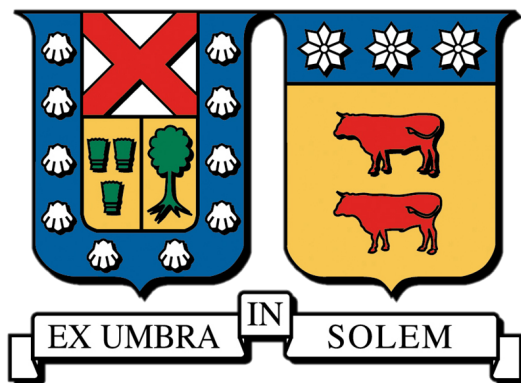


UNIVERSIDAD TÉCNICA FEDERICO SANTA MARÍA

DEPARTAMENTO DE INGENIERÍA QUÍMICA Y AMBIENTAL



Validación de Medidas Espectroscópicas de Muestras
de Vino Utilizando Inteligencia Artificial

JULLIANNE DE JESÚS RIVEROS CARMONA

MEMORIA PARA OPTAR AL TÍTULO DE INGENIERO/A CIVIL QUÍMICO/A

Guía: Dr. Iván Cornejo
Co-referentes: Dr. Juan Armijo Mancilla
Dr. Felipe Angel Figueroa

AGOSTO-2025



CONSTANCIA DE VALIDACIÓN Y CONFIDENCIALIDAD DE MONOGRAFÍA A REPOSITORIO ACADÉMICO

1.- IDENTIFICACIÓN DEL TRABAJO ACADÉMICO

Tipo de monografía (marcar una opción): ☒ Memoria o trabajo de título; ☐ Tesis de Postgrado;

Título del trabajo: Validación de Medidas Espectroscópicas de Muestras de vino Utilizando Inteligencia Artificial.

Nombre del candidato(a): Jullianne de Jesús Riveros Carmona

Carrera / Grado: Ingeniería Civil Química

Campus: Santiago San Joaquín ; **Departamento:** Ingeniería Química y Ambiental

2.- VALIDACIÓN DEL PROFESOR GUÍA/DIRECTOR DE TESIS

Yo, Iván Cornejo García, en mi calidad de profesor(a) guía/director(a) del trabajo académico mencionado anteriormente **DEJO CONSTANCIA** que:

- He revisado esta versión del documento y corresponde a la versión final aprobada del trabajo.
- El trabajo cumple con los requisitos académicos y de formato establecidos por la institución

3.- EVALUACIÓN DE CONFIDENCIALIDAD POR PROPIEDAD INDUSTRIAL

☒ El trabajo **NO contiene información que amerite confidencialidad** y puede ser publicado de inmediato en repositorio con acceso abierto.

☐ El trabajo **CONTIENE** información con potenciales implicancias de propiedad industrial o intelectual y requiere un periodo de confidencialidad (embargo) por:

☐ 6 meses; ☐ 12 meses; ☐ 2 años; ☐ 3 años; ☐ 5 años; ☐ 10 años

Fundamentación de la necesidad de confidencialidad (obligatorio si se solicita embargo):

4.- FIRMAS

Profesor(a) guía o director(a) de memoria o tesis:

Fecha: 09/10/2025

; Firma:

Estudiante o Candidato(a):

Fecha: 09/10/2025

; Firma:

Este formulario debe ser insertado como página 2 de la memoria o tesis, completado y firmado por estudiante y profesor(a) antes de la entrega en portal PRISMA de Biblioteca USM.

Resumen

La presente memoria de título se enmarca en el proyecto FONDEF ID23I10024 y tiene por objetivo general desarrollar un método automatizado basado en medidas espectroscópicas e inteligencia artificial para la evaluación rápida y precisa de características del vino. Esto se abordó mediante la integración de espectroscopía visible, análisis físico-químico y técnicas de aprendizaje automático supervisado, con énfasis en la clasificación de cepas de vino a partir de un volumen limitado de datos y prototipos de bajo costo.

La metodología se estructuró en seis etapas: calibración de prototipos, recolección de muestras, medición espectral, caracterización fisicoquímica y procesamiento de datos. Se evaluó la reproducibilidad instrumental, se diseñaron protocolos de captura estandarizados y se generó una base de datos con 36 muestras físicas representativas de cepas provenientes de Viña El Escorial.

En la etapa de modelamiento, se exploraron dos configuraciones de entrada: espectros crudos y combinados con variables fisicoquímicas, pero ninguno tuvo un resultado favorable de manera independiente. Por ende, se implementó un árbol de clasificación jerárquica organizado por niveles de separación y clasificación de cepas específicas. El modelo final, compuesto por submodelos con algoritmos supervisados, preprocesamiento de análisis de componentes principales (PCA) y/o análisis discriminante lineal (LDA), escaladores y *permutation feature importance* (PFI), alcanzó un accuracy del 80% utilizando únicamente espectros (modelo Spec) y un 93% al incorporar propiedades fisicoquímicas (modelo Full), destacando el impacto del $\%v/v^*$ como variable con mayor poder discriminante, complementada levemente por la conductividad, seguido de $^{\circ}Brix^*$ y pH en la jerarquía de importancias.

Asimismo, se identificaron limitaciones del uso de ruido artificial (gaussiano y *salt and pepper*) como método de *data augmentation*, dada la complejidad matricial del vino y la alteración de agrupaciones naturales de cepas. Se concluyó que la estructura espectral y la interacción entre componentes del vino dificultan la extrapolación realista de nuevos datos sintéticos sin degradar el rendimiento del modelo.

A pesar de la necesidad de optimización y generalización, este trabajo sienta las bases para futuras validaciones en terreno y escalamiento de volumen de datos, orientado al fortalecimiento de herramientas analíticas accesibles para la industria del vino chileno.

Abstract

This thesis is part of the FONDEF ID23I10024 project and aims to develop an automated method based on visible spectroscopy and artificial intelligence for wine grape variety classification using a limited volume of data. The methodology followed six stages: prototype calibration, sample collection and measurement, physicochemical characterization, and modeling. A database of 36 representative wine samples from Viña El Escorial was created. A hierarchical classification tree of supervised models was implemented, incorporating preprocessing with principal component analysis (PCA) and/or linear discriminant analysis (LDA), along with scalers and permutation feature importance (PFI). The model achieved 80% accuracy using only spectral data (*Spec* model), and 93% when incorporating physicochemical variables (*Full* model). Limitations were identified in using artificial noise as a data augmentation method due to the spectral complexity of wine. This work lays the groundwork for future field validations and scalability, contributing accessible analytical tools to the Chilean wine industry.

Agradecimientos

Agradezco primeramente a mis profesores de memoria: Iván Cornejo, Felipe Angel y Juan Armijo, quienes confiaron en mí al darme esta oportunidad de trabajar con ellos en este gran proyecto. Gracias además a sus grupos de trabajo con los cuales compartí muchos momentos gratos. Agradezco también a Gabriela Garrido y Pablo Jaramillo por el apoyo y guía constante en este camino.

A la mafia, gracias por no dejarme sola pese a la distancia, por hacerme saber que siempre estarán para mí, por las recibidas en sus casas cuando no tenía dónde llegar y por sobre todo siempre tratarme igual que siempre a pesar de no vernos durante meses, incluso años. Gracias también a Gaspar, que siempre me hace saber que está ahí aunque hay veces que realmente pasamos meses sin hablar. Tampoco dejo de lado a mis amigos que he hecho en la universidad, Valentina, Pía, Anakarina, Benjamín, Paula, Danna, Valeska, Sebastián, Carolina, quienes me apoyaron en momentos difíciles, me escucharon, me enseñaron y a los cuales les debo mucho en el avance que tuve en estos siete años. Agradezco aún más a quienes me cuidaron cuando me enfermaba o me ayudaron a cuidar a mi Waffle cuando tenía que viajar. Y podría seguir con la lista de gente que conocí en la universidad a los cuales les tengo gratitud, porque aunque algunas personas pasaron por mi vida fugazmente o nunca logré tener un mayor contacto fuera de la institución, el único hecho de haber compartido un recuerdo feliz conmigo, ya los hace dueños de un pedacito de mi corazón.

Scarlet, a quien podría dedicarle una página entera si pudiera, la persona que siempre estuvo ahí en mis peores momentos, la que me apoyó incondicionalmente y de la cual agradezco no solo a ella, sino a la vida o el destino por haber cruzado nuestros caminos. Y no me podía olvidar de mi madre Claudia, quien siempre se mantuvo a mi lado durante todo el camino, fue la primera en confiar en mi potencial desde pequeña. La persona que me dio todo para que pudiese estudiar y crecer como persona, así como profesional. Finalmente, quiero conmemorar y dedicar este trabajo a mi angelito: mi gatita Waffle, a la cual le prometí compartir este momento en persona, pero que, lamentablemente, no pudo esperarme más para partir al cielo. Compartimos noches desveladas donde yo estudiaba y tú me acompañabas en mis piernas, me despertabas en las mañanas, y te acurrucabas conmigo para dormir. Eras mi fuente de luz, aún te sigo extrañando y sé que me acompañas en cada paso que doy, y cada uno es dedicado a ti.

Tabla de Contenidos

1	Introducción	1
1.1	La vitivinicultura como sector estratégico	1
1.2	El control de calidad en la producción de vino	2
1.2.1	Camino del color del vino	3
1.2.2	Regulación varietales y blending de cepas	4
1.3	Impacto del cambio climático en la vitivinicultura	4
1.4	Desafíos y necesidad de respuestas rápidas	5
1.5	Motivación	6
2	Objetivos	8
2.1	Objetivo General	8
2.2	Objetivos Específicos	8
2.3	Alcances del Proyecto	8
3	Marco Teórico	10
3.1	Espectroscopía	10
3.1.1	Espectro Electromagnético	10
3.2	Espectroscopía UV-Visible (UV-Vis)	11
3.2.1	Transmitancia	11
3.2.2	Absorbancia	12
3.3	Espectroscopía en la Industria Vitivinícola	13
3.4	Inteligencia Artificial aplicada al análisis espectroscópico	13
3.5	Datasets	15
3.6	Normalización de los Espectros	16
3.7	Arbol de modelos y Submodelos	16
3.7.1	Organización Jerárquica del Árbol de Modelos	17
3.7.2	Submodelos: Componentes del Árbol de Modelos	17
3.7.3	Pipeline	18

3.8	Reducción Dimensional: PCA y LDA en Espectroscopía	19
3.8.1	Análisis de Componentes Principales (PCA)	19
3.8.2	Análisis Lineal de Discriminantes (LDA)	20
3.9	Aumentación de Datos (Data Augmentation) en Modelos IA	21
3.10	Miniaturización de Espectrómetros	22
3.11	Evaluación del Rendimiento de Modelos de IA	23
3.11.1	Precisión (Accuracy)	24
3.11.2	F1 score	24
3.11.3	Matriz de Confusión	25
3.12	Overfitting o Sobreajuste	26
3.13	Estrategias de Validación de Modelos	26
3.13.1	Validación Cruzada K-Fold	27
3.13.2	Validación Cruzada Leave-One-Out (LOO-CV)	27
3.13.3	Separación por Muestras Físicas	27
4	Materiales y Métodos	29
4.1	Equipos y Herramientas Generales	30
4.1.1	Materiales para Medición de Variables Físicoquímicas	30
4.2	Recolección y Almacenamiento de Muestras de Vino	30
4.3	Impacto en el modelamiento de IA debido a la baja cantidad de datos	32
4.3.1	Aumentación de datos por ruido	32
4.3.2	Clasificación jerárquica	33
4.3.3	Inclusión de propiedades físicoquímicas	33
4.4	Recolección de Datos	33
4.5	Balance de conjunto de prueba	34
4.6	Distribución Espectral Cepas de vino	35
4.7	Estrategias de Ajuste y Optimización	36
4.8	Aumentación de datos	36
4.9	Evaluación y Validación del Modelo	37
5	Resultados	39
5.1	Construcción y organización del conjunto de datos (Datasets)	39
5.1.1	Separación conjunto entrenamiento y prueba	41
5.1.2	Cálculo variables teóricas	41
5.1.3	Dataset Espectral (Spec)	41
5.1.4	Dataset Full	42
5.1.5	Inyección Opcional de Ruido Sintético	42

5.1.6 Organización Final	43
5.2 Evaluación Preliminar Modelos Supervisados Individuales	44
5.2.1 Algoritmos utilizados en las pruebas iniciales	44
5.3 Diseño y Organización Arbol de Modelos	46
5.3.1 Nodo Inicial: MPNS-11 (11 Cepas)	47
5.3.2 Nodo SM-8 (Syrah-Malbec)	48
5.3.3 Nodo CSPSPV-7	49
5.3.4 Nodo CFMSZ-ID	50
5.4 Organización interna de los archivos y ensamblaje en el árbol	53
5.4.1 Estructura ml_tree_V3	54
5.5 Resultados submodelos internos (nodos) del árbol de modelos	54
5.5.1 MPNS-11	54
5.5.2 MPNS-ID	55
5.5.3 SM-8	56
5.5.4 CSPSPV-ID	57
5.5.5 CSPSPV-7	57
5.5.6 CFMSZ-ID	58
5.5.7 MS-ID	58
5.5.8 Análisis General	59
5.6 Resultados árbol de modelos-ml_tree_v3	59
5.6.1 Matriz de confusión modelo <i>Spec</i>	60
5.6.2 Matriz de confusión modelo <i>Full</i>	61
5.6.3 Análisis general árbol de modelos <i>Spec</i> y <i>Full</i>	62
5.7 Reducción de variables fisicoquímicas	63
5.8 Resultados modelos M1	63
5.8.1 Análisis General modelo M1	66
5.9 Resultados modelos M2	68
5.9.1 Análisis General Modelo M2	70
5.10 Jerarquía de variables fisicoquímicas	71
5.11 Efecto de la inyección de ruido sintético sobre la agrupación de cepas en LDA	73
5.12 Automatización de agrupación de cepas para los nodos del árbol de modelos	75
5.12.1 Solución basada en agrupación automática por matriz de confusión	75
5.12.2 Evaluación de la automatización de agrupación de cepas para los nodos del árbol de modelos	77

6 Conclusiones y Recomendaciones	79
6.1 Recomendaciones	81
Referencias	82
A Anexos	I
A.1 Anexo A: Matrices de Confusión Modelos <i>Spec</i> y <i>Full</i>	I
A.1.1 Anexo A.1: Matrices de Confusión MPNS-11 y MPNS-ID	I
A.1.2 Anexo A.2: Matrices de Confusión SM-8	III
A.1.3 Anexo A.3: Matrices de Confusión CSPSPV-ID y CSPSPV-7	IV
A.1.4 Anexo A.4: Matrices de Confusión CFMSZ-ID y MS-ID	VI
A.2 Anexo B: Matrices de Confusión del Modelo M1	VIII
A.2.1 Anexo B.1: Matrices de Confusión MPNS-11 y MPNS-ID	IX
A.2.2 Anexo B.2: Matrices de Confusión SM-8	X
A.2.3 Anexo B.3: Matrices de Confusión CSPSPV-ID y CSPSPV-7	X
A.2.4 Anexo B.4: Matrices de Confusión CFMSZ-ID y MS-ID	XI
A.3 Anexo C: Matrices de Confusión del Modelo M2	XII
A.3.1 Anexo C.1: Matrices de Confusión MPNS-11 y MPNS-ID	XIII
A.3.2 Anexo C.2: Matrices de Confusión SM-8	XIV
A.3.3 Anexo C.3: Matrices de Confusión CSPSPV-ID y CSPSPV-7	XIV
A.3.4 Anexo C.5: Matrices de Confusión CFMSZ-ID y MS-ID	XV

Lista de Tablas

3.1	Ejemplo de una Matriz de Confusión	25
4.1	Cepas y cantidad de muestras obtenidas de Viñedo El Escorial	31
4.2	Ejemplo estructura dataset utilizado.	34
5.1	Accuracy y F1 <i>score</i> obtenidos inicialmente por cada modelo de inteligencia artificial supervisado individual, utilizando los dos tipos de datasets.	45
5.2	Métricas de evaluación nodo MPNS-11 submodelos separadores Full y Spec.	55
5.3	Métricas de evaluación nodo MPNS-ID submodelos clasificadores <i>Full</i> y <i>Spec.</i>	56
5.4	Métricas de evaluación nodo SM-8 submodelos separadores <i>Full</i> y <i>Spec.</i>	56
5.5	Métricas de evaluación nodo CSPSPV-ID submodelos separadores <i>Full</i> y <i>Spec.</i>	57
5.6	Métricas de evaluación nodo CSPSPV-7 submodelos <i>Full</i> y <i>Spec.</i>	57
5.7	Métricas de evaluación nodo CFMSZ-ID submodelos clasificadores Full y Spec.	58
5.8	Métricas de evaluación nodo MS-ID submodelos clasificadores <i>Full</i> y <i>Spec.</i>	58
5.9	Resultados globales del árbol de modelos para un dataset <i>Spec</i> y <i>Full</i>	59

Lista de Figuras

1.1	Evolución de la Producción de Vinos años 1998–2024 (SAG, 2024).	2
1.2	Racimo de uva País con daño por temperatura elevada. (Zurita-Silva et al., 2020).	5
3.1	Espectro electromagnético con énfasis en la región visible (400–700 nm)(Nuñez et al. (2020))	11
3.2	Cuatro tipos de aplicaciones de aprendizaje profundo en espectroscopía Raman, según el formato de salida del modelo(Luo et al. (2022)).	15
3.3	Gráfico de cargas de PC1 vs. PC2 obtenido mediante PCA sobre las propiedades físico-químicas de excipientes y formulaciones de granulación. Los puntos verdes representan descriptores originales, y los círculos resaltan agrupaciones según la contribución de cada descriptor al primer y segundo componente principal (Wang et al. (2021)).	20
3.4	Diagramas de dispersión 2D de las dos componentes más importantes para la reducción de dimensionalidad de los coeficientes MFCC extraídos del conjunto SVD: (a) PCA muestra agrupamientos solapados entre Healthy, Dysphonia, Laryngitis y Reinke’s Edema. (b) LDA mejora notablemente la separación de clases, indicando componentes más discriminantes para cada patología (Bhowmik et al. (2021)).	21
4.1	Prototipo actual de espectrómetro móvil.	29
4.2	Visualización 3D y 2D de agrupación espectral de las 15 cepas de vino mediante transformación LDA.	36
5.1	Visualización 3D y 2D de agrupación espectral de 11 cepas de vino mediante transformación LDA para dataset V3.	46
5.2	Visualización 3D y 2D de agrupación espectral de 8 cepas de vino resultantes luego de la separación del nodo MPNS-11, mediante transformación LDA.	48
5.3	Visualización 3D y 2D de agrupación espectral de 7 cepas de vino resultantes luego de la separación del nodo SM-8, mediante transformación LDA.	49

5.4	Visualización 3D y 2D de agrupación espectral de 4 cepas de vino resultantes luego de la separación del nodo CSPSPV-7, mediante transformación LDA.	51
5.5	Visualización 3D y 2D de agrupación espectral de las cepas Syrah y Malbec mediante transformación LDA.	52
5.6	Organización y diseño visual detallado del árbol de modelos creado.	53
5.7	Matriz de confusión árbol completo para dataset <i>Spec V3</i> de 11 cepas. En la diagonal principal se indican las muestras correctamente clasificadas para cada cepa, mientras que los valores fuera de ella muestran las mal clasificaciones entre cepas.	60
5.8	Matriz de confusión árbol completo para dataset <i>Full V3</i> de 11 cepas. En la diagonal principal se indican las muestras correctamente clasificadas para cada cepa, mientras que los valores fuera de ella muestran las mal clasificaciones entre cepas.	61
5.9	Visualización 3D y 2D de agrupación espectral de las 4 cepas de vino del nodo CFMSZ-ID con propiedades fisicoquímicas agregadas, mediante transformación LDA.	62
5.10	Matriz de confusión del árbol completo para el modelo V3, utilizando componentes M1 con el dataset de 11 cepas. En la diagonal principal se indican las muestras correctamente clasificadas para cada cepa, mientras que los valores fuera de ella muestran las malas clasificaciones entre cepas.	65
5.11	Dominio del espectro de cepas de vinos según las 4 variables fisicoquímicas utilizadas.	67
5.12	Dominio del espectro de cepas de vinos según la variable fisicoquímica %v/v*.	68
5.13	Matriz de confusión árbol completo para dataset M2 V3 de 11 cepas. En la diagonal principal se indican las muestras correctamente clasificadas para cada cepa, mientras que los valores fuera de ella muestran las mal clasificaciones entre cepas.	69
5.14	Dominio del espectro de cepas de vinos según las variable fisicoquímica conductividad.	70
5.15	Dominio del espectro de cepas de vinos según las variable fisicoquímica °Brix*.	72
5.16	Dominio del espectro de cepas de vinos según las variable fisicoquímica pH.	72
5.17	Comparación entre la distribución original sin ruido (izquierda) y la distribución resultante tras la inyección de ruido gaussiano (derecha), proyectadas mediante LDA para las cepas Malbec, Syrah, Syrah-Malbec y Zinfandel.	73
5.18	Comparación entre la distribución original sin ruido (izquierda) y la distribución resultante tras la inyección de ruido <i>salt and pepper</i> (derecha), proyectadas mediante LDA para las cepas Malbec, Syrah, Syrah-Malbec y Zinfandel.	74
5.19	Dendrograma de agrupación jerárquica de cepas seleccionadas para prueba, creado a partir de la matriz de confusión del modelo. Las longitudes de las ramas reflejan la “distancia de confusión” entre cepas.	76

A.1	Matriz de confusión del nodo MPNS-11 del modelo árbol base <i>Spec</i> . En la diagonal principal se indican las muestras correctamente clasificadas para cada cepa, mientras que los valores fuera de ella muestran las mal clasificaciones entre cepas.	I
A.2	Matriz de confusión del nodo MPNS-ID del modelo árbol base <i>Spec</i> . En la diagonal principal se indican las muestras correctamente clasificadas para cada cepa, mientras que los valores fuera de ella muestran las mal clasificaciones entre cepas.	II
A.3	Matriz de confusión del nodo MPNS-11 del modelo árbol base <i>Full</i> . En la diagonal principal se indican las muestras correctamente clasificadas para cada cepa, mientras que los valores fuera de ella muestran las mal clasificaciones entre cepas.	II
A.4	Matriz de confusión del nodo MPNS-ID del modelo árbol base <i>Full</i> . En la diagonal principal se indican las muestras correctamente clasificadas para cada cepa, mientras que los valores fuera de ella muestran las mal clasificaciones entre cepas.	III
A.5	Matriz de confusión del nodo SM-8 del modelo árbol base <i>Spec</i> . En la diagonal principal se indican las muestras correctamente clasificadas para cada cepa, mientras que los valores fuera de ella muestran las mal clasificaciones entre cepas.	III
A.6	Matriz de confusión del nodo SM-8 del modelo árbol base <i>Full</i> . En la diagonal principal se indican las muestras correctamente clasificadas para cada cepa, mientras que los valores fuera de ella muestran las mal clasificaciones entre cepas.	IV
A.7	Matriz de confusión del nodo CSPSPV-ID del modelo árbol base <i>Spec</i> . En la diagonal principal se indican las muestras correctamente clasificadas para cada cepa, mientras que los valores fuera de ella muestran las mal clasificaciones entre cepas.	IV
A.8	Matriz de confusión del nodo CSPSPV-7 del modelo árbol base <i>Spec</i> . En la diagonal principal se indican las muestras correctamente clasificadas para cada cepa, mientras que los valores fuera de ella muestran las mal clasificaciones entre cepas.	V
A.9	Matriz de confusión del nodo CSPSPV-ID del modelo árbol base <i>Full</i> . En la diagonal principal se indican las muestras correctamente clasificadas para cada cepa, mientras que los valores fuera de ella muestran las mal clasificaciones entre cepas.	V
A.10	Matriz de confusión del nodo CSPSPV-ID del modelo árbol base <i>Full</i> . En la diagonal principal se indican las muestras correctamente clasificadas para cada cepa, mientras que los valores fuera de ella muestran las mal clasificaciones entre cepas.	VI
A.11	Matriz de confusión del nodo CFMSZ-ID del modelo árbol base <i>Spec</i> . En la diagonal principal se indican las muestras correctamente clasificadas para cada cepa, mientras que los valores fuera de ella muestran las mal clasificaciones entre cepas.	VI
A.12	Matriz de confusión del nodo MS-ID del modelo árbol base <i>Spec</i> . En la diagonal principal se indican las muestras correctamente clasificadas para cada cepa, mientras que los valores fuera de ella muestran las mal clasificaciones entre cepas.	VII

A.13 Matriz de confusión del nodo CFMSZ-ID del modelo árbol base <i>Full</i> . En la diagonal principal se indican las muestras correctamente clasificadas para cada cepa, mientras que los valores fuera de ella muestran las mal clasificaciones entre cepas.	VII
A.14 Matriz de confusión del nodo MS-ID del modelo árbol base <i>Full</i> . En la diagonal principal se indican las muestras correctamente clasificadas para cada cepa, mientras que los valores fuera de ella muestran las mal clasificaciones entre cepas.	VIII
A.15 Matriz de confusión del nodo MPNS-11 del modelo M1. En la diagonal principal se indican las muestras correctamente clasificadas para cada cepa, mientras que los valores fuera de ella muestran las mal clasificaciones entre cepas.	IX
A.16 Matriz de confusión del nodo MPNS-ID del modelo M1. En la diagonal principal se indican las muestras correctamente clasificadas para cada cepa, mientras que los valores fuera de ella muestran las mal clasificaciones entre cepas.	IX
A.17 Matriz de confusión del nodo SM-8 del modelo M1. En la diagonal principal se indican las muestras correctamente clasificadas para cada cepa, mientras que los valores fuera de ella muestran las mal clasificaciones entre cepas.	X
A.18 Matriz de confusión del nodo CSPSPV-ID del modelo M1. En la diagonal principal se indican las muestras correctamente clasificadas para cada cepa, mientras que los valores fuera de ella muestran las mal clasificaciones entre cepas.	X
A.19 Matriz de confusión del nodo CSPSPV-7 del modelo M1. En la diagonal principal se indican las muestras correctamente clasificadas para cada cepa, mientras que los valores fuera de ella muestran las mal clasificaciones entre cepas.	XI
A.20 Matriz de confusión del nodo CFMSZ-ID del modelo M1. En la diagonal principal se indican las muestras correctamente clasificadas para cada cepa, mientras que los valores fuera de ella muestran las mal clasificaciones entre cepas.	XI
A.21 Matriz de confusión del nodo MS-ID del modelo M1. En la diagonal principal se indican las muestras correctamente clasificadas para cada cepa, mientras que los valores fuera de ella muestran las mal clasificaciones entre cepas.	XII
A.22 Matriz de confusión del nodo MPNS-11 del modelo M2. En la diagonal principal se indican las muestras correctamente clasificadas para cada cepa, mientras que los valores fuera de ella muestran las mal clasificaciones entre cepas.	XIII
A.23 Matriz de confusión del nodo MPNS-ID del modelo M2. En la diagonal principal se indican las muestras correctamente clasificadas para cada cepa, mientras que los valores fuera de ella muestran las mal clasificaciones entre cepas.	XIII
A.24 Matriz de confusión del nodo SM-8 del modelo M2. En la diagonal principal se indican las muestras correctamente clasificadas para cada cepa, mientras que los valores fuera de ella muestran las mal clasificaciones entre cepas.	XIV

A.25 Matriz de confusión del nodo CSPSPV-ID del modelo M2. En la diagonal principal se indican las muestras correctamente clasificadas para cada cepa, mientras que los valores fuera de ella muestran las mal clasificaciones entre cepas.	XIV
A.26 Matriz de confusión del nodo CSPSPV-7 del modelo M2. En la diagonal principal se indican las muestras correctamente clasificadas para cada cepa, mientras que los valores fuera de ella muestran las mal clasificaciones entre cepas.	XV
A.27 Matriz de confusión del nodo CFMSZ-ID del modelo M2. En la diagonal principal se indican las muestras correctamente clasificadas para cada cepa, mientras que los valores fuera de ella muestran las mal clasificaciones entre cepas.	XV
A.28 Matriz de confusión del nodo MS-ID del modelo M2. En la diagonal principal se indican las muestras correctamente clasificadas para cada cepa, mientras que los valores fuera de ella muestran las mal clasificaciones entre cepas.	XVI

Nomenclatura

Símbolo	Significado, unidades
A	Absorbancia, adimensional
c	Concentración de la sustancia, mol/L
C	Conductividad eléctrica de la muestra, mS
$I_c(\lambda)$	Lectura espectral de la muestra, adimensional
$I_w(\lambda)$	Lectura blanco analítico, adimensional
l	Longitud del camino óptico, cm
P	Potencia transmitida por la muestra, W
P_0	Potencia incidente (blanco), W
pH	Potencial de hidrógeno (medida de acidez), adimensional
$T(\lambda)$	Transmitancia normalizada por blanco, adimensional
$T_{max}(\lambda)$	Transmitancia normalizada por máximo, adimensional
X	Variable de entrada de modelos
$^{\circ}Brix$	Grados Brix sobreestimados por efecto del alcohol, $^{\circ}Brix$
$\%v/v^*$	Estimación del volumen de alcohol (mezcla refractada), %
λ	Longitud de onda, nm
σ	Desviación estandar, adimensional
ϵ	Coefficiente de absorptividad, $L \cdot mol^{-1} \cdot cm^{-1}$
max_{λ}	Máximo valor de longitudde onda, nm
Siglas	Significado
IA	Inteligencia Artificial
ML	Aprendizaje Automático (Machine Learning)
VIS	Luz Visible (longitud de onda en el rango visible)
in situ	En el lugar (del latín, "en el sitio")
FP	Falsos positivos (predicción positiva con clase real negativa)
FN	Falsos negativos (predicción negativa con clase real positiva)
TN	Verdaderos negativos (predicción negativa con clase real negativa)
TP	Verdaderos positivos (predicción positiva con clase real positiva)

Capítulo 1

Introducción

1.1 La vitivinicultura como sector estratégico

La vitivinicultura es una de las actividades agrícolas más antiguas y relevantes a nivel global, destacándose no solo por su impacto en la producción económica, sino también por su influencia en el desarrollo rural, el turismo enológico y la identidad cultural de diversas regiones. En particular, Chile ha consolidado su posición como uno de los principales exportadores de vino en el mundo, gracias a su diversidad de cepas, condiciones climáticas favorables y altos estándares de calidad en la producción. Internacionalmente, del vino chileno se valora su relación calidad-precio y su diversidad de cepas, como Cabernet Sauvignon, Carmenère y Syrah (Statista (2024)). Gracias a esto, la industria vitivinícola chilena ha experimentado un crecimiento en la producción de vino en los últimos años, según datos del Servicio Agrícola y Ganadero (SAG). Esta tendencia se observa en la Figura 1.1, donde se representa la evolución histórica de la producción nacional de vinos entre los años 1998 y 2024, diferenciando entre vinos con denominación de origen (D.O.), sin D.O., y aquellos elaborados con uva de mesa. La curva total muestra un incremento sostenido desde inicios de los años 2000 hasta alcanzar un máximo en 2013, seguido de fluctuaciones con una leve tendencia a la baja hacia 2024.

Evolución de la Producción de Vinos años 1998–2024 (Hl.)

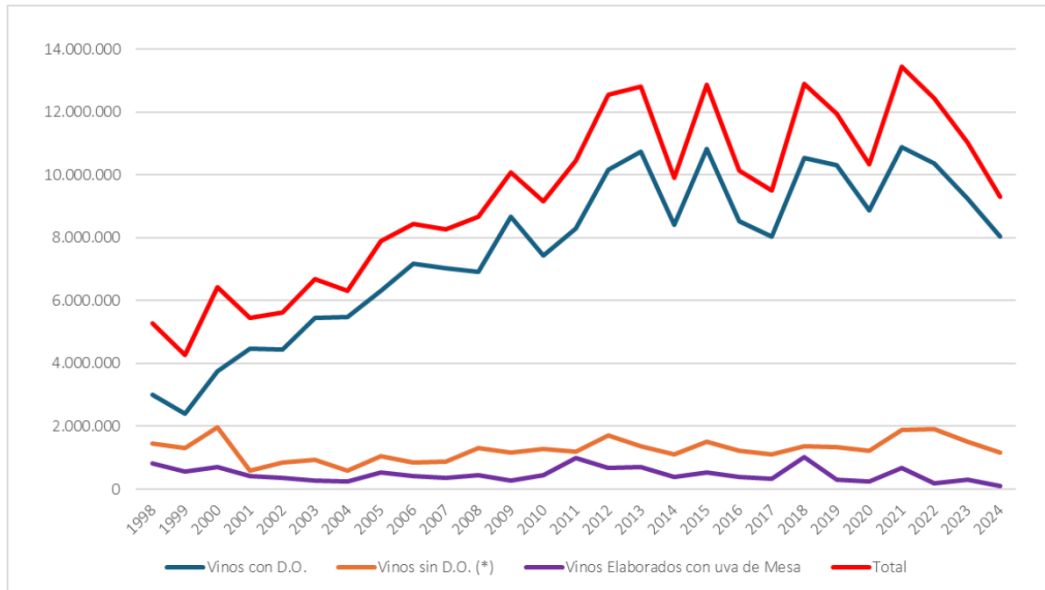


Figura 1.1: Evolución de la Producción de Vinos años 1998–2024 (SAG, 2024).

Según la Organización Internacional de la Viña y el Vino (OIV,2024), Chile ocupa el cuarto lugar entre los mayores exportadores de vino del mundo, siendo superado sólo por países europeos de amplia trayectoria en materias vitivinícolas, como Francia, España e Italia (Odepa (2024b)). Esto representa una parte significativa de la economía nacional, aportando un 0,5% al PIB nacional (Statista (2024)). Según datos de la Oficina de Estudios y Políticas Agrarias (Odepa (2024a)), en enero de 2024, las exportaciones totales de vino alcanzaron 64,7 millones de litros, por un total de USD 138 millones FOB (Free On Board).

1.2 El control de calidad en la producción de vino

Una de las áreas más críticas dentro de la industria vitivinícola es el control de calidad, ya que el proceso de elaboración del vino es altamente complejo y depende de múltiples variables que afectan su calidad final. Desde la selección de las uvas hasta la fermentación y el embotellado, cada etapa debe ser supervisada cuidadosamente para garantizar un producto consistente y competitivo en el mercado internacional.

Los enólogos juegan un papel fundamental en este proceso, ya que son responsables de evaluar y monitorear parámetros como la maduración de la uva, la fermentación, el perfil aromático, el equilibrio químico del vino, su acidez y azúcar, lo que es crucial para decidir el mejor momento para la cosecha y otros pasos clave en el proceso de vinificación y posterior embotellado (Ponce

(2019)). Tradicionalmente, esta evaluación se ha basado en catas sensoriales, que son degustaciones organizadas donde se evalúan las características organolépticas del vino. Los parámetros involucrados en la evaluación sensorial de los vinos son el color, el aroma y el sabor, siendo el color del vino una de las principales características organolépticas que se utilizan para establecer su calidad y aceptabilidad en el mercado (Almanza-Merchán et al. (2015)).

Para complementar su respuesta, los enólogos recurren a distintas técnicas de análisis químicos para evaluar la composición físico-química del vino. Entre las técnicas más utilizadas se encuentra la espectrofotometría UV-Vis, la cual permite medir, de manera no destructiva, la concentración de compuestos fenólicos y antocianinas, esenciales para determinar el color y la estabilidad del vino (Gullifa et al. (2023)). Asimismo, se emplean métodos más avanzados como la cromatografía de gases (CG) y la cromatografía líquida de alta resolución (HPLC), los cuales facilitan la identificación y cuantificación de compuestos volátiles y no volátiles que influyen en el aroma y sabor del vino (Bindon et al. (2014)).

1.2.1 Camino del color del vino

El color de un vino no es solo una característica estética: es un indicador de contenido fenólico, evolución aromática y estabilidad (García (2017)). El denominado "camino del color" corresponde al seguimiento de la evolución del pigmento del vino, desde la formación inicial hasta su maduración y embotellado. Para ello, se emplean instrumentos espectrofotométricos que miden la absorbancia (A) a 420 nm, 520 nm y 620 nm, lo que permite calcular la intensidad y tonalidad del color mediante la suma $A_{420} + A_{520} + A_{620}$ y la relación A_{420}/A_{520} (Hensel et al. (2022)).

La Organización Internacional de la Viña y el Vino (OIV) recomienda, además, el uso de sistemas colorimétricos objetivos como el modelo CIE Lab* y el método de Glories, que permiten cuantificar parámetros como luminosidad, croma y matiz. Estas herramientas facilitan una descripción precisa del color a lo largo del tiempo y entre distintos lotes (Fan et al. (2023)).

Si bien estos métodos entregan información precisa y detallada sobre la composición del vino, presentan ciertas limitaciones prácticas, particularmente en términos de accesibilidad, tiempo de análisis y costos operativos. La mayoría requiere equipos especializados, reactivos químicos y laboratorios equipados, lo cual dificulta su implementación en contextos donde se requiere una evaluación rápida. Además, su uso implica una demanda considerable de tiempo para la obtención de resultados, lo que retrasa la toma de decisiones en el proceso enológico. En viñas de menor escala o con recursos limitados, la necesidad de realizar extracciones y manipulaciones de muestra introduce una variabilidad considerable, lo que compromete la confiabilidad y limita su aplicación en terreno (Donais and Vandenabeele (2021)).

1.2.2 Regulación varietales y blending de cepas

Para que un vino pueda ser etiquetado con el nombre de una variedad específica, como Cabernet Sauvignon o Carménère, la legislación chilena exige que al menos el 75% del contenido corresponda a la cepa declarada. Esta norma está establecida en el Decreto N.º 464 del Ministerio de Agricultura, que regula el uso de denominaciones de origen y designaciones varietales en el etiquetado de vinos en Chile (Servicio Agrícola y Ganadero (SAG) (nd)).

En el contexto chileno, muchos productores utilizan esta flexibilidad normativa para complementar el vino con pequeñas proporciones de otras variedades. Esta práctica de mezcla, conocida como *blending*, se emplea para equilibrar aspectos como el color, la estructura tánica y el perfil sensorial. Según Cáceres-Mella et al. (2014), mezclar cepas tintas —como Cabernet Sauvignon, Carmenère y Merlot— permite armonizar la composición fenólica y mejorar características como la astringencia y el cuerpo, sin comprometer la tipicidad del vino. A nivel técnico, Bindon et al. (2017) respaldan que las diferencias en la arquitectura celular de las bayas influyen en la extracción de taninos y antocianos, lo que justifica el uso de cepas con mayor capacidad extractiva para corregir deficiencias cromáticas o estructurales en el vino final.

Estudios como los de Bustos et al. (2004) demuestran que las cepas tintas chilenas presentan diferencias significativas en su perfil fenólico, lo cual sustenta técnicamente su uso en mezclas diseñadas para alcanzar un objetivo sensorial específico. Este tipo de intervención no solo responde a requerimientos organolépticos, sino también a criterios de producción y mercado, ya que permite ajustar el volumen final a partir de partidas adquiridas a terceros, manteniendo al mismo tiempo el cumplimiento legal para el etiquetado varietal.

1.3 Impacto del cambio climático en la vitivinicultura

A pesar de su expansión, la vitivinicultura se enfrenta a crecientes desafíos producto del cambio climático. Este fenómeno ha comenzado a impactar las condiciones de producción y los ciclos de maduración tanto de la uva como del vino.

Una de las principales consecuencias del cambio climático que afecta a esta industria es el aumento sostenido de las temperaturas promedio, lo cual impacta directamente en el ciclo de crecimiento de la vid (planta que produce uvas) y la composición química de las uvas. Se ha evidenciado que el exceso de temperatura provoca un desequilibrio en la acumulación de antocianinas y otros polifenoles, compuestos responsables del color en uvas tintas. En la Figura 1.2 se muestra un racimo de uva País con daño atribuido a temperaturas elevadas, donde se observa un desarrollo desigual de las bayas, producto del estrés térmico (Zurita-Silva et al. (2020)).



Figura 1.2: Racimo de uva País con daño por temperatura elevada. (Zurita-Silva et al., 2020).

Por otro lado, el cambio climático también ha generado alteraciones en la disponibilidad de agua para el cultivo de la vid. En algunas regiones, el aumento de la temperatura ha venido acompañado de una reducción en las precipitaciones, lo que genera estrés hídrico en las vides y afecta el rendimiento de los viñedos. En otras zonas, se ha observado un incremento de lluvias en períodos críticos, lo que aumenta el riesgo de enfermedades fúngicas y afecta la calidad de la cosecha (de Ingenieros Agrónomos Enólogos (ANIAE)).

La adaptación a estas condiciones climáticas ha llevado a muchos productores a modificar sus prácticas agrícolas, migrando sus cultivos hacia zonas con condiciones más favorables o incrementando los períodos de riego (Zurita-Silva et al., 2020). No obstante, estos cambios generan nuevos desafíos, como la necesidad de ajustar técnicas de cultivo y la gestión de enfermedades emergentes.

1.4 Desafíos y necesidad de respuestas rápidas

Con la creciente variabilidad climática, la toma de decisiones se ha vuelto más compleja para los enólogos debido a una mayor presión en la creciente demanda de decisiones más rápidas y precisas, lo cual influye sobre la disponibilidad y la capacidad de los enólogos para manejar todas

las etapas del proceso de vinificación. La utilización de métodos tradicionales resulta insuficiente para adaptarse a las nuevas condiciones y exigencias de la industria como consecuencia del cambio climático. Por otro lado, la práctica del blending, si bien es una herramienta técnica eficaz para ajustar el perfil sensorial, implica una toma de decisiones compleja basada en el conocimiento del comportamiento organoléptico y químico de múltiples cepas. Esto requiere información precisa sobre su concentración de polifenoles, taninos y antocianos, así como su compatibilidad sensorial, lo que no siempre está disponible en tiempo real para los productores y enólogos. En el caso de viñas pequeñas o medianas, que a menudo recurren a la compra de partidas externas para complementar su producción, esta falta de información dificulta el trabajo de los enólogos, quienes deben tomar decisiones con un alto grado de incertidumbre. Esto afecta directamente su capacidad para predecir el comportamiento de las mezclas, comprometiendo la coherencia sensorial del vino entre lotes y aumentando el riesgo de perder la tipicidad de la cepa declarada. Esto genera una evidente necesidad de desarrollar e implementar soluciones tecnológicas que permitan una gestión más eficiente y adaptativa de los recursos y procesos vitivinícolas.

1.5 Motivación

Ante esta realidad, surge la motivación para desarrollar un sistema que pueda complementar y, en algunos casos, mejorar las herramientas tradicionales disponibles para los enólogos. Tecnologías como la espectroscopía visible (VIS), combinadas con herramientas de inteligencia artificial (IA) surgen como una herramienta valiosa para el apoyo del trabajo de los enólogos, ofreciendo un potencial para mejorar la eficiencia en el monitoreo del vino y facilitar la toma de decisiones en el viñedo y la bodega.

La espectroscopía permite la obtención de datos precisos, los cuales se pueden correlacionar con las características ópticas de los componentes del vino sin necesidad de recurrir a métodos destructivos o invasivos, además de requerir la mínima preparación de la muestra y su bajo costo (Cozzolino et al. (2008)).

Complementariamente, la inteligencia artificial facilita el procesamiento y análisis de grandes volúmenes de datos espectrales y transformarlos en información útil y accionable. Los algoritmos de aprendizaje automático permiten identificar patrones y realizar predicciones sobre la evolución del vino, optimizando el proceso de vinificación, reduciendo la incertidumbre y mejorando la toma de decisiones (Luo et al. (2022)).

Por lo tanto, la combinación de estas tecnologías permitiría realizar una toma no destructiva de datos, los cuales podrían ser analizados en tiempo real tanto por el enólogo como por un algoritmo de inteligencia artificial. Esto facilitaría una respuesta rápida y oportuna frente a los cambios en las condiciones de producción. Asimismo, se espera apoyar al trabajo de enólogos enfocados a

viñas pequeñas o medianas, al momento de tomar decisiones rápidas sobre mezclas y compra de partidas complementarias. La posible falta de información inmediata sobre el color, la composición fenólica y el cumplimiento varietal dificulta asegurar la tipicidad del vino y la coherencia entre lotes. Frente a ello, esta herramienta permitiría evaluar en terreno, de forma precisa, el perfil cromático y fenólico del vino, optimizando tanto el seguimiento del camino del color como la gestión eficiente de los blends conforme a las exigencias normativas vigentes.

El presente estudio, realizado bajo el contexto del proyecto FONDEF ID23I10024, propone la integración de estas tecnologías en un sistema portátil de espectroscopía VIS con algoritmos de IA que pueda realizar mediciones en tiempo real directamente en el viñedo o durante el proceso de vinificación. Este dispositivo permitirá a los enólogos y viticultores planificar con anticipación y tomar decisiones informadas, ajustando los procesos de producción del vino para adaptarse a las condiciones cambiantes y asegurar la calidad del producto.

Capítulo 2

Objetivos

2.1 Objetivo General

Desarrollar un método automatizado basado en medidas espectroscópicas e inteligencia artificial para la evaluación rápida y precisa de las características espectroscópicas del vino, a partir de un limitado volumen de datos.

2.2 Objetivos Específicos

- i. Parametrizar las características espectroscópicas del vino que se relacionan con sus características físico-químicas.
- ii. Implementar estrategias de aumentación de datos (*data augmentation*) para ampliar el conjunto de muestras disponibles, mejorando la capacidad de generalización y robustez del modelo.
- iii. Desarrollar un set de datos base para el entrenamiento del modelo de aprendizaje automático enfocados en la identificación de cepas de vino.
- iv. Entrenar al modelo de inteligencia artificial para analizar los datos espectroscópicos y predecir la cepa del vino, evaluando su precisión mediante validaciones internas realizadas sobre el conjunto de datos disponible.

2.3 Alcances del Proyecto

Este proyecto se centra en el desarrollo y validación de un modelo de inteligencia artificial para la clasificación de cepas de vino utilizando espectroscopia visible. Los alcances del trabajo están

limitados por los recursos disponibles, incluyendo un conjunto reducido de datos espectroscópicos de cepas de vino específicas, lo que implica que el análisis y los resultados son aplicables principalmente en el contexto laboratorio. Los modelos de clasificación desarrollados se validaron únicamente a nivel de laboratorio sin extenderse a escenarios industriales. Además, el proyecto no incluye un análisis sobre los recursos computacionales necesarios.

Capítulo 3

Marco Teórico

3.1 Espectroscopía

La espectroscopía es una técnica analítica que estudia la interacción entre la radiación electromagnética y la materia. A través de esta interacción, es posible obtener información sobre la composición química, la estructura molecular y otras propiedades físicas de una muestra (Skoog et al. (2007)).

Este método se basa en la capacidad de las moléculas para absorber, emitir o reflejar la luz en longitudes de onda específicas. Una fuente de radiación incide sobre la muestra, y se mide la luz absorbida, transmitida o emitida. Esta información permite inferir propiedades cualitativas y cuantitativas del analito, es decir, la sustancia o compuesto químico de interés presente en la muestra, ya que cada sustancia presenta una huella espectral característica (Skoog et al. (2007)).

3.1.1 Espectro Electromagnético

El espectro electromagnético comprende las radiaciones electromagnéticas ordenadas según su longitud de onda, frecuencia o energía. Estas radiaciones corresponden a combinaciones de campos eléctricos y magnéticos alternos que se propagan en forma de ondas. Cuando una molécula absorbe luz (energía), ocurre una transición desde un estado energético fundamental (E1) a uno excitado (E2). Cada molécula presenta estados excitados o bandas características únicas, lo que permite identificarla mediante sus espectros de absorción. El espectro electromagnético se divide en regiones que abarcan un intervalo continuo de longitudes de onda o frecuencias (Martelo Vidal (2014b)). En este trabajo, el interés se centra en la región del espectro visible, comprendida aproximadamente entre 400 y 700 nanómetros, ya que es en este rango donde se perciben los colores por el ojo humano. Las regiones adyacentes a la zona visible son el ultravioleta (UV), por debajo de los 400 nm, y el infrarrojo cercano (NIR), por encima de los 700 nm, ambas también utilizadas en análisis espectroscópicos avanzados. Estas zonas resultan particularmente relevantes en aplicaciones

analíticas basadas en espectroscopía, debido a su compatibilidad con sensores portátiles y la presencia de señales espectrales asociadas a compuestos fenólicos y pigmentos presentes en el vino (Nuñez et al. (2020)).

A continuación, se presenta la distribución clásica del espectro electromagnético, destacando la región visible:

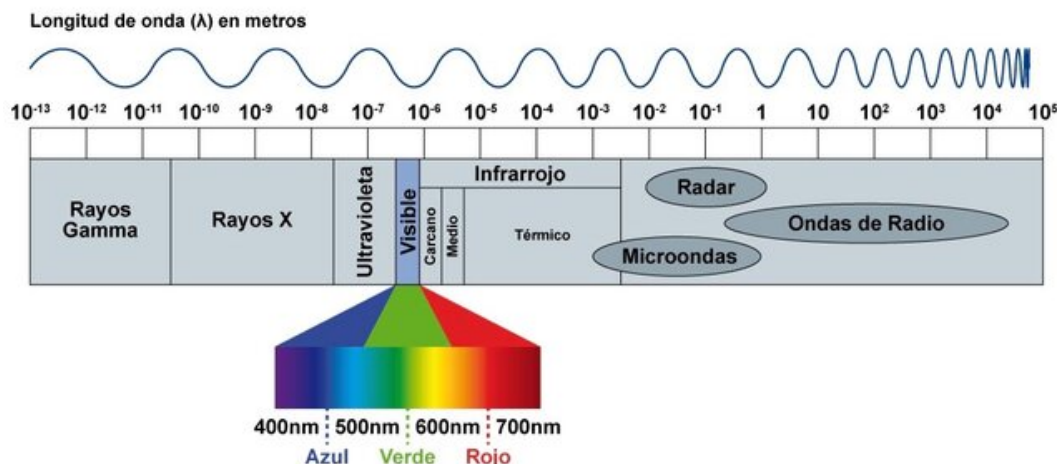


Figura 3.1: Espectro electromagnético con énfasis en la región visible (400–700 nm) (Nuñez et al. (2020))

3.2 Espectroscopía UV-Visible (UV-Vis)

La espectroscopía UV-Visible es una técnica analítica empleada para estudiar la interacción de la luz con las moléculas de una muestra, en el rango de longitudes de onda correspondientes a la luz ultravioleta (UV) y visible (Vis). En este rango, las moléculas absorben luz a distintas longitudes de onda según sus características electrónicas y vibracionales.

La cantidad de luz absorbida por una muestra está directamente relacionada con la concentración de los compuestos presentes, lo que permite realizar análisis tanto cualitativos como cuantitativos (Skoog et al. (2007)). Esta técnica se basa en la ley de Lambert-Beer, que establece una relación lineal entre la absorbancia y la concentración de un analito en solución, bajo condiciones ideales de homogeneidad y baja concentración.

3.2.1 Transmitancia

La transmitancia (T) es una medida de la proporción de luz que atraviesa una muestra. Se define como la razón entre la intensidad de luz que emerge de la muestra y la intensidad de luz incidente. Su valor se expresa en porcentaje, entre 0% y 100%. Una transmitancia alta indica que la muestra

deja pasar la mayor parte de la luz incidente (Skoog et al. (2007)).

Para calcular la transmitancia teórica por blanco, se divide la intensidad espectral de la muestra por la del blanco de referencia, longitud a longitud (esto se hace siempre comparando la misma celda y el mismo sensor). En este proyecto, se trabajó con 18 longitudes de onda específicas dentro del espectro visible, por lo que esta fórmula fue aplicada individualmente a cada una (Skoog et al. (2007)):

$$T(\lambda) = \frac{I_c(\lambda)}{I_w(\lambda)} \quad (3.1)$$

Donde $T(\lambda)$ corresponde a la transmitancia teórica, $I_c(\lambda)$ es la lectura espectral de la muestra o señal incidente y $I_w(\lambda)$ es la lectura blanca o señal del blanco.

También se puede calcular una forma alternativa de transmitancia, que fue desarrollada experimentalmente durante este trabajo y resultó ser útil para ciertos análisis. Esta transmitancia normalizada al máximo no corresponde a una formulación con base teórica establecida en la literatura, sino que se propuso como una técnica de normalización práctica. Se define como:

$$T_{max}(\lambda) = \frac{I_c(\lambda)}{\max_{\lambda} I_w(\lambda)} \quad (3.2)$$

3.2.2 Absorbancia

La absorbancia (A), que indica cuánta luz es absorbida por la muestra, se relaciona con la transmitancia mediante la siguiente expresión logarítmica (Skoog et al. (2007)):

$$A = -\log(T(\lambda)) \quad (3.3)$$

A medida que la muestra absorbe más luz, la transmitancia disminuye y la absorbancia aumenta. Esta última es fundamental en espectroscopía, ya que se relaciona linealmente con la concentración de una sustancia según la ley de Beer-Lambert:

$$A = \epsilon cl \quad (3.4)$$

Donde:

- A es la absorbancia.
- ϵ Coeficiente de absorción $L \cdot \text{mol}^{-1} \cdot \text{cm}^{-1}$.
- c concentración de la sustancia mol/L .
- l es la longitud del camino óptico a través de la muestra en cm .

Esta relación permite cuantificar compuestos presentes en soluciones líquidas. En el caso del vino, esta técnica se aplica comúnmente para determinar la concentración de compuestos fenólicos (como antocianinas y taninos), claves para la estabilidad y color del producto final.

3.3 Espectroscopía en la Industria Vitivinícola

En el análisis enológico moderno, la espectroscopía UV-Visible (UV-Vis) se ha consolidado como una herramienta esencial para la caracterización química de los vinos, gracias a su rapidez, precisión y bajo costo operativo. Esta técnica permite obtener información cuantitativa sobre compuestos relevantes, como los polifenoles, los cuales influyen directamente en propiedades sensoriales del vino, como el color, el aroma y la astringencia (sensación de sequedad y aspereza en boca) (García (2017)).

La espectroscopía UV-Vis se ha utilizado eficientemente para estimar el contenido fenólico total mediante métodos como el índice de Folin-Ciocalteu e IPT280. Además, permite determinar parámetros cromáticos relevantes, como el espacio CIELab y los parámetros Glories. Estos análisis han sido útiles para clasificar vinos según su variedad, envejecimiento u origen geográfico (García (2017)).

A pesar de sus ventajas, simplicidad, velocidad y versatilidad, esta técnica presenta limitaciones en la selectividad cuando se aplica a matrices complejas como el vino, debido a la superposición espectral entre distintos compuestos. No obstante, la aplicación de técnicas estadísticas multivariadas ha permitido superar estas restricciones, ampliando su uso en el análisis multicomponente enología (García (2017)).

Estudios como el de Cozzolino et al. (2008) han demostrado el potencial de esta técnica para analizar la concentración de diversos elementos en vinos australianos, resaltando su capacidad para entregar información precisa y detallada sobre la composición del vino. Además, su combinación con algoritmos de inteligencia artificial ha evidenciado ser una herramienta poderosa para mejorar la precisión de las predicciones en el análisis enológico.

3.4 Inteligencia Artificial aplicada al análisis espectroscópico

La inteligencia artificial (IA) se ha posicionado como una herramienta clave en el análisis químico moderno, especialmente cuando se integra con técnicas espectroscópicas. Según Cubo et al. (2021), la IA comprende una serie de métodos computacionales que permiten simular procesos humanos como el aprendizaje, la clasificación y la predicción, siendo especialmente útil para analizar grandes volúmenes de datos espectrales. Esto ha permitido superar las limitaciones de los métodos tradicionales, entregando mayor rapidez, precisión y eficiencia en la detección e identificación de

compuestos complejos.

En este contexto, una de las ramas más relevantes es el aprendizaje automático, o *machine learning* (ML), el cual permite desarrollar algoritmos capaces de encontrar patrones y relaciones en los datos sin necesidad de programación explícita para cada tarea (Cubo et al. (2021)). En espectroscopía, los algoritmos de ML se han aplicado exitosamente en el preprocesamiento de señales, reducción de ruido, extracción de características y clasificación de muestras (Mengstu et al. (2024)).

Dentro del aprendizaje automático, el *deep learning* o aprendizaje profundo destaca por utilizar redes neuronales con múltiples capas para modelar relaciones no lineales complejas en grandes conjuntos de datos. Aunque su implementación requiere una mayor capacidad computacional y volúmenes significativos de datos, ha demostrado ser especialmente eficaz en tareas como el reconocimiento de patrones espectrales y la mejora de la precisión en modelos de clasificación supervisada (Manual Nuevas Tecnologías, 2023).

Uno de los enfoques más utilizados es el aprendizaje supervisado, donde se requiere un conjunto de datos etiquetados para entrenar el modelo. Estos algoritmos establecen relaciones entre características de entrada (como intensidades espectrales) y clases específicas, permitiendo predecir nuevas muestras. Entre los algoritmos más comunes destacan las máquinas de vectores de soporte (SVM), redes neuronales artificiales (ANN), Random Forest, Naive Bayes y K-Nearest Neighbors (KNN) (Pan et al. (2022)).

En los últimos años, la combinación de espectroscopía con inteligencia artificial (IA) ha revolucionado los métodos de análisis no destructivos en alimentos y bebidas, permitiendo abordar matrices complejas como jugos, aceites, productos lácteos y, particularmente, el vino. Esta integración ha facilitado la interpretación de espectros complejos mediante modelos supervisados y no supervisados, permitiendo predecir variables como la composición química, detectar adulteraciones o clasificar productos según su origen geográfico y variedad (Mengstu et al. (2024); Pan et al. (2022); Luo et al. (2022)).

Mengstu et al. (2024) realizaron una revisión exhaustiva del uso de espectroscopía junto a aprendizaje automático en el procesamiento de alimentos, destacando el rol clave de la espectroscopía NIR y Raman en tareas como clasificación de variedades, identificación de origen y análisis de contenido. Los modelos más utilizados en su estudio fueron PLS-LDA (37%), SVM (19%) y PCA (15%), lo que evidencia la prevalencia de enfoques supervisados y técnicas de reducción de dimensionalidad. Aunque este trabajo entrega una visión actualizada y sólida del estado del arte en el sector agroalimentario, también advierte que el uso de modelos *deep learning* sigue siendo limitado y poco explorado en aplicaciones reales, lo cual representa una oportunidad para futuras investigaciones.

Por su parte, Pan et al. (2022) analizaron el uso combinado de espectroscopía Raman con

redes neuronales convolucionales (CNN), demostrando cómo el deep learning puede superar limitaciones tradicionales en la interpretación espectral. Su estudio demuestra que estos modelos pueden adaptarse a variaciones sutiles en las señales, mejorando la precisión sin necesidad de preprocesamiento intensivo. Sin embargo, también se señala que su implementación práctica aún presenta desafíos, debido a la alta demanda computacional y la necesidad de grandes volúmenes de datos etiquetados para el entrenamiento.

Luo et al. (2022) presentaron una revisión crítica sobre deep learning en espectroscopía Raman, analizando más de 60 estudios en matrices como alimentos, productos farmacéuticos y materiales. Clasificaron los modelos según su función (clasificación, regresión o detección de anomalías) y destacaron arquitecturas prometedoras como las redes neuronales recurrentes (RNN) y los autoencoders. Su trabajo incluyó una figura que muestra el flujo de trabajo típico en estos modelos (Figura 3.2), con aplicaciones en preprocesamiento, clasificación, regresión y realce de señales.

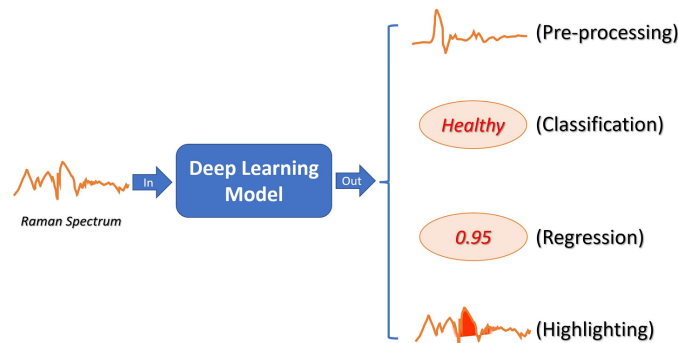


Figura 3.2: Cuatro tipos de aplicaciones de aprendizaje profundo en espectroscopía Raman, según el formato de salida del modelo(Luo et al. (2022)).

Como fortaleza, el deep learning permite automatizar tareas como la selección de características y el preprocesamiento, eliminando pasos intermedios. No obstante, los autores también advierten sobre la falta de estandarización en los protocolos experimentales y de validación cruzada, lo que afecta la reproducibilidad de los modelos entre estudios (Luo et al. (2022)).

En resumen, los estudios revisados evidencian que la integración entre espectroscopía e inteligencia artificial constituye una herramienta poderosa y en expansión, especialmente en la caracterización de alimentos y bebidas; aún enfrenta desafíos prácticos para su implementación a gran escala.

3.5 Datasets

En el contexto del aprendizaje automático y los estudios experimentales, el concepto de dataset (conjunto de datos) es central para la construcción y validación de modelos. De acuerdo con Rosli

et al. (2016), un dataset proveniente de mediciones experimentales se define como “una colección de elementos que incluye uno o más valores de medición”. Estos valores están asociados a métricas que describen características cuantificables de entidades dentro de una población de estudio. A su vez, se requieren metadatos (datos adicionales que ayudan a interpretar otros datos) de medición, definidos como “información sobre las entidades y las métricas”, para garantizar la correcta interpretación del contenido y evitar ambigüedades en el análisis de los datos.

En esta memoria, se hace uso de datasets estructurados generados a partir de mediciones espectroscópicas, las cuales son transformadas y organizadas en tablas que permiten su uso como entrada para modelos de clasificación supervisada. La correcta estructuración de estos datos, incluyendo etiquetas, valores y metadatos contextuales, resulta esencial para asegurar la validez de los resultados obtenidos.

3.6 Normalización de los Espectros

La normalización permite escalar los datos espectrales a un nivel comparable, minimizando los efectos de intensidad asociados a factores no deseados como el tamaño de partícula, la dispersión o diferencias instrumentales. Una técnica común es la Corrección del Efecto Multiplicativo de la Dispersión (MSC), que ajusta cada espectro frente a un espectro de referencia (generalmente el espectro máximo o un blanco), mediante una línea de mínimos cuadrados. Esta corrección es eficaz frente a desplazamientos espectrales por variaciones en el tamaño de partícula, aunque requiere recalcular parámetros si cambia el conjunto de espectros (Zhou et al. (2022); Martelo Vidal (2014a)).

En este proyecto, se empleará la normalización MSC utilizando como blanco un espectro de referencia inicial. Se evaluaron distintos blancos, como agua y etanol, mediante pruebas con el prototipo de mini espectrómetro, sin observarse diferencias en los valores registrados.

3.7 Arbol de modelos y Submodelos

Un árbol de modelos es una estructura jerárquica compuesta por modelos independientes que permite resolver problemas complejos de forma progresiva. Esta estructura divide el problema global en subproblemas más pequeños, los cuales son abordados con modelos especializados en cada etapa. Este enfoque es útil para tareas de clasificación o predicción con bajo volumen de datos o cuando las clases presentan características muy similares, lo que dificulta su diferenciación (Silla and Freitas (2011); An et al. (2021)).

El objetivo principal de un árbol de modelos es que cada submodelo se enfoque en un aspecto específico de la tarea, facilitando la gestión de problemas complejos mediante una división modular.

En lugar de entrenar un único modelo para toda la clasificación, se utilizan submodelos en distintos niveles del árbol, cada uno con una función particular, lo que permite optimizar los resultados generales. Esta estrategia mejora tanto la precisión como la eficiencia del sistema, haciendo que sea más escalable y comprensible (Han and Guo (2024); An et al. (2021)).

Sin embargo, uno de sus principales desafíos es que, si bien su organización jerárquica puede mejorar la clasificación en conjuntos complejos, un árbol con demasiados niveles o submodelos puede volverse difícil de interpretar. Además, cada submodelo puede requerir entrenamiento y gestión por separado, aumentando la complejidad computacional y el tiempo de procesamiento (Silla and Freitas (2011); Han and Guo (2024)).

3.7.1 Organización Jerárquica del Árbol de Modelos

La estructura del árbol se organiza en niveles, donde cada nivel representa una etapa de predicción o separación. En cada uno, el modelo toma una decisión que reduce el número de clases a considerar en el siguiente nivel. Este enfoque jerárquico permite mejorar la eficiencia y escalabilidad del proceso de clasificación en contextos donde se requiere simplicidad interpretativa o cuando se trabaja con conjuntos de datos de tamaño limitado o con clases muy similares entre sí (Maimon and Rokach (2014)).

3.7.2 Submodelos: Componentes del Árbol de Modelos

Los submodelos son los componentes individuales de un árbol de modelos encargados de tareas específicas de clasificación o predicción. En contextos de inteligencia artificial supervisada, cada submodelo se entrena con datos etiquetados y se orienta a predecir un aspecto particular de la muestra. Tienen un propósito único y están adaptados a la tarea que resuelven (Silla and Freitas (2011)).

Existen dos tipos de submodelos que pueden usarse de forma independiente o combinada para lograr una clasificación más precisa y eficiente: submodelos separadores y clasificadores (An et al. (2021)).

Submodelos Separadores

Dividen los datos en grupos más homogéneos y se aplican comúnmente en los primeros niveles del árbol para establecer categorías amplias.

Submodelos Clasificadores

Asignan etiquetas específicas a los datos ya separados por los submodelos anteriores. Se emplean en niveles más profundos para clasificaciones más detalladas.

El diseño modular de los submodelos permite flexibilidad, ya que se pueden actualizar partes del sistema sin reentrenar todo el modelo. Esto resulta útil cuando los datos cambian o se expanden. Sin embargo, el uso excesivo de submodelos puede aumentar los requisitos computacionales y generar inconsistencias jerárquicas si no se gestionan adecuadamente las dependencias entre niveles (An et al. (2021); Han and Guo (2024)).

Investigaciones previas destacan ventajas y desafíos de este enfoque. Silla and Freitas (2011) realizaron una revisión sobre clasificación jerárquica aplicada a distintos dominios (como textos, bioinformática e imágenes), concluyendo que este enfoque mejora la modularidad, la interpretabilidad y la precisión al respetar relaciones padre-hijo entre clases. No obstante, también identificaron desventajas como la propagación de errores entre niveles y la complejidad del entrenamiento al usar múltiples clasificadores.

An et al. (2021) implementaron un modelo jerárquico para detectar enfermedades en plantas. Utilizaron un esquema de separación y posterior clasificación, lo que facilitó trabajar con conjuntos de datos reducidos y clases similares. Su enfoque mejoró significativamente la precisión respecto a modelos planos, destacando la reutilización de conocimiento y la escalabilidad. Sin embargo, también se enfrentaron a desafíos como el costo computacional y la dificultad de mantener consistencia entre niveles profundos.

Por su parte, Han and Guo (2024) propusieron un sistema jerárquico para clasificar enfermedades en hojas leñosas. Alcanzaron una precisión del 85% utilizando una estructura de árbol entrenada por niveles. Su modelo fue eficiente para clasificar clases interrelacionadas, aunque dependía fuertemente de la disponibilidad de etiquetas bien definidas en cada nivel, lo que limitaba su aplicabilidad en contextos sin taxonomías claras.

3.7.3 Pipeline

En el aprendizaje automático, un pipeline es una forma de organizar y automatizar todas las etapas que se realizan desde que se recibe un conjunto de datos hasta que se entrena un modelo y se obtienen resultados. Un pipeline funciona como una línea de montaje: los datos entran por un extremo, pasan por una serie de pasos definidos (como limpieza, transformación, entrenamiento del modelo, etc.), y al final se obtiene una predicción o resultado. Esta estructura permite que todo el proceso se realice de forma reproducible, ordenada y consistente (Graßberger et al. (2023b); Graßberger et al. (2023a)). Técnicamente, es un algoritmo importable que permite esta secuencia de funciones.

Un pipeline no solo organiza estos pasos, sino que permite repetir el proceso automáticamente sobre distintos datos, sin necesidad de programar todo desde cero cada vez (Grafberger et al. (2023b)).

3.8 Reducción Dimensional: PCA y LDA en Espectroscopía

El análisis de grandes volúmenes de datos espectrales puede resultar complejo debido a su alta dimensionalidad. Por ejemplo, un espectro UV-Vis puede contener entre 200 y 800 puntos de absorbancia, lo que genera demasiadas variables para un modelo de IA, especialmente si el número de muestras es bajo. Para abordar esto, se aplican técnicas de reducción dimensional que permiten disminuir el número de variables sin perder información clave.

Este proceso busca simplificar los datos, conservando la información relevante para mejorar el rendimiento de los modelos. En espectroscopía, dos técnicas comunes para este fin son el Análisis de Componentes Principales (PCA) y el Análisis Lineal Discriminante (LDA) (Martelo Vidal (2014a)).

3.8.1 Análisis de Componentes Principales (PCA)

El PCA es una técnica no supervisada que reduce la dimensionalidad de los datos al proyectarlos en un conjunto de componentes ortogonales que capturan la mayor variabilidad posible. Cada componente principal (PC) es una combinación lineal de las variables originales, ordenadas según la cantidad de varianza que explican. El primer componente representa la dirección de mayor variabilidad, seguido por el segundo, y así sucesivamente (Abdi and Williams (2010)).

Esto permite representar, por ejemplo, un espectro con cientos de variables mediante solo dos o tres componentes principales, sin pérdida significativa de información. Esta reducción facilita la visualización de agrupaciones y diferencias entre muestras, lo que resulta útil en el análisis exploratorio de datos.

En el presente trabajo, PCA permite proyectar las muestras de vino en un espacio reducido, donde se agrupan naturalmente según su cepa, facilitando así una clasificación inicial y revelando relaciones estructurales entre las muestras sin necesidad de supervisión previa.

En espectroscopía, ha sido aplicado exitosamente para clasificar muestras de vino, como en el estudio de Cozzolino et al. (2008), donde se logró reducir la cantidad de datos espectrales y resaltar patrones relevantes. Wang et al. (2021) aplicaron PCA para reducir 18 variables (como densidades, tamaños de partícula y porosidades) a solo dos componentes que explicaron el 56,1% de la varianza. El primer componente agrupó variables relacionadas con tamaño y fluidez, mientras que el segundo reunió descriptores de densidad y porosidad. Esta reducción sirvió como base para modelos posteriores (como PLS) en estudios de comportamiento reológico de muestras. A continuación, se

presentan gráficas PCA obtenidas en este trabajo, útiles para visualizar agrupaciones entre cepas.

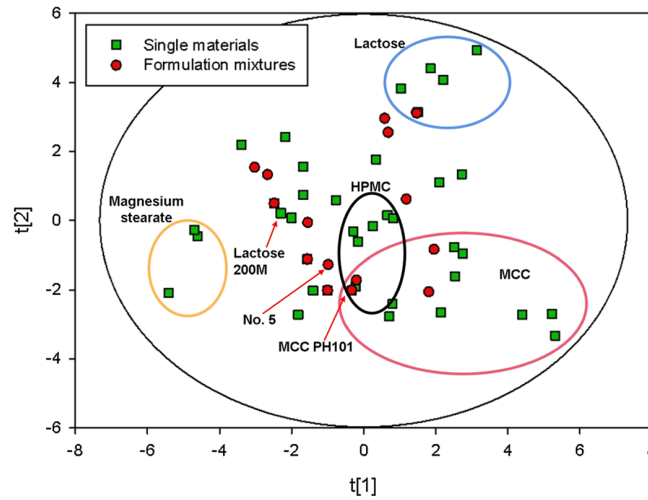


Figura 3.3: Gráfico de cargas de PC1 vs. PC2 obtenido mediante PCA sobre las propiedades físico-químicas de excipientes y formulaciones de granulación. Los puntos verdes representan descriptores originales, y los círculos resaltan agrupaciones según la contribución de cada descriptor al primer y segundo componente principal (Wang et al. (2021)).

3.8.2 Análisis Lineal de Discriminantes (LDA)

El Análisis Lineal de Discriminantes (LDA) es una técnica supervisada que requiere conocer la clase a la que pertenece cada muestra para entrenar el modelo. Su objetivo es encontrar combinaciones lineales de características espectrales que mejor separen las distintas clases (Martínez and Kak (2001)). A diferencia de PCA, que se enfoca en maximizar la variabilidad total, LDA busca maximizar la separación entre clases, lo que lo hace especialmente útil en tareas de clasificación (Bhowmik et al. (2021)).

En el análisis de cepas de vino, se ha demostrado la utilidad de LDA para separar clases con características espectrales similares. García (2017) aplicó esta técnica en combinación con espectroscopía UV-Vis para mejorar la clasificación de vinos. Por otro lado, Cetó et al. (2013) la utilizaron junto a una lengua electrónica para discriminar tipos de cerveza, alcanzando una tasa de clasificación del 81,9% y una especificidad promedio del 96,4%.

No obstante, LDA presenta limitaciones, como su sensibilidad a la multicolinealidad y la necesidad de que las clases tengan matrices de covarianza similares. Si esta condición estadística no se cumple, su aplicabilidad puede verse restringida.

Bhowmik et al. (2021) compararon PCA y LDA sobre 13 coeficientes cepstrales en frecuencia Mel (MFCC), demostrando que LDA lograba una mejor separación entre clases de voz. Este resultado respaldó su uso previo al entrenamiento de modelos K-Nearest Neighbors en diagnósticos

de enfermedades vocales.

A continuación, se presenta una de las gráficas resultantes de este trabajo, como ejemplo de visualización de un gráfico LDA.

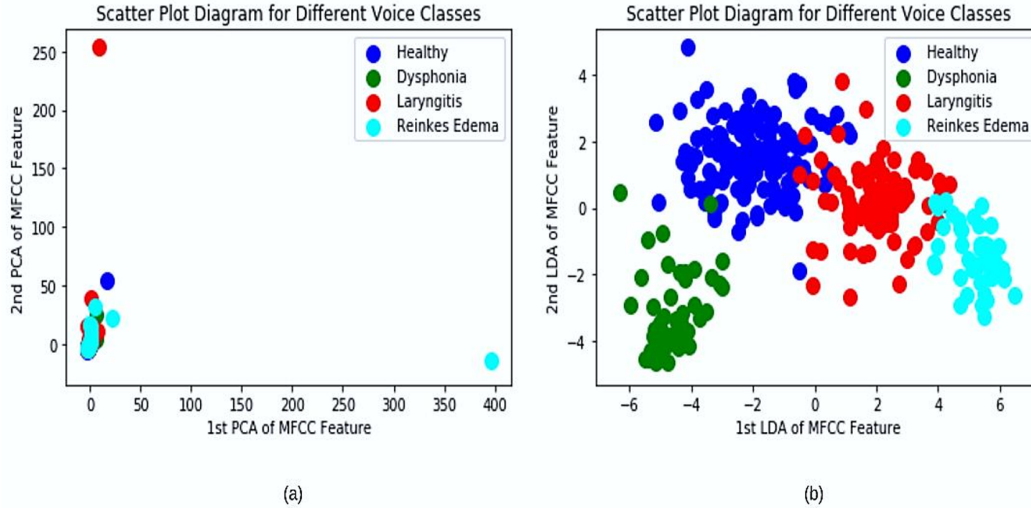


Figura 3.4: Diagramas de dispersión 2D de las dos componentes más importantes para la reducción de dimensionalidad de los coeficientes MFCC extraídos del conjunto SVD: (a) PCA muestra agrupamientos solapados entre Healthy, Dysphonia, Laryngitis y Reinke's Edema. (b) LDA mejora notablemente la separación de clases, indicando componentes más discriminantes para cada patología (Bhowmik et al. (2021)).

3.9 Aumentación de Datos (Data Augmentation) en Modelos IA

Una de las principales dificultades al trabajar con modelos de inteligencia artificial es la escasez de datos, lo que puede afectar su capacidad de aprendizaje. En estos casos, la técnica de aumentación de datos (*data augmentation*) se presenta como una solución efectiva para generar nuevas muestras artificiales a partir de datos existentes, mejorando así el desempeño del modelo sin necesidad de recolectar nuevos datos reales (Zhou et al. (2022)).

Desde una perspectiva teórica, la aumentación actúa como una forma de regularización espectral implícita, ya que modifica la distribución de los valores propios de la matriz de covarianza. Esto puede beneficiar tanto a modelos subparametrizados como sobreparametrizados, aunque requiere un diseño cuidadoso para evitar un aumento del sesgo (Lin et al. (2024); Zhou et al. (2022)). Entre las técnicas más comunes destacan (Zhou et al. (2022)):

- Ruido gaussiano, que simula variaciones naturales sumando una perturbación que consiste en sumar ruido aleatorio con distribución normal a cada punto del espectro. La desviación

estándar (σ del ruido se ajusta entre un porcentaje del valor original para cada espectro.

- Ruido *salt and pepper*, el cual introduce valores atípicos en un pequeño porcentaje de puntos del espectro, reemplazándolos por los valores mínimo y máximo posibles. Imita defectos impulsivos o errores puntuales en la adquisición de datos.
- *Cutout*, *random masking* e interpolación entre muestras, que ocultan o modifican segmentos de datos.
- Alteraciones controladas en condiciones experimentales (como temperatura o iluminación), útiles para enriquecer el espacio de entrada sin modificar la clase subyacente (Lin et al. (2024); Luo et al. (2022)).

No obstante, la efectividad de estas técnicas depende del contexto. En aplicaciones clínicas, por ejemplo, se requiere que las transformaciones sean fisiológicamente plausibles. Zhou et al. (2022) advierten que ciertas alteraciones, como la modificación extrema de amplitud o filtros de color, pueden degradar el rendimiento al introducir características irreales.

El estudio de Lin et al. (2024) propone un marco formal para evaluar el impacto espectral de estas técnicas, mostrando que algunas reducen más la varianza que el sesgo, lo que mejora la generalización en clasificación (aunque no siempre en regresión). Por su parte, Zhou et al. (2022) demostraron que transformaciones suaves, como *flip* horizontal (inversión de la señal en el eje temporal) o PCA sobre MFCC (coeficientes cepstrales en las frecuencias de Mel, una representación común en análisis de señales acústicas o espectrales), respetan mejor las restricciones fisiológicas en espectrogramas cardíacos.

Finalmente, en el ámbito de la espectroscopía Raman, Luo et al. (2022) realizaron una revisión exhaustiva donde identifican la aumentación de datos como una estrategia clave para mejorar el rendimiento de modelos en escenarios con escasez de muestras. Destacan que, si se diseñan adecuadamente, estas transformaciones permiten una mejor generalización de redes neuronales profundas. No obstante, una mala selección puede introducir sesgos o distorsiones que comprometan la calidad del aprendizaje.

3.10 Miniaturización de Espectrómetros

La miniaturización de espectrómetros se ha convertido en una de las principales tendencias en el desarrollo de instrumentación analítica, permitiendo su integración en aplicaciones portátiles, móviles y de campo. Este avance ha sido posible gracias al desarrollo de tecnologías ópticas, microelectrónicas y MEMS (sistemas microelectromecánicos), que han permitido reducir significativamente el tamaño de los dispositivos sin comprometer su rendimiento (Guan et al. (2023)).

A pesar de su diseño compacto, los espectrómetros miniaturizados mantienen componentes esenciales: fuente de luz (generalmente LEDs o lámparas halógenas), sistemas de separación espectral (rejillas de difracción, filtros Fabry-Pérot o interferómetros Michelson miniaturizados) y detectores (como fotodiodos InGaAs o CMOS). Esta configuración ha permitido su aplicación en espectroscopía NIR (infrarrojo cercano) para agricultura, alimentos, farmacología y monitoreo ambiental, ofreciendo análisis no destructivos, rápidos y en tiempo real (Beć et al. (2021)).

Una de sus principales ventajas es la portabilidad, que facilita su uso en terreno. Además, su bajo consumo energético y la posibilidad de operar de forma autónoma o conectados a smartphones los hace más accesibles que los espectrómetros de banco. También permiten análisis "in situ" (en el sitio) que agilizan la toma de decisiones (Beć et al. (2021)).

Sin embargo, presentan limitaciones. Su resolución espectral suele ser reducida, lo que dificulta la diferenciación de señales cercanas. También pueden verse afectados por interferencias ambientales. Además, presentan una sensibilidad reducida en condiciones de baja concentración, lo cual limita su capacidad de detección para compuestos presentes a niveles traza. Algunos modelos tienen restricciones en el rango espectral, lo cual limita su uso en análisis más amplios (Guan et al. (2023)).

Beć et al. (2021) realizaron una revisión detallada sobre espectrómetros NIR miniaturizados, destacando cómo tecnologías como interferómetros Michelson integrados en MEMS, filtros variables lineales (LVF) y máscaras de Hadamard han permitido una reducción efectiva de tamaño sin perder funcionalidad. Señalan que sensores como NeoSpectra, MicroNIR o SCiO operan eficientemente en el rango de 900–2500 nm y pueden ser usados en el bolsillo o integrados a teléfonos. No obstante, requieren mayor supervisión que los equipos de laboratorio, debido a la necesidad de calibración frecuente y sensibilidad a variaciones espectrales.

Por su parte Guan et al. (2023) analizaron el impacto de estos dispositivos en nuevas áreas de aplicación, destacando su uso en monitoreo de salud, evaluación de biomarcadores y tecnologías móviles. También advierten sobre limitaciones como sensibilidad reducida o susceptibilidad a interferencias, lo que ha impulsado el desarrollo de estrategias de compensación digital y reconstrucción espectral mediante inteligencia artificial.

3.11 Evaluación del Rendimiento de Modelos de IA

Evaluar el rendimiento de los modelos de IA es fundamental para determinar qué tan bien están funcionando frente a una tarea específica. Esto se logra mediante métricas cuantitativas que permiten comparar diferentes enfoques y optimizar el desempeño del sistema. En contextos con conjuntos de datos reducidos, como es común en espectroscopía, estas métricas aseguran que el modelo pueda generalizar correctamente a nuevos datos. En este trabajo, se utilizarán principalmente la precisión y el F1 *score*, junto con la matriz de confusión, para evaluar el desempeño de los modelos. Estas

herramientas no solo cuantifican aciertos y errores, sino que también facilitan la interpretación de resultados en contextos que requieren alta confiabilidad (Gardeli et al. (2025)).

3.11.1 Precisión (Accuracy)

La precisión mide la proporción de predicciones correctas realizadas por el modelo, es decir, cuántas de las muestras etiquetadas como positivas lo eran realmente. Se calcula como el número total de predicciones correctas dividido por el total de predicciones (Han and Guo (2024)).

$$Accuracy = \frac{Predicciones\ correctas}{Total\ de\ predicciones} = \frac{TP + TN}{Total} \quad (3.5)$$

Donde:

- TP (True Positives) = verdaderos positivos (casos positivos clasificados como positivos).
- TN (True Negative) = verdaderos negativos (casos negativos clasificados como negativos).

Esta métrica es útil cuando las clases están equilibradas, es decir, cuando las distintas clases tienen aproximadamente el mismo número de muestras, y se desea minimizar los falsos positivos (Chicco and Jurman (2020)). Por ejemplo, en sistemas de detección de fraude, una baja precisión podría significar que se etiquetan erróneamente muchas transacciones legítimas como fraudulentas.

3.11.2 F1 score

El F1 *score* combina precisión y recuperación (*recall*) en una sola métrica, calculada como la media armónica entre ambas. Es particularmente útil cuando se trabaja con clases desbalanceadas, ya que penaliza fuertemente los valores extremos (Han and Guo (2024)).

$$F1\ score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (3.6)$$

Donde:

- Precision = $\frac{TP}{TP+FP}$ → Casos positivos clasificados correctamente entre todos los clasificados como positivos
- Recall = $\frac{TP}{TP+FN}$ → Casos positivos clasificados correctamente entre todos los casos realmente positivos.

Un F1 *score* alto indica que el modelo tiene tanto alta precisión como alta sensibilidad; si una de ellas es baja, el F1 será bajo, incluso si la otra es excelente.

Esto lo convierte en una métrica relevante para escenarios donde se necesita un equilibrio entre falsos positivos y falsos negativos (Chicco and Jurman (2020)). Por ejemplo, en la clasificación de cepas con pocas muestras, un modelo con alta precisión pero baja sensibilidad podría ignorar las clases minoritarias. El *F1 score* permite evaluar si el modelo realmente está reconociendo todas las clases presentes.

3.11.3 Matriz de Confusión

La matriz de confusión es una herramienta que permite evaluar detalladamente el rendimiento de un modelo de clasificación. Indica cuántas veces el modelo predijo correctamente o confundió una clase con otra, ayudando no solo como métrica, sino también como guía explicativa para detectar fallas estructurales en el modelo (Manai et al. (2024)).

Estructura de la Matriz de Confusión

Si se tiene un problema con dos clases: Positiva y Negativa. La matriz de confusión tiene 4 elementos principales:

Tabla 3.1: Ejemplo de una Matriz de Confusión

	Predicción Positiva	Predicción Negativa
Real Positiva	True Positives (TP)	False Negatives (FN)
Real Negativa	False Positives (FP)	True Negatives (TN)

- True Positives (TP): El modelo predijo positivo y la clase real era positiva.
- False Positives (FP): El modelo predijo positivo pero la clase real era negativa.
- True Negatives (TN): El modelo predijo negativo y la clase real era negativa.
- False Negatives (FN): El modelo predijo negativo pero la clase real era positiva.

Manai et al. (2024) emplean esta matriz para identificar fallos específicos del modelo y proponer mejoras en la selección de características, ajuste de parámetros y arquitectura, logrando mejoras notables en métricas como precisión, *recall* y *F1-score*. En su experimento con el conjunto UNSW-NB15, la precisión aumentó de 81,9% a 89,6% tras analizar los errores con la matriz de confusión. El estudio destaca su utilidad para ajustar el modelo de forma guiada y mejorar la transparencia en sistemas críticos, considerando los *trade-offs* entre falsos positivos y negativos según la aplicación.

[Zhou et al. \(2022\)](#), por su parte, aplican estas métricas a la clasificación de señales espectrales. Mediante el análisis de diferentes arquitecturas de redes neuronales y técnicas de aumentación, utilizan la matriz de confusión para identificar errores sistemáticos entre clases específicas. Esto es clave en contextos con clases desbalanceadas o de baja representación, donde la precisión por sí sola puede ser engañosa.

3.12 Overfitting o Sobreajuste

Según [Salman and Liu \(2019\)](#), el sobreajuste puede entenderse como una especie de "adicción" del modelo a los detalles específicos del conjunto de entrenamiento, incluyendo sus errores, inconsistencias o incluso ruido aleatorio. El modelo no solo captura las relaciones verdaderas, sino que también incorpora fluctuaciones no representativas del fenómeno real que se quiere modelar. Como consecuencia, aunque el error en el entrenamiento continúa disminuyendo, el error en el conjunto de validación comienza a aumentar: este es el punto en que se reconoce que el modelo ha empezado a sobreajustarse.

En su análisis, los autores observan que el modelo comienza a estabilizar su error en muestras correctamente clasificadas, pero incrementa su pérdida en aquellas difíciles o mal clasificadas. Este comportamiento indica que el modelo empieza a esforzarse demasiado por "complacer" a casos atípicos o ambiguos, afectando su aprendizaje global.

[Ghojogh and Crowley \(2019\)](#) complementan esta visión al describir el sobreajuste como resultado de un desequilibrio entre la complejidad del modelo y la cantidad o calidad de los datos. Un modelo muy complejo con pocos datos tiende a aprender relaciones que no se sostienen fuera del conjunto de entrenamiento. Como consecuencia, funciona muy bien con los datos ya vistos, pero falla al enfrentar datos nuevos, perdiendo capacidad de generalización.

3.13 Estrategias de Validación de Modelos

Las estrategias de validación permiten estimar la capacidad de un modelo de inteligencia artificial para generalizar a datos nuevos no vistos durante el entrenamiento. Su objetivo es verificar que el modelo no se ajuste en exceso a los datos de entrenamiento (evitando el sobreajuste u overfitting) y que sea capaz de mantener un buen desempeño ante nuevas muestras ([Pérez-Planells et al. \(2015\)](#); [Granados-Tavera et al. \(2019\)](#)).

Entre las técnicas más empleadas se encuentran los métodos de validación cruzada, siendo K-Fold (*fold* siendo subconjuntos en los que se divide el conjunto de datos para validación cruzada) y Leave-One-Out las utilizadas en este proyecto. Dada la limitada cantidad de datos disponibles, también se considera una validación basada en la separación por muestras físicas.

3.13.1 Validación Cruzada K-Fold

La validación K-Fold consiste en dividir el conjunto de datos en K subconjuntos de tamaño similar. En cada iteración, uno de los subconjuntos se utiliza como conjunto de validación y los K-1 restantes para entrenamiento. El proceso se repite K veces, utilizando cada bloque como validación una vez, y el rendimiento final se obtiene promediando los resultados de todas las iteraciones (Pérez-Planells et al. (2015)).

En su estudio, Pérez-Planells et al. (2015) compararon K-Fold con el método hold-out en la evaluación de parámetros biofísicos sobre datos hiperespectrales. Concluyeron que K-Fold con $K = 10$ entrega estimaciones más estables y representativas, ya que todos los datos se usan para entrenamiento y validación, a diferencia del método hold-out, que puede generar resultados inconsistentes debido a la aleatoriedad en la partición.

3.13.2 Validación Cruzada Leave-One-Out (LOO-CV)

LOO-CV es un caso especial de validación cruzada donde el número de subconjuntos equivale al total de muestras. En cada iteración se usa una sola muestra como validación y el resto para entrenamiento. Esta técnica maximiza el uso de los datos disponibles, lo que la hace adecuada en contextos con muy pocas muestras. Sin embargo, es más demandante computacionalmente, ya que el modelo se entrena tantas veces como muestras existan (Granados-Tavera et al. (2019)).

El estudio de Granados-Tavera et al. (2019), enfocado en modelos QSAR (*Quantitative Structure-Activity Relationship*, una técnica que predice la actividad de compuestos químicos a partir de su estructura), comparó distintas técnicas de validación, incluyendo LOO-CV. El autor señaló que este método puede generar resultados optimistas si el modelo no generaliza bien, ya que tiene una varianza alta. Sin embargo, también permite comparar configuraciones de modelos de forma objetiva, usando métricas como el coeficiente de determinación (R^2) o el error cuadrático medio (RMSE), lo que ayuda a elegir la configuración más adecuada.

3.13.3 Separación por Muestras Físicas

Esta técnica se basa en utilizar conjuntos de datos recolectados en diferentes momentos o condiciones experimentales, evitando que una misma muestra sea empleada tanto para el entrenamiento como para la evaluación. De este modo, se asegura que el modelo generalice correctamente a nuevas observaciones, sin estar influenciado por información previamente vista durante el entrenamiento. Este enfoque busca reducir el sobreajuste (overfitting) y entregar una evaluación más honesta del rendimiento del modelo (Zhang and Wang (2023)).

Esta estrategia resulta especialmente útil cuando se trabaja con datos reales y en contextos donde se desea una estimación del desempeño más cercana a las condiciones de uso final. No

obstante, su aplicación puede presentar desafíos logísticos, sobre todo cuando las muestras son escasas, ya que separar adecuadamente los datos puede afectar la representatividad o la distribución de las clases (Zhang and Wang (2023); Feng et al. (2021)).

Este enfoque ha sido aplicado en estudios recientes, como el de (Feng et al. (2021)), quienes utilizaron muestras recolectadas en diferentes momentos para validar modelos de clasificación de alimentos fermentados. En su estudio, evitaron deliberadamente que una misma muestra se usara tanto para entrenar como para validar, lo que les permitió evaluar el desempeño del modelo bajo condiciones más realistas y con mayor generalización.

En este proyecto, debido a la baja disponibilidad de datos y a la naturaleza experimental del estudio, se optó por aplicar validación mediante separación por muestras físicas. Adicionalmente, se complementó esta validación con una validación cruzada tipo *k-folds*, permitiendo una evaluación más robusta del rendimiento del modelo al aprovechar diferentes subconjuntos de datos en cada iteración.

Capítulo 4

Materiales y Métodos

Este proyecto tiene como objetivo desarrollar y validar una plataforma basada en Inteligencia Artificial (IA) para la detección analítica de moléculas en matrices líquidas y sólidas en tiempo real, utilizando un espectrómetro visible de tamaño reducido (MESPECTRO-VIS). Para evaluar la reproducibilidad y repetitividad de las mediciones espectroscópicas, se implementó un diseño experimental con 4 prototipos de espectrómetro portátil, nominalmente idénticos, todos basados en el sensor óptico SparkFun AS7265X y empleando la misma celda óptica. Esta celda corresponde a una estructura circular de vidrio, en la cual se introduce la muestra líquida.

Para observar la disposición de la instrumentación, véase la Figura 4.1. Este enfoque permite analizar las variaciones inducidas por las celdas y verificar la consistencia entre los equipos.



Figura 4.1: Prototipo actual de espectrómetro móvil.

Es importante destacar que esta metodología fue desarrollada bajo la supervisión de Pablo Jaramillo  (Jaramillo (2025)), Ingeniero en Computación, quien diseñó la arquitectura de archivos utilizada en las experimentaciones, análisis y conclusiones posteriores.

4.1 Equipos y Herramientas Generales

- 4 mini espectrómetro integrados con un módulo de sensor óptico AS7265X que cubre el rango de longitud de onda de 410nm a 940nm con 18 canales (longitudes de onda con ancho de 20 nm aprox).
- 4 computadores portátiles (siempre conectados a la corriente) con software Espectro IA, instalado en cada computador. Además de 4 cables USB a micro USB.
- 2 celdas circulares de vidrio, con su respectivo porta celda de plástico.

Además, se implementó un proceso de evaluación de consistencia entre sensores, comparando las lecturas bajo condiciones similares para asegurar la fiabilidad de los datos recolectados. Se seleccionaron aquellos sensores y celdas con mejor comportamiento espectral para las mediciones finales. Este paso fue crucial debido a la variabilidad inherente en sensores y componentes ópticos.

4.1.1 Materiales para Medición de Variables Fisicoquímicas

- 2 refractómetros, uno escalado para °Brix y otro para °Alcohol (v/v).
- pHmetro digital con electrodo de sonda (modelo AD1030, de la marca Ltech Chile).
- Conductímetro portátil con electrodo de sonda (modelo, de la marca Ltech Chile).
- Soluciones de calibración pH de 4,01, 7,01 y 10,1.
- Solución de calibración conductímetro de 1413 $\mu\text{S}/\text{cm}$.

4.2 Recolección y Almacenamiento de Muestras de Vino

Las muestras utilizadas en este estudio provienen directamente del viñedo El Escorial, ubicado en San Felipe, región de Valparaíso, el cual colabora activamente con el proyecto. Este viñedo, de menor escala en comparación con productores industriales, limita la cantidad de muestras disponibles para el análisis. Se recolectaron muestras de diversas cepas, y en algunos casos solo se obtuvo una muestra física por cepa.

El proceso de recolección garantiza condiciones homogéneas: todas las muestras fueron tomadas de forma estandarizada y etiquetadas adecuadamente para su posterior análisis. Las mediciones se realizaron en un laboratorio controlado, y no se contempló la variabilidad asociada a condiciones ambientales como temperatura, iluminación o entorno “in situ” (en el sitio).

Tras la toma de datos, las muestras fueron almacenadas a 4°C en frascos ámbar de vidrio, con el fin de evitar la degradación de compuestos sensibles y preservar su integridad hasta el análisis químico. Cada frasco fue etiquetado con la cepa, tiempo en barril y momento de recolección, asegurando la trazabilidad de los resultados.

A continuación, se presenta una tabla con las cepas y el número de muestras recolectadas en El Escorial. Cada cepa fue recolectada por duplicado, resultando en dos frascos por variedad, y cada muestra corresponde a un volumen aproximado de 20 mL.

Tabla 4.1: Cepas y cantidad de muestras obtenidas de Viñedo El Escorial

Cepa de Vino	N° de muestras
Syrah	6
Malbec	5
Petite Syrah	4
Zinfandel	4
Syrah - Malbec	3
Cabernet Franc	2
Muscat	2
Petite Verdot	2
Sangiovese	2
Cabernet Sauvignon	1
Carginan	1
Carmenere	1
Chenin Blanc	1
Petite Noir	1
Temopranillo	1

Cómo se observa en la Tabla [4.1](#), algunas cepas disponen de una sola muestra física. En el contexto de algoritmos de inteligencia artificial, este volumen de datos es limitado para alimentar el modelo, ya que cada muestra aporta únicamente un conjunto de lecturas espectrales. Por lo tanto, el tamaño del dataset crece en proporción directa (1:1) al número de muestras físicas, y no a la cantidad de mediciones replicadas.

El análisis empírico con modelos supervisados evidenció que esta escasez de muestras reduce considerablemente la capacidad del modelo para aprender características generales de cada tipo de vino. Por ello, se optó por incorporar lecturas de vinos comerciales específicamente de las cepas Syrah y Malbec, adquiridos en el mercado con el objetivo de aumentar la representación de estas variedades que presentaban mayores dificultades de separación en el modelo.

No obstante, estos vinos industriales no son necesariamente representativos de los vinos artesanales del viñedo, y podrían introducir sesgos en el modelo. En la elaboración industrial se incorporan prácticas como la mezcla de lotes, procesos de homogeneización, y la adición de aditivos

enológicos para garantizar un perfil constante y una producción uniforme. Estas intervenciones pueden alterar significativamente los marcadores químicos del vino en comparación con aquellos producidos bajo mínima intervención artesanal (Laffort (2021)).

Finalmente, las muestras actuales provienen de la etapa de maduración (previa al embotellamiento). Se espera, de manera condicional y dentro de una posible línea de trabajo futura asociada al proyecto FONDEF, realizar una segunda campaña de muestreo durante la vendimia en momentos estratégicos, con el fin de capturar distintos estados de madurez del vino. Esta planificación se enmarca en el contexto del Proyecto FONDEF actual y sus proyecciones, aunque su ejecución dependerá de la continuidad y extensión del proyecto.

4.3 Impacto en el modelamiento de IA debido a la baja cantidad de datos

El bajo volumen de datos disponibles para el entrenamiento de modelos impone múltiples restricciones en la selección de algoritmos, validación del rendimiento y diseño general del sistema de clasificación. Por ello, se descarta el uso de modelos complejos como redes neuronales profundas, que, si bien poseen alta capacidad de representación, requieren grandes volúmenes de datos y presentan alto riesgo de sobreajuste.

Esta problemática se agrava cuando la distribución de clases es desbalanceada o algunas cepas están representadas por un número muy reducido de muestras. La validación cruzada también se vuelve menos confiable en este contexto, ya que cada subconjunto de datos tiene un alto impacto en la métrica final, dificultando una evaluación precisa del rendimiento del modelo en datos no vistos.

Además, con clases desbalanceadas se complica la definición de fronteras entre cepas, especialmente cuando algunas tienen muchas más muestras que otras, lo que puede sesgar el modelo hacia aquellas con mayor representación. Análisis exploratorios como PCA muestran que ciertas cepas forman grupos espectrales solapados, lo que indica una alta similitud entre ellas y dificulta su distinción precisa.

Para mitigar los efectos de esta limitación, se adoptaron tres estrategias específicas:

4.3.1 Aumentación de datos por ruido

Se aplicó ruido gaussiano y *salt and pepper* a los espectros, generando variaciones artificiales que simulan nuevas mediciones sin alterar significativamente la información original. Esto incrementa la diversidad del dataset y preserva las características esenciales de cada muestra.

4.3.2 Clasificación jerárquica

En lugar de entrenar un único modelo global, se implementó un árbol de clasificación con etapas intermedias, lo que permite abordar el problema de forma modular y mejorar la precisión. Si bien este enfoque introduce dificultades adicionales (como entrenar y validar múltiples submodelos), es útil para reducir la complejidad del problema. No obstante, implica mayor carga computacional, riesgo de propagación de errores entre etapas y desafíos en la definición de criterios de agrupamiento, por lo que su uso no es común en la literatura. Aun así, en contextos específicos como este, resulta beneficioso.

4.3.3 Inclusión de propiedades fisicoquímicas

Se incorporaron variables como pH, °Brix* y conductividad, que complementan la información espectral y mejoran la capacidad del modelo para distinguir entre cepas, aportando mayor robustez y precisión en las predicciones.

4.4 Recolección de Datos

Tras la recolección y almacenamiento, cada muestra fue introducida en una celda de medición conectada al prototipo del mini espectrómetro de rango visible desarrollado en este proyecto. La medición se realizó mediante el software asociado, que registró los valores de "count" por longitud de onda y los almacenó en archivos `.csv`.

Cabe destacar que el dispositivo no entrega absorbancia ni transmitancia directamente, sino valores de intensidad cruda (*count*), que luego se transforman mediante ecuaciones teóricas (ver Apartado 3).

Se distinguieron dos tipos de muestras:

- **Blancos:** necesarios para calcular absorbancia o transmitancia; medidos tres veces por cada dispositivo.
- **Reales:** muestras de vino, medidas una única vez en cada dispositivo.

Inicialmente, los archivos CSV presentaban formato transpuesto, lo que generó inconsistencias. Para resolverlo, se desarrolló una clase en Python que automatiza la lectura, corrige el formato, elimina datos nulos y verifica la integridad de los archivos, estandarizando así el flujo de procesamiento posterior.

Tabla 4.2: Ejemplo estructura dataset utilizado.

Etiqueta 1	Etiqueta 2	Etiqueta 3
Data	Data	Data
Data	Data	Data

4.5 Balance de conjunto de prueba

Inicialmente se evaluó la posibilidad de aplicar balanceo también al conjunto de entrenamiento. Sin embargo, se observó que reducir la variabilidad perjudicaba la robustez y generalización de los submodelos. Por tanto, se decidió aplicar balance únicamente al conjunto de prueba, manteniendo la distribución original durante el entrenamiento.

Este balance se implementó mediante un procedimiento de poda (`prune_groups`), siguiendo la siguiente lógica:

1. Detección de grupos: Se identifica cada categoría (cepa) en el conjunto de prueba y se calcula el número de muestras por grupo.
2. Determinación del tamaño objetivo: Se toma como referencia el grupo con menor cantidad de muestras, que define el nuevo tamaño para todas las clases (salvo que se especifique otro valor).
3. Eliminación aleatoria: Para cada grupo con más muestras que el mínimo, se eliminan aleatoriamente (sin reemplazo) muestras hasta igualar el tamaño objetivo.
4. Propagación a datos correlativos: Si existen tablas o *DataFrames* vinculados (como espectros o variables fisicoquímicas), se aplica el mismo filtro para mantener la consistencia entre índices.

Este enfoque garantiza que el conjunto de prueba tenga una distribución equitativa entre clases, evitando que la evaluación favorezca cepas más abundantes. Al mismo tiempo, el conjunto de entrenamiento conserva toda la información original sin perder variabilidad ni introducir sesgos o correlaciones artificiales.

4.6 Distribución Espectral Cepas de vino

A continuación se presenta uno de los pretratamientos más utilizados en el trabajo para el diseño y organización de modelos. Este corresponde a un archivo organizado en módulos que permiten leer, procesar y visualizar datos espectrales de muestras de vino proyectados en espacios de baja dimensión usando LDA, en 2D y 3D.

Este archivo fue desarrollado por mi supervisor Pablo Jaramillo, ingeniero de computación, para su implementación en el análisis del comportamiento de cepas de vino mediante técnicas de reducción de dimensionalidad supervisada. A diferencia de métodos como PCA, aquí se emplea LDA, ya que el objetivo principal es maximizar la separabilidad entre todas las cepas de vino para encontrar las diferencias reales entre los grupos. El flujo de trabajo se puede describir en los siguientes pasos:

1. El archivo importa bases de datos que contienen información espectral, etiquetas de las cepas, y eventualmente propiedades físico-químicas de los vinos (pH, conductividad eléctrica (mS), grados Brix sobrestimados por alcohol ($^{\circ}\text{Brix}^*$) y una estimación de refracción de mezcla alcohólica($\%v/v^*$)). También integra funciones auxiliares para transformar los espectros en diferentes escalas, como transmitir la señal a escala de transmitancia máxima o transmitancia teórica.
2. Luego, permite alternar entre varios modos de preprocesamiento espectral (counts crudos, transmitancia por máximos, o transmitancia por blancos) y decidir si incluir o no las propiedades físico-químicas como parte de las variables de entrada. Además, posibilita filtrar el conjunto de cepas a visualizar, incluyendo o excluyendo muestras comerciales según la necesidad.
3. Posteriormente implementa dos modelos LDA, uno con dos componentes (para visualización 2D) y otro con tres componentes (para visualización 3D). LDA encuentra las proyecciones lineales que maximizan la separación entre las distintas cepas (clases).
4. Utilizando la biblioteca Plotly, genera gráficos de dispersión (scatter plots) en 2D y 3D donde cada punto representa una muestra de vino proyectada según las componentes obtenidas por LDA. Los colores y leyendas permiten identificar fácilmente las distintas cepas incluidas. El usuario puede interactuar mediante consola para agregar o eliminar cepas, alternar modos de preprocesamiento, y actualizar las visualizaciones dinámicamente. Todos los cambios se reflejan en tiempo real en un archivo HTML que se abre automáticamente en el navegador.

En la Figura [4.2](#), presentada a continuación, se muestran los gráficos en 2D y 3D obtenidos al ejecutar el código de visualización. Esto se hizo considerando las 15 cepas de vinos utilizadas en el proyecto.

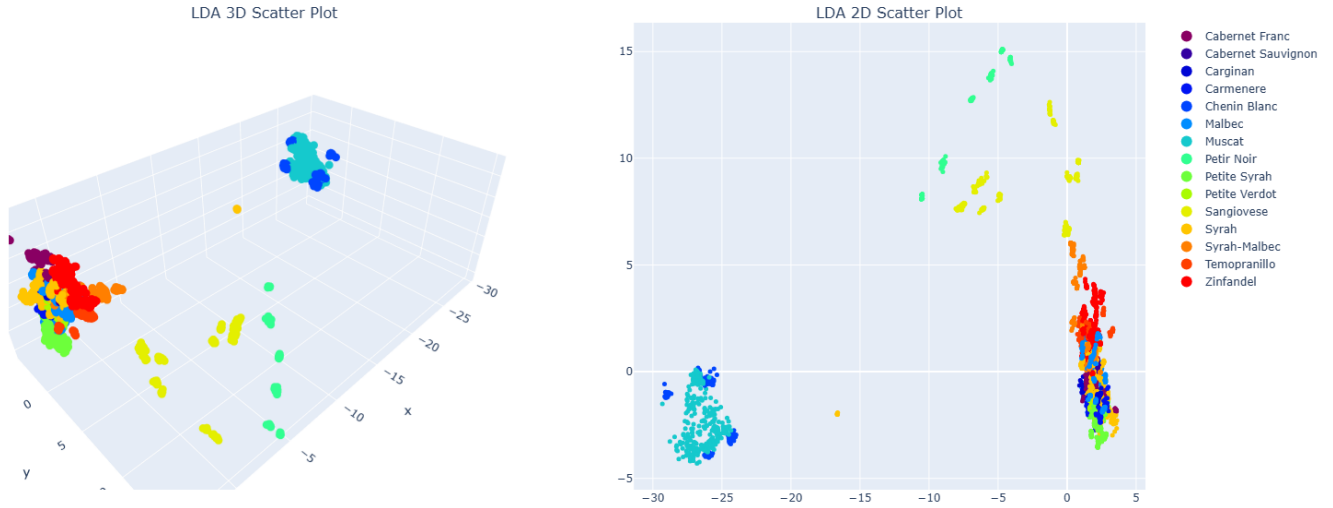


Figura 4.2: Visualización 3D y 2D de agrupación espectral de las 15 cepas de vino mediante transformación LDA.

De esta forma, al usar esta herramienta se busca explorar visualmente la estructura de los datos, observar solapamientos o separaciones entre cepas bajo distintos preprocesamientos y configuraciones. Además, se espera detectar posibles sesgos o problemas de representatividad en las muestras. Es especialmente útil para decidir estrategias de clasificación, identificar necesidad de aumentación de datos y comunicar visualmente los resultados a públicos técnicos y no técnicos.

4.7 Estrategias de Ajuste y Optimización

Durante la etapa de entrenamiento del modelo y sus respectivos submodelos, se aplicaron estrategias de ajuste y optimización destinadas a maximizar el rendimiento frente a un conjunto de datos reducido y desbalanceado. Una parte clave de este proceso fue la selección de hiperparámetros, realizada de forma experimental a partir de los resultados obtenidos en las evaluaciones de desempeño.

Esta búsqueda sistemática permitió identificar configuraciones óptimas para parámetros como la profundidad en árboles de decisión, el número de vecinos en KNN y el valor de penalización en SVM. Con ello, se minimizó el error de clasificación, se evitó el sobreajuste y se mejoró la capacidad de generalización de los modelos.

4.8 Aumentación de datos

Como se explicó en la sección 4.5, la escasez de muestras motivó la implementación de diversas estrategias de aumentación de datos para mejorar la capacidad de generalización del modelo.

- **Aumentación por ruido computacional:** Se introdujeron perturbaciones artificiales en los espectros mediante ruido gaussiano y *salt and pepper*, simulando condiciones de medición variables sin alterar sustancialmente las características del vino. El ruido gaussiano mostró mejor desempeño en pruebas internas; su desviación estándar (σ) se ajustó entre 0,5% y 5% de la desviación estándar del espectro original, definiendo así el nivel de perturbación. En el caso de *salt and pepper*, se reemplazó un pequeño porcentaje de puntos (entre 1% y 3%) por los valores mínimo y máximo observados.
- **Aumentación por experimentos de laboratorio:** Se realizaron mediciones repetidas de las mismas muestras bajo condiciones ligeramente distintas, utilizando distintos sensores y placas de medición. Aunque no genera nuevas muestras, esta técnica permite evaluar la consistencia entre mediciones y seleccionar aquellas más confiables.
- **Aumentación con vinos comerciales:** Para ampliar el dataset y mejorar la representación de cepas que presentaban mayores problemas de separación, se incorporaron vinos comerciales de las variedades Syrah y Malbec. Como se dijo en la sección anterior, estas cepas fueron específicamente seleccionadas debido a su escasa representación en la base de datos original. Aunque estos vinos no reproducen fielmente las características de los provenientes del viñedo El Escorial, se utilizaron como complemento en las primeras etapas del entrenamiento del modelo, contribuyendo a mejorar su capacidad de diferenciación en fases iniciales del ajuste.

4.9 Evaluación y Validación del Modelo

Dada la estructura jerárquica de árbol de modelos, la evaluación se realiza de forma general y específica: cada submodelo se valida por separado antes de integrarse en el sistema final. Esta separación permite evaluar la efectividad en cada nivel, identificar posibles debilidades en etapas tempranas y facilitar ajustes antes de alcanzar el modelo completo. Por ello, cada vez que se emplean los modelos, se reportan métricas tanto para los submodelos como para el árbol completo.

Para medir el rendimiento, se utiliza *accuracy* (exactitud) como variable principal, cuya definición varía según el área de aplicación. En este trabajo, se adopta la presentada en la (3.5), por ser pertinente a los datos utilizados y a los objetivos planteados.

Sin embargo, la precisión por sí sola no es suficiente en contextos con clases desbalanceadas, ya que puede enmascarar un bajo desempeño en las clases menos representadas. Para mitigar este sesgo, se aplicó balanceo solo en la etapa de entrenamiento, evitando la sobre-representación de cepas con más muestras y favoreciendo una mejor generalización del modelo.

Aun así, para evaluar de forma más justa el rendimiento entre clases, se incorporó el *F1 score* (ecuación 3.6), que combina precisión y sensibilidad. Este enfoque permite una evaluación más

profunda sobre si el modelo logra un desempeño equitativo entre cepas.

Además, se utilizó la matriz de confusión para analizar las predicciones respecto a las etiquetas reales, identificando errores comunes entre cepas similares. Esta información es clave para ajustar o reestructurar los submodelos del árbol y considerar técnicas adicionales de validación.

En resumen, la metodología desarrollada permitió enfrentar efectivamente las principales limitaciones del proyecto, como la baja cantidad de datos y la similitud espectral entre cepas. Mediante el uso de modelos jerárquicos supervisados, estrategias de validación y técnicas de aumentación, se logró construir un modelo robusto y adaptable.

En síntesis, la metodología desarrollada permite enfrentar de manera efectiva las principales limitaciones del proyecto, como la baja cantidad de datos y la similitud espectral entre cepas. A través de un enfoque supervisado, el diseño de un árbol de modelos jerárquico y el uso de técnicas de validación y aumentación, se logra construir un modelo robusto y adaptable. Si bien se obtuvieron resultados que se consideran satisfactorios, se pueden identificar oportunidades de mejora, especialmente mediante la incorporación de más datos y la exploración de modelos más complejos como redes neuronales.

Capítulo 5

Resultados

En esta sección se presentan los principales resultados obtenidos tras entrenar y evaluar los modelos propuestos.

Es importante destacar que esta metodología fue desarrollada bajo la supervisión de Pablo Jaramillo  (Jaramillo (2025)), Ingeniero en Computación, quien es el autor de los archivos bajo los cuales se realizan las experimentaciones, el análisis y las posteriores conclusiones propias.

5.1 Construcción y organización del conjunto de datos (Datasets)

Dada la limitada cantidad de muestras disponibles y la necesidad de contar con al menos dos réplicas físicas por clase para entrenar y validar los modelos, se trabajó con distintas versiones del conjunto de datos. Para el análisis exploratorio se utilizaron 15 cepas con el fin de observar su distribución individual; mientras que para entrenar los modelos se emplearon 11 cepas, excluyendo Carignan, Carménère, Tempranillo y Chenin Blanc.

Para organizar estos datos, se empleó el script `make_data.py`, encargado de generar de forma sistemática los conjuntos necesarios para entrenar y evaluar los modelos de identificación y clasificación de cepas de vino, tanto para nodos individuales como para el árbol jerárquico completo. Este script asocia las mediciones espectrales, denominadas data Spec (espectros por longitud de onda de cada muestra), con sus respectivas etiquetas de cepa y con un conjunto de parámetros físicoquímicos y derivados, agrupados bajo el nombre de data Full. Esta última incluye información complementaria como pH, grados Brix, transmitancia y color CIE-LAB, además de datos sintéticos generados con ruido cuando es requerido. La explicación detallada del contenido y estructura de los datasets *Spec* y *Full* será presentada en una subsección posterior.

Inicialmente, el script verifica si existen los archivos base que contienen los espectros crudos

(por ejemplo, `wine_data.csv`) y las lecturas de referencia o blancos (`white_data.csv`). Si no están disponibles, los genera a partir de carpetas con archivos de mediciones. Cada archivo incluye, por muestra, los valores de intensidad en una serie por longitud de onda principalmente del espectro visible (del 410 nm al 940 nm), además de metadatos como identificador de muestra, número de toma (“take”), celda del espectrómetro, sensor utilizado y una marca de filtrado. La consolidación se realiza con la librería Pandas para lectura y concatenación de hojas de cálculo.

Una vez verificada la existencia de los archivos base, se cargan los DataFrames y se aplica un primer filtrado sobre las muestras comerciales (según el nombre del archivo) y las marcadas como “filtradas”. En el caso de muestras comerciales, se eliminan aquellas cuyo identificador comienza con la letra “c”. Luego, cada identificador alfanumérico se transforma en un valor numérico único, de modo que que ‘c12’ (vino comercial) o ‘25’ (vino experimental) queden representados por enteros distintos, evitando ambigüedades al trabajar con etiquetas y fusiones de tablas.

Paralelamente, se integran dos fuentes de información complementaria: (a) un archivo que relaciona cada identificador (label) con la cepa y su estado de fortificación, y (b) una tabla con los parámetros fisicoquímicos medidos (pH, grados Brix, grado alcohólico y conductividad). Ambas se integran con el DataFrame principal mediante operaciones de merge por columnas claves, de forma que cada fila espectral adquiere, además, su información complementaria.

El script permite además aplicar criterios de filtrado según cepas o identificadores específicos, y definir qué subconjuntos de vinos se usarán en cada versión experimental. También permite agrupar cepas en un mismo valor de etiqueta conjunta (por ejemplo, uniendo Chenin Blanc, Muscat y Sangiovese en el nodo A del árbol), reduciendo la complejidad de problemas de clasificación o de pocas muestras. Todas las variables categóricas relevantes (número de toma, celda, sensor, cepa, etc.) se codifican numéricamente en valores consecutivos, lo que facilita su uso en modelos de aprendizaje automático.

Finalmente, una vez filtrado y codificado el conjunto de datos, se divide en tres bloques principales:

- **wines_raw**: contiene solo los espectros (longitudes de onda).
- **wines_labels**: contiene etiquetas y metadatos complementarios.
- **wine_properties**: contiene propiedades fisicoquímicas (pH, conductividad, %v/v* y °Brix*), necesaria para trabajar cuando se desea clasificar modelos Full (data de espectros más propiedades fisicoquímicas).

Para las muestras de referencia (**white**), se sigue el mismo procedimiento, generando los bloques **white_raw** y **white_labels**. Esto permite definir con claridad qué datos se usarán como entrada (espectros) y cuáles como salida o información adicional para el análisis.

5.1.1 Separación conjunto entrenamiento y prueba

En esta etapa, el script divide las muestras en dos subconjuntos: entrenamiento y prueba. Esta partición se realiza de forma estratificada por cepa, con el objetivo de asegurar una representación equilibrada de clases en ambos conjuntos. Por defecto, el 70% de las muestras se asigna al conjunto de entrenamiento y el 30% restante al de prueba.

Adicionalmente, es posible sobrescribir manualmente esta partición para ciertos identificadores, forzando que algunas muestras específicas se asignen a uno u otro subconjunto. Esto es útil, por ejemplo, para asegurar que los vinos comerciales permanezcan siempre en el conjunto de entrenamiento.

Como salida, esta partición genera dos conjuntos de índices, para entrenamiento `train_labels` y el otro para prueba `test_labels`, que se utilizan para extraer los subconjuntos correspondientes desde `wines_raw`, formando así los DataFrames `wine_train` y `wine_test`.

Cabe destacar que, si en etapas anteriores se optó por no eliminar muestras marcadas como filtradas, dichas muestras pueden mantenerse en entrenamiento para evitar su pérdida durante la separación, dependiendo del valor de la bandera `drop_filtered_on_test`.

5.1.2 Cálculo variables teóricas

Una vez definidos los subconjuntos de entrenamiento y prueba, se calculan para cada muestra las variables teóricas: transmitancia (teórica y por máximos) y color CIE-LAB. Para ello, se construye un código que reconoce las longitudes de onda del espectro (del 410 nm al 940 nm) y convierte vectores de intensidad en vectores de transmisión o coordenadas de color.

La transmitancia teórica se obtiene según la ecuación (3.1), mientras que la transmitancia por máximos se calcula con la ecuación (3.2). Ambas matrices se guardan por separado para entrenamiento y prueba.

Luego, se genera el color CIE-LAB a partir del espectro de transmitancia teórica. El vector de transmitancia se transforma en coordenadas L, a y b siguiendo la definición del espacio de color CIE-LAB. El resultado se almacena en un DataFrame de tres columnas por muestra, que se guarda como `cielab_train.csv` y `cielab_test.csv`.

Esta representación de color se incluye como variable auxiliar, ya que estudios previos han demostrado que los parámetros CIE-LAB permiten diferenciar características de cepas con alta precisión (García (2017)). Estas variables se incluyen en datasets *Spec* y *Full*, descritos a continuación.

5.1.3 Dataset Espectral (Spec)

Este conjunto contiene únicamente las señales espectrales obtenidas con el mini espectrofotómetro mostrado en la Figura 4.1, que opera en el rango del espectro visible. Este dispositivo registra 18

longitudes de onda, expresadas en nanómetros (nm): 410, 435, 460, 485, 510, 535, 560, 585, 610, 645, 680, 705, 730, 760, 810, 860, 900 y 940.

Cada señal corresponde al valor de intensidad espectral (counts) medido en cada una de estas longitudes de onda y constituye la base del análisis espectroscópico para la identificación de cepas de vino. Junto a esta señal se incorpora su correspondiente blanco, necesario para transformar los valores de intensidad en transmitancia o absorbancia. En total, este dataset incluye 36 variables de entrada y los modelos que utilizan solo señales espectrales se denominan “Modelos Spec”.

5.1.4 Dataset Full

Este conjunto combina las 18 señales espectrales descritas previamente con cuatro propiedades fisicoquímicas medidas experimentalmente en cada muestra: pH, conductividad eléctrica (mS), grados Brix sobrestimados (°Brix*) y una estimación de refracción de mezcla alcohólica (%v/v*). Estas propiedades fueron seleccionadas por su facilidad de medición y su relevancia para caracterizar químicamente las matrices del vino. Se incorporaron como variables complementarias para evaluar su aporte a los modelos de clasificación.

Inicialmente, ambos grupos de datos (espectros y propiedades fisicoquímicas) se almacenan en un DataFrame unificado que se gestiona mediante funciones de filtrado y procesamiento. A partir de este, se generan subconjuntos organizados en carpetas según el tipo de modelo a emplear. Así, los modelos que incorporan propiedades fisicoquímicas a la data de espectros se conocen como “Modelos Full”, y estos tienen como entrada 40 variables.

5.1.5 Inyección Opcional de Ruido Sintético

En caso de querer aumentar la cantidad de datos de entrenamiento y robustecer el modelo, el script incluye una opción de generación de ruido sintético. Esta puede configurarse para aplicar ruido gaussiano o *salt and pepper*. Cuando esta función se activa, se clonan muestras del conjunto de entrenamiento y se les aplica el tipo de ruido seleccionado. Luego, se recalculan sus valores de transmitancia y color CIE-LAB. Estas muestras ruidosas se anexan al final de los DataFrames originales, junto con sus etiquetas y propiedades fisicoquímicas correspondientes.

El ruido gaussiano introduce perturbaciones continuas distribuidas normalmente alrededor de los valores originales, lo que simula errores sistemáticos suaves o variaciones instrumentales. En cambio, el ruido salt and pepper genera alteraciones abruptas en puntos aislados, representando errores puntuales o fallos de lectura en ciertos sensores. La incorporación de ambos tipos de ruido permite al modelo aprender a tolerar tanto desviaciones suaves como anomalías más extremas.

Así, este nuevo conjunto final de entrenamiento incluye tanto los espectros originales como sus réplicas alteradas, lo que busca mejorar la capacidad de generalización del modelo y su tolerancia

ante pequeñas variaciones experimentales.

5.1.6 Organización Final

Finalmente, el script guarda todos los resultados siguiendo una estructura de carpetas organizada por versión. En la raíz de cada subcarpeta (por ejemplo, V1/MPSSZ-ID, V2/MPSSZ-ID o V3/MPSSZ-ID), se almacena un conjunto de archivos bajo el nombre común de ALL-ID, que agrupa la totalidad de las muestras con sus señales espectrales, propiedades fisicoquímicas, etiquetas de cepa y demás variables complementarias. Desde este conjunto se derivan los siguientes subconjuntos específicos:

1. **rich_wine_data.csv**: archivo con todos los datos originales ("*rich*"), incluyendo espectros, etiquetas y propiedades.
2. **wine_labels.csv**: archivo con las etiquetas completas de las muestras.
3. **transmittance_by_max.csv** y **transmittance_by_white.csv**: archivos de transmitancia normalizada, calculada según dos criterios distintos.
 - (a) transmitancia por blanco corresponde a la definición teórica clásica de transmitancia espectral, obtenida dividiendo la intensidad de la muestra por la intensidad de un blanco.
 - (b) transmitancia normalizada al máximo es una medida experimental desarrollada durante este trabajo, permite comparar espectros normalizados en condiciones instrumentales variables. Ambas versiones se utilizaron para evaluar el impacto de la normalización en el desempeño de los modelos.
4. **physical_properties.csv**: contiene los parámetros fisicoquímicos (pH, conductividad, °Brix*, etc.).
5. **wine_train.csv**, **wine_test.csv**, **labels_train.csv**, **labels_test.csv**: subconjuntos de entrenamiento y prueba de las mediciones crudas y sus etiquetas.
6. **transmittance_max_train.csv**, **transmittance_max_test.csv**, **transmittance_white_train.csv**, **transmittance_white_test.csv**, **cielab_train.csv**, **cielab_test.csv**: subconjuntos de transmitancia y color CIE-LAB, divididos para entrenamiento y prueba.
7. **properties_train.csv**, **properties_test.csv**: archivos de propiedades fisicoquímicas separados por conjunto.

Esta organización permite mantener un registro claro y reproducible de los datos generados para cada versión del experimento, así como de los submodelos, ya que cada carpeta contiene exclusivamente los archivos relevantes para cada nodo, facilitando la trazabilidad entre el flujo de datos, los modelos utilizados y sus respectivos conjuntos de entrada y salida.

5.2 Evaluación Preliminar Modelos Supervisados Individuales

Es importante mencionar que el proyecto tiene un enfoque de algoritmos de aprendizaje supervisado, dado que el objetivo principal consiste en asociar cada muestra espectral con su correspondiente cepa de vino. Métodos no supervisados, como DBSCAN o k-means, fueron descartados, ya que los datos no presentaban estructuras densas ni agrupaciones naturales que facilitaran la identificación de grupos homogéneos de cepas, como puede verse en la Figura 5.1.

Previo a la construcción del árbol jerárquico de modelos, se realizaron experimentos exploratorios con distintos algoritmos de clasificación supervisada. Esta etapa preliminar tuvo como finalidad evaluar si era posible optimizar un modelo individual capaz de clasificar la totalidad de las cepas disponibles. Para ello, se trabajó con las dos configuraciones principales de entrada, mencionadas anteriormente: data puramente espectral (Spec) y data espectral enriquecida con propiedades (Full). Esto permite comparar el desempeño de los modelos al variar la cantidad y tipo de variables de entrada.

5.2.1 Algoritmos utilizados en las pruebas iniciales

En esta etapa preliminar, las variables de entrada, ya fueran del dataset *Spec* o *Full*, fueron seleccionadas de forma automatizada para cada modelo individual. La selección se basó en la métrica Permutation Feature Importance (PFI), disponible en la biblioteca *scikit-learn*, que cuantifica la relevancia de cada variable midiendo la caída de rendimiento (por ejemplo, en accuracy y/o F1) al permutar sus valores. Cuanto mayor es la caída, mayor es la importancia de la variable.

Los modelos evaluados en esta etapa se muestran a continuación, estos se ordenaron según su *accuracy* en el modelo *Full*, de mayor a menor:

Tabla 5.1: Accuracy y F1 *score* obtenidos inicialmente por cada modelo de inteligencia artificial supervisado individual, utilizando los dos tipos de datasets.

Algoritmo	Accuracy Modelo Spec	Accuracy Modelo Full
Logistic Regression	58%	60%
K-Nearest Neighbors	57%	60%
Gradient Boosting	55%	57%
Gaussian Naive Bayes (GaussianNB)	45%	53%
Support Vector Classifier (SVC)	53%	53%
Random Forest (RF)	51%	51%
Decision Tree (DT)	49%	50%

A pesar de probar diversos algoritmos y realizar ajustes de hiperparámetros, ningún modelo individual logró superar consistentemente el 60% de accuracy o el 58% de F1 *score*, incluso al incorporar variables fisicoquímicas o técnicas complementarias como PCA, LDA y la propia PFI. Este bajo rendimiento se atribuye, en gran parte, a la baja cantidad de muestras y al desbalance de clases, donde algunas cepas estaban sobrerrepresentadas, introduciendo sesgos que afectaban la generalización del modelo.

También se evaluaron transformaciones adicionales, como la transmitancia teórica, coordenadas CIELAB y normalización por media y desviación estándar. Sin embargo, ninguna de estas estrategias aportó una separación clara entre clases, evidenciando que el problema de clasificación presentaba fronteras difusas e intrínsecamente difíciles de resolver, atribuida a las similitudes espectrales entre algunas cepas analizadas.

Dado este escenario, se descartó el uso de redes neuronales profundas y demás modelos de aprendizaje profundo. Aunque son efectivas en problemas complejos, su aplicación requiere grandes volúmenes de datos para evitar sobreajuste. Como se muestra en la Tabla 4.1, muchas cepas tienen pocas muestras, lo que hace inviable este tipo de enfoques sin comprometer la capacidad de generalización del sistema.

En consecuencia, los resultados obtenidos motivaron la búsqueda de una alternativa más especializada. Se decidió emplear una estrategia jerárquica, estructurando un árbol de modelos supervisados para resolver el problema de clasificación en etapas progresivas. Esta solución ofrece los siguientes beneficios:

1. Dividir progresivamente las cepas en grupos más homogéneos.
2. Optimizar cada submodelo para distinguir únicamente dentro de su grupo asignado.
3. Minimizar el sesgo por desbalance y la varianza debida a la escasez de datos.

Este enfoque jerárquico permite un procesamiento escalonado, donde cada nodo del árbol resuelve un subproblema con menos clases, facilitando el entrenamiento y mejorando la precisión general.

Si bien este esquema no implica necesariamente una mejora en eficiencia computacional durante el entrenamiento, ya que cada submodelo debe entrenarse por separado y de forma secuencial, sí permite cierta eficiencia en la etapa de inferencia, ya que cada modelo predictivo (*predict*) solo se aplica a un subconjunto reducido de clases.

5.3 Diseño y Organización Arbol de Modelos

Como primer paso para organizar las cepas de vino en un sistema jerárquico, se proyectaron todas las muestras del conjunto de entrenamiento utilizando LDA, con el objetivo de analizar el resultado visual de la herramienta. Esta etapa se realizó sobre el dataset V3, que contiene 11 de las 15 cepas recolectadas.

El espacio generado por LDA (véase Figura 5.1) revela que algunas cepas se agrupan claramente sin solaparse, mientras que otras forman nubes difusas o se superponen. Este patrón justifica el bajo rendimiento de clasificadores multinomiales aplicados directamente a todas las clases, y complementa la idea de que es preferible segmentar el problema. Las proyecciones utilizadas se construyeron a partir de espectros transformados a transmitancia por máximos (según la ecuación (3.2)), donde cada punto se coloreó según la etiqueta de cepa correspondiente.

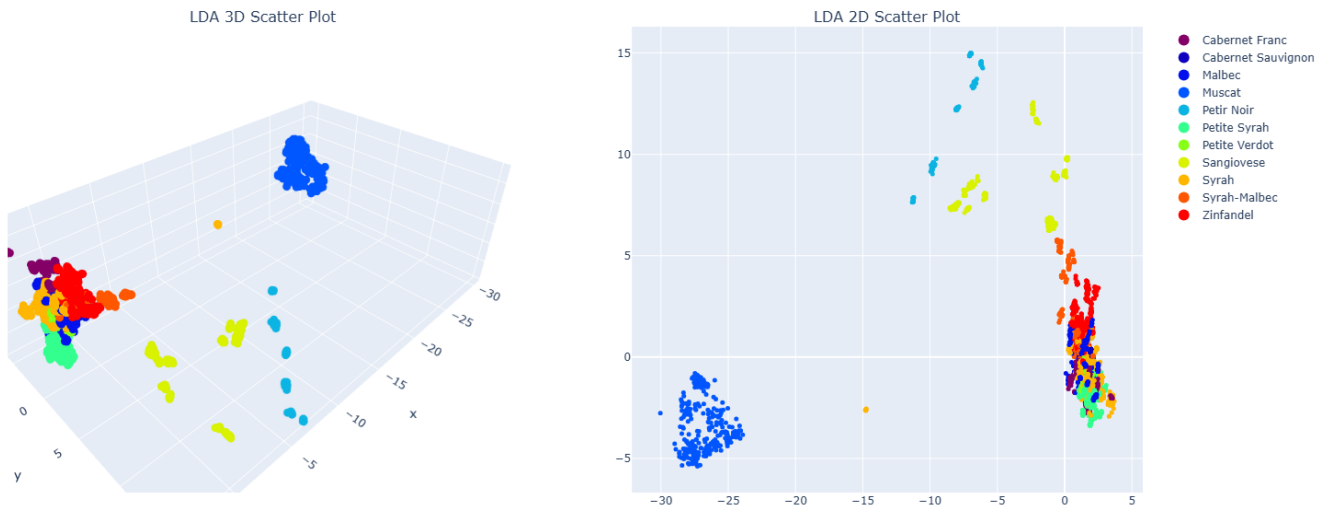


Figura 5.1: Visualización 3D y 2D de agrupación espectral de 11 cepas de vino mediante transformación LDA para dataset V3.

Con base en esta observación, se adoptó una estrategia de tipo *coarse-to-fine* (de lo general a lo específico). En lugar de entrenar un único modelo multinomial, se construyó un árbol jerárquico de modelos. En este esquema, cada nodo actúa como filtro al identificar subconjuntos de cepas fáciles de distinguir y etiquetarlos de manera definitiva dentro de ese subconjunto. Este enfoque

escalonado reduce el ruido de clases no relevantes en cada etapa, mejora la precisión global del sistema y permite aprovechar mejor la estructura latente del espacio LDA.

A continuación, se describen los elementos clave que conforman cada nodo del árbol:

- Criterio espacial en el espacio LDA (expresado en las figuras): Se utiliza la distribución de puntos proyectados por cepa, representados con colores para identificar subconjuntos claramente diferenciables.
- Módulos de código encargados de cada nodo: Las clases "**separator**" que detectan si una muestra pertenece a un cierto subconjunto de cepas. Las clases "**classifier**" que, habiendo recibido ese subconjunto, deciden la etiqueta final (o, en el caso de Syrah-Malbec, la etiqueta "provisional" al nodo SM).

5.3.1 Nodo Inicial: MPNS-11 (11 Cepas)

Como se evidencia en la Figura [5.1](#) cuando se proyectan las 11 cepas en el espacio 2D y 3D, se observan tres grandes conglomerados de cepas, pero ninguno de ellos está completamente aislado ni exento de solapamientos internos. Aunque hay ciertos puntos concentrados en una misma zona, entre ellos todavía existe cierta dispersión que dificulta trazar una frontera nítida que separe cada cepa en forma inequívoca.

En la Figura 5.1 se observa la proyección de las 11 cepas en los espacios LDA 2D y 3D. En ambas proyecciones se observa que algunas variedades se agrupan en regiones muy alejadas unas de otras. Por ejemplo, las muestras de Muscat aparecen en el extremo izquierdo. De forma similar, Pinot Noir (etiquetada como "Petir Noir" por error de nombre en la medición) forma un grupo claramente aislado en la zona media-alta de los ejes, y Sangiovese ocupa una región distinta. Este comportamiento indica que, para esas cepas, las características espectrales usadas se diferencian lo suficiente como para que el LDA las separe sin apenas confusión.

En contraste, un grupo de variedades aparece concentrado en una misma "nube" de puntos, tanto en el espacio 3D como en la proyección 2D. Esto sugiere que sus espectros son muy similares en las dimensiones discriminantes principales, dificultando su separación directa.

Dada esta complejidad, se diseñó el nodo MPNS-11, que agrupa todas las cepas inicialmente y divide el problema en dos etapas mediante dos objetos: `MPNS_separator` y `MPNS_classifier`.

MPNS_separator

Este módulo actúa como filtro para identificar si una muestra pertenece a las cepas **Muscat**, **Petite Noir** o **Sangiovese**, o no. Para ello: Cómo funciona:

1. Recibe como entrada el DataFrame completo, sea *Spec* o Full.

2. Aplica un pipeline (escalado + PCA + clasificador) entrenado para distinguir las tres cepas de interés del resto.
3. Predice si cada muestra pertenece o no al subconjunto MPNS.
4. Separa el DataFrame en:
 - **MPNS**: filas clasificadas como Muscat, Petite Noir o Sangiovese.
 - **others_MPNS**: el resto de las muestras.

Si MPNS no está vacío, esas filas se envían a:

MPNS_classifier

Este segundo módulo toma únicamente las muestras clasificadas como MPNS y les asigna la etiqueta final: “Muscat”, “Petite Noir” o “Sangiovese”. Utiliza un pipeline más fino (por ejemplo, escalado + PCA + SVM multiclase), entrenado solo con estas tres cepas.

Su salida es un nuevo DataFrame con la columna **wine_strain** ya definida, lo que permite que estas muestras no pasen al resto del árbol.

5.3.2 Nodo SM-8 (Syrah-Malbec)

Como se muestra en la Figura [5.2](#) los puntos naranjas, correspondientes a Syrah-Malbec, aparecen proyectados en una región relativamente aislada del acoplamiento más grande. Sin embargo, su distribución dispersa, con puntos extendidos en zonas vecinas, dificulta definir una frontera clara. Dado esto, se decide crear un nodo específico para separar y clasificar únicamente esta cepa.

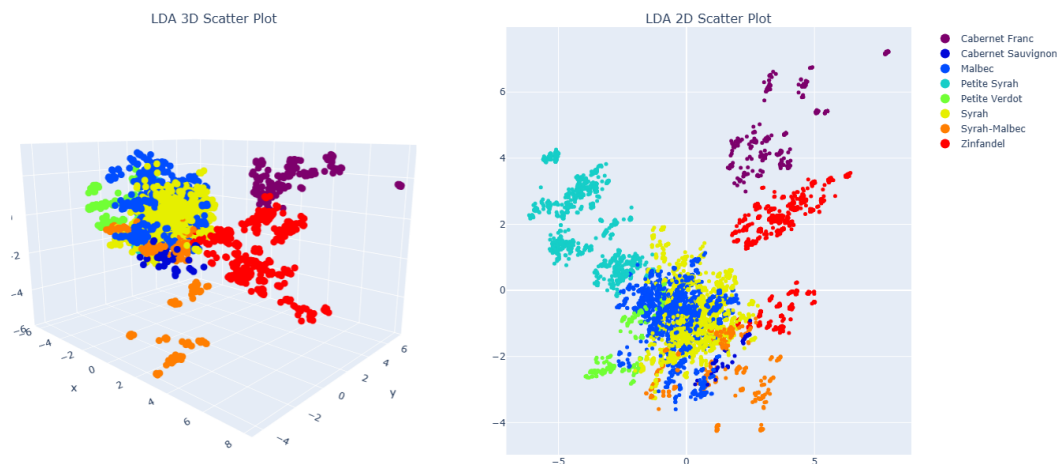


Figura 5.2: Visualización 3D y 2D de agrupación espectral de 8 cepas de vino resultantes luego de la separación del nodo MPNS-11, mediante transformación LDA.

SM_separator

Este módulo recibe como entrada el DataFrame `others_MPNS` (todas las filas no clasificadas como MPNS). Su propósito es detectar si una fila pertenece a la región que agrupa Syrah-Malbec y etiquetarla directamente.

Para ello, se aplica un pipeline entrenado para distinguir entre “Syrah-Malbec” y “otras cepas”, utilizando como referencia las proyecciones LDA y coordenadas asociadas. El modelo devuelve un arreglo que clasifica cada muestra como “SM” o “no SM”, dividiendo así el conjunto en:

- **SM**: filas identificadas como Syrah-Malbec.
- **others_SM**: filas restantes.

Las muestras SM se etiquetan directamente como `wine_strain = 'Syrah-Malbec'`, quedando fuera del árbol como nodo hoja. Las muestras no clasificadas como SM (`others_SM`) continúan su procesamiento en el siguiente nodo bajo el nombre `others_rest`.

5.3.3 Nodo CSPSPV-7

Tras las separaciones realizadas por los nodos anteriores, la Figura 5.3 muestra la proyección LDA de 7 cepas restantes. En esta etapa se identifican tres cepas con cierta separación espacial: Cabernet Sauvignon, Petite Syrah y Petite Verdot, mientras que el resto aún forma una nube densa y solapada.

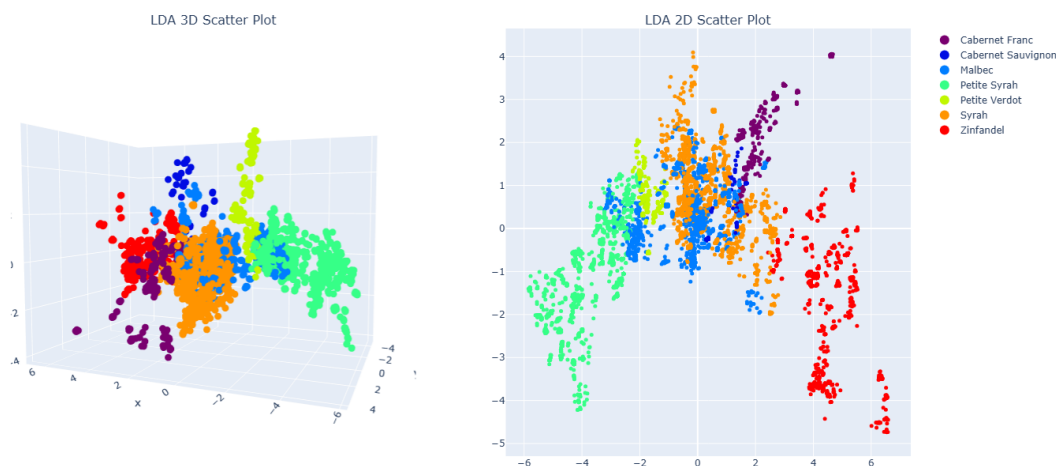


Figura 5.3: Visualización 3D y 2D de agrupación espectral de 7 cepas de vino resultantes luego de la separación del nodo SM-8, mediante transformación LDA.

Aquí se nota que:

- Cabernet Sauvignon se ubica desplazado hacia la parte superior derecha.
- Petite Syrah forma un subclúster denso en la zona izquierda-central.
- Petite Verdot aparece próximo a Petite Syrah, aunque reconocible.
- El grupo restante (Cabernet Franc, Malbec, Syrah, Zinfandel) sigue mezclado, aunque algunas clases como Cabernet Franc o Zinfandel comienzan a desplazarse hacia otras zonas.

Con base en estas observaciones, se decide que CSPSPV-7 se encargue de identificar en bloque las tres cepas mejor definidas (Cabernet Sauvignon, Petite Syrah y Petite Verdot), dejando el resto para un nodo posterior (CFMSZ-ID).

CSPSPV_separator

Este módulo filtra las muestras que podrían pertenecer a una de las tres cepas mencionadas. Recibe como entrada el DataFrame `others_rest` (filas que no fueron clasificadas por nodos anteriores) y aplica un pipeline entrenado para distinguir si una muestra corresponde a alguna de las tres clases que conforman el grupo CSPSPV. A partir de esto se obtiene una predicción binaria que divide el conjunto en:

- CSPSPV: muestras consideradas parte del nodo (Cabernet Sauvignon, Petite Syrah o Petite Verdot).
- CFMSZ: muestras restantes, que se almacenan para la siguiente etapa.

Las muestras de CSPSPV se envían al siguiente módulo clasificador.

CSPSPV_classifier

Este clasificador recibe las muestras prefiltradas por el `CSPSPV_separator` y les asigna una de tres etiquetas: “Cabernet Sauvignon”, “Petite Syrah” o “Petite Verdot”. Para ello, entrena un modelo multiclase exclusivamente con esas tres cepas, y devuelve un DataFrame con la columna `wine_strain` etiquetada.

Las muestras clasificadas se extraen del flujo como nodos hoja. El resto de las cepas (Cabernet Franc, Malbec, Syrah, Zinfandel) se acumulan en el grupo `CFMSZ` y se envían al clasificador final.

5.3.4 Nodo CFMSZ-ID

Para el nodo final se trabaja con las cuatro cepas restantes: Cabernet Franc, Zinfandel, Malbec, Syrah. Como se muestra en la Figura [5.4](#), la proyección LDA evidencia que Cabernet Franc aparece

relativamente aislado en la parte superior del gráfico, Zinfandel con un comportamiento similar a la derecha aunque con algunos de sus extremos cercanos a Syrah, mientras que Malbec y Syrah forman una nube densa y fuertemente acoplada.

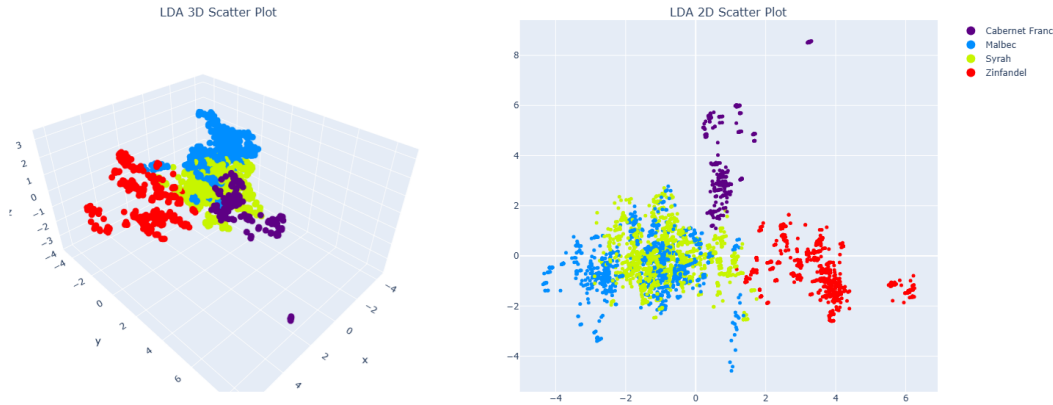


Figura 5.4: Visualización 3D y 2D de agrupación espectral de 4 cepas de vino resultantes luego de la separación del nodo CSPSPV-7, mediante transformación LDA.

Dada esta distribución, se decidió resolver este último tramo en dos pasos:

- `CFMSZ_classifier` etiqueta con confianza aquellas muestras que claramente son “Cabernet Franc” o “Zinfandel”. Es decir, se optimiza un modelo tal que la predicción se equivoque, a propósito y con alta tasa de confusión, en la agrupación “Malbec–Syrah”.
- Aquellas muestras que caen en la zona de incertidumbre “Malbec y Syrah” se derivan a un clasificador binario específico (`MS_classifier`), entrenado únicamente para distinguir entre esas dos cepas estrechamente relacionadas.

Esta estrategia permitió cerrar el árbol sin añadir más niveles, evitando un aumento innecesario de complejidad y reduciendo la carga computacional. Su elección se justificó porque, tras diversos experimentos con este subconjunto, representó el mejor compromiso entre simplicidad estructural y rendimiento de clasificación.

CFMSZ_classifier

Este clasificador recibe el conjunto CFMSZ (muestras no clasificadas por `CSPSPV_classifier`) e intenta asignar una etiqueta entre cuatro clases: *Cabernet Franc*, *Malbec*, *Syrah* y *Zinfandel*. Para ello, realiza una predicción multiclase utilizando un modelo entrenado sobre este subconjunto. Todas las muestras reciben una etiqueta correspondiente a una de estas cuatro cepas. Sin embargo, a nivel de sistema, se asume que las predicciones de Cabernet Franc y Zinfandel son confiables, mientras que las muestras clasificadas como Malbec o Syrah se consideran inciertas debido a su

solapamiento espectral y a la alta tasa de confusión observada durante el entrenamiento. A partir de esto, se aplica una lógica condicional:

- Las muestras etiquetadas como *Cabernet Franc* o *Zinfandel* se consideran clasificadas con alta seguridad y se asignan directamente a nodos hoja.
- Las muestras etiquetadas como Malbec o Syrah se derivan a un clasificador binario posterior (`MS_classifier`) especializado en resolver esa ambigüedad.

De este modo, `CFMSZ_classifier` actúa como un primer filtro general, y la interpretación de sus resultados se ajusta según el nivel de confianza empírico observado por clase.

MS-ID

Dado que Malbec y Syrah comparten una distribución espectral muy similar como se evidencia en la Figura 5.5, se entrena un clasificador binario dedicado a distinguir exclusivamente entre estas dos cepas.

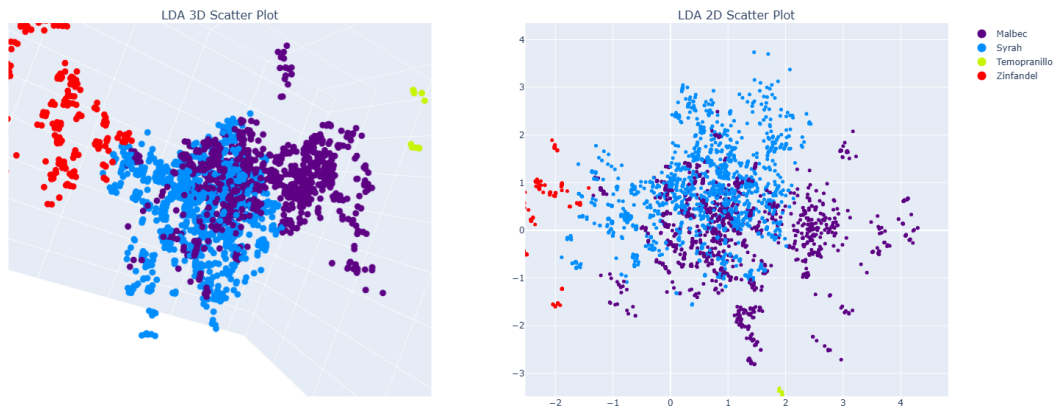


Figura 5.5: Visualización 3D y 2D de agrupación espectral de las cepas Syrah y Malbec mediante transformación LDA.

Este submodelo, denominado `MS_classifier`, recibe como entrada el subconjunto MS (muestras ambiguas según CFMSZ) y ejecuta un pipeline en dos etapas:

- Preprocesamiento: Incluye PCA de baja dimensión, escalado robusto y selección de características espectrales relevantes.
- Clasificador binario: Entrenado específicamente para separar Malbec de Syrah.

Al restringirse a solo dos clases, este modelo puede especializarse en capturar las diferencias sutiles que los clasificadores generales no lograron distinguir.

Finalmente, se devuelve un DataFrame con la etiqueta `wine_strain` asignada como Malbec o Syrah, completando así la clasificación de todas las muestras del sistema.

5.4 Organización interna de los archivos y ensamblaje en el árbol

A continuación se presenta la estructura final del árbol de modelos, junto con los submodelos asociados a cada agrupación de cepas. En el esquema visual complementario, se usan tres colores para distinguir niveles:

- Amarillo: agrupaciones principales de separación por nivel del árbol.
- Celeste: carpetas de submodelos dentro de cada nivel.
- Lila: grupos finales de cepas dentro de cada dataset previamente definido.

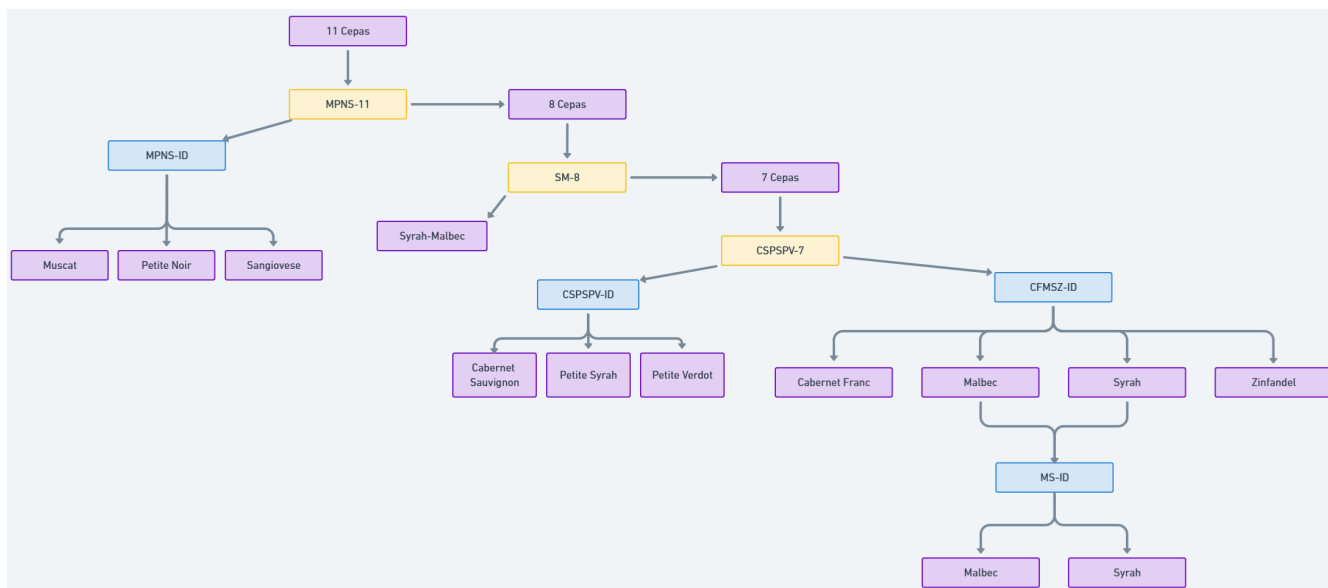


Figura 5.6: Organización y diseño visual detallado del árbol de modelos creado.

El archivo principal que organiza el flujo de entrenamiento y testeo se denomina `ml_tree_v3.py`, donde **V3** representa la versión del arreglo final de cepas con datos disponibles para prueba. Este archivo central coordina la ejecución completa del árbol a través de una clase principal: `WineStrainPredictorV3`. La estructura de archivos se organiza como sigue:

- `ml_tree_v3.py`: clase principal que orquesta todos los nodos del árbol.
- `mpns.py`: define `MPNS_separator` y `MPNS_classifier`.
- `sm.py`: define `SM_separator`.
- `cspspv.py`: define `CSPSPV_separator` y `CSPSPV_classifier`.

- `cfmsz.py`: define `CFMSZ_classifier` ((el cual etiqueta las cuatro cepas CF, M, S y Z; el sistema interpreta con alta confianza solo CF y Z, como se explica en la sección correspondiente) y `MS_classifier`.

5.4.1 Estructura `ml_tree_V3`

Dentro de `ml_tree_v3.py`, la clase `WineStrainPredictorV3` está estructurada en torno a tres componentes principales:

1. Instanciación de módulos En el método `_init_`, se instancian los distintos módulos del árbol, cada uno de los cuales carga internamente su pipeline y modelo entrenado. Cada módulo acepta el argumento `fragmented`, que indica si se desea cargar los modelos de manera fraccionada, es decir, por separado, para permitir su exportación individual en la implementación del sistema en el servidor de la aplicación. Esta opción agiliza el despliegue del modelo y reduce la carga computacional al evitar la carga simultánea de todos los submodelos.

2. Método `predict_core(X, out=[])` Este método central ejecuta el flujo completo de inferencia sobre un conjunto de entrada `X`, propagando las muestras a través de los distintos niveles del árbol. Al finalizar, el parámetro `out` contiene una lista de `DataFrames`, cada uno correspondiente a una de las cepas finales.

3. Concatenación final Una vez procesados todos los nodos del árbol, `predict_core` concatena los `DataFrames` generados en `out`, de modo que cada fila original de `X` quede finalmente etiquetada con su cepa correspondiente.

En este flujo, es importante mencionar los dos tipos de módulos creados para poder predecir las cepas de vino: separadores y clasificadores. Cada par separador / clasificador está encapsulado en un archivo propio. Esta modularidad no solo permite escalar o modificar partes específicas del árbol sin alterar el sistema completo, sino que facilita la incorporación de nuevas cepas o características, manteniendo la trazabilidad y la eficiencia en los flujos de trabajo.

5.5 Resultados submodelos internos (nodos) del árbol de modelos

5.5.1 MPNS-11

Este separador corresponde al nodo inicial del árbol de 11 cepas, encargado de dividir las clases más alejadas del núcleo acoplado, actuando como una primera gran frontera. Las imágenes completas

de las matrices de confusión que representan este comportamiento se encuentran en las Figuras [A.2](#) y [A.4](#) del Anexo [A.1](#).

Tabla 5.2: Métricas de evaluación nodo MPNS-11 submodelos separadores Full y Spec.

Versión	Accuracy	F1
Spec	90 %	90 %
Full	99 %	99 %

A partir de la Tabla [5.2](#) se observa que, al utilizar únicamente las señales espectrales (versión *Spec*), el modelo logra capturar gran parte de la separación entre las 11 cepas, alcanzando valores de *accuracy* y F1 de 90%. Si bien el error global de aproximadamente 10% podría parecer elevado, se considera aceptable dado que estas métricas se mantienen sobre el umbral del 90%, utilizando únicamente información espectral.

No obstante, al incorporar variables físico-químicas como el grado alcohólico (%v/v*), °Brix*, pH y conductividad, se produce una mejora sustancial. En su versión *Full*, el modelo eleva su *accuracy* y F1 hasta el 99%. Este salto cuantitativo demuestra el gran valor informativo de dichas variables para resolver confusiones entre cepas cuyos espectros resultan similares. De hecho, en el modo *Spec* se detectaron 59 falsos negativos, los cuales fueron reducidos a 8 en *Full*, eliminando por completo los falsos positivos. Esto evidencia que la frontera de decisión del modelo se torna mucho más precisa al integrar esta información complementaria.

Desde una perspectiva de implementación, este resultado justifica plenamente el coste adicional de medir variables físico-químicas, especialmente en un sistema monolítico de 11 clases. Según los análisis de importancia de variables vía Permutation Feature Importance (PFI), estas propiedades muestran una relevancia crítica en cepas difíciles de clasificar.

5.5.2 MPNS-ID

Este clasificador dentro del nodo de separación inicial se considera clasificador binario, ya que en la versión actual la cepa Sangiovese no contenía datos suficientes para la validación, pero sí tuvo para el entrenamiento, por ende se decide entrenar para mejorar robustez, pero no para pruebas, por ende no hay validación para esa cepa específica, y reduce el identificador a uno de tipo binario. Las imágenes completas de las matrices de confusión para cada submodelo del árbol se pueden ver en las Figuras [A.2](#) y [A.4](#) presentes en Anexo [A.1](#).

Tabla 5.3: Métricas de evaluación nodo MPNS-ID submodelos clasificadores *Full* y *Spec*.

Modelo	Accuracy	F1
Spec	100%	100 %
Full	100 %	100 %

La tabla anterior demuestra que la forma clara y separada de clústeres espectrales de estas clases comparadas a las demás cepas. Este comportamiento trivial en términos de clasificación es también confirmado por la herramienta PFI, que asigna una importancia cero a todas las variables físicoquímicas (%v/v*, °Brix*, pH, conductividad), indicando que no aportan valor adicional al proceso de decisión en este nodo. Esto implica que la separación puede lograrse de manera completamente confiable utilizando únicamente señales espectrales, principalmente porque las cepas involucradas corresponden a vinos de distinto color, lo cual reduce el coste computacional y operativo del sistema al evitar la necesidad de medir variables adicionales. En conjunto, estos resultados evidencian que el nodo MPNS-ID cumple el rol de un "gancho temprano" altamente eficiente.

5.5.3 SM-8

Este nodo corresponde exclusivamente al separador de la cepa "Syrah-Malbec" respecto de las ocho restantes. Las imágenes completas de las matrices de confusión asociadas a este nodo se pueden ver en las Figuras [A.5](#) y [A.6](#) en Anexo [A.1](#).

Tabla 5.4: Métricas de evaluación nodo SM-8 submodelos separadores *Full* y *Spec*.

Modelo	Accuracy	F1
Spec	97%	97 %
Full	97 %	97 %

De este set de métricas se evidencia que hay un alto rendimiento con *Spec*, es decir, con solo el espectro puro se alcanzan valores de *accuracy* y F1 de 97 % , sin generar falsos positivos (0 muestras clasificadas erróneamente como *Syrah-Malbec*). Este comportamiento es deseable en clasificadores binarios intermedios, ya que prioriza evitar contaminar las siguientes ramas del árbol, aunque implique tolerar algunos falsos negativos (12 muestras en este caso). Esto indica una separación espectral clara y un modelo robusto con baja tasa de error crítica.

Por otro lado, el modelo *Full* tiene una mejora marginal ya que sólo corrige 1 falso negativo adicional, pero al costo de introducir un nuevo error (1 muestra mal clasificada como *Syrah-Malbec*), lo que representa un riesgo al generar error de arrastre para siguientes nodos. PFI indica que las

variables físicoquímicas tienen baja importancia en la zona, por lo que su inclusión aporta poco valor, por lo que en un entorno de recursos limitados se puede prescindir de ellas en este submodelo.

5.5.4 CSPSPV-ID

Este nodo separa las cepas "Cabernet Sauvignon", "Petite Syrah" y "Petite Verdot" del resto. Las matrices de confusión se evidencian en las Figuras [A.7](#) y [A.9](#) en Anexo [A.1](#). Los resultados se presentan en la siguiente tabla:

Tabla 5.5: Métricas de evaluación nodo CSPSPV-ID submodelos separadores *Full* y *Spec*.

Modelo	Accuracy	F1
Spec	100%	100 %
Full	100%	100 %

Las métricas evidencian una separación perfecta (Accuracy y F1: 100%) tanto en modelos *Full* como *Spec*. Tal como en MPNS-ID, las cepas forman subnubes espectralmente aisladas, lo que se traduce en un caso trivial. La herramienta PFI asigna importancia cero a las variables físicoquímicas, por lo que no aportan mejora alguna en este punto. El límite de separación está completamente definido por las señales espectrales.

5.5.5 CSPSPV-7

Una vez separadas, se clasifica entre "Cabernet Sauvignon", "Petite Syrah" y "Petite Verdot". A diferencia del nodo anterior, aquí sí se evidencian errores: la Figura [A.8](#) (Anexo [A.1](#)) muestra 53 falsos positivos y 75 falsos negativos, lo que implica que, aunque la agrupación global fue sencilla, la diferenciación entre estas tres cepas presenta cierta complejidad. A pesar de ello, los valores de *accuracy* (92%) y *F1 score* (92%) siguen siendo satisfactorios (Tabla [5.6](#)). Al igual que en el nodo anterior, la inclusión de variables físicoquímicas no mejora el rendimiento, ya que *Spec* y *Full* obtienen resultados idénticos. Se concluye que la frontera espectral ya encapsula la información necesaria para esta separación, haciendo redundante cualquier variable adicional.

Tabla 5.6: Métricas de evaluación nodo CSPSPV-7 submodelos *Full* y *Spec*.

Modelo	Accuracy	F1
Spec	92%	92 %
Full	92%	92 %

5.5.6 CFMSZ-ID

En este nodo se clasifica entre "Cabernet Franc", "Malbec", "Syrah" y "Zinfandel".

Tabla 5.7: Métricas de evaluación nodo CFMSZ-ID submodelos clasificadores Full y Spec.

Modelo	Accuracy	F1
Spec	66%	67%
Full	82%	83 %

De la tabla anterior se observan mejoras relevantes al integrar variables fisicoquímicas. El modelo *Full* supera ampliamente al modelo *Spec* en su evaluación. Esta mejora se explica por la incorporación de variables clave según PFI, que corrigen la confusión inicial entre Cabernet Franc y Malbec/Syrah, elevando el acierto desde 119 de 160 a 160 de 160 (ver Figuras [A.11](#) y [A.13](#) en Anexo [A.1](#)).

Sin embargo, persisten errores significativos para Zinfandel: en *Spec*, se confunde con Malbec en 49 casos y con Syrah en 10 (realmente Zinfandel clasificados erróneamente); en *Full*, aunque se reduce el error, aún hay 33 confusiones con Malbec y 19 con Syrah. Esto sugiere que, pese a contar con variables adicionales, el espectro de Zinfandel comparte características con Malbec, y su separación total requiere atención adicional. Por ello, podría considerarse, para un futuro, generar otro subnodo binario específico ("Zinfandel vs resto") antes de MS-ID, o bien revisar los arrastres de error en el árbol completo.

En suma, *Full* es la opción preferente para este nodo: disminuye el error restante en casi un tercio y logra una diferenciación más clara entre Cabernet Franc y las cepas restantes. La inversión en variables fisicoquímicas resulta justificada.

5.5.7 MS-ID

Finalmente, se evalúa el clasificador binario final de las cepas "Malbec" y "Syrah".

Tabla 5.8: Métricas de evaluación nodo MS-ID submodelos clasificadores *Full* y *Spec*.

Modelo	Accuracy	F1
Spec	79%	79%
Full	100%	100 %

Con la data de solo espectro, el modelo falla en un 21% de los casos (ver Figuras [A.12](#) y [A.14](#) en Anexo [A.1](#)), lo cual se refleja con alta superposición espectral en el espacio LDA mostrada con anterioridad(ver Figura [5.2](#)). Esto hace que muchas muestras sean indistinguibles entre sí

usando solo las 18 señales originales. El modelo *Full*, en cambio, aprovecha las propiedades fisicoquímicas para separar sin ambigüedad ambas cepas. Gracias a ello, logra un 100% de *accuracy* y F1, evidenciando que estas variables capturan diferencias sutiles que el espectro puro no revela.

En consecuencia, para el nodo MS-ID es crítico contar con todas las variables. La integración de propiedades adicionales no solo mejora el rendimiento cuantitativo, sino que soluciona completamente la ambigüedad espectral.

5.5.8 Análisis General

Los submodelos “fáciles” (MPNS-ID, CSPSPV-ID) alcanzan métricas óptimas usando únicamente espectro, gracias a que PFI selecciona longitudes de onda altamente discriminativas. Los nodos intermedios (SM-8, CSPSPV-7) también mantienen un desempeño aceptable con *Spec* ($> 90\%$), y aunque *Full* mejora marginalmente, la elección de agregar o no datos fisicoquímicos queda abierta según necesidad.

En cambio, para submodelos “difíciles” (MPNS-11, CFMSZ-ID, MS-ID), el uso de propiedades fisicoquímicas mejora significativamente las métricas, justificando su incorporación.

La robustez de *Spec* en nodos “fáciles” confirma que PFI selecciona atributos genuinos, no artefactos, incluso bajo datos reales y balanceados. Cada nodo parte con un conjunto amplio (18 longitudes + 4 propiedades), y la Permutation Feature Importance reduce la dimensionalidad reteniendo solo predictores relevantes. Así, los nodos espectrales conservaron longitudes clave, mientras que aquellos *Full*-dependientes incorporaron propiedades fisicoquímicas críticas o el total de ellas.

Así, la arquitectura modular (separadores + clasificadores), combinada con PFI, logra un equilibrio eficiente entre precisión y complejidad. No se observa sobreajuste: los modelos *Spec* alcanzan 100% y los *Full* corrigen debilidades puntuales, evidenciando que la validación cruzada y el balance en training fueron efectivos.

5.6 Resultados árbol de modelos-ml_tree_v3

Los resultados globales del árbol de modelos se presentan a continuación:

Tabla 5.9: Resultados globales del árbol de modelos para un dataset *Spec* y *Full*.

Modelo	Accuracy	F1
Spec	80%	81%
Full	93%	93%

5.6.1 Matriz de confusión modelo *Spec*

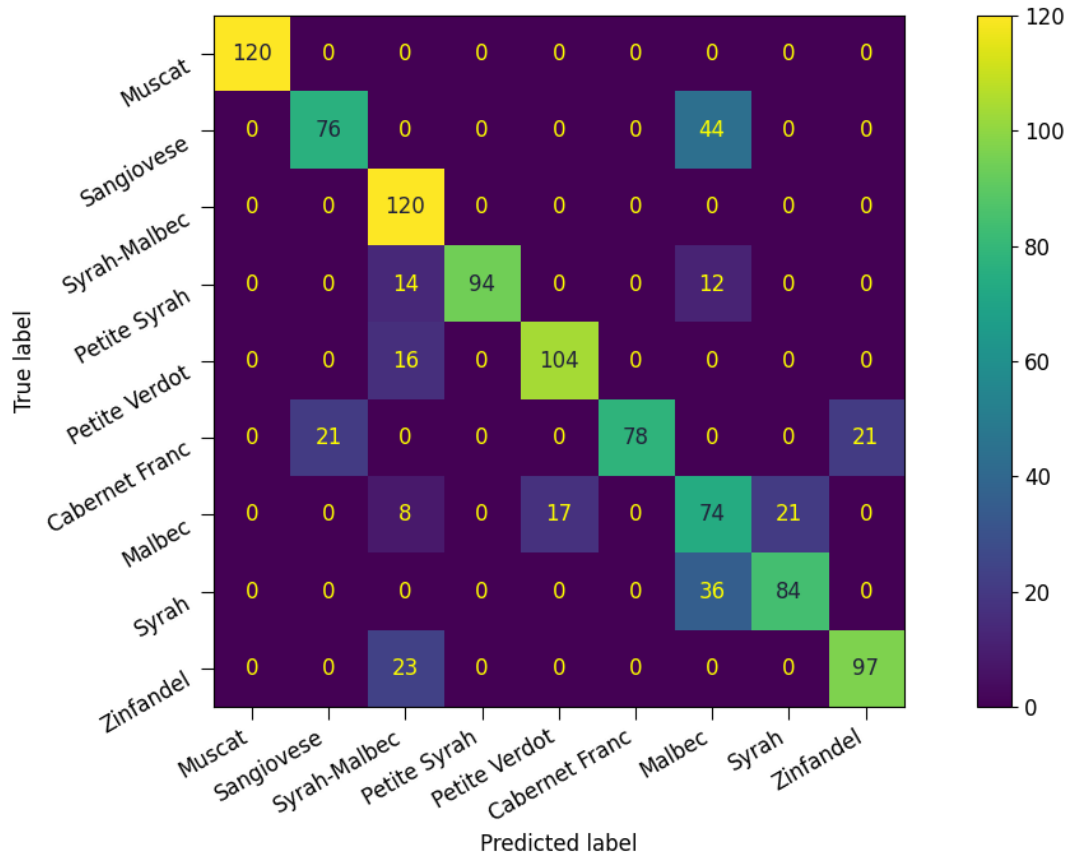


Figura 5.7: Matriz de confusión árbol completo para dataset *Spec* V3 de 11 cepas. En la diagonal principal se indican las muestras correctamente clasificadas para cada cepa, mientras que los valores fuera de ella muestran las mal clasificaciones entre cepas.

El árbol completo, entrenado exclusivamente con las 18 longitudes de onda, alcanza un 80% de accuracy y un 81% F1 global. La matriz de confusión evidencia que clases con clusters espectrales nítidos —como Muscat, Syrah-Malbec y Zinfandel— logran desempeños perfectos o casi perfectos (Muscat y SM 120 aciertos de 120, Zinfandel 97 de 120). En cambio, Sangiovese (76 coincidencias de 120) y Cabernet Franc (78 aciertos de 120) presentan graves solapamientos: Sangiovese se confunde 44 veces con Malbec y Cabernet Franc se dispersa entre Sangiovese, Petite Verdot y Zinfandel.

Los nodos intermedios también muestran deterioro: Petite Syrah alcanza 94 coincidencias de 120, con 14 errores hacia SM y 12 hacia PV; Petite Verdot obtiene 104 de 120, con 16 errores hacia PS. A pesar de su separación temprana, Malbec y Syrah solo logran 74 aciertos de 120 y 84 de 120, respectivamente. Estos resultados refuerzan que, aunque la señal espectral explica gran parte de la discriminación, no es suficiente para resolver completamente los solapamientos entre cepas cercanas.

5.6.2 Matriz de confusión modelo Full

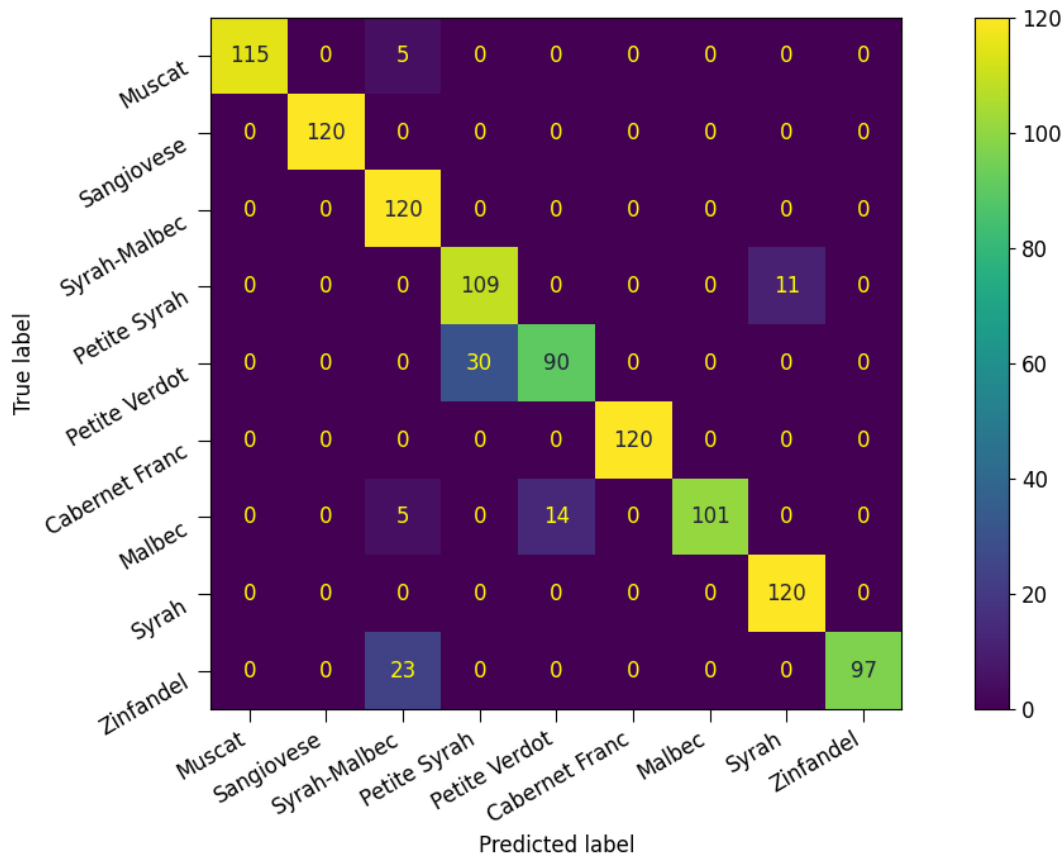


Figura 5.8: Matriz de confusión árbol completo para dataset Full V3 de 11 cepas. En la diagonal principal se indican las muestras correctamente clasificadas para cada cepa, mientras que los valores fuera de ella muestran las mal clasificaciones entre cepas.

Al incorporar las cuatro variables adicionales (%v/v*, °Brix*, pH y conductividad), la *performance* global sube a 92% de accuracy y F1. Recorriendo la matriz de confusión, se observa que casi todas las clases se acercan a 100% de aciertos individuales, exceptuando:

- Muscat baja levemente a 115 aciertos de 120, con 5 muestras mal clasificadas como Syrah-Malbec. Este desvío puede interpretarse como un “daño colateral” al expandir la zona de influencia del nodo SM.
- Zinfandel mantiene el valor de Spec, con 23 errores hacia Syrah-Malbec. Pese a este desbalance, sigue siendo una clase relativamente estable.
- Petite Verdot baja su valor de éxito a 90 de 120 debido a 30 falsos negativos hacia SM.
- Petite Syrah mejora su valor a 109 aciertos, con 11 errores hacia Zinfandel. Esta mejora sugiere una delimitación más clara respecto a su nodo hermano (Zin).

- Malbec asciende a 101 casos correctos de 120, con 5 errores hacia SM y 14 hacia Petite Verdot, superando significativamente a Spec.

Este salto en las clases más conflictivas muestra que las propiedades físico-químicas aportan las sutilezas que el espectro puro no distingue. Esto muestra un impacto directo en la corrección de confusiones críticas, como las de Sangiovese con Malbec, y de Syrah con Malbec, que ahora se eliminan completamente. Aun cuando existen algunos errores, este modelo presenta un balance global más robusto, capaz de sostener altos niveles de precisión incluso en etapas intermedias.

Esta clarificación del modelo puede explicarse a partir de la proyección LDA de los espectros de las cepas, con propiedades fisicoquímicas incluidas, para el último y más confuso nodo del árbol: CFMSZ-ID (Figura 5.9). Aquí, al incluir variables fisicoquímicas, los clústeres se reorganizan y delimitan con mayor nitidez. Esta redistribución permite desacoplar completamente a Malbec y Syrah, mostrando una separación clara entre ambas, lo que antes no era posible con espectro puro.

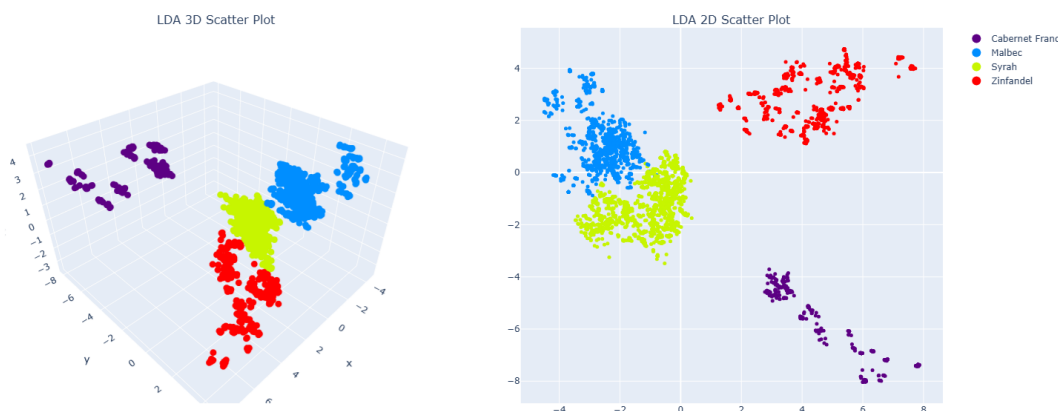


Figura 5.9: Visualización 3D y 2D de agrupación espectral de las 4 cepas de vino del nodo CFMSZ-ID con propiedades fisicoquímicas agregadas, mediante transformación LDA.

5.6.3 Análisis general árbol de modelos Spec y Full

La comparación entre ambos modelos completos muestra que las longitudes de onda permiten identificar con alta precisión los grupos naturalmente separados (Muscat, Syrah-Malbec, Zinfandel), logrando aislamientos consistentes con alto porcentaje de acierto. Sin embargo, en las zonas de mayor solapamiento espectral, como Sangiovese vs Malbec, Cabernet Franc vs Petit Verdot/Zinfandel y Malbec vs Syrah, el modelo *Spec* exhibe caídas críticas de rendimiento, con clases mal clasificadas incluso en su totalidad o con precisiones intermedias, resultando con valores de métricas del 80%.

Esto confirma que, pese a que los submodelos puedan tener métricas elevadas de forma aislada, su integración en el árbol completo genera errores de arrastre y combinaciones débiles en nodos sensibles, incapaces de resolver las ambigüedades espectrales presentes.

La incorporación de variables fisicoquímicas (%v/v*, °Brix*, pH y conductividad) permite estabilizar esos nodos críticos, en particular MPNS-11, CFMSZ-ID y MS-ID, aumentando la evaluación general a un 92% y eliminando las confusiones más severas. Esta arquitectura híbrida y modular, reforzada por PFI, permite mantener el uso de espectro puro en nodos fáciles, reservando las variables adicionales solo donde son necesarias. Con ello, se logra un modelo eficiente, robusto y de bajo costo computacional, sin necesidad de sobreajustar el flujo de datos. Sin embargo, es importante considerar que, debido a su diseño supervisado, el modelo asume que la distribución de los datos futuros sigue una estructura similar a la observada durante el entrenamiento. Esto implica que, si el conjunto de datos disponible no es lo suficientemente representativo, el rendimiento del sistema podría verse afectado ante nuevas muestras que presenten variaciones no contempladas.

5.7 Reducción de variables fisicoquímicas

Para evaluar el impacto de las variables fisicoquímicas en la capacidad discriminativa del sistema jerárquico de clasificación, se definieron dos familias de modelos intermedios: M1 y M2.

En los modelos M1, se combinan las 18 señales espectrales con una única variable fisicoquímica (%v/v*, °Brix*, pH o conductividad). Esto permite evaluar de forma aislada el aporte de cada propiedad al desempeño de los nodos del árbol, identificando si alguna de ellas genera por sí sola una mejora significativa en la separación de clases frente al modelo spec. Además, se busca determinar en qué nodos estas variables resultan más críticas.

Este análisis tiene un fin práctico: servir como base de diseño para situaciones en que solo se dispone de una medición de laboratorio, ayudando a establecer cuál debería priorizarse.

Los modelos M2, en cambio, integran las 18 señales espectrales junto a dos variables fisicoquímicas, explorando todas las combinaciones posibles. Esto permite estudiar la sinergia o redundancia entre propiedades: si se complementan, deberían mejorar el rendimiento respecto a su uso individual; si no aportan nueva información, pueden incluso degradarlo por solapamiento de datos.

5.8 Resultados modelos M1

A partir de las pruebas realizadas con cuatro versiones del modelo M1 (una por cada variable), se observó que ninguna alcanzó por sí sola el rendimiento del modelo Full, que incorpora todas las propiedades fisicoquímicas. No obstante, la inclusión exclusiva de %v/v* fue la que generó las mejoras más significativas. Por ejemplo, mientras que las versiones con pH o °Brix* apenas superaban el 85% (véase matrices de confusión de cada submodelo M1 en Anexo [A.2](#)) de exactitud en nodos como MS-ID, los modelos con %v/v* lograban precisiones notoriamente superiores, especialmente en casos donde el espectro por sí solo no ofrecía una separación suficientemente

robusta. Esto sugiere que esta variable introduce una separación intermedia efectiva, que recupera buena parte de la capacidad del modelo completo sin necesidad de medir todas las variables. A continuación, se detallan los principales saltos de desempeño observados al comparar los nodos de Spec, M1 ($\%v/v^*$) y Full, excluyendo nodos donde *Spec* ya era perfecto o donde Full no ofrecía mejoras adicionales.

- **MPNS-11:** Pasa de 90% (Spec) a 96% (M1), acercándose al 99% de Full. El aumento de 6 puntos evidencia cómo $\%v/v^*$ reduce las confusiones iniciales entre las 11 cepas.
- **SM-8:** M1 alcanza 99%, superando en 2 puntos a Full. El modelo con $\%v/v^*$ es más preciso para aislar específicamente el bloque Syrah-Malbec.
- **CFMSZ-ID:** M1 sube de 66% a 76%, un salto de 10 puntos que posiciona a $\%v/v^*$ como la variable de mayor impacto en este nodo crítico.
- **MS-ID:** M1 alcanza 93%, desde 79% en Spec, recuperando más de la mitad de la zona de solapamiento sin necesidad de otras propiedades. Full logra el 100%.

Por otro lado, en el análisis global del árbol, el modelo M1 con $\%v/v^*$ logra un *accuracy* y F1 del 87%, inferior al 99% de Full, pero superior a *Spec* en varias zonas problemáticas. Esto confirma que $\%v/v^*$ permite resolver muchas de las clases difíciles y delimita con claridad dónde su efecto es limitado.

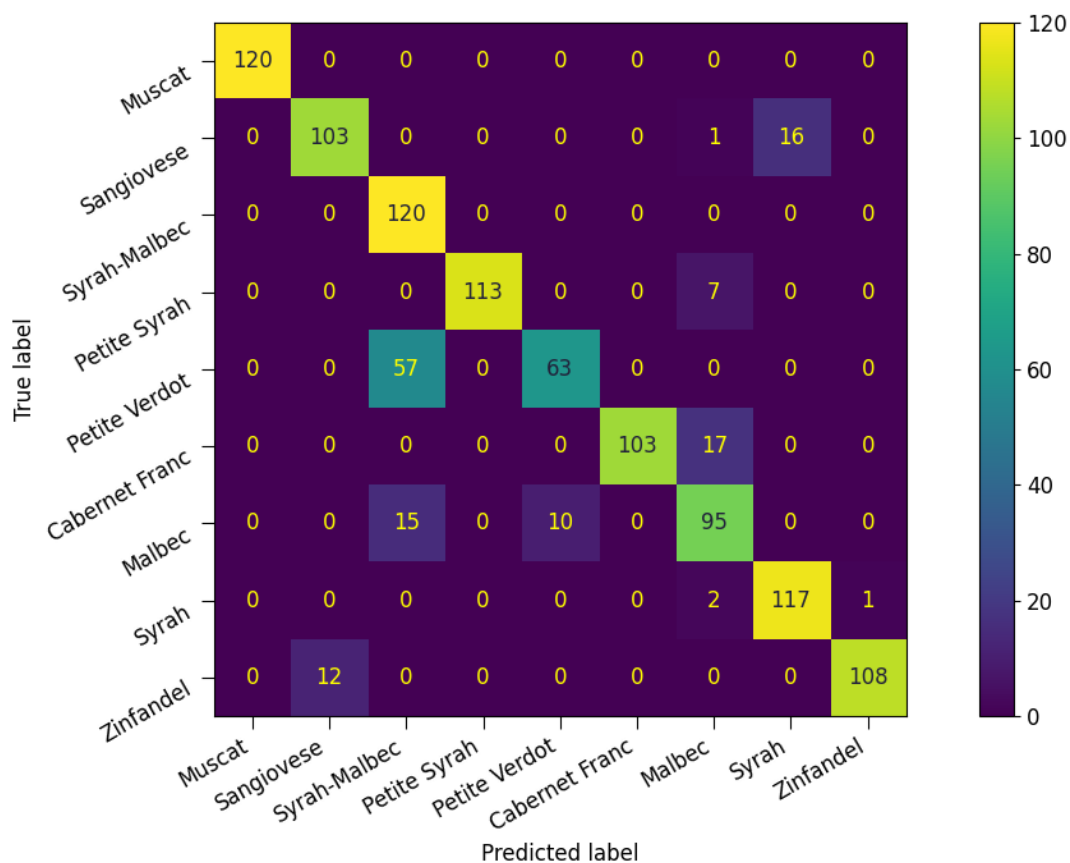


Figura 5.10: Matriz de confusión del árbol completo para el modelo V3, utilizando componentes M1 con el dataset de 11 cepas. En la diagonal principal se indican las muestras correctamente clasificadas para cada cepa, mientras que los valores fuera de ella muestran las malas clasificaciones entre cepas.

La matriz de confusión para M1 ($\%v/v^*$) con las 11 cepas permite observar comportamientos particulares:

- Sangiovese no presenta errores en Full, pero en M1 se confunde 14 veces: 13 con Syrah y 1 con Malbec. Aunque $\%v/v^*$ atenúa la fuerte confusión vista en Spec, no logra separar completamente a Sangiovese de cepas con mayor contenido alcohólico.
- Petite Syrah presenta 12 errores como Malbec. M1 no logra diferenciarla del todo.
- Petite Verdot añade 54 errores al bloque Syrah-Malbec y alcanza solo 66 aciertos, frente a los 87 de Full. Aunque $\%v/v^*$ mejora parcialmente el resultado, es insuficiente para igualar el desempeño completo.
- Cabernet Franc logra 105 de 120 aciertos, con 15 errores hacia Malbec. En Full se obtiene 100% de precisión, lo que indica que $\%v/v^*$ corrige parcialmente el problema, pero no completamente.

- Malbec acierta 88 veces, y se confunde 12 veces con SM y 20 con Petite Verdot. Full mejora con 97 aciertos.
- Syrah logra 117/120 aciertos (vs. 120 en Full), con errores menores hacia Malbec y Zinfandel.
- Zinfandel alcanza 111 aciertos, con 9 errores hacia SM. Full obtiene 120 de 120.

5.8.1 Análisis General modelo M1

El modelo M1, al incorporar únicamente la variable $\%v/v^*$ junto a las señales espectrales, logra mejoras sustanciales frente a Spec. Sin embargo, persisten confusiones en nodos como CFMSZ-ID, donde Malbec, Syrah y Zinfandel se solapan, mientras que el binario MS-ID logra clasificaciones casi perfectas. Esto indica que $\%v/v^*$ resuelve parte importante del solapamiento Malbec–Syrah, pero no corrige por completo los casos límite, lo que sugiere que una sola variable no es suficiente para alcanzar el rendimiento de Full. Surge entonces la duda de si una combinación con otra propiedad podría mejorar aún más el modelo.

Cuando se analiza un conjunto amplio de muestras, se identifican patrones sistemáticos asociados a cada cepa, reflejados como desplazamientos en los ejes del espacio LDA (Figura 5.11). En particular, el eje de $\%v/v^*$ muestra un mayor rango dinámico, generando subconjuntos de muestras claramente separados. Esto contrasta con las otras variables evaluadas (pH, conductividad y $^{\circ}\text{Brix}^*$), que presentan rangos más estrechos y un alto nivel de solapamiento entre cepas, lo que disminuye su capacidad discriminativa cuando se usan de forma aislada. Por ejemplo, varias variedades comparten un pH entre 3,3 y 3,7, o una conductividad cercana a 2,2 mS/cm, lo que limita su utilidad para separar clases complejas.

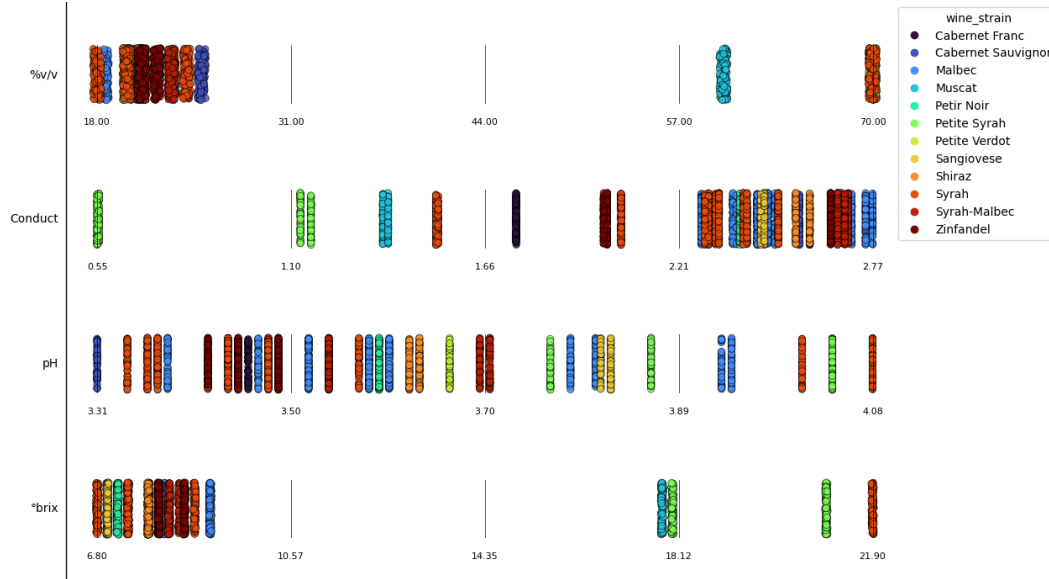


Figura 5.11: Dominio del espectro de cepas de vinos según las 4 variables fisicoquímicas utilizadas.

La estimación de refracción de mezcla alcohólica ($\%v/v^*$) refleja, además, variables integradas como el grado de madurez de la uva, concentración de azúcares fermentables y eficiencia fermentativa. Cada cepa presenta patrones únicos de acumulación de azúcares: por ejemplo, Malbec tiende a alcanzar menores niveles finales de alcohol que Syrah, producto de diferencias en la cinética de maduración y su perfil de compuestos como péptidos y polifenoles.

Como muestra la Figura 5.12, $\%v/v^*$ logra resolver uno de los principales problemas observados en Spec: el acoplamiento entre Malbec y Syrah. Se observa una zona crítica (aprox. $\%v/v^* = 18$) donde ambas cepas coexisten sin separación clara; sin embargo, también hay puntos Syrah en torno a $\%v/v^* = 70$ que permiten su separación efectiva. Esto hace que $\%v/v^*$ sea una variable altamente informativa y potencialmente suficiente como único descriptor jerárquico en ciertas regiones del árbol.

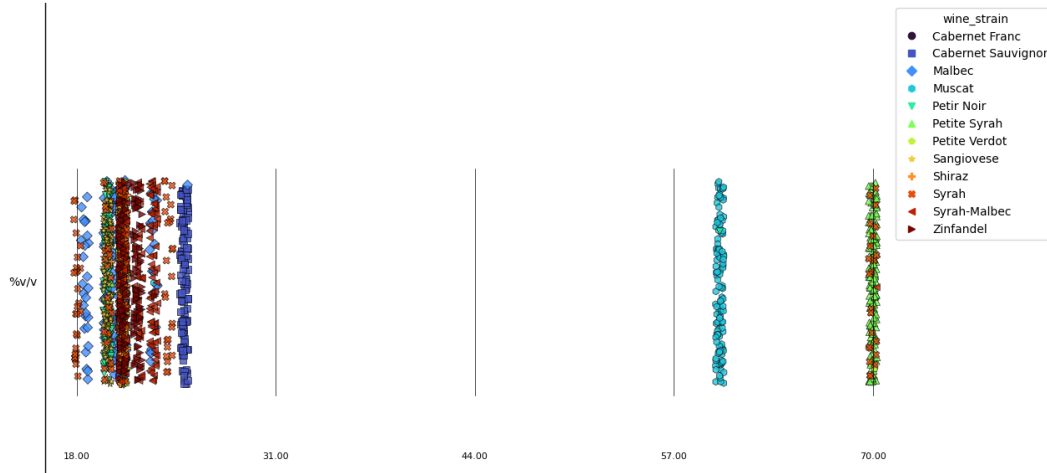


Figura 5.12: Dominio del espectro de cepas de vinos según la variable fisicoquímica $\%v/v^*$.

5.9 Resultados modelos M2

Para analizar la interacción entre variables fisicoquímicas, se definió el modelo M2, que combina las 18 señales espectrales con dos propiedades de las cuatro disponibles. El objetivo fue comprobar si una dupla seleccionada podía recuperar gran parte de la ganancia del modelo Full, superando a M1 sin necesidad de usar las cuatro variables.

Tras probar todas las combinaciones, se identificó que la pareja con mejor rendimiento global fue $\%v/v^* + \text{conductividad}$. No obstante, cuando se analizan los nodos del árbol, aunque se ven mejoras con respecto a Spec, la evaluación de M2 no alcanza los valores de Full, y en algunos casos iguala los valores de M1 (véase matrices de confusión de cada submodelo M2 en Anexo [A.3](#)), lo cual resalta la hipótesis de que la señal de conductividad está correlacionada con $\% v/v^*$, de modo que su valor incremental es menor en submódulos.

La evaluación global del árbol M2, el cual no necesariamente iguala la evaluación independiente de los nodos, como se ha visto en resultados anteriores, resulta en un *accuracy* global de 89% ($F1 = 89\%$), valor intermedio entre M1 (87%) y Full (91%).

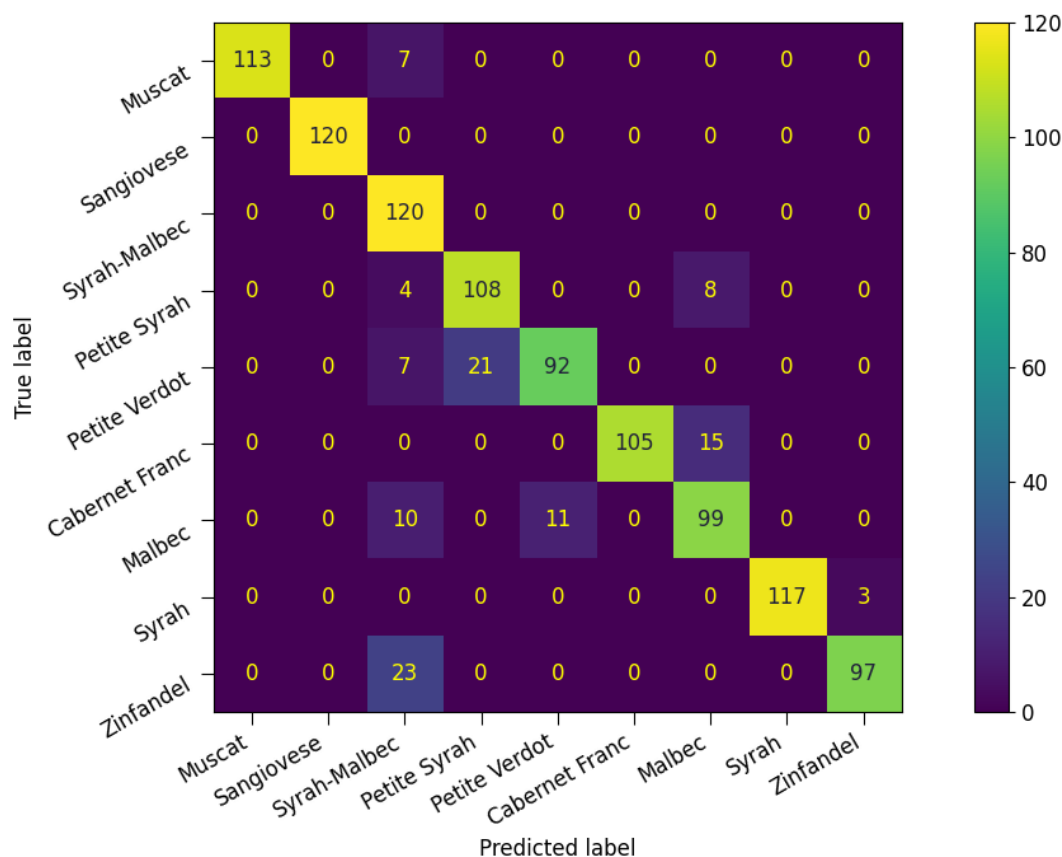


Figura 5.13: Matriz de confusión árbol completo para dataset M2 V3 de 11 cepas. En la diagonal principal se indican las muestras correctamente clasificadas para cada cepa, mientras que los valores fuera de ella muestran las mal clasificaciones entre cepas.

En general, añadir conductividad aporta separación adicional en ciertos casos, pero su efecto es menor que usar el conjunto completo de variables. Además, se puede notar que, dado el poco o nulo aumento que se tuvo con respecto a M1

Sin embargo, para separar por completo Malbec vs Syrah, la conductividad no logra explotar las pequeñas diferencias que sí se capturan agregando todas las variables, por lo que M2 no alcanza el 100% de Full.

De la matriz de confusión del árbol se puede notar que:

- **Muscat:** Tiene 7 errores hacia Syrah-Malbec. M1 no cometía estos errores; aquí la inclusión de conductividad ensancha la frontera SM-8 y atrapa algunas muestras erróneamente.
- **Sangiovese:** M1 se confundía con Syrah, M2 corrige esta confusión. $\%v/v^* +$ conductividad logra una separación más precisa en este caso.
- **Petite Syrah:** Resulta en 8 errores hacia Malbec (vs los 12 de M1). M2 mejora ligeramente el solapamiento Petit Syrah–Malbec.

- **Syrah:** Presenta 3 errores a comparación de los 2 en M1. Conductividad introduce un pequeño ruido adicional hacia Zinfandel.
- **Zinfandel:** 23 errores hacia SM (vs 9 en M1). El nodo se ve afectado por una frontera más laxa, lo que reduce precisión. La combinación de estas dos variables atrapa demasiadas muestras en SM.

5.9.1 Análisis General Modelo M2

Los resultados de M2 muestran que esta configuración recupera todas las confusiones de Sangiovese que persistían en M1, además de reducir parcialmente los errores en Petite Syrah, aunque sin corregir errores persistentes en cepas como Petite Verdot, Cabernet Franc y Malbec, y amplía las confusiones en Zinfandel y Muscat, especialmente en el nodo SM.

Para profundizar, la Figura 5.14 muestra la distribución espectral según conductividad. Allí se observa que cepas como Zinfandel y Cabernet Franc tienden a valores más altos, mientras que Petite Verdot y Petite Syrah se ubican a la izquierda del eje. Esta separación reduce la confusión en esas clases. Lo mismo ocurre con Muscat y Sangiovese, consideradas clases “fáciles”, cuya separación ya era clara incluso con únicamente espectros.

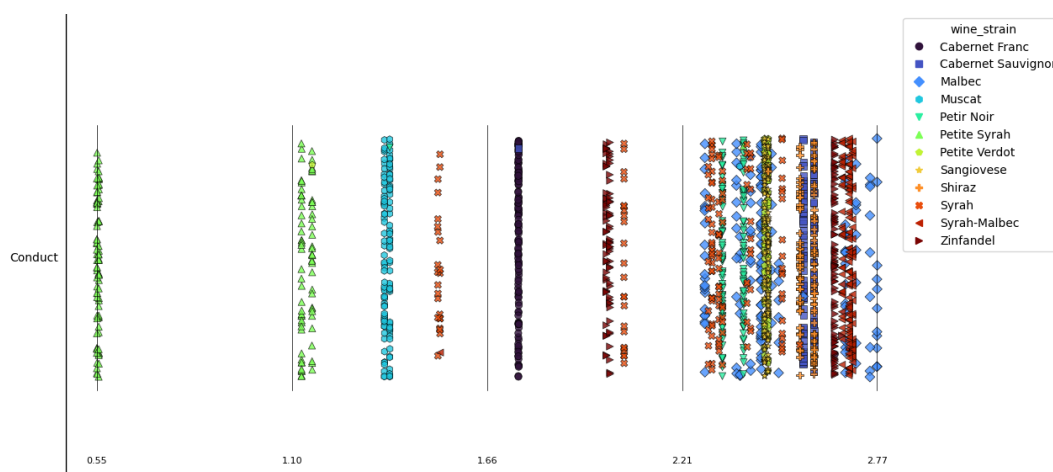


Figura 5.14: Dominio del espectro de cepas de vinos según las variable fisicoquímica conductividad.

No obstante, Syrah y Malbec aún se mezclan en una franja central, con un ligero desplazamiento de Malbec hacia valores más bajos. Syrah-Malbec, al ser un vino ensamblado de ambas cepas, forma un clúster intermedio que abarca zonas de ambas distribuciones. Esto explica por qué el separador SM-8 identifica bien ese bloque mixto, pero luego, las muestras reales de Syrah y Malbec quedan demasiado próximas entre sí, generando errores en las ramas siguientes del árbol.

Zinfandel, por su parte, amplía su solapamiento con SM (18 muestras atrapadas). El nodo SM-8 no discrimina bien el extremo inferior de conductividad, donde Zinfandel se ubica. Además,

existen regiones donde se solapan simultáneamente Malbec, Syrah y Syrah-Malbec, lo que agrava la ambigüedad en ambas dimensiones ($\%v/v^*$ y conductividad).

Esto revela que un clasificador basado solo en espectro + dos propiedades no es suficiente para resolver todos los casos límite. Para ello, sería necesario añadir un tercer o cuarto eje, como pH y/o $^{\circ}\text{Brix}^*$. Aun así, la conductividad aporta un eje de dispersión adicional relevante, especialmente en combinación con $\%v/v^*$. Permite descubrir diferencias sutiles (como entre Sangiovese y Malbec) que M1 no captaba. Sin embargo, su rango dinámico es menor en zonas críticas como Malbec/Syrah, por lo que su aporte es valioso pero insuficiente para alcanzar el rendimiento de Full. No obstante, M2 se aproxima significativamente.

5.10 Jerarquía de variables fisicoquímicas

A partir de los experimentos con los modelos M1 y M2, se estableció una jerarquía clara de las cuatro propiedades fisicoquímicas evaluadas, en función de su impacto en la clasificación jerárquica de cepas. Este orden considera tanto las mejoras cuantitativas (*accuracy*, F1 en nodos críticos) como la capacidad de cada variable para “despejar” regiones densamente pobladas del espacio de características.

pH, conductividad eléctrica (mS), grados Brix sobrestimados ($^{\circ}\text{Brix}^*$) y una estimación de refracción de mezcla alcohólica($\%v/v^*$)

La estimación de refracción de mezcla alcohólica($\%v/v^*$) destaca como la más influyente, al elevar el rendimiento de *Spec* (79,35%) hasta 87,04% en M1 y acercarse al valor de Full (91,20%). Además, permite resolver gran parte de los solapamientos entre Sangiovese vs Malbec y Cabernet Franc vs Malbec. En la Figura 5.12 se observa cómo esta variable separa claramente cepas como Muscat, Petite Verdot y Sangiovese de las regiones densas Malbec-Syrah-Syrah-Malbec, gracias a su amplio rango dinámico. La conductividad ocupa el segundo lugar. Su combinación con $\%v/v^*$ en M2 permite alcanzar un 89,26% de *accuracy* (F1 = 89,29%), corrigiendo errores que M1 no resolvía, como los de Sangiovese. La Figura 5.14 muestra cómo esta variable aporta una dispersión adicional, especialmente útil en cepas difíciles (Malbec, Syrah, Zinfandel), aunque su efecto es menor al estar parcialmente correlacionada con $\%v/v^*$.

El grado Brix sobrestimado ($^{\circ}\text{Brix}^*$) ocupa la tercera posición. Su efecto en M1 fue reducido (ganancia global <4%), y en M2 no superó a la combinación $\%v/v^*$ + conductividad. Como se observa en la Figura 5.15, su rango estrecho (6,8 a 10,5) genera solapamientos importantes entre cepas. Aun así, aporta cierta separación en valores extremos: Muscat en el rango inferior y Petite Syrah en el superior.

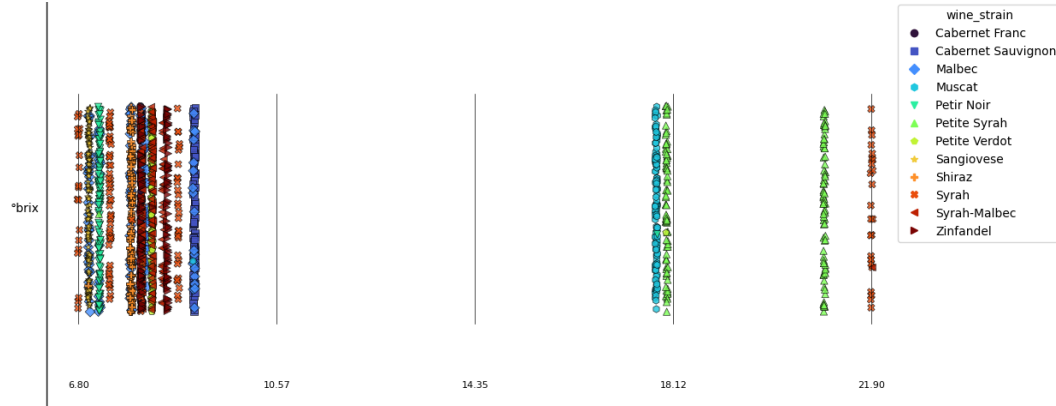


Figura 5.15: Dominio del espectro de cepas de vinos según las variable fisicoquímica °Brix*.

Finalmente, el pH se posiciona como la variable menos relevante. Con un rango entre 3,3 y 4,1, presenta un gran solapamiento entre cepas como Cabernet Franc, Malbec, Syrah y Sangiovese (véase Figura 5.16). Solo algunas cepas, como Cabernet Sauvignon (pH $\approx$ 3,89) o Syrah-Malbec ($\approx$ 4,08), muestran desplazamientos, aunque en clases ya bien diferenciadas. Su principal ventaja radica en la facilidad y rapidez de medición.

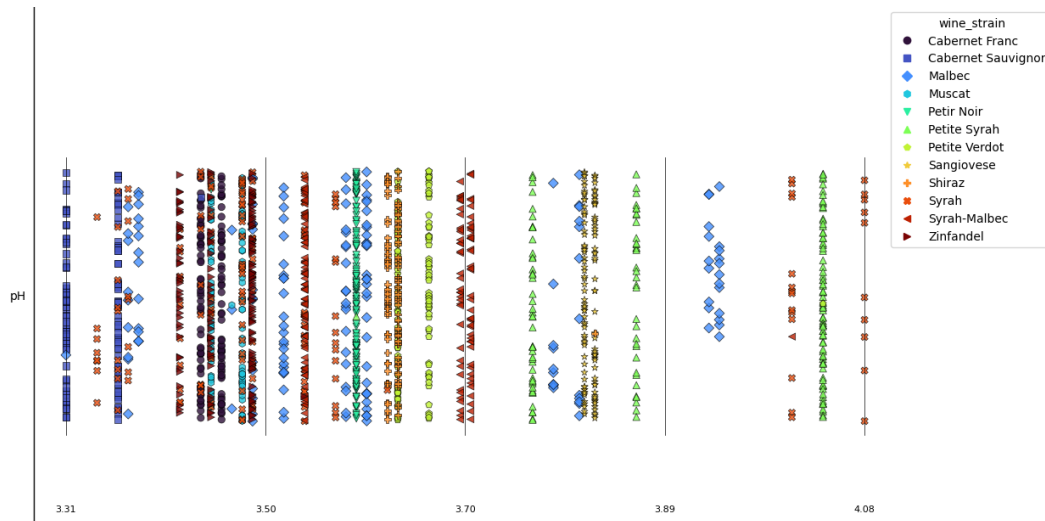


Figura 5.16: Dominio del espectro de cepas de vinos según las variable fisicoquímica pH.

En base a estos resultados, se propone el siguiente orden de prioridad para mediciones fisicoquímicas:

1. **%v/v***: por su alta capacidad para separar zonas críticas con solapamientos intensos.
2. **Conductividad**: por su efecto complementario y su aporte en nodos intermedios.
3. **°Brix***: útil en extremos, aunque limitado en zonas centrales.

4. **pH**: solo recomendable en contextos con recursos limitados o como complemento de caracterización.

Este orden permite optimizar recursos de laboratorio, priorizando variables que entregan mayor ganancia clasificatoria, y considerando la incorporación de otras solo cuando se requiera una separación más fina entre cepas.

5.11 Efecto de la inyección de ruido sintético sobre la agrupación de cepas en LDA

Las figuras presentadas a continuación muestran las alteraciones en la agrupación de cepas (Malbec, Syrah, Syrah-Malbec y Zinfandel) en la representación obtenida mediante Análisis Discriminante Lineal (LDA) al introducir ruido sintético tipo gaussiano (b) y *salt and pepper* (c). Estas alteraciones son analizadas comparativamente respecto a la configuración original, sin ruido, (a) para destacar los fenómenos observados claramente.

En ambos tipos de ruido, la distribución original se distorsiona significativamente. Malbec, que inicialmente se agrupaba con Syrah, pierde esta relación al aumentar su proximidad a Syrah-Malbec, incluso con niveles bajos de ruido (1–5%).



Figura 5.17: Comparación entre la distribución original sin ruido (izquierda) y la distribución resultante tras la inyección de ruido gaussiano (derecha), proyectadas mediante LDA para las cepas Malbec, Syrah, Syrah-Malbec y Zinfandel.



Figura 5.18: Comparación entre la distribución original sin ruido (izquierda) y la distribución resultante tras la inyección de ruido *salt and pepper* (derecha), proyectadas mediante LDA para las cepas Malbec, Syrah, Syrah-Malbec y Zinfandel.

De la Figura 5.17 se evidencia que el ruido gaussiano introduce variaciones continuas que suavizan las fronteras espectrales, en particular entre Malbec y Syrah-Malbec. Dado el dominio limitado de las muestras, el modelo LDA interpreta erróneamente esta variabilidad como similitud inherente, generando agrupaciones inesperadas. Así, pequeñas perturbaciones terminan modificando drásticamente la representación espectral.

Por otro lado, el ruido *salt and pepper* (Figura 5.18) inserta valores extremos puntuales en el espectro, generando rupturas abruptas en las agrupaciones. Malbec y Syrah, que antes estaban separadas por diferenciación suficiente, se ven disgregadas. Nuevamente, Syrah se reagrupa con Syrah-Malbec, esta vez por la redistribución forzada de puntos "anómalos", lo que genera una representación menos realista. Este efecto revela que un conjunto pequeño de muestras no puede resistir adecuadamente perturbaciones espectrales extremas.

Para entender el mal comportamiento que tuvo el ruido sintético para las muestras, clave entender que el vino presenta una matriz espectral compleja y variable, influida por múltiples factores físicos y químicos como cepa, condiciones climáticas de cosecha, tiempo en barrica, temperatura de almacenamiento, entre otros (García (2017)). Cada vino tiene un perfil espectral único, cuya representación adecuada requiere un conjunto amplio y representativo de datos.

En este contexto, la inyección de ruido busca simular variabilidad, pero cuando se aplica sobre un dominio pequeño o poco representativo, no logra emular la complejidad real del vino. El modelo LDA, en respuesta, reestructura las agrupaciones para acomodar estos patrones artificiales, degradando la separación entre clases.

Por tanto, aunque el ruido sintético puede ser útil en otros contextos, en este caso particular, con una matriz altamente variable y bajo volumen de datos, no aporta valor para ampliar el dominio ni

mejorar la clasificación. Al contrario, puede inducir patrones ajenos al comportamiento real del vino y conducir a conclusiones erróneas o a una disminución significativa de la capacidad discriminativa del modelo.

5.12 Automatización de agrupación de cepas para los nodos del árbol de modelos

Durante la validación del modelo jerárquico de clasificación de cepas de vino, se observó una disminución significativa en su desempeño al incorporar nuevos datos de las mismas cepas, pero con ligeras variaciones en condiciones de almacenamiento, como temperatura o tiempo en bodega. Esta incorporación redujo el rendimiento del sistema hasta un 60%, evidenciando una limitante crítica del enfoque original: la estructura del árbol se definió de forma estática, basada en la visualización de proyecciones LDA (2D y 3D) construidas sobre un volumen reducido de datos.

Este carácter estático implica que el modelo refleja solo las relaciones observadas en el conjunto inicial, sin considerar la variabilidad natural de la agrupación de una cepa ante nuevos contextos espectrales. No contempla interacciones dinámicas asociadas a nuevas mediciones o a factores relevantes en matrices complejas como el vino, lo que limita su capacidad de generalización.

Como consecuencia, el modelo no puede garantizar que las agrupaciones observadas inicialmente se mantendrán al integrar más datos, ni anticipar cómo cambiarán con nueva información. Esta falta de adaptabilidad compromete su robustez ante cambios experimentales o nuevos lotes de muestras.

Frente a esta problemática, surge la necesidad de explorar soluciones que doten al sistema de mayor flexibilidad y capacidad de adaptación, para mejorar su rendimiento y mantener su utilidad ante actualizaciones continuas. A continuación, se presenta la solución propuesta para abordar esta limitación.

5.12.1 Solución basada en agrupación automática por matriz de confusión

Para resolver la falta de robustez del árbol de modelos ante actualizaciones de datos, se desarrolló una solución que agrupa las cepas de forma automática y dinámica, en función de cómo un modelo de clasificación inicial tiende a confundirse entre ellas, evitando depender de una organización visual o manual (como la basada en LDA). La estrategia, descrita a continuación, se inspira en el trabajo de [Xiong \(2012\)](#), quien propone construir estructuras jerárquicas a partir de matrices de confusión de clasificadores planos para mejorar el rendimiento global del sistema de clasificación.

Esta metodología permite que la estructura del árbol se ajuste a las dificultades reales del modelo para separar cepas similares, haciéndolo más robusto frente a nuevos datos o cambios experimentales.

El proceso comienza con una función que materializa las confusiones del clasificador. Para ello, se define un modelo base (SVM con normalización y eliminación de variables de baja varianza) y se aplica validación cruzada estratificada en K-folds. En cada iteración, el modelo entrena con una partición y predice en test, acumulando una matriz de confusión general que refleja, "fold a fold", qué cepas se confunden con mayor frecuencia.

Luego, esta matriz se normaliza fila a fila, convirtiendo los errores en una matriz de similitud, que se transforma en una matriz de distancias simétricas. Se aplica el método de Ward para generar un dendrograma jerárquico que grafica la proximidad entre cepas según su índice de confusión (Figura 5.19). A partir de este dendrograma, se elige el número óptimo de grupos/clusters, buscando el mejor rendimiento.

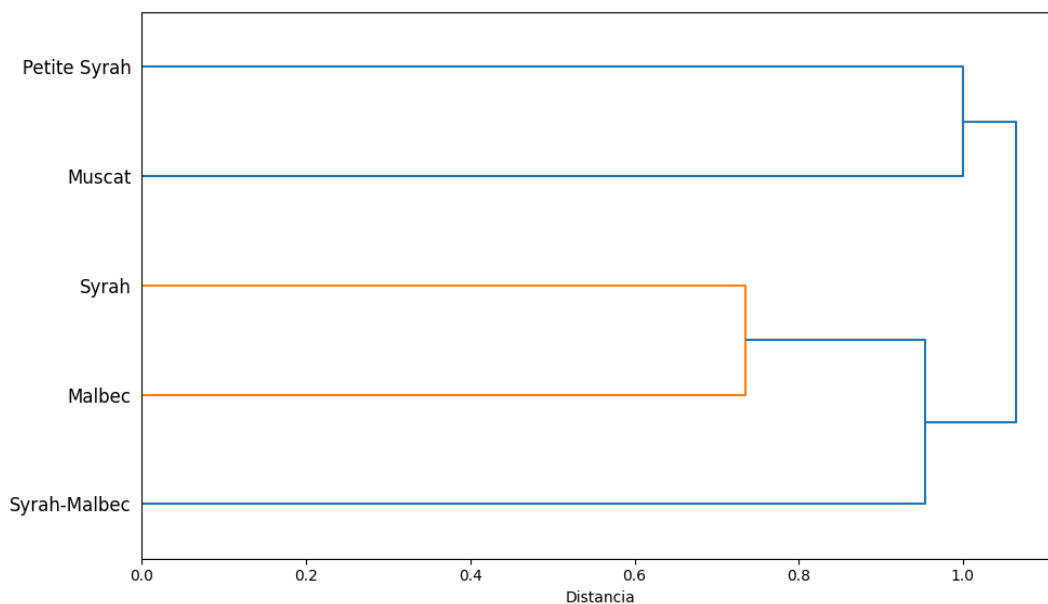


Figura 5.19: Dendrograma de agrupación jerárquica de cepas seleccionadas para prueba, creado a partir de la matriz de confusión del modelo. Las longitudes de las ramas reflejan la “distancia de confusión” entre cepas.

Una vez definidos los grupos, se crea una función que asigna cada cepa a un grupo. Esta estructura se comunica mediante un diccionario, que permite reetiquetar las muestras originales con su nuevo grupo. Por ejemplo, si Malbec y Petite Verdot son difíciles de distinguir, podrían quedar en el mismo grupo.

Con estas nuevas etiquetas, se reescribe la tarea de clasificación: ahora el modelo no predice la cepa individual, sino el grupo de confusión. Sobre esta nueva variable objetivo se entrena

un segundo modelo (con escalado, normalización y regresión logística), sometido nuevamente a validación cruzada. Se calculan métricas como *accuracy* y F1 para verificar la capacidad del modelo de capturar correctamente estas “familias” de cepas.

Para consolidar el aprendizaje de los distintos folds (subconjuntos en los que se divide el conjunto de datos para validación cruzada), se implementa un VotingClassifier con pesos suaves, ensamblando los clasificadores entrenados durante la creación de la matriz de distancias simétricas, sin necesidad de reentrenamiento. Esto permite reducir la varianza de predicción y aumentar la estabilidad ante nuevos datos. En la fase de test, se compara su rendimiento con el de un clasificador individual, observando mejoras en métricas clave.

En conjunto, esta implementación garantiza que la estructura jerárquica se derive de las confusiones reales del modelo, en lugar de una segmentación estática. Esto permite recalculas las agrupaciones cada vez que se incorporan nuevos datos (por ejemplo, cepas con distintos tiempos de barrica), adaptándose automáticamente. El sistema logra así una clasificación más robusta y generalizable frente a la heterogeneidad de la matriz espectral, resolviendo la degradación del rendimiento al actualizar el conjunto de datos.

5.12.2 Evaluación de la automatización de agrupación de cepas para los nodos del árbol de modelos

Tras aplicar el código de agrupación automática, se construyó una matriz de confusión acumulada entre cepas, donde cada fila representa la cepa real y cada columna la predicción del modelo base durante validación cruzada. A partir de esta matriz, se calculó una métrica de distancia (1 – similitud normalizada por fila) y se realizó clustering jerárquico, identificando tres grupos óptimos:

- **Grupo 1:** Malbec, Syrah, Syrah-Malbec
- **Grupo 2:** Muscat
- **Grupo 3:** Petite Syrah

Se utilizó este subconjunto de cinco cepas porque fue con el que se obtuvo el mejor rendimiento del modelo bajo esta estrategia automática, sin necesidad de ajustes adicionales. Esto permitió evaluar cuál era el límite máximo de cepas que podían agruparse manteniendo un desempeño alto (sobre 95% de evaluación en todas las ejecuciones), antes de que comenzara a disminuir la precisión del sistema.

Este esquema refleja agrupaciones consistentes entre cepas que el clasificador tiende a confundir, constituyendo así una base sólida para establecer nuevos niveles jerárquicos de clasificación. No obstante, se reconoce que esta propuesta aún tiene espacio para mejoras y ajustes, los cuales no

fueron abordados debido a las limitaciones de tiempo del presente trabajo, quedando abierta la posibilidad de continuar su desarrollo en investigaciones futuras.

En su estado actual, la solución demuestra un potencial sólido para organizar dinámicamente árboles jerárquicos de clasificación, adaptándose a nuevas muestras o condiciones. Sin embargo, aún requiere más desarrollo y validación externa para integrarse como un componente estable dentro del sistema completo de automatización de tu memoria de título.

Capítulo 6

Conclusiones y Recomendaciones

Se desarrolló y validó un método automatizado que integra espectroscopía visible con algoritmos de inteligencia artificial, orientado a la clasificación rápida, no destructiva y precisa de cepas de vino. El sistema se basó en un dispositivo portátil equipado con el sensor AS7265X y una metodología computacional capaz de adquirir, normalizar y analizar los espectros, aplicar preprocesamiento, reducción de dimensionalidad, aumentación de datos y entrenar un árbol jerárquico de modelos. Este pipeline completo logró transformar datos crudos de conteo en predicciones de cepa en tiempo real, con un rendimiento que alcanzó un accuracy del 80% para *Spec* y 93% para *Full* al aplicar el modelo jerárquico. Estos resultados son altamente relevantes considerando el reducido tamaño del dataset y la diversidad de cepas involucradas.

El escenario *Spec* incluye 36 variables de entrada (las 18 longitudes de onda y sus respectivos 18 blancos normalizados), mientras que *Full* incorpora además 4 variables fisicoquímicas, totalizando 40 entradas por muestra. En ambos casos, la trazabilidad y separación estratificada de los datos permitió mantener consistencia en la validación cruzada. A través del análisis visual con LDA, se observó que la inclusión de estas variables fisicoquímicas permite una mejor separación de los clústeres entre cepas que comparten espectros similares, fortaleciendo la capacidad discriminativa del modelo.

Además de la clasificación espectral, se evaluó el impacto de las propiedades fisicoquímicas del vino mediante una jerarquización de variables aplicada en modelos supervisados M1 y M2. El modelo M1, construido únicamente con la variable $\%v/v^*$, alcanzó un rendimiento destacado, evidenciando que esta es la variable de mayor poder discriminativo entre cepas. El modelo M2 incorporó además la conductividad, lo que permitió una mejora marginal en la clasificación, deduciendo que esta no era más que un complemento de la información entregada por $\%v/v^*$, pero sin modificar significativamente la estructura general del agrupamiento. Para comprender mejor la influencia de estas variables, se realizó un análisis LDA utilizando las cuatro propiedades fisicoquímicas, el cual mostró una separación más clara entre clústeres en comparación con el análisis espectral puro.

A partir de este análisis, se estableció un orden de importancia que posiciona a %v/v* como la variable más influyente, seguida por conductividad, °Brix* y, en último lugar, pH.

En cuanto a las estrategias de aumentación de datos, se implementó ruido computacional controlado (gaussiano entre 0,5% y 5% σ , y *salt and pepper* entre 1% y 3% de puntos). Sin embargo, se evidenció que esta técnica no logra extrapolar la agrupación natural de cepas. El análisis LDA mostró que el ruido alteraba la estructura de los clústeres: Malbec, que originalmente se acoplaba con Syrah, pasó a acoplarse con Syrah-Malbec. Esto revela que la simple agregación de ruido no es suficiente frente a la complejidad de la matriz composicional del vino, influenciada por factores como el año de cosecha, condiciones climáticas, fermentación, tiempo en barrica y otras variables críticas. Por tanto, se concluye que la robustez ante nuevas muestras requiere representar esa variabilidad real mediante un muestreo más amplio y no únicamente mediante aumentación estadística.

El conjunto de datos base fue generado con el script `make_data.py`, que automatiza la conexión entre espectros, blancos y metadatos fisicoquímicos. Este flujo permitió implementar tanto el escenario *Spec* como *Full* con estructura estandarizada, cálculo de color CIELAB, normalización y filtrado de muestras, garantizando la integridad del análisis.

El modelo final consistió en un árbol jerárquico compuesto por cuatro nodos (MPNS-ID, SM-8, CSPSPV-7 y CFMSZ-ID), entrenados con *pipelines* ajustados a la complejidad de cada subconjunto. Se aplicaron técnicas como escalado, reducción de dimensionalidad (PCA y LDA), y clasificadores como *SVM* y *Random Forest*. Esta arquitectura logró un *accuracy* unitario del 100 % en nodos claramente separados, 88 % de *accuracy* global con el modelo *Spec* y 93 % en el escenario *Full*. La estrategia *coarse-to-fine* demostró ser superior a los modelos monolíticos, ya que permitió adaptar la complejidad del modelo a cada etapa de clasificación.

Sin embargo, al incorporar nuevas muestras de las mismas cepas pero con distinto tiempo en barrica, el rendimiento del modelo jerárquico cayó al 60 %, evidenciando que su estructura estática no generaliza ante variaciones reales no vistas. Esta limitación refleja los desafíos de implementación industrial, donde la variabilidad es parte del producto y no puede capturarse completamente en los datos iniciales. Para abordar este desafío, se propuso una metodología dinámica basada en *clustering* jerárquico sobre la matriz de confusión acumulada, que permite redefinir la jerarquía según las confusiones reales observadas. Esta solución permite adaptar el sistema a condiciones cambiantes, con un enfoque más resiliente y representativo del entorno productivo.

En síntesis, este trabajo establece las bases para el desarrollo de herramientas portátiles de espectroscopía e inteligencia artificial aplicadas a la industria vitivinícola, demostrando que es posible alcanzar altos niveles de precisión con infraestructura accesible y modelos bien diseñados.

6.1 Recomendaciones

- Ampliar el volumen de datos, especialmente de cepas menos representadas, para fortalecer la generalización y explorar modelos de *deep learning*.
- Optimizar el prototipo físico, mejorando la resolución espectral, la relación señal/ruido y la robustez mecánica para uso en terreno.
- Validar el prototipo móvil periódicamente en viñedos bajo condiciones reales (temperatura, iluminación) para ajustar algoritmos y robustecer el sistema.
- Implementar algoritmos de auto-calibración basados en blancos y referencias internas, para mantener la estabilidad de las mediciones en el tiempo.
- Mantener un enfoque experimental flexible, capaz de incorporar nuevas fuentes de variabilidad y ajustar dinámicamente las estrategias de agrupamiento y clasificación, garantizando robustez, generalización y aplicabilidad real del sistema.

Referencias

- Abdi, H. and Williams, L. J. (2010). Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459.
- Almanza-Merchán, P. J., Reyes, A. J., Ayala, M. L., Balaguera, W., and Serrano-Cely, P. A. (2015). Evaluación sensorial del vino artesanal de uva isabella (*vitis labrusca* l.). *Ciencia y Agricultura*, 12(2):71–81.
- An, G., Akiba, M., Omodaka, K., Nakazawa, T., and Yokota, H. (2021). Hierarchical deep learning models using transfer learning for disease detection and classification based on small number of medical images. *Scientific reports*, 11(1):4250.
- Beć, K. B., Grabska, J., and Huck, C. W. (2021). Principles and applications of miniaturized near-infrared (nir) spectrometers. *Chemistry–A European Journal*, 27(5):1514–1532.
- Bhowmik, S., Hasan, M., and Hakim, M. A. (2021). A dimensionality reduction based efficient multiple voice disease recognition scheme using mel-frequency cepstral coefficients and k-nearest neighbors algorithm. In *Proceedings of the International Conference on Big Data, IoT, and Machine Learning: BIM 2021*, pages 301–313. Springer.
- Bindon, K., Kassara, S., and Smith, P. A. (2017). Towards a model of grape tannin extraction under wine-like conditions: the role of suspended mesocarp material and anthocyanin concentration. *Australian Journal of Grape and Wine Research*, 23(1):22–32.
- Bindon, K. A., Kassara, S., Cynkar, W. U., Robinson, E. M., Scrimgeour, N., and Smith, P. A. (2014). Comparison of extraction protocols to determine differences in wine-extractable tannin and anthocyanin in *vitis vinifera* l. cv. shiraz and cabernet sauvignon grapes. *Journal of Agricultural and Food Chemistry*, 62(20):4558–4570.
- Bustos, M., Duarte-Mermoud, M., Beltrán, N., Salah, S., Loyola, E., Peña-Neira, A., and Jalocha, J. (2004). Clasificación de vinos chilenos usando un enfoque bayesiano. *Viticultura y Enología Profesional*, 90:63–70.

- Cáceres-Mella, A., Peña-Neira, Á., Avilés-Gálvez, P., Medel-Marabolí, M., del Barrio-Galán, R., López-Solís, R., and Canals, J. M. (2014). Phenolic composition and mouthfeel characteristics resulting from blending chilean red wines. *Journal of the Science of Food and Agriculture*, 94(4):666–676.
- Cetó, X., Gutiérrez-Capitán, M., Calvo, D., and Del Valle, M. (2013). Beer classification by means of a potentiometric electronic tongue. *Food chemistry*, 141(3):2533–2540.
- Chicco, D. and Jurman, G. (2020). The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21:1–13.
- Cozzolino, D., Kwiatkowski, M., Damberg, R., Cynkar, W., Janik, L., Skouroumounis, G., and Gishen, M. (2008). Analysis of elements in wine using near infrared spectroscopy and partial least squares regression. *Talanta*, 74(4):711–716.
- Cubo, E., Mir Rivera, P., and Sánchez Ferro, Á. (2021). Manual sen de nuevas tecnologías en trastornos del movimiento. *Ediciones SEN: Madrid, Spain*.
- de Chile, P. U. C. (2025). Mención viticultura y enología. Último acceso: 13 marzo 2025.
- de Ingenieros Agrónomos Enólogos (ANIAE), A. (2023). Informe de previsión de vendimia, 2023. Technical report, Oficina de Estudios y Políticas Agrarias (ODEPA).
- Donais, M. K. and Vandenabeele, P. (2021). Portable spectroscopy for on-site and in situ archaeology studies. *Portable Spectroscopy and Spectrometry*, pages 523–544.
- Fan, S., Zhang, Y., and Li, Y. (2023). A new approach for quantitative classification of red wine color from the perspective of micro and macro levels. *Fermentation*, 9(6):519.
- Feng, L., Wu, B., Zhu, S., He, Y., and Zhang, C. (2021). Application of visible/infrared spectroscopy and hyperspectral imaging with machine learning techniques for identifying food varieties and geographical origins. *Frontiers in Nutrition*, 8:680357.
- García, L. (2017). Análisis del contenido polifenólico en vinos tintos mediante lengua electrónica, ftir y uv-vis. *Valladolid*.
- Gardeli, C., Sykioti, S., Exarchos, G., Koliatsou, M., Andritsos, P., and Panagou, E. Z. (2025). The differentiation of extra virgin olive oil from other olive oil categories based on ftir spectroscopy and random forest. *Applied Sciences*, 15(3):1061.
- Ghojogh, B. and Crowley, M. (2019). The theory behind overfitting, cross validation, regularization, bagging, and boosting: tutorial. *arXiv preprint arXiv:1905.12787*.

- Grafberger, S., Groth, P., and Schelter, S. (2023a). Automating and optimizing data-centric what-if analyses on native machine learning pipelines. *Proceedings of the ACM on Management of Data*, 1(2):1–26.
- Grafberger, S., Guha, S., Groth, P., and Schelter, S. (2023b). mlwhatif: What if you could stop re-implementing your machine learning pipeline analyses over and over? *Proceedings of the VLDB Endowment*, 16(12):4002–4005.
- Granados-Tavera, K., Tilvez, E. A., and Ahumado-Monterrosa, M. (2019). Modelado de las relaciones cuantitativas estructura-actividad (qsar) de análogos del tipifarnib con actividad antichagásica. *Información tecnológica*, 30(1):3–14.
- Guan, Q., Lim, Z. H., Sun, H., Chew, J. X. Y., and Zhou, G. (2023). Review of miniaturized computational spectrometers. *Sensors*, 23(21):8768.
- Gullifa, G., Barone, L., Papa, E., Giuffrida, A., Materazzi, S., and Risoluti, R. (2023). Portable nir spectroscopy: The route to green analytical chemistry. *Frontiers in Chemistry*, 11:1214825.
- Han, D. and Guo, C. (2024). Automatic classification of ligneous leaf diseases via hierarchical vision transformer and transfer learning. *Frontiers in Plant Science*, 14:1328952.
- Hensel, M., Di Nonno, S., Mayer, Y., Scheiermann, M., Fahrner, J., Durner, D., and Ulber, R. (2022). Specification and simplification of analytical methods to determine wine color. *Processes*, 10(12):2707.
- Jaramillo, P. (2025). Asesoría técnica y desarrollo de algoritmos para detección espectroscópica. <https://orcid.org/0009-0006-6202-4907>. Comunicación personal.
- Laffort (2021). *USA Handbook – Technical Enology Manual*. Laffort Oenologie. Accedido en junio de 2025.
- Lin, C.-H., Kaushik, C., Dyer, E. L., and Muthukumar, V. (2024). The good, the bad and the ugly sides of data augmentation: An implicit spectral regularization perspective. *Journal of Machine Learning Research*, 25(91):1–85.
- Luo, R., Popp, J., and Bocklitz, T. (2022). Deep learning for raman spectroscopy: a review. *Analytica*, 3(3):287–301.
- Maimon, O. Z. and Rokach, L. (2014). *Data mining with decision trees: theory and applications*, volume 81. World scientific.

- Mamede, R., Pereira, F., and Aires-de Sousa, J. (2021). Machine learning prediction of uv-vis spectra features of organic compounds related to photoreactive potential. *Scientific Reports*, 11(1):23720.
- Manai, E., Mejri, M., and Fattahi, J. (2024). Confusion matrix explainability to improve model performance: Application to network intrusion detection. In *2024 10th International Conference on Control, Decision and Information Technologies (CoDIT)*, pages 1–5. IEEE.
- Martelo Vidal, M. J. (2014a). *Desarrollo de métodos rápidos basados en espectroscopía UV/VIS/NIR para el análisis de vinos*. PhD thesis, Universidade de Santiago de Compostela.
- Martelo Vidal, M. J. (2014b). *Desarrollo de métodos rápidos basados en espectroscopía UV/VIS/NIR para el análisis de vinos*. PhD thesis, Universidade de Santiago de Compostela.
- Martinez, A. M. and Kak, A. C. (2001). Pca versus lda. *IEEE transactions on pattern analysis and machine intelligence*, 23(2):228–233.
- Martínez-Pinilla, O., Martínez-Lapuente, L., Guadalupe, Z., and Ayestarán, B. (2012). Sensory profiling and changes in colour and phenolic composition produced by malolactic fermentation in red minority varieties. *Food research international*, 46(1):286–293.
- Mengstu, M., Taner, A., and Duran, H. (2024). Spectroscopy and machine learning in food processing survey. In *BIO Web of Conferences*, volume 85, page 01022. EDP Sciences.
- Nuñez, C. et al. (2020). *Fusión de imágenes visible e infrarroja utilizando morfología matemática multiescala*. PhD thesis, Facultad de Ciencias y Tecnología Universidad Autónoma de Asunción Tesis
- Odepa (2024a). Boletín del vino. Technical report, Oficina de Estudios y Políticas Agrarias, Gobierno de Chile.
- Odepa (2024b). Vinos. Technical report, Oficina de Estudios y Políticas Agrarias, Gobierno de Chile.
- of Manchester, U. (2024). Introducción a la espectroscopia molecular. In *Coursera*.
- Pan, L., Zhang, P., Daengngam, C., Peng, S., and Chongcheawchamnan, M. (2022). A review of artificial intelligence methods combined with raman spectroscopy to identify the composition of substances. *Journal of Raman Spectroscopy*, 53(1):6–19.
- Pérez-Planells, L., Delegido, J., Rivera-Caicedo, J. P., and Verrelst, J. (2015). Análisis de métodos de validación cruzada para la obtención robusta de parámetros biofísicos. *Revista de teledetección*, 44(-):55–65.

- Ponce, C. (2019). Monitoreo y control en la producción del cultivo de uva de vino (*vitis vinífera*) variedades malbec, merlot y tempranillo en el viñedo el lobo. Master’s thesis, Universidad Politécnica del Bicentenario.
- Rosli, M. M., Tempero, E., and Luxton-Reilly, A. (2016). What is in our datasets? describing a structure of datasets. In *Proceedings of the Australasian Computer Science Week Multiconference*, pages 1–10.
- SAG (2024). Informe final cosecha 2024. Technical report, Servicio Agrícola y Ganadero, Gobierno de Chile.
- Salman, S. and Liu, X. (2019). Overfitting mechanism and avoidance in deep neural networks. *arXiv preprint arXiv:1901.06566*.
- Servicio Agrícola y Ganadero (SAG) (n.d.). Ley n° 18.455 y sus reglamentos: Decreto n° 78, decreto n° 464, decreto n° 521. Disponible por contacto: vinas@sag.gob.cl.
- Silla, C. N. and Freitas, A. A. (2011). A survey of hierarchical classification across different application domains. *Data mining and knowledge discovery*, 22:31–72.
- Skoog, D. A., Holler, F. J., Nieman, T. A., et al. (2007). *Principios de análisis instrumental*, volume 6. McGraw-Hill Madrid.
- Statista (2024). La industria del vino en chile - datos estadísticos. Technical report, Statista Research Department.
- Wang, Z., Cao, J., Li, W., Wang, Y., Luo, G., Qiao, Y., Zhang, Y., and Xu, B. (2021). Using a material database and data fusion method to accelerate the process model development of high shear wet granulation. *Scientific Reports*, 11(1):16514.
- Xiong, Y. (2012). Building text hierarchical structure by using confusion matrix. In *2012 5th International Conference on BioMedical Engineering and Informatics*, pages 1250–1254. IEEE.
- Zhang, Y. and Wang, Y. (2023). Machine learning applications for multi-source data of edible crops: A review of current trends and future prospects. *Food Chemistry: X*, 19:100860.
- Zhou, G., Chen, Y., and Chien, C. (2022). On the analysis of data augmentation methods for spectral imaged based heart sound classification using convolutional neural networks. *BMC medical informatics and decision making*, 22(1):226.
- Zurita-Silva, A., Díaz, I., Verdugo, N., and Mejía, N. (2020). Desafíos de la viticultura en un escenario de cambio climático. *Revista Campo & Tecnología INIA*, 4(16):38–41.

Apéndice A

Anexos

A.1 Anexo A: Matrices de Confusión Modelos *Spec* y *Full*

Para facilitar la impresión y mantener la claridad en los nodos secundarios del árbol de modelos, en las matrices de confusión se emplean identificadores numéricos (por ejemplo, 0, 1, 2, 3 o 1 y 2) en lugar de los nombres completos de las cepas. De este modo, se evita la saturación de texto en espacios reducidos y se asegura una visualización limpia y legible.

A.1.1 Anexo A.1: Matrices de Confusión MPNS-11 y MPNS-ID

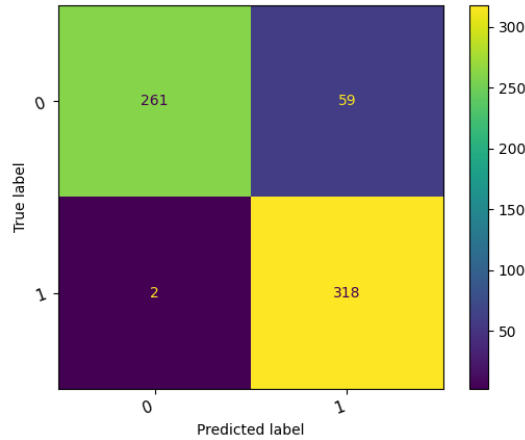


Figura A.1: Matriz de confusión del nodo MPNS-11 del modelo árbol base *Spec*. En la diagonal principal se indican las muestras correctamente clasificadas para cada cepa, mientras que los valores fuera de ella muestran las mal clasificaciones entre cepas.

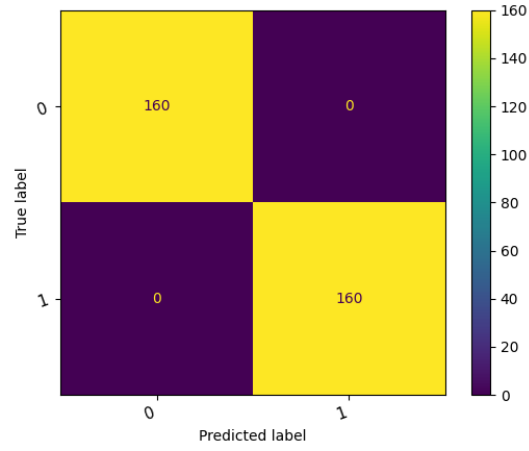


Figura A.2: Matriz de confusi3n del nodo MPNS-ID del modelo 3rbol base *Spec*. En la diagonal principal se indican las muestras correctamente clasificadas para cada cepa, mientras que los valores fuera de ella muestran las mal clasificaciones entre cepas.

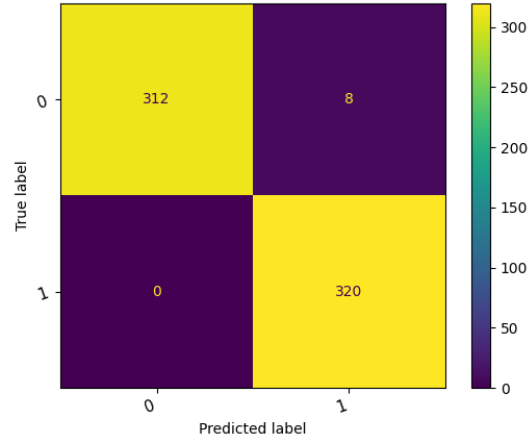


Figura A.3: Matriz de confusi3n del nodo MPNS-11 del modelo 3rbol base *Full*. En la diagonal principal se indican las muestras correctamente clasificadas para cada cepa, mientras que los valores fuera de ella muestran las mal clasificaciones entre cepas.

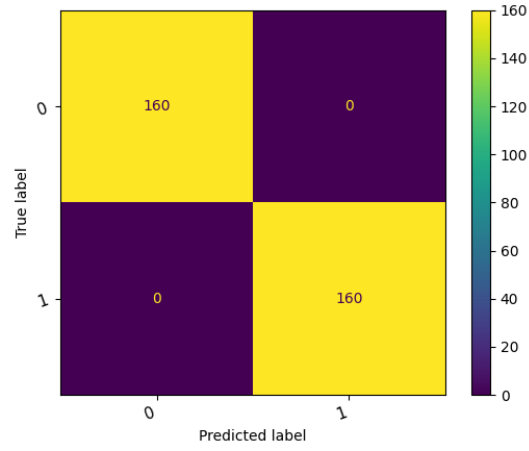


Figura A.4: Matriz de confusión del nodo MPNS-ID del modelo árbol base *Full*. En la diagonal principal se indican las muestras correctamente clasificadas para cada cepa, mientras que los valores fuera de ella muestran las mal clasificaciones entre cepas.

A.1.2 Anexo A.2: Matrices de Confusión SM-8

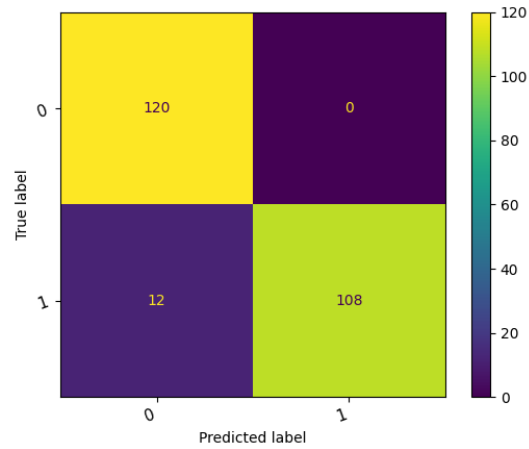


Figura A.5: Matriz de confusión del nodo SM-8 del modelo árbol base *Spec*. En la diagonal principal se indican las muestras correctamente clasificadas para cada cepa, mientras que los valores fuera de ella muestran las mal clasificaciones entre cepas.

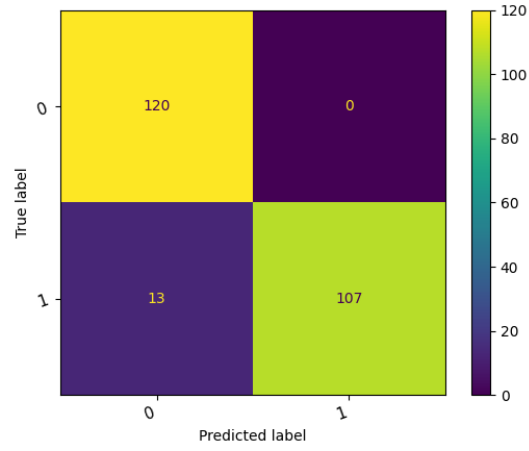


Figura A.6: Matriz de confusión del nodo SM-8 del modelo árbol base *Full*. En la diagonal principal se indican las muestras correctamente clasificadas para cada cepa, mientras que los valores fuera de ella muestran las mal clasificaciones entre cepas.

A.1.3 Anexo A.3: Matrices de Confusión CSPSPV-ID y CSPSPV-7

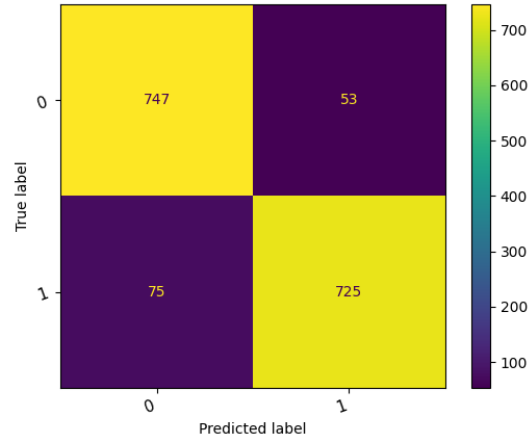


Figura A.7: Matriz de confusión del nodo CSPSPV-ID del modelo árbol base *Spec*. En la diagonal principal se indican las muestras correctamente clasificadas para cada cepa, mientras que los valores fuera de ella muestran las mal clasificaciones entre cepas.

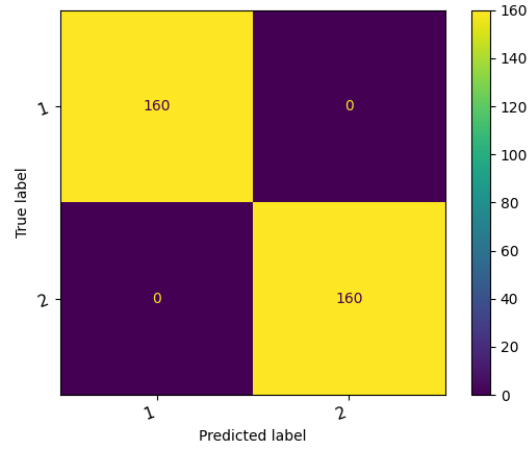


Figura A.8: Matriz de confusión del nodo CSPSPV-7 del modelo árbol base *Spec*. En la diagonal principal se indican las muestras correctamente clasificadas para cada cepa, mientras que los valores fuera de ella muestran las mal clasificaciones entre cepas.

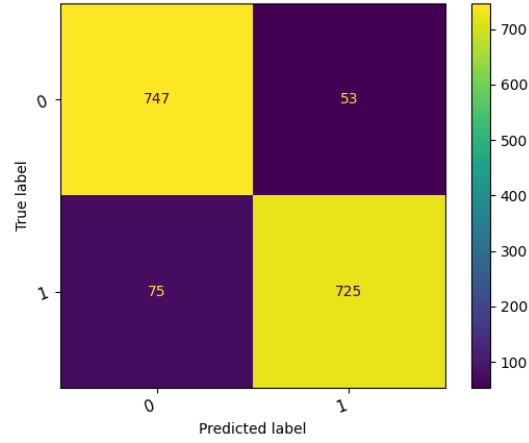


Figura A.9: Matriz de confusión del nodo CSPSPV-ID del modelo árbol base *Full*. En la diagonal principal se indican las muestras correctamente clasificadas para cada cepa, mientras que los valores fuera de ella muestran las mal clasificaciones entre cepas.

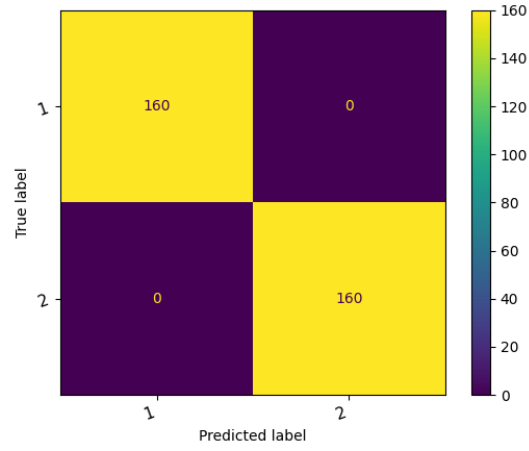


Figura A.10: Matriz de confusión del nodo CSPSPV-ID del modelo árbol base *Full*. En la diagonal principal se indican las muestras correctamente clasificadas para cada cepa, mientras que los valores fuera de ella muestran las mal clasificaciones entre cepas.

A.1.4 Anexo A.4: Matrices de Confusión CFMSZ-ID y MS-ID

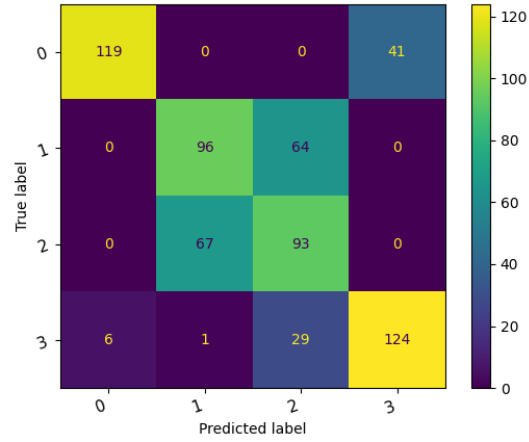


Figura A.11: Matriz de confusión del nodo CFMSZ-ID del modelo árbol base *Spec*. En la diagonal principal se indican las muestras correctamente clasificadas para cada cepa, mientras que los valores fuera de ella muestran las mal clasificaciones entre cepas.

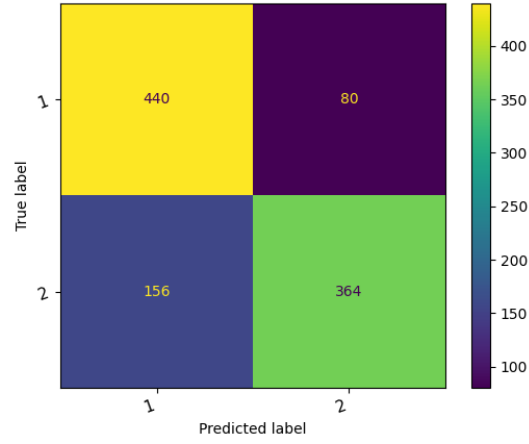


Figura A.12: Matriz de confusión del nodo MS-ID del modelo árbol base *Spec*. En la diagonal principal se indican las muestras correctamente clasificadas para cada cepa, mientras que los valores fuera de ella muestran las mal clasificaciones entre cepas.

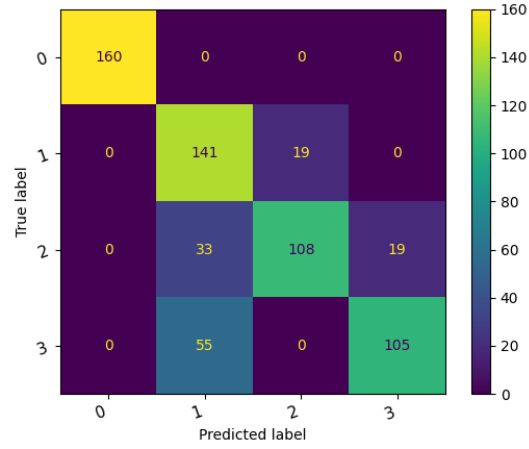


Figura A.13: Matriz de confusión del nodo CFMSZ-ID del modelo árbol base *Full*. En la diagonal principal se indican las muestras correctamente clasificadas para cada cepa, mientras que los valores fuera de ella muestran las mal clasificaciones entre cepas.

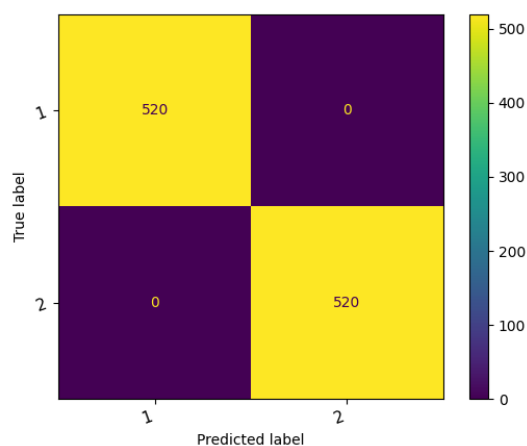


Figura A.14: Matriz de confusión del nodo MS-ID del modelo árbol base *Full*. En la diagonal principal se indican las muestras correctamente clasificadas para cada cepa, mientras que los valores fuera de ella muestran las mal clasificaciones entre cepas.

A.2 Anexo B: Matrices de Confusión del Modelo M1

Para facilitar la impresión y mantener la claridad en los nodos secundarios del árbol de modelos, en las matrices de confusión se emplean identificadores numéricos (por ejemplo, 0, 1, 2, 3 o 1 y 2) en lugar de los nombres completos de las cepas. De este modo, se evita la saturación de texto en espacios reducidos y se asegura una visualización limpia y legible.

A.2.1 Anexo B.1: Matrices de Confusión MPNS-11 y MPNS-ID

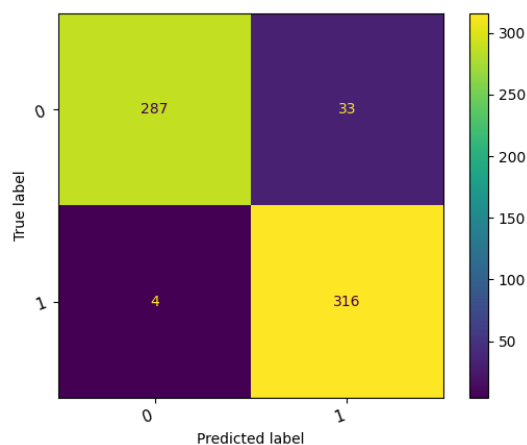


Figura A.15: Matriz de confusión del nodo MPNS-11 del modelo M1. En la diagonal principal se indican las muestras correctamente clasificadas para cada cepa, mientras que los valores fuera de ella muestran las mal clasificaciones entre cepas.

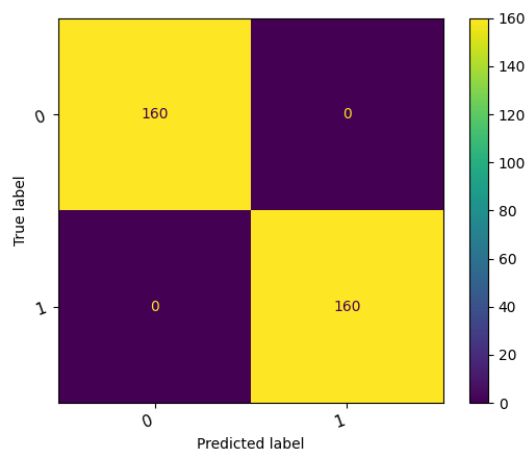


Figura A.16: Matriz de confusión del nodo MPNS-ID del modelo M1. En la diagonal principal se indican las muestras correctamente clasificadas para cada cepa, mientras que los valores fuera de ella muestran las mal clasificaciones entre cepas.

A.2.2 Anexo B.2: Matrices de Confusión SM-8

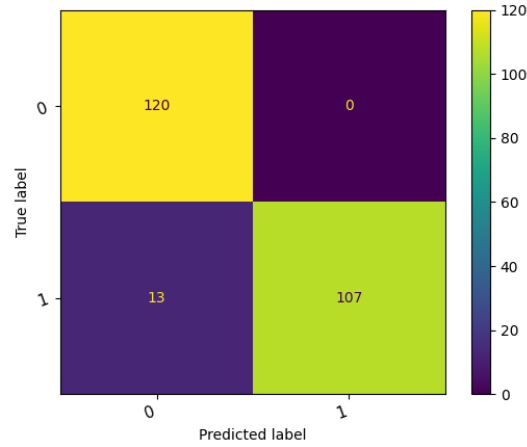


Figura A.17: Matriz de confusión del nodo SM-8 del modelo M1. En la diagonal principal se indican las muestras correctamente clasificadas para cada cepa, mientras que los valores fuera de ella muestran las mal clasificaciones entre cepas.

A.2.3 Anexo B.3: Matrices de Confusión CSPSPV-ID y CSPSPV-7

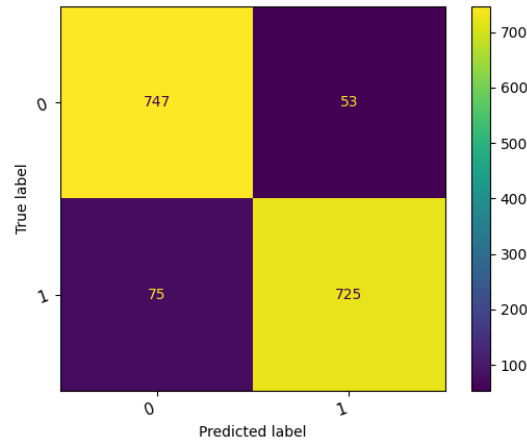


Figura A.18: Matriz de confusión del nodo CSPSPV-ID del modelo M1. En la diagonal principal se indican las muestras correctamente clasificadas para cada cepa, mientras que los valores fuera de ella muestran las mal clasificaciones entre cepas.

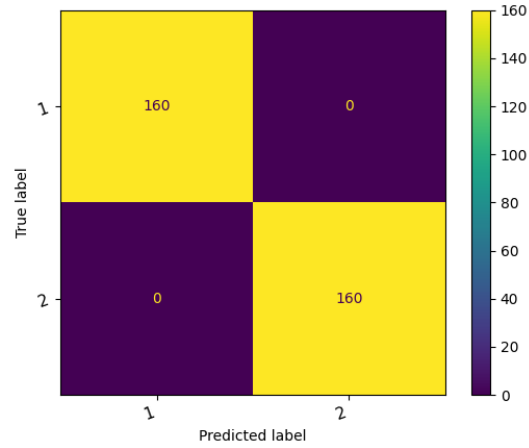


Figura A.19: Matriz de confusión del nodo CSPSPV-7 del modelo M1. En la diagonal principal se indican las muestras correctamente clasificadas para cada cepa, mientras que los valores fuera de ella muestran las mal clasificaciones entre cepas.

A.2.4 Anexo B.4: Matrices de Confusión CFMSZ-ID y MS-ID

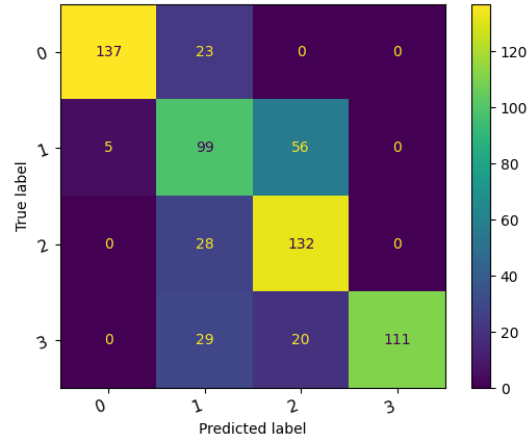


Figura A.20: Matriz de confusión del nodo CFMSZ-ID del modelo M1. En la diagonal principal se indican las muestras correctamente clasificadas para cada cepa, mientras que los valores fuera de ella muestran las mal clasificaciones entre cepas.

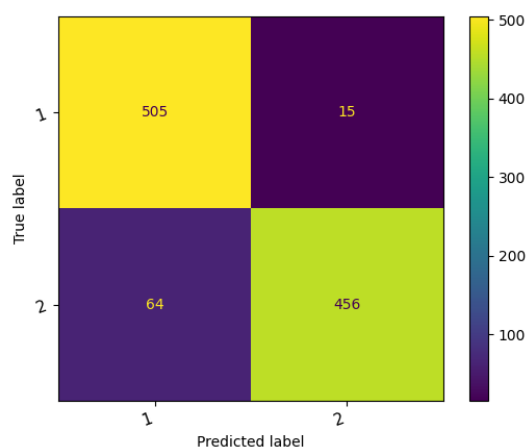


Figura A.21: Matriz de confusión del nodo MS-ID del modelo M1. En la diagonal principal se indican las muestras correctamente clasificadas para cada cepa, mientras que los valores fuera de ella muestran las mal clasificaciones entre cepas.

A.3 Anexo C: Matrices de Confusión del Modelo M2

Para facilitar la impresión y mantener la claridad en los nodos secundarios del árbol de modelos, en las matrices de confusión se emplean identificadores numéricos (por ejemplo, 0, 1, 2, 3 o 1 y 2) en lugar de los nombres completos de las cepas. De este modo, se evita la saturación de texto en espacios reducidos y se asegura una visualización limpia y legible.

A.3.1 Anexo C.1: Matrices de Confusión MPNS-11 y MPNS-ID

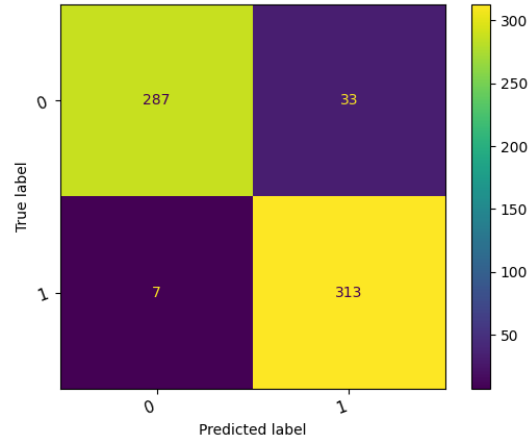


Figura A.22: Matriz de confusión del nodo MPNS-11 del modelo M2. En la diagonal principal se indican las muestras correctamente clasificadas para cada cepa, mientras que los valores fuera de ella muestran las mal clasificaciones entre cepas.

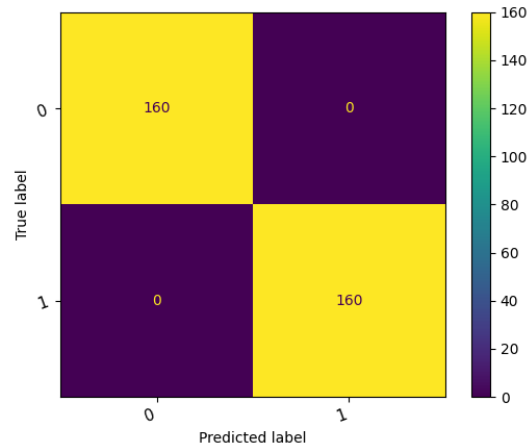


Figura A.23: Matriz de confusión del nodo MPNS-ID del modelo M2. En la diagonal principal se indican las muestras correctamente clasificadas para cada cepa, mientras que los valores fuera de ella muestran las mal clasificaciones entre cepas.

A.3.2 Anexo C.2: Matrices de Confusión SM-8

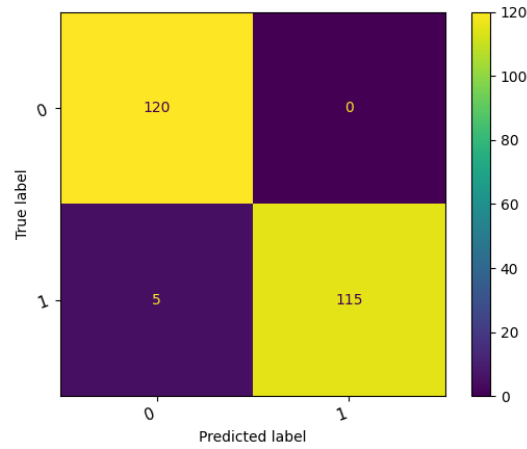


Figura A.24: Matriz de confusión del nodo SM-8 del modelo M2. En la diagonal principal se indican las muestras correctamente clasificadas para cada cepa, mientras que los valores fuera de ella muestran las mal clasificaciones entre cepas.

A.3.3 Anexo C.3: Matrices de Confusión CSPSPV-ID y CSPSPV-7

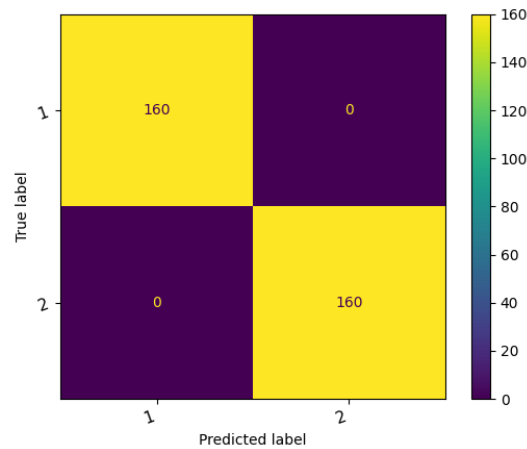


Figura A.25: Matriz de confusión del nodo CSPSPV-ID del modelo M2. En la diagonal principal se indican las muestras correctamente clasificadas para cada cepa, mientras que los valores fuera de ella muestran las mal clasificaciones entre cepas.

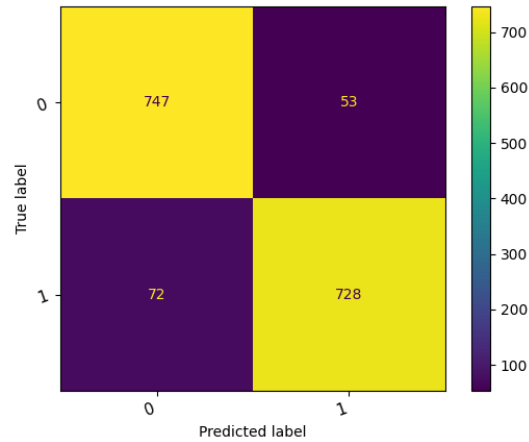


Figura A.26: Matriz de confusión del nodo CSPSPV-7 del modelo M2. En la diagonal principal se indican las muestras correctamente clasificadas para cada cepa, mientras que los valores fuera de ella muestran las mal clasificaciones entre cepas.

A.3.4 Anexo C.5: Matrices de Confusión CFMSZ-ID y MS-ID

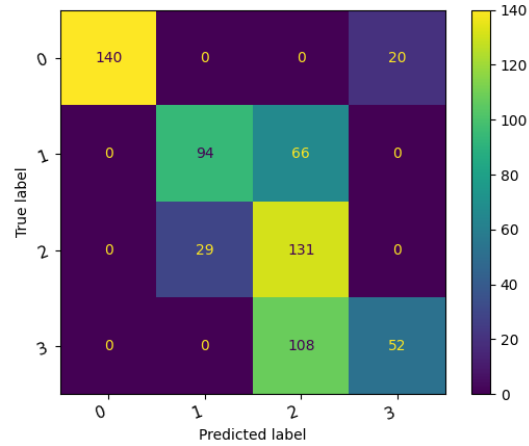


Figura A.27: Matriz de confusión del nodo CFMSZ-ID del modelo M2. En la diagonal principal se indican las muestras correctamente clasificadas para cada cepa, mientras que los valores fuera de ella muestran las mal clasificaciones entre cepas.

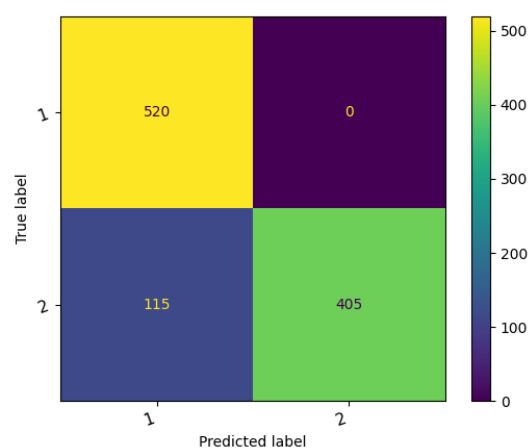


Figura A.28: Matriz de confusión del nodo MS-ID del modelo M2. En la diagonal principal se indican las muestras correctamente clasificadas para cada cepa, mientras que los valores fuera de ella muestran las mal clasificaciones entre cepas.