0,8826 y 3,6832 para diferentes añadas	R	Una sola variedad y pequeño número de bayas (6) por muestra *Touriga Franca*: 608	PCA supera al resto de técnicas (técnicas globales no lineales, Kernel PCA, M*ultilayer Autoencoders,* LLE y *Laplacian Eigenmaps*) NNs y SVM ofrecen buenos resultados. Ninguna destaca en términos globales
3.2.4 ***2021***	Predicción del contenido de azúcar con distintos métodos de Aprendizaje Automático	RR, PLS, NN y 1D CNN Imágenes Hiperespectrales (380 - 1028 nm)	SG	Azúcar TF = 7,87 - 30,26 TN = 6,95 - 29,66 TB = 5,48 - 29,95	R	Varias variedades y pequeño número de bayas por muestra *Touriga Franca*: 1748 *Touriga Nacional*: 454 *Tinta Barroca*: 463	1D CNN supera al resto de metodologías estudiadas, ofreciendo todos los métodos predicciones favorables
3.2.5 ***2021***	Compara 16 métodos de regresión lineal y no lineal para predecir el contenido de azúcar, pH y antocianina de las uvas	L&NL-PAC con 16 métodos distintos. Destaca RR Imágenes Hiperespectrales (380 - 1028 nm)	MCS SNV Normalización SG + 1ª derivada* SG + 2ª derivada	Azúcar 0,998 – 1,060 pH 0,168 – 0,175 Antocianinas 19,773 – 19,961	R	Una sola variedad y pequeño número de bayas (6) por muestra *Touriga Franca*: 240	L&NLPAC permitió la identificación de los mejores métodos de regresión, así como longitudes de onda más contribuyentes a la variabilidad de cada parámetro enológico
3.2.6 ***2019***	Medición del nivel de azúcar con espectroscopía a partir de niveles de antocianinas, flavonoides y sólidos solubles	Destaca PLS-DA Espectro-radiómetro Vis-NIR (350 - 2500 nm)	PLSR+SNV PCR+SNV	Antocianina *Syrah* 16,90 | 16,79 *Cabernet Sauvignon* 13,56 | 13,66 *General* 16,74 | 18,46	R	Una sola añada y pequeño número de bayas por muestra *Cabernet Sauvignon*: 576 *Syrah:* 432	Atributos de calidad PCR, PLSR y MLR ofrecen precisiones ≥ 90% para 2 atributos y ≥70% para el restante Tipo de vino PLS-DA obtiene el mejor índice de predicción para diferenciar las categorías de vino: 93,15 %
3.2.7 ***2018***	Predecir la concentración de antocianinas, pH y el contenido de azúcar en bayas	SVR + Núcleo de base radial gaussiana + *n-fold-cross-validation* Imágenes Hiperespectrales (380 – 1028 nm)	Normalizado	Azúcar 3,19 pH 0,25 Antocianinas 35,6 mg·L^{-1}	R	Variedad TF, cuatro añadas y pequeño número de bayas (6) por muestra *Touriga Franca*: 552 *Touriga Nacional*: 60 *Tinta Barroca*: 84	Los mejores resultados de R^2 obtenidos por el modelo fueron 0.89, 0.81 y 0,90, con valores RMSE de 35,6 mg·L^{-1}, 0,25 y 3,19 ◦Brix, para concentración de antocianinas, índice de pH y contenido de azúcar, respectivamente
3.2.8 ***2017***	Monitorizar la calidad de las uvas utilizando predicciones de contenido de azúcar basadas en imágenes hiperespectrales	PLS NN Imágenes Hiperespectrales (380 – 1028 nm)	Normalizado	Azúcar 2012 | 2013 0,94 | 1,34 0,96 | 1,36	R	Modelo con una añada y pequeño número de bayas por muestra. También probado con una añada diferente *Touriga Franca: 324*	PLSR y NNs muestran un rendimiento similar. Ambos métodos son capaces de predecir el contenido de azúcar
3.2.9 ***2017***	Estimación de pH y antocianinas mediante imágenes hiperespectrales y NN	ANN Imágenes Hiperespectrales (380 – 1028 nm)	-	*TF - TN – TB* pH 0,191 - 0,17 – 0,176 Antocianina 22.1 - 23.2 - 51,3 mg·L^{-1}	R	Modelo con 2 añadas y pequeño número de bayas (6) por muestra *Touriga Franca*: 240+81 *Touriga Nacional:* 60 *Tinta Barroca:* 84	Resultados favorables en cuanto a pH de todas las variedades y antocianinas TF y TN. Mala generalización para antocianinas TB
3.2.10 ***2016***	Analizar las bayas mediante espectroscopía y PLS	PLS Espectrómetro de matriz de diodos Vis-NIR (350 – 850 nm)	Original Normalizado SG SNV	Azúcar | pH 0,65 | 0,05 0,87 |0,05 0,65 | 0,05 1,83 | 0,08	R	Una añada + 3 variedades y gran nº de bayas por muestra *Cabernet Sauvignon: 79* *Cabernet Franc: 163* *Syrah: 74*	A partir de los datos entrenados se probó sobre los datos de la siguiente cosecha, obteniendo resultados favorables

3.2.11 ***2015***	Capacidad de imágenes hiperespectrales para predecir cambios en el contenido de antocianinas en uvas de vino durante la maduración	PLSR PLSR+SVM + K-fold-cross-validation Imág. Hiperesp. Con Espectrógrafo (900 – 1700 nm)	Original Savitzky–Golay (SG) SNV Multiplicative scatter correction (MSC) 1st derivate 2nd derivate SG (PLSR+SVM)	Antocinina 0,015 mg·g^{-1} 0,013 mg·g^{-1} 0,013 mg·g^{-1} 0,022 mg·g^{-1} 0,041 mg·g^{-1} 0,028 mg·g^{-1} 0,005 mg·g^{-1}	R	Modelo con una añada y gran número de bayas por muestra (50 bayas) 120 muestras	Los mejores resultados se obtuvieron empleando PLSR + SVM
3.2.12 ***2015***	Medición no destructiva y simultánea de varios parámetros enológicos utilizando datos hiperespectrales	NN + PCA + *7-fold-cross-validation* Imágenes Hiperespectrales (380 - 1028 nm)	Normalización	Azúcar = 0,95 pH = 0,18 Antocianina = 14 mg·L^{-1}	R	Modelo con una añada y pequeño número de bayas (6) por muestra *Touriga Franca*: 240	El conjunto de prueba (30 muestras) reveló valores RMSE de 0,18, 0,95 °Brix y 14 mg/l para pH, azúcares y contenido de antocianinas
3.2.13 ***2014***	Detección de parámetros de las uvas para medir la maduración	MLPS Imág. Hiperesp. sensor hiperesp. NIS (900 – 1700)	Original MSC SNV	Azúcar 1,37 1,610 1,890 pH = 0,180	R	Modelo con una añada y gran cantidad de bayas por muestra Uva roja: 99 Uva blanca:114	Los resultados obtenidos presentan un buen potencial para una detección rápida y económica de parámetros en uvas intactas
3.2.14 ***2014***	Medir el contenido de azúcar con dos NN distintas	NN + *n-fold-cross-validation* Imágenes Hiperespectrales (380 - 1028 nm)	-	Azúcar 1,09 1,185	R	Modelo entrenado con años diferentes 240 muestras (2012) 84 muestras (2013)	Ambas redes entrenadas presentan desempeños similares. Los valores RMSE de 2013 son ligeramente más altos
3.2.15 ***2011***	Espectroscopía NIR + MLPS para optimizar la calidad el proceso de maduración	MPLS LOCAL Espectro-fotómetro de matriz de diodos NIR (380 – 1700 nm)	1st derivate 2nd derivate 1st derivate 2nd derivate	Azúcar = 1,69 pH = 0,17 Azúcar = 1,32 pH = 0,15	R	Modelo con tres añadas - 2006,2007 y 2008- gran nº de bayas por muestra y 25 variedades de uva blanca y roja 363 muestras	MLSP ofrece buenos resultados de medición de parámetros relacionados con el azúcar (r^2 = 0,94 para los parámetros SSC y contenido de azúcar reductor) y una buena precisión para los parámetros relacionados con la acidez (r^2 entre 0,73 y 0,87) LOCAL destaca sobre MLSP
3.2.16 ***2011***	Estimación de la concentración de antocianina de la uva utilizando datos hiperespectrales	*AdaBoost* + PCA Imágenes Hiperespectrales (400 – 1000 nm)	-	Antocianina 14,6 – 20,2	R	Una añada (2009) y una variedad *Cabernet Sauvignon: 46*	Redes neuronales + *AdaBoost* ofrece grandes resultados frente a PLS. Los resultados no siempre mejoran cuando se agrega una nueva Red Neuronal empleando *AdaBoost*
3.2.17 ***2010***	Predicción del contenido de sólidos solubles (SSC) y el pH de las uvas	GA-LS-SVM PLS Espectro-radiómetro Vis-NIR (325 – 1075 nm)	-	pH 0,1257 0,1487	R	Modelo con 3 variedades y un pequeño número de bayas por muestra 439 muestras	Los modelos establecidos en base a longitudes de onda efectivas superaron a los que utilizan datos espectrales completos
3.2.18 ***2008***	Caracterización de la madurez de las uvas empleando espectroscopía visible vs de fluorescencia frontal	PLS Espectro-fluorímetro (250 – 310 nm)	-	Antocianina 1,51 mg·g^{-1}	R	Modelo con una variedad *Cabernet Franc: 170*	Las fechas de maduración fueron predichas de forma correcta por ambos espectros
3.2.19 ***2007***	Predicción de concentración de antocianinas	PLS PLS-ANN LOCAL Espectrómetro Vis-NIR (400 – 2500 nm)	SG	Antocianina 0,16 mg·g^{-1} 0,18 mg·g^{-1} 0,23 mg·g^{-1} (peor)	R	Modelo de varias añadas y 9 variedades de uva	Usando el modelo híbrido ANN+PLS se reduce la necesidad de calibrar el modelo al cambiar de añada al predecir la concentración de antocianinas, aun teniendo un error algo superior a PLS
3.2.20 ***2005***	Determinar el contenido de sólidos solubles e identificar diferentes variedades y orígenes de uvas mediante tecnología NIR	PLS Refractómetro NIR (800-500 nm)	SG (1st and 2nd derivate)	Sólidos solubles 0,65 y 1,09 1,270 – 2,160	R	Modelo de una añada para una y dos variedades y pequeño número de bayas por muestra	La tecnología NIR es adecuada para determinar el contenido de sólidos solubles, aunque requiere un modelo de calibración para cada variedad

Tabla 23. Tabla global artículos de espectroscopía aplicada a la predicción de variables enológicas

	OBJETIVO	MÉTODO IA	PRE PROCESADO	RMSE *root-mean-square error of predictions* AZÚCAR EN ◦BRIX	R/T	MUESTRAS TOTALES	RESULTADOS
3.3.1 2019	Identificar y diferenciar variedades de vino	SVM CNN	SG Logaritmo MSC SNV 1ª derivada 2ª derivada	-	R	Una sola añada 35.833 espectros de hojas de 626 plantas de 64 variedades	SVM TN: 63,02% Non-TN: 81,90% CNN TF: 93,82% Non-TF: 91,63%
3.3.2 2018	Diferenciar tipos de plantas presentes en los cultivos de vid	*AdaBoost* + RWNN *Random Forest* SVM	-	-	R	10 variedades, una añada 1200 muestras o 37200 muestras con variantes de reflectancia	El porcentaje de clasificación al separar las 10 variedades varió entre 41,7 y 70,8% *AdaBoost* + RWNN brindó mejores resultados Con 37200 muestras los resultados mejoran
3.3.3 2013	Clasificar variedades de vid a partir de espectroscopia foliar	PLS + MCCV	-	-	R	3 variedades de una añada 300 hojas	92,0%, 94,2% y 94,6% de correcta clasificación para *Tempranillo, Garnacha y Cabernet Sauvignon*
3.3.4 2008	Diferenciar 5 variedades de vino tinto	BPNN	-	-	T	Modelo de 5 variedades 175 muestras	ICA + BPNN es una optimización de métodos tradicionales de reconocimiento de patrones
3.3.5 2008	Clasificar vinos de arroz de diferentes añadas	PLS+BNNN	-	-	T	Modelo de una variedad y 3 añadas *Kuaijishan: 240 muestras*	PLS + BPNN lograron índices de discriminación del 96,67%. Se distinguen regiones espectrales relevantes para componentes principales
3.3.6 2008	Detección de 6 tipos de ácidos en el vino tinto	ICA-NNR (con BP-ANN)	-	-		Modelo de una variedad y añada *Cabernet Sauvignon: 44*	ICA-NNR tiene ventajas tanto en el coeficiente de correlación como en el error estándar de calibración frente PCR y PLS
3.3.7 2008	Discriminar variedades de vino tinto	LS- SVM	-	-	T	Modelo con 5 variedades y una añada 175 muestras	Se logró una tasa de reconocimiento del 94 % con un error predictivo ±0,1, y un 100% con umbral de error predictivo ±0,2

Tabla 24. Tabla global artículos de espectroscopía aplicada a distintos objetivos

3.5 Gráfico con algoritmos más relevantes

En el punto que acontece se adjuntan unos gráficos de las técnicas mencionadas en los artículos comprendidos en este tercer punto del estudio de Inteligencia Artificial aplicada en el sector vitícola.

El Gráfico 5 pretende listar los algoritmos de Aprendizaje Automático óptimos para predecir las variables enológicas que afectan de forma inmediata a la calidad del vino. La mayor parte de los estudios se centran en estimar los valores de uno o varios de los siguientes parámetros: contenido de azúcar, pH y concentración de antocianinas.

Este gráfico añade un punto al algoritmo denominado por los investigadores como mejor modelo de predicción. El eje de abscisas corresponde al número de veces en el que dicho algoritmo ha proporcionado los mejores resultados en un estudio dado. Además, se incluye una distinción en el eje de abscisas de los preprocesados que han requerido los espectros para lograr los resultados más favorables. De esta forma no sólo se obtiene una visión esquemática de los algoritmos que brindan resultados favorables, sino que también se adjunta información sobre los tratamientos que requerirá emplear técnicas de

espectroscopía para ello. El eje de ordenadas muestra todos los algoritmos que han logrado destacar en, al menos, uno de los documentos de análisis.

Dado que el tratamiento mediante algoritmos IA se realiza sobre los espectros obtenidos mediante espectroscopía, detectar buenas prácticas de combinación entre ambas ramas de trabajo es de gran interés para encontrar futuras líneas de investigación. Es por esta razón por la que se adjunta el Gráfico 6 mostrando de forma particular y visual cuales han sido las mejores prácticas de preprocesamiento para las técnicas destacables en la predicción de los atributos presentes en el vino. Como la Tabla 24 contiene artículos con objetivos de índole variada, no se realizan comparaciones entre los resultados obtenidos por dichos artículos.

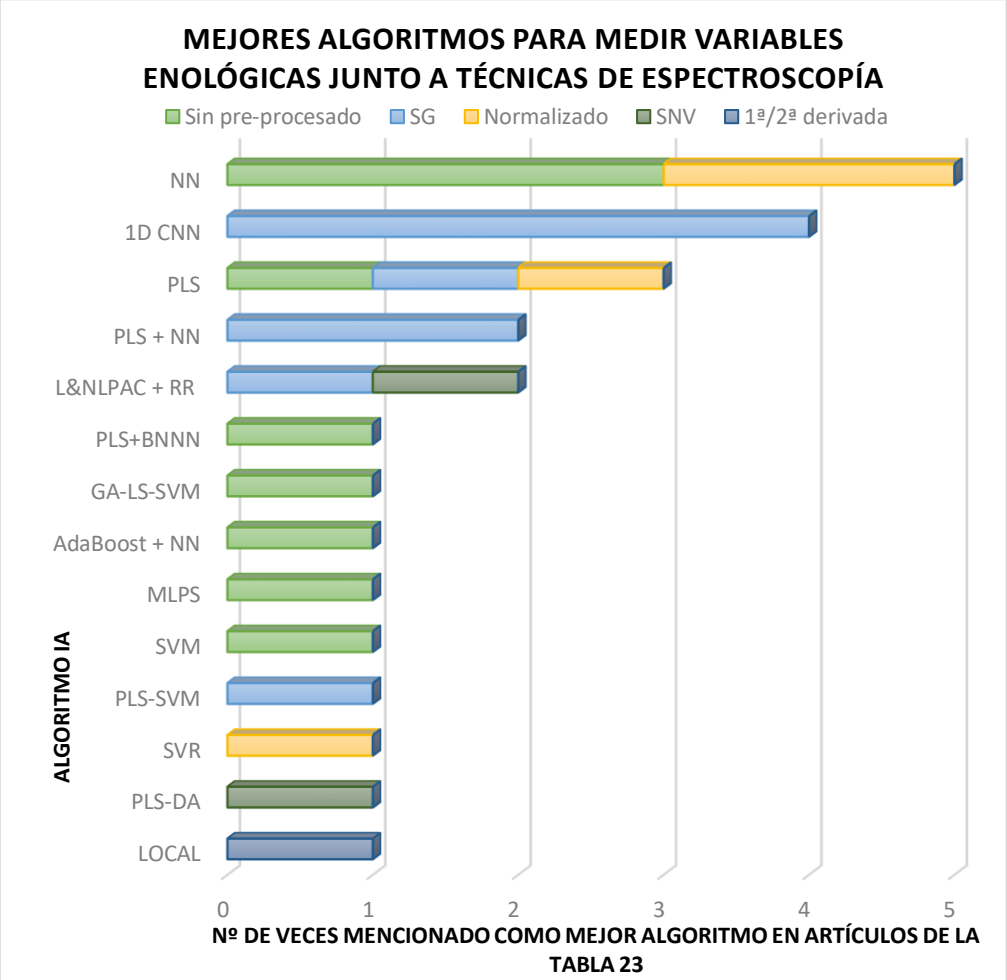

Gráfico 5. Algoritmos IA empleados junto a técnicas espectrales. Se distingue el tipo de preprocesamiento aplicado por cada algoritmo

LAS REDES NEURONALES Y PLS DESTACAN DE FORMA CLARA A LA HORA DE OBTENER MEDICIONES FAVORABLES SOBRE PARÁMETROS ENOLÓGICOS.

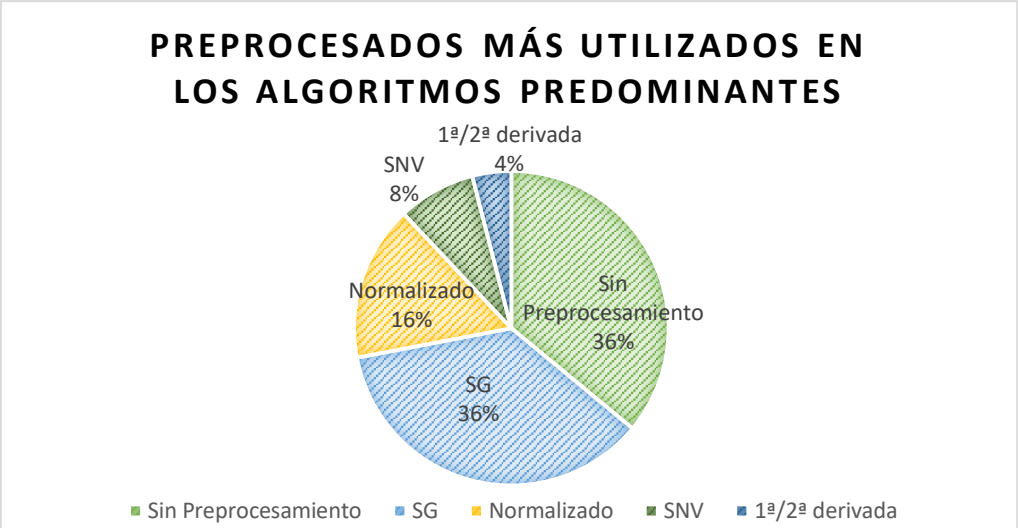

Gráfico 6. Preprocesados de espectros aplicados previa aplicación de los algoritmos de Inteligencia Artificial

LOS ALGORITMOS EMPLEADOS EN ESTA RECOPILACIÓN OBTIENEN LOS MEJORES RESULTADOS, GENERALMENTE, EMPLEANDO ESPECTROS SIN NINGÚN TIPO DE TRATAMIENTO, APLICANDO UN SUAVIZADO *SAVITZKY-GOLAY (SG)* O NORMALIZANDO.

3.6 Conclusiones

La espectroscopía, en especial las imágenes hiperespectrales, ha resultado ser una gran aliada a la hora de desempeñar la ardua labor de medir parámetros enológicos durante los procesos de elaboración del vino.

Los equipos más utilizados en el estado actual de la técnica operan entre 400 y 1000 nm (cubriendo rangos visibles y de infrarrojo cercano). Aunque la longitud de onda del infrarrojo cercano por encima de 1000 nm puede contener bandas de absorción proporcionando información adicional, los equipos capaces de operar en longitudes de onda por encima de 1000 nm son mucho más caros.

Las CÁMARAS HIPERESPECTRALES llevan integrados espectrómetros capaces de registrar información química de manera muy precisa y en tiempo real, así como dispositivos de iluminación que facilitan la adquisición de imágenes, evitando además requerir de múltiples dispositivos electrónicos. Estas facultades hacen de las cámaras hiperespectrales un gran utensilio para INCLUIR ESPECTROSCOPÍA en estudios, A UN PRECIO MÁS ASEQUIBLE.

En comparación con el análisis químico tradicional, los sistemas descritos en este bloque basados, principalmente, en imágenes hiperespectrales provenientes de cámaras hiperespectrales, mitigan los costos de adquisición de información y NO DESTRUYE LAS UVAS, pero crea la necesidad de utilizar modelos robustos que puedan extraer conocimiento y aprender de los patrones en los espectros.

Las técnicas hiperespectrales combinadas con el Aprendizaje Automático han cobrado importancia en los últimos años para medir los parámetros enológicos correlacionados con el estado de maduración de las uvas, realizar análisis y predicciones sin la necesidad de emplear el producto final: el vino.

4 Bases de Datos relevantes y su Importancia en los Estudios

Como hemos podido ver a lo largo de los puntos redactados hasta ahora, la importancia de usar una Base de Datos adecuada es crucial para obtener unos resultados útiles y favorables.

4.1 Base de Datos UC Irvine (UCI)

Con cierta recurrencia, la Base de Datos empleada para realizar las mediciones y predicciones de calidad es la **BBDD UCI (UC Irvine)** disponible online: http://archive.ics.uci.edu/ml/index.php.

Las siglas UCI corresponden a la Universidad de California en Urvine. El repositorio de Aprendizaje Automático de UCI es una colección de Bases de Datos -622 *datasets*- de múltiples categorías, accesibles y descargables por cualquier usuario para el análisis empírico de algoritmos de Aprendizaje Automático. Las principales características que hacen de esta Base de Datos un gran recurso para los investigadores se nombran a continuación:

- Su gran accesibilidad, ya que es información descargable en la web.
- De carácter gratuito y público.
- Consta de 2 variedades diferentes de vino: tinto y blanco.
- Cuenta con un número considerable de muestras ya calificadas según su calidad; a saber, 1599 instancias de vino tinto y 4898 muestras de vino blanco, con calidades numeradas de 0 a 10.
- Ambos conjuntos de datos contienen 11 variables fisicoquímicas - acidez fija, acidez volátil, ácido cítrico, azúcar residual, cloruros, dióxido de azufre libre, dióxido de azufre total, densidad, pH, sulfatos y alcohol-, lo que permite categorizar el desempeño de algoritmos IA en cuanto a identificación de patrones y variables influyentes que afectan a la calidad del vino.
- Su gran aceptación permite realizar comparaciones entre los estudios de diferentes autores de forma más significativa.

POR ESTAS RAZONES, LA BASE DE DATOS UCI HA SIDO LA ELEGIDA POR LA MAYOR PARTE DE AUTORES QUE HAN REALIZADO ESTUDIOS RELACIONADOS CON LA CALIDAD DEL VINO.

La carpeta descargable incluye diferentes documentos Excel, según el tipo de vino, así como diversos archivos de texto plano (de extensión *.txt*) con comentarios de los autores sobre el contenido y el uso de la ristra de datos. Observamos un ejemplo de muestras de vino tinto en la Figura 36.

	A	B	C	D	E	F	G	H	I	J	K	L
1	fixed acidity	volatile acid	citric acid	residual sug	chlorides	free sulfur d	total sulfur c	density	pH	sulphates	alcohol	quality
2	7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5
3	7.8	0.88	0	2.6	0.098	25	67	0.9968	3.2	0.68	9.8	5
4	7.8	0.76	0.04	2.3	0.092	15	54	0.997	3.26	0.65	9.8	5
5	11.2	0.28	0.56	1.9	0.075	17	60	0.998	3.16	0.58	9.8	6
6	7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5
7	7.4	0.66	0	1.8	0.075	13	40	0.9978	3.51	0.56	9.4	5
8	7.9	0.6	0.06	1.6	0.069	15	59	0.9964	3.3	0.46	9.4	5
9	7.3	0.65	0	1.2	0.065	15	21	0.9946	3.39	0.47	10	7
10	7.8	0.58	0.02	2	0.073	9	18	0.9968	3.36	0.57	9.5	7
11	7.5	0.5	0.36	6.1	0.071	17	102	0.9978	3.35	0.8	10.5	5
12	6.7	0.58	0.08	1.8	0.097	15	65	0.9959	3.28	0.54	9.2	5
13	7.5	0.5	0.36	6.1	0.071	17	102	0.9978	3.35	0.8	10.5	5
14	5.6	0.615	0	1.6	0.089	16	59	0.9943	3.58	0.52	9.9	5
15	7.8	0.61	0.29	1.6	0.114	9	29	0.9974	3.26	1.56	9.1	5
16	8.9	0.62	0.18	3.8	0.176	52	145	0.9986	3.16	0.88	9.2	5
17	8.9	0.62	0.19	3.9	0.17	51	148	0.9986	3.17	0.93	9.2	5
18	8.5	0.28	0.56	1.8	0.092	35	103	0.9969	3.3	0.75	10.5	7
19	8.1	0.56	0.28	1.7	0.368	16	56	0.9968	3.11	1.28	9.3	5
20	7.4	0.59	0.08	4.4	0.086	6	29	0.9974	3.38	0.5	9	4
21	7.9	0.32	0.51	1.8	0.341	17	56	0.9969	3.04	1.08	9.2	6
22	8.9	0.22	0.48	1.8	0.077	29	60	0.9968	3.39	0.53	9.4	6
23	7.6	0.39	0.31	2.3	0.082	23	71	0.9982	3.52	0.65	9.7	5
24	7.9	0.43	0.21	1.6	0.106	10	37	0.9966	3.17	0.91	9.5	5

winequality-red

Figura 36. Muestras de vino tinto pertenecientes a la BBDD UCI

La Figura 36 nos muestra los índices asociados a cada muestra de uva diferenciando los distintos atributos a estudiar, es decir, las 11 variables fisicoquímicas. Según estos valores y la opinión de expertos enólogos, cada muestra fue categorizada y corresponde a una calidad diferente. De igual forma ocurre con la BBDD asociada a las muestras de vino blanco.

La mayor parte de las muestras rondan en torno a *calidad* = 5 o 6, categoría denominada *media;* siendo clases minoritarias las muestras de calidad *baja* o *alta*. Es por esta razón por la cual muchos autores deciden realizar *preprocesamientos* que permitan contar con datos más uniformes y menos sesgados.

Como curiosidad, según [3], artículo que predice la calidad del vino haciendo uso de la BBDD UCI donde los autores examinan cuidadosamente los datos para asegurarse

de que no existe ninguna anomalía, se detectan 24 instancias duplicadas para el vino tinto y 937 instancias duplicadas para vino blanco. Sin embargo, no se consideran como anomalía porque todas las características y los valores de las etiquetas son exactamente iguales.

FUTUROS ESTUDIOS PODRÁN REALIZARSE SOBRE LA BASE DE DATOS UCI EMPLEANDO DIVERSOS ALGORITMOS NO UTILIZADOS SOBRE ESTA BBDD HASTA LA FECHA, CON EL FIN DE MEJORAR LOS ÍNDICES DE PREDICCIÓN OBTENIDOS EN LOS ESTUDIOS PRESENTES EN EL PUNTO 2 DE ESTE DOCUMENTO.

4.2 Otras Bases de Datos

Aunque la Base de Datos por excelencia ha sido la BBDD UCI, otros autores han optado por emplear Bases de Datos con otras características -recordemos que la BBDD UCI categoriza muestras según su calidad y variables fisicoquímicas, por lo que será inútil a la hora de realizar estudios de otra índole-, o bien formar sus propias Bases de Datos privadas para realizar sus estudios.

LA MAYOR PARTE DE LOS ESCRITOS QUE HAN REQUERIDO EL USO DE UNA BASE DE DATOS ESPECÍFICA CREADA POR LOS AUTORES O ENÓLOGOS ASOCIADOS AL STUDIO, TIENE CARÁCTER PRIVADO Y NO ES ACCESSIBLE.

Uno de estos casos corresponde a los artículos que incluyen detección mediante sensores MOS presentes en las E-Nose, o todos aquellos que incluyen espectroscopía, ya que utilizan métodos específicos de obtención de imágenes hiperespectrales.

5 Conclusiones Globales y Líneas Futuras

5.1 Conclusiones Globales

Este Trabajo Fin de Grado analiza principalmente la aplicación de técnicas de Inteligencia Artificial, mayoritariamente *Machine Learning*, como método de mejora en los procesos de fabricación del vino, medición de parámetros enológicos o predicción de factores relevantes en las fases de elaboración de esta bebida alcohólica con tanta importancia a nivel nacional y mundial.

En primer lugar, se estudian modelos que incluyen Aprendizaje Automático como medio para predecir y/o mejorar la calidad del vino. Los artículos desarrollados hasta la fecha se centran en aplicar algoritmos IA para encontrar variables relevantes que afecten a la calidad del brebaje y/o entrenar algoritmos que ofrezcan índices de exactitud favorables para realizar predicciones sobre los ámbitos deseados. Por tanto, el éxito de dichos dos objetivos recae principalmente en:

- Encontrar algoritmos cuyas tasas de **PREDICCIÓN** alcancen valores próximos al **100%** con mínima tasa de error. Depende en gran medida de la Base de Datos.
- Emplear técnicas de **SELECCIÓN DE VARIABLES** que mejoren las estadísticas propias del algoritmo empleado.
- Utilizar **BASES DE DATOS** con número adecuado de instancias que permita lograr buenas cifras en fases de entrenamiento.

Lograr dichos objetivos no es trivial. El gran número de algoritmos desarrollados y las posibles combinaciones entre ellos puede suponer un impedimento. Además, tratándose de un sector influyente a nivel económico, muchos de los estudios son privados y las Bases de Datos disponibles son pocas.

El segundo bloque de este documento analiza investigaciones que incluyen espectroscopía como técnica añadida para obtener espectros de las uvas y recopilar así información de forma no destructiva y sin requerir el producto final ni su malgaste.

A pesar de que, empleando espectroscopía, la Base de Datos tendrá las características deseadas por el investigador y no hay más impedimento que la propia adquisición de imágenes con equipo especializado (lo cual implica altos costes y previa organización por parte de los impulsores del estudio), también es un ámbito limitado por la inmensidad de modelos disponibles. Puede ocurrir que el preprocesado de los espectros obtenidos no sea adecuado y las predicciones no alcancen tasas aceptables, o bien que la técnica de reducción de dimensionalidad no aporte datos con los que el algoritmo IA pueda trabajar, o incluso que técnicas de selección de características y componentes principales no brinden resultados coherentes una vez aplicado el algoritmo ML.

Estas limitaciones han provocado que los estudios realizados sobre este sector no adquiriesen índices de precisión demasiado elevados, llegando incluso a desestimar algunas metodologías. A medida que avanzamos en el tiempo, contando con una mayor globalización de los datos y mejoras sustanciales en el campo de la Inteligencia Artificial y las tecnologías, los resultados obtenidos por los investigadores fueron mejorando, dicha información se toma como base para estudios futuros, los cuales consiguen mejorar las tasas de los artículos antecesores hasta el punto de alcanzar índices de precisión de clasificación del 100% en estudios recientes empleando AdaBoost [1] tanto con selección de características como sin aplicar ninguna técnica previa. Un claro ejemplo de este hecho son los Árboles de Decisión, *Decision Tree*, cuyos resultados no superaban índices de precisión muy superiores al 60% en los artículos más antiguos presentes en este documento empleando, por ejemplo, C4.5 [21]. En cambio, recientes estudios muestran éxitos prometedores empleando el algoritmo J48 [2] tanto para tareas relacionadas con la calidad, como combinadas con técnicas de espectroscopía.

Por tanto, a medida que avanzan los tiempos y se obtienen mejoras de los modelos previos sobre los cuales basamos nuestros estudios, mejoramos de forma sustancial también la forma de aplicar dicho conocimiento y de transmitir dicha información, aspecto clave para un desarrollo sostenible y exponencial hacia técnicas 100% efectivas ya sea en el sector vitícola como en cualquier otro sector.

5.2 Líneas Futuras

Enlazando con las conclusiones desarrolladas en el punto anterior, los grandes avances en la investigación dependen fundamentalmente de la transmisión del conocimiento entre expertos.

Basándose en las tendencias actuales y en las mejoras sustanciales de los resultados que ofrecen los algoritmos se pueden desarrollar modelos cuyo principal objetivo sea mejorar las tasas ya obtenidas, ya sea empleando algoritmos punteros en el campo de la Inteligencia Artificial, o bien **VARIANDO FACTORES INFLUYENTES** en el desempeño del propio algoritmo. Son muchas las posibilidades viables disponibles, como puede ser la implementación o no implementación de técnicas de selección de características, por ejemplo, PCA, emplear funciones de activaciones diferentes en el caso de las Redes Neuronales o variar el *kernel* si se utiliza SVM para encontrar hiperplanos más adecuados, incluso alterar el número de muestras presentes en los conjuntos de entrenamiento y de prueba.

Además, muchas líneas de investigación optan por buscar **MODELOS HÍBRIDOS** que combinen las ventajas de distintos algoritmos o bien compensando los inconvenientes que presente alguno de ellos.

Son infinitas las oportunidades que la Inteligencia Artificial aplicada a la enología nos brinda, pero para poderlas llevar a cabo es esencial contar con la ayuda de los enólogos y las bodegas. Muchas bodegas desconocen las posibilidades que la Inteligencia Artificial puede aportar a su campo y, por ello, son reacios a implantar metodologías IA o incluso a publicar información sobre sus viñedos. A día de hoy es necesario ampliar la variedad de **BASES DE DATOS** sólidas que permitan realizar distintos experimentos hallando múltiples patrones para distintos objetivos, y dicho cometido puede llevarse a cabo únicamente con la ayuda de expertos de la vid.

Las tendencias futuras de aplicación de metodologías y técnicas IA en un campo completamente transversal como es la enología residen en la **UNIÓN**: fusión de algoritmos, enlace entre expertos de sectores diferentes y un vínculo fuerte entre los investigadores de una misma línea de trabajo.

6 Sumario de Algoritmos

AdaBoost, *Adaptive Boosting*: Meta-algoritmo de clasificación estadística. Se puede utilizar junto otros algoritmos de aprendizaje para mejorar el rendimiento. La salida de los otros algoritmos se combina en una suma ponderada que representa la salida final del clasificador potenciado [55].

Análisis Discriminante: Modelo en el que las variables puedan predecir el grupo más adecuado en el que se debe incluir una muestra. El análisis discriminante es usado tanto para determinar las variables que discriminan entre dos o más grupos constituidos de forma natural, como para determinar las variables que contribuyen a la mejor predicción entre grupos [82].

- **Lineal (LDA) / Lineal Paso a Paso (SLDA)**
  - **Distancia de Mahalanobis:** Distancia que determina la similitud entre dos variables aleatorias multidimensionales [57].
- **Factorial (FDA)**
- **Cuadrático (QDA)**

Correlación de Pearson -a partir del coeficiente de correlación de Pearson-: Índice que mide el grado de relación de dos variables siempre y cuando ambas sean cuantitativas y continuas.

DTC o DT, *Decision Tree Clasifier*: Un árbol de decisión es un algoritmo de aprendizaje supervisado no paramétrico, que se utiliza tanto para tareas de clasificación como de regresión. Tiene una estructura de árbol jerárquica, que consta de un nodo raíz, ramas, nodos internos y nodos hoja [58].

- ***Boosted* C5.0:** El algoritmo C5.0 tiene un método especial para mejorar su tasa de precisión, llamado impulso o boosting. Funciona construyendo múltiples modelos en una secuencia.
- **C4.5:** Puede utilizar la ganancia de información o las proporciones de ganancia para evaluar los puntos de división dentro de los árboles de decisión.
  - **J48:** Algoritmo de clasificación basado en el algoritmo C4.5, empleado para generar árboles de decisión con base en el grupo de datos de entrenamiento [8].
- **CART:** Abreviatura de "árboles de clasificación y regresión". Utiliza la impureza de Gini para identificar el atributo ideal para la división. La impureza de Gini mide la frecuencia con la que se clasifica incorrectamente un atributo elegido al azar. Cuando se evalúa usando la impureza de Gini, un valor más bajo es más ideal.
- ***Extra Trees* o *Extremely Randomized Trees*:** Proceso extremadamente aleatorio en el que se crean múltiples árboles de decisión al azar y luego se combina los resultados de cada árbol para encontrar la respuesta final.
- **ID3, *Iterative Dichotomiser* 3:** Este algoritmo aprovecha la entropía y la ganancia de información como métricas para evaluar las divisiones de candidatos [58].
- **GBDT, *Gradient Boosting Decision Tree*:** técnica de Aprendizaje Automático para optimizar el valor predictivo de un modelo a través de pasos sucesivos en el proceso de aprendizaje [59].
- **PART:** Evita la generalización precipitada, y usa los mismos mecanismos que el C4.5 para construir un árbol.
- **RPART:** Clasificador Aprendizaje Automático [58].

***Fuzzy Logic* o Lógica Difusa):** Lógica matemática que permite tomar decisiones de distinta intensidad en función de grados intermedios de cumplimiento de una premisa.

- ***Fuzzy Cognitive Maps*:** Estructuras de gráficos difusos para representar el razonamiento causal [60].
  - **Entrenados con *Non Linear Hebbian Learning***
- **FMCDM *-Fuzzy Logic Multi Criteria Decision Making-*:** Modelo que realiza selección de variables y ofrece resultados sobre dichas variables [15].
- **FIR*, Fuzzy Inductive Reasoning*:** Modelo cualitativo, no paramétrico y superficial basado en lógica difusa. Tiene la capacidad de describir sistemas que no pueden ser descritos fácilmente por métodos clásicos de matemáticas o estadística [22].
- **FIS, *Fuzzy Inference System*:** Sistema que utiliza funciones de pertenencia difusa *-fuzzy membership functions-* para tomar una decisión [61].
- **GFS:** *Genetic-Fuzzy Systems* es un sistema difuso aumentado por un proceso de aprendizaje basado en computación evolutiva, que incluye algoritmos genéticos, programación genética y estrategia evolutiva.
  - **GFS-GPG-R:** Sistema basado en operadores gramaticales de programación genética.
  - **MOGUL-IRLHC-R:** Enfoque iterativo de aprendizaje de reglas que utiliza el paradigma MOGUL, pero en este caso el objetivo es aprender bases restringidas de conocimiento de tipo Mamdani -método basado en mín.-máx.- aproximadas a partir de ejemplos.
  - **MOGUL-TSK-R:** Metodología para la obtención de sistemas genéticos basados en reglas difusas bajo el enfoque de aprendizaje de reglas iterativas [22].

GA, *Genetic Algorithm* o Algoritmo Genético: Algoritmos de optimización, búsqueda y aprendizaje inspirados en los procesos de evolución natural y evolución genética, con el fin de resolver problemas de optimización [62].

***Gradient Boosting*:** Técnica para el análisis de la regresión y problemas de clasificación estadística, a partir de un modelo predictivo, típicamente, árboles de decisión. Construye el modelo de forma escalonada y los generaliza permitiendo la optimización arbitraria de una función de pérdida diferenciable [63].

Generalización apilada/acumulada: Tipo de meta-aprendizaje debido a la transferencia de información de nivel base al meta-nivel. La capacidad de aprender de los aprendices base (*base-learners*) ayuda a reforzar las fortalezas y debilidades de cada aprendiz base para una mayor precisión en el metanivel [17].

ICA, Análisis de componentes independientes: Método computacional que sirve para separar una señal multivariante en subcomponentes aditivos suponiendo que la señal de origen tiene una independencia estadística y es no-Gausiana [64].

- **FastICA**

Jack-*Knife*: Técnica para corregir el sesgo de estimación.

JRip: Algoritmo de Aprendizaje Automático que identifica las clases creando un conjunto de reglas [65].

KNN *-K-closest neighbour* o *Kth nearest neighbour method-*: Clasificador de aprendizaje supervisado no paramétrico, que utiliza la proximidad para hacer clasificaciones o predicciones sobre la agrupación de un punto de datos individual [66].

KStar: Clasificador basado en variables determinadas por alguna función de similitud. Se diferencia de otros aprendizajes en que usa una función de distancia basada en entropía [10].

L&NL-PAC: Marco de análisis predictivo lineal y no lineal para evaluar el rendimiento de predicción de diferentes métodos de regresión, cubriendo los principales métodos de Aprendizaje Automático que están completamente integrados con los enfoques de preprocesamiento espectral más comunes [32].

***Laplacian Eigenmaps*:** Algoritmo de reducción de la dimensión no-lineal basado en la matriz Laplaciana.

***Levenberg-Marquardt*:** Algoritmo matemático que resuelve problemas de mínimos cuadrados no lineales.

LightGBM: Algoritmo de gradiente rápido, efectivo para resolver tareas de clasificación y regresión que ocupa menos memoria y obtiene una mejor predicción que XGBoost [5].

LLE, *Local Linear Embedding*: Método de búsqueda de la proyección de menor dimensión de los datos que preserve las distancias dentro de los vecindarios locales [67].

MF-DCCA, *Multifractal Detrended Cross-Correlation Analysis*: Algoritmo que investiga las relaciones dinámicas entre las variables de las muestras [5].

MSE, Error cuadrático medio: Criterio de evaluación más usado para problemas de regresión.

***Naïve Bayes*:** Algoritmo clasificador probabilístico simple con fuerte suposición de independencia. Maneja valores continuos y discretos para hacer predicciones probabilísticas. Es altamente escalable e insensible a las características irrelevantes. Pero tiene una limitación, si el conjunto de entrenamiento es demasiado grande, otros clasificadores que apliquen validación cruzada de forma recurrente sobre el conjunto de datos ofrecerán mejores resultados [10] [68].

- ***Gaussian Naïve Bayes***

PCA, *Principal Component Analysis*: Técnica utilizada para describir un conjunto de datos en términos de nuevas variables («componentes») no correlacionadas. Los componentes se ordenan por la cantidad de varianza original que describen [69].

- **PCA-CG, algoritmo Guterman–Boger:** Estima el número de neuronas en la capa oculta además de evitar caer en mínimos locales.
- **Kernel PCA**
- **Multivía PCA**

***Percentage Split*:** Tipo de test WEKA que permite evaluar la calidad del clasificador según lo bien que clasifique un porcentaje de los datos.

RF, *Random Forest*: Algoritmo de Aprendizaje Automático que combina la salida de múltiples árboles de decisión para llegar a un único resultado. Su facilidad de uso y flexibilidad han impulsado su adopción, ya que maneja problemas de clasificación y regresión [70].

RFE, *Recursive Feature Elimination*: Técnica de reducción de variables que minimiza la complejidad del modelo eliminando características una por una hasta que queda la cantidad óptima de características [71].

Redes neuronales (NN o ANN)

- **AlexNet**
- ***Autoencoders / Multilayer Autoencoders*:** Aprender codificaciones eficientes de datos no etiquetados (aprendizaje no supervisado) [72].
- **BFGS *Quasi-Newton***
- **CNN, Convolucionales**
  - **1D CNN:** Empleadas en texto y señales de una sola dimensión.
  - **MCNN, *CNN-based multitask model***
- **RWNN, de peso aleatorio**
- **BPNN o BP-ANN, de retropropagación**
  - **MBPNN, *Multitask model based on BPNN***
  - ***Resilent Backpropagation* o Retropropagación Resistente**
- **GoogLeNet**
- **Inception v3**
- **INNN, *Intervals' Number, IN – NN***
- **Residuales**
  - **ResNet-50,**
  - **ResNet-101**
- **SqueezeNet**
- **DNN, *Deep Neural Model*, Redes Neuronales Profundas**
- **MLP, *Multi-layer Neural Network o Multilayer Perceptron*:** También llamado Perceptrón Multicapa, es una Red Neuronal Artificial formada por múltiples capas.

Regression o Regresión

- **LR, Logistic Regression:** El algoritmo de regresión logística tiene como principal aplicación los problemas de clasificación binaria.
- **MLR, Lineal/Multiple Regression:** Modelo estadístico versátil para evaluar las relaciones entre un destino continuo y los predictores [73].
- **PLS o PLSR, Regresión de Mínimos Cuadrados Parciales:** Método quimiométrico cuyo objetivo es predecir un conjunto de variables dependientes, Y, a partir de un conjunto de variables independientes X [35].
  - **PLS de intervalo (i-PLS)**
  - **MPLS, Multivía PLS o Regresión de Mínimos Cuadrados Modificada**
- **LOCAL:** Algoritmo predictivo que opera buscando y seleccionando de una biblioteca (basada en el conjunto de entrenamiento) las muestras más similares espectralmente a la muestra que se va a predecir [42].
- **RR, *Ridge Regression:*** Método de estimación de los coeficientes de modelos de regresión múltiple en escenarios donde las variables independientes están altamente correlacionadas [74].

SA, *Simulated Annealing*: También llamado Recocido Simulado, es un algoritmo de búsqueda metaheurística para problemas de optimización global; con objetivo encontrar una buena aproximación al valor óptimo de una función en un espacio de búsqueda grande [75].

SGDC, *Stochastic Gradient Decision Classifier*: Algoritmo de optimización y clasificación utilizado para encontrar los valores de los parámetros de una función que minimice la función de coste.

SIMCA, *Soft Independent Modelling of Class Analogy*: Técnica de clasificación.

SMOTE: Método de sobremuestreo para abordar problemas de desequilibrio de datos volviendo a muestrear el espacio de datos para crear puntos sintéticos de la clase menos dominante.

SOM, Mapas Autoorganizados: Algoritmo para clasificar las observaciones por medio de gráficos organizados.

- **Con corrección de Laplace**
- **De Kohonen**
- **GNB, *Gaussian Naïve Bayes***

SVM, *Support Vector Machine:* Algoritmo de aprendizaje supervisado utilizado en muchos problemas de clasificación y regresión. El objetivo es encontrar un hiperplano que separe de la mejor forma posible dos clases diferentes de puntos de datos [76].

- **SVM de intervalo**
- **LS-SVM, Máquina de Vectores de Soporte de Mínimos Cuadrados**

Validación cruzada *k-fold*: Las técnicas de validación cruzada dejando uno afuera o k iteraciones son muy populares para evaluar el rendimiento de los algoritmos de clasificación [77].

XGB

- **XGBoost:** Algoritmo de Aprendizaje Automático basado en un árbol de decisiones y utiliza un marco de potenciación de gradientes.

Preprocesamientos Imágenes Hiperespectrales [48]

MSC, *Multiplicative Scattering Correction:* MSC tiene como objetivo **eliminar** los efectos de **dispersión aditivos y multiplicativos no deseados** de los datos. Esto implica ajustar cada espectro de reflectancia individual, R_j, al espectro de datos de entrenamiento promedio, R_{mean}, y luego corregir el espectro individual con los coeficientes de ajuste, b_0 y b_1, siguiendo las ecuaciones:

$$\boldsymbol{R_j = b_0 + b_1 R_{mean} + error} \qquad\qquad \boldsymbol{MSC_j = \frac{R_j - b_0}{b_1}}$$

donde MSC_j es la corrección de dispersión multiplicativa para cada espectro.

Normalización

SG, *Savitzky-Golay:* El filtro SG consiste en reemplazar cada punto del espectro por el ajuste de mínimos cuadrados con un polinomio, en este caso de primer orden, de los puntos circundantes. **Se aplica para suavizar los espectros.** Esto es equivalente a reemplazar cada punto (R_i) por una combinación lineal (SG_i) de sí mismo y sus puntos circundantes a través de la ecuación:

$$SG_i = \sum_n C_n R_n$$

donde los Cn son coeficientes que proporcionan el ajuste de mínimos cuadrados polinómicos. El índice *n* incluye el punto *i* y algún número predefinido de puntos circundantes.

SNV, *Standard Normal Variate:* El SNV de un espectro, SNV_j, se obtiene restando de cada espectro de muestra, R_j, su valor medio, $mean_j$, y dividiendo el resultado por su desviación estándar, std_j, según la ecuación:

$$\boldsymbol{SNV_j = \frac{R_j - mean_j}{std_j}}$$

Es similar a MSC, SNV no usa información de un espectro de datos de entrenamiento promedio.

La derivada de primer orden, R(1), tiene la capacidad de eliminar los efectos aditivos, mientras que la derivada de segundo orden, R(2), elimina los efectos tanto aditivos como multiplicativos.

7 Bibliografía

[1] P. T. P. O. K. P. W. a. K. D. Bhardwaj, «A machine learning application in wine quality prediction,» Machine Learning with Applications, vol. 8, nº 100261, 2022.

[2] P. R. J. M. e. a. Koranga M, «Analysis of white wineusing machine learning algorithms,» Mater Today Proc. , vol. 46, nº 11087-11093, 2021.

[3] C.-W. W. C.-H. C. Terry Hui-Ye Chiu, «A Hybrid Wine Classification Model for Quality,» p. pp. 430–438, 2021.

[4] R. P. S. P. K. S. B. Kushalatha.M.R., «MACHINE LEARNING APPROACH FOR ATTRIBUTE IDENTIFICATION AND QUALITY PREDICTION OF RED WINE,» International Journal of Creative Research Thoughts (IJCRT), vol. 9, nº 2320-2882, 2021.

[5] K. L. G.-z. J. Chao Ye, «A new red wine prediction framework using machine learning,» 2020.

[6] A. K. S. B. C. Bipul Shaw, «Wine Quality Analysis Using Machine,» 2020.

[7] K. A. a. N. M. S. Kumar, «Red Wine Quality Prediction Using Machine Learning Techniques,» de International Conference on Computer Communication and Informatics (ICCCI -2020), Coimbatore, INDIA, 2020.

[8] S. G. A. O. D. C.-G. Luisa F. Galeano-Arias, «Análisis de calidad del vino por medio de técnicas de,» Información Tecnológica, vol. 32, nº 1, pp. 17-26, 2021.

[9] C. S. W. X. Shuhao Zhang, «Research on Red Wine Quality Based on Data Visualization,» de 2020 3rd International Conference on Artificial Intelligence and Big Data (ICAIBD), 2020.

[10] S. P, «Wine Quality Prediction Using Data Mining,» de 2019 1st International Conference on Advanced Technologies in Intelligent Control, Environment, Computing & Communication Engineering (ICATIECE), 2019.

[11] D.-K. K. Garima Agrawal, «Wine Quality Classification with Multilayer Perceptron,» International Journal of Internet, Broadcasting and Communication, vol. 10, nº 2, pp. 25-30, 2018.

[12] «Prediction of Quality for Different Type of Wine based on Different Feature Sets Using Supervised Machine Learning Techniques,» ICACT Transactions on Advanced Communications Technology (TACT), vol. 7, nº 3, pp. 1122-1127, 2018.

[13] A. A. A.-A. K. L. H. ,. J. T. L. M. S. Satyabrata Aich, «A Classification Approach with Different Feature Sets to Predict the Quality of Different Types of Wine using Machine Learning Techniques,» International Conference on Advanced Communications Technology(ICACT), pp. pp. 1-2, 2018.

[14] A. T. a. R. Sehrawat, «Wine Quality Detection through Machine Learning Algorithms,» de 2018 International Conference on Recent Innovations in Electrical, Electronics & Communication Engineering (ICRIEECE), 2018.

[15] C. S. K. A. T. B. I. P. S. K. Y. K. Sofoklis Petropoulos, «Fuzzy logic tool for wine quality classification,» Agriculture, vol. 142, nº Part B, pp. 552-562, 2017.

[16] Y. Gupta, «Selection of important features and predicting wine quality using machine learning techniques,» Procedia Computer Science, vol. 125, pp. 305-312, 2017.

[17] D. M. Taylor Larkin, «Statistical Analysis and Data Mining: The ASA Data Science Journal,» Statistical Analysis and Data Mining: The ASA Data Science Journal, vol. 13, nº 5, pp. 451-464, 2017 (published 2020).

[18] T. X. F. M. Gongzhu Hu, «Classification of Wine Quality with Imbalanced Data,» 2016 IEEE International Conference on Industrial Technology (ICIT), pp. 1712-1217, 2016.

[19] A. A. Yesim Er, «The Classification of White Wine and Red Wine According to Their Physicochemical Qualities,» Advanced Technology & Science 2013, International Journal of Intelligent Systems and Applications in Engineering (IJISAE), vol. 4, pp. 23-26, 2016.

[20] K. B. a. P. P. G. V. P. Groumpos, «A new mathematical modelling approach for viticulture and winemaking using fuzzy cognitive maps,» 2016 ELEKTRO, pp. 57-61, 2016.

[21] J. P. a. K. K. S. Lee, «Assessing wine quality using a decision tree,» 2015 IEEE International Symposium on Systems Engineering (ISSE), pp. 176-178, 2015.

[22] M. F. E. A. Nebot À., «Modeling Wine Preferences from Physicochemical Properties using Fuzzy Techniques,» de In Proceedings of the 5th International Conference on Simulation and Modeling Methodologies, Technologies and Applications (SIMULTECH-2015),, 2015.

[23] A. M. N. R. F. J. A. M. P. Appalasamy, «Classification-based Data Mining Approach for Quality Control in Wine Production,» Journal of Applied Sciences, vol. 12, nº 6, pp. 598-601, 2012.

[24] P. S. a. A. N. S. Shanmuganathan, «Data Mining Techniques for Modelling Seasonal Climate Effects on Grapevine Yield and Wine Quality,» 2010 2nd International Conference on Computational Intelligence, Communication Systems and Networks, pp. 84-89, 2010.

[25] Q. P. Haishan Tian, «Data mining application for upgrading quality of wine production,» The 2010 International Conference on Apperceiving Computing and Intelligence Analysis Proceeding, pp. 109-111, 2010.

[26] J. N. J. S. M. D. J. M. P. N. JORGE RIBEIRO, «Wine Vinification prediction using Data Mining tools,» COMPUTING and COMPUTATIONAL INTELLIGENCE, pp. 78-85, 2009.

[27] A. C. F. A. T. M. J. R. Paulo Cortez, «Modeling wine preferences by data mining from physicochemical properties,» Decision Support Systems, vol. 47, nº 4, pp. 547-553, 2009.

[28] V. G. a. P. Melo-Pinto, «Towards robust Machine Learning models for grape ripeness assessment,» 2021 18th International Joint Conference on Computer Science and Software Engineering (JCSSE), pp. 1-5, 2021.

[29] A. M.-F. P. M.-P. V. Gomes, «Application of hyperspectral imaging and deep learning for robust prediction of sugar and ph levels in wine grape berries,» Sensors, vol. 21, nº 10, p. 3459, 2021.

[30] P. M.-P. Rui Silva, «A review of different dimensionality reduction methods for the prediction of sugar content from hyperspectral images of wine grape berries,» Applied Soft Computing, vol. 113, 2017.

[31] R. M. R.-M. F. M.-F. A. M.-P. P. Gomes V, «Prediction of Sugar Content in Port Wine Vintage Grapes Using Machine Learning and Hyperspectral Imaging,» Processes, vol. 9, nº 7, p. 1241, 2021.

[32] R. R. R. M. M.-F. A. M.-P. P. Gomes V, «Determination of Sugar, pH, and Anthocyanin Contents in Port Wine Grape Berries through Hyperspectral Imaging: An Extensive Comparison of Linear and Non-Linear Predictive Methods,» Applied Sciences, vol. 11, nº 21, p. 10319, 2021.

[33] N. F. O. M. M. S. F. R. P. R. B. J. T. M. Danieldos Santos Costa, «Development of predictive models for quality and maturation stage attributes of wine grapes using vis-nir reflectance spectroscopy,» Postharvest Biology and Technology, vol. 150, pp. 166-178, 2019.

[34] G. V. M.-F. A. M.-P. P. Silva R, «Using Support Vector Regression and Hyperspectral Imaging for the Prediction of Oenological Parameters on Different Vintages and Varieties of Wine Grape Berries,» Remote Sensing, vol. 10, nº 2, p. 312, 2018.

[35] A. M. F. A. F. P. M.-P. Véronique M. Gomes, «Comparison of different approaches for the prediction of sugar content in new vintages of whole Port wine grape berries using hyperspectral imaging,» Computers and Electronics in Agriculture, vol. 140, pp. 244-254, 2017.

[36] A. F. P. M.-L. L. P. A. M. F. P. M.-P. Véronique Gomes, «Characterization of neural network generalization in the determination of pH and anthocyanin content of wine grape in new vintages and varieties,» Food Chemistry, vol. 218, pp. 40-46, 2017.

[37] R. B. B. A. G. R. Michael Fadock, «Visible-Near Infrared Reflectance Spectroscopy for Nondestructive Analysis of Red Wine Grapes,» American Journal of Enology and Viticulture, vol. 67, pp. 38-46, 2016.

[38] F. Z. J. N. X. L. Z. Z. S. Y. Shanshan Chen, «Predicting the anthocyanin content of wine grapes by NIR hyperspectral imaging,» Food Chemistry, vol. 172, pp. 788-793, 2015.

[39] C. F. A. M.-F. A. M.-F. P. L. d. C. P. M.-P. Armando M. Fernandes, «Brix, pH and anthocyanin content determination in whole Port wine grape berries by hyperspectral imaging and neural networks,» Computers and Electronics in Agriculture, vol. 115, pp. 88-96, 2015.

[40] J. M. H.-H. F. J. R.-P. F. J. H. Julio Nogales-Bueno, «Determination of technological maturity of grapes and total phenolic compounds of grape skins in red and white cultivars during ripening by near infrared hyperspectral image: A preliminary approach,» Food Chemistry, vol. 152, pp. 586-591, 2014.

[41] A. M. F. A. F. P. M.-P. V. M. Gomes, «Determination of sugar content in whole Port Wine grape berries combining hyperspectral imaging with neural networks methodologies,» 2014 IEEE Symposium on Computational Intelligence for Engineering Solutions (CIES), pp. 188-193, 2014.

[42] V. D. P.-M. M.-I. L. M.-T. S. González-Caballero, «Optimization of NIR Spectral Data Management for Quality Control of Grape Bunches during On-Vine Ripening,» Sensors, vol. 11, nº 6, pp. 6109-6124, 2011.

[43] P. O. J. P. M. A. A. O. V. F. M. J. C. P. M.-P. Armando Manuel Fernandes, «Determination of anthocyanin concentration in whole grape skins using hyperspectral imaging and adaptive boosting neural networks,» Journal of Food Engineering, vol. 105, nº 2, pp. 216-226, 2011.

[44] D. W. Y. H. Fang Cao, «Soluble solids content and pH prediction and varieties discrimination of grapes based on visible–near infrared spectroscopy,» Computers and Electronics in Agriculture, vol. 71, nº 1, pp. S15-S18, 2010.

[45] E. D. D. B. C. M. D. S. F. J. Marine Le Moigne, «Front face fluorescence spectroscopy and visible spectroscopy coupled with chemometrics have the potential to characterise ripening of Cabernet Franc grapes,» Analytica Chimica Acta, vol. 621, nº 1, pp. 8-18, 2008.

[46] D. C. R. D. W. C. M. G. L. J. Janik, «The prediction of total anthocyanin concentration in red-grape homogenates using visible-near-infrared spectroscopy and artificial neural networks,» Analytica Chimica Acta, vol. 594, nº 1, pp. 107-118, 2007.

[47] C. J. S. A. Ignacio Arana, Journal of Near Infrared Spectroscopy, vol. 13, nº 6, 2005.

[48] A. B. U. J. E.-D. J. C. J. S. P. M.-P. Armando M. Fernandes, «Grapevine variety identification using "Big Data" collected with miniaturized spectrometer combined with support vector machines and convolutional neural networks,» Computers and Electronics in Agriculture, vol. 163, 2019.

[49] A. U. J. E.-D. J. S. J. C. P. M.-P. ArmandoFernandes, «Assessment of grapevine variety discrimination using stem hyperspectral data and AdaBoost of random weight neural networks,» Applied Soft Computing, vol. 72, pp. 140-155, 2018.

[50] A. M. F. B. M. J. T. P. M.-P. Maria P. Diago, «Identification of grapevine varieties using leaf spectroscopy and partial least squares,» Computers and Electronics in Agriculture, vol. 99, pp. 7-13, 2013.

[51] Y. H. Y. W. Guifang Wu, «Discrimination of Varieties of Red Wines Based on Independent Component,» 2008 Congress on Image and Signal Processing, pp. 272-276, 2008.

[52] F. C. L. W. Y. H. Fei Liu, «Discrimination of Rice Wine Age Using Visible and Near Infrared Spectroscopy Combined with BP Neural Network,» 2008 Congress on Image and Signal Processing, pp. 267-271, 2008.

[53] M. L. Limin Fang, «Detection of Six Kinds of Acid in Red Wine with Infrared Spectroscopy Based on FastICA and Neural Network,» Proceedings of 2008 3rd International Conference on Intelligent System and Knowledge Engineering, pp. 856-861, 2008.

[54] L. W. a. Y. H. Fei Liu, «Application of least squares support vector machines for discrimination of red wine using visible and near infrared spectroscopy,» 2008 3rd International Conference on Intelligent System and Knowledge Engineering, pp. 1002-1006, 2008.

[55] «Wikipedia contributors. AdaBoost. In Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=AdaBoost&oldid=1110063122».

[56] Library, «Evaluación de indicadores de la calidad del aceite de oliva virgen: fortalezas, debilidades y oportunidades,» de https://1library.co/article/tratamiento-matem%C3%A1tico-datos-materiales-m%C3%A9todos-muestras-reactivos.qmr2wr7y, pp. 103-107.

[57] P. Mahalanobis, «On the generalised distance in statistics,» Proceedings of the National Institute of Science of India, pp. 49-55, 1936.

[58] «IBM - IBM Analítica - ¿Qué es un árbol de decisión?,» [En línea]. Available: https://www.ibm.com/es-es/topics/decision-trees#:~:text=Un%20%C3%A1rbol%20de%20decisi%C3%B3n%20es,nodos%20internos%20y%20nodos%20hoja.

[59] C3.ai, «C3.ai - Gradient-Boosted Decision Trees (GBDT),» [En línea]. Available: https://c3.ai/glossary/data-science/gradient-boosted-decision-trees-gbdt/.

[60] B. Kosko, «Fuzzy cognitive maps,» International Journal of Man-Machine Studies, vol. 24, nº 1, pp. 65-75, 1986.

[61] A. R. A. Hossain, «13.19 - Sensor-Controlled Intelligent Vehicle Systems: Demand and Needs for a Global Automotive Landscape,» Comprehensive Materials Processing, vol. 13, pp. 473-497, 2014.

[62] «BIOINFORMÁTICA - Algoritmos Genéticos I. Conceptos Básicos,» 2014. [En línea]. Available: https://sci2s.ugr.es/sites/default/files/files/Teaching/GraduatesCourses/Bioinformatica/Tema%2006%20-%20AGs%20I.pdf.

[63] «Colaboradores de Wikipedia, "Gradient boosting," Wikipedia, La enciclopedia libre,» [En línea]. Available: https://es.wikipedia.org/w/index.php?title=Gradient_boosting&oldid=141528519 .

[64] X. D. C. a. J. Z. Y. Wang, «Study on electronic nose recognition method with fast independent component analysis and neural network,» CHINESE JOURNAL OF SENSORS AND ACTUATORS, vol. 20, pp. 38-41, 2007.

[65] J. M. J. H.-T. O. C.-B. B. H.-O. Daniel Alarcón-Narváez, «Clasificadores basados en reglas y selección de atributos para el diagnóstico clínico de subtipos del Síndrome de Guillain-Barré,» Research in Computing Science, vol. 149, nº 8, p. 41–53 , 2020.

[66] «IBM - Analítica - ¿Qué es el algoritmo de k vecinos más cercanos?,» [En línea]. Available: https://www.ibm.com/es-es/topics/knn.

[67] L. K. S. Sam T. Roweis, «Nonlinear Dimensionality Reduction by Locally Linear Embedding,» Science, vol. 290, nº 5500, pp. 2323-2326, 2000.

[68] O. D. C. L. P. Rodolfo Mosquera, «Máquinas de Soporte Vectorial, Clasificador Naïve Bayes y Algoritmos Genéticos para la Predicción de Riesgos Psicosociales en Docentes de Colegios Públicos Colombianos,» Información tecnológica, vol. 29, nº 6, 2018.

[69] P. R. Peres-Neto, D. A. Jackson y K. M. Somers, «How many principal components? stopping rules for determining the number of non-trivial axes revisited,» Computational Statistics & Data Analysis , vol. 49, nº 4, pp. 974-997, 2004.

[70] I. C. Education, «IBM - IBM Cloud Learn Hub - What is Random Forest?,» 2020. [En línea]. Available: https://www.ibm.com/cloud/learn/random-forest.

[71] B. T., «Powerful Feature Selection with Recursive Feature Elimination (RFE) of Sklearn,» 2021. [En línea]. Available: https://towardsdatascience.com/powerful-feature-selection-with-recursive-feature-elimination-rfe-of-sklearn-23efb2cdb54e.

[72] M. A. Kramer, «Nonlinear Principal Component Analysis Using Autoassociative Neural Networks,» AIChE Journal, vol. 37, nº 2, p. 233–243, 1991.

[73] I. C. A. 11.1.x, «IBM - Regresión lineal múltiple,» 2021. [En línea]. Available: https://www.ibm.com/docs/es/cognos-analytics/11.1.0?topic=tests-multiple-linear-regression.

[74] D. E. Hilt y D. W. Seegrist, «Ridge, a computer program for calculating ridge regression estimates,» 1977.

[75] S. Kirkpatrick, C. D. Gelatt y M. P. Vecchi, «Optimization by Simulated Annealing,» Science, vol. 220, nº 4598, pp. 671-680, 1983.

[76] «MathWorks - Support Vector Machine (SVM) - Hiperplanos óptimos como límites de decisión,» [En línea]. Available: https://es.mathworks.com/discovery/support-vector-machine.html.

[77] M. Leticia Laura-Ochoa, «Evaluación de Algoritmos de Clasificación utilizando Validación Cruzada,» de 17th LACCEI International Multi-Conference for Engineering, Education, and Technology, Jamaica, 2019.