



UNIVERSIDAD AUTÓNOMA CHAPINGO

DEPARTAMENTO DE INGENIERÍA AGROINDUSTRIAL

POSGRADO EN CIENCIA Y TECNOLOGÍA AGROALIMENTARIA

**CLASIFICACIÓN DE MEZCALES MEDIANTE INTELIGENCIA
ARTIFICIAL Y ANÁLISIS SENSORIAL DESCRIPTIVO**

TESIS

Que como requisito parcial para obtener el grado de:

MAESTRO EN CIENCIA Y TECNOLOGÍA AGROALIMENTARIA

Presenta:

ANGEL ARTURO JUÁREZ MOLINA

Bajo la supervisión de:

ELEAZAR AGUIRRE MANDUJANO, DR



**Maestría en
Ciencia y Tecnología
Agroalimentaria**



Chapingo, Estado de México, diciembre de 2025

CLASIFICACIÓN DE MEZCALES MEDIANTE INTELIGENCIA ARTIFICIAL Y ANÁLISIS SENSORIAL DESCRIPTIVO

Tesis realizada por **Angel Arturo Juárez Molina** bajo la supervisión del Comité Asesor indicado, aprobada por el mismo y aceptada como requisito parcial para obtener el grado de:

MAESTRO EN CIENCIA Y TECNOLOGÍA AGROALIMENTARIA

DIRECTOR:



DR. ELEAZAR AGUIRRE MANDUJANO

CO-DIRECTOR:



DR. ELEAZAR AGUIRRE ANAYA

ASESOR:



DR. ARTEMIO PÉREZ LÓPEZ

CONTENIDO

LISTA DE CUADROS	vi
LISTA DE FIGURAS	vii
DEDICATORIAS	viii
AGRADECIMIENTOS	ix
DATOS BIOGRÁFICOS	x
RESUMEN GENERAL	xi
GENERAL ABSTRACT	xii
1. INTRODUCCIÓN GENERAL	1
1.1. Literatura citada	4
2. REVISIÓN DE LITERATURA SOBRE LAS BASES FISCOQUÍMICAS, SENSORIALES Y COMPUTACIONALES PARA LA CLASIFICACIÓN DEL MEZCAL POR ESPECIE DE AGAVE	6
2.1. El mezcal y su variabilidad biológica	7
2.2. Perfil fisicoquímico del mezcal	9
2.2.1. Alcoholes superiores	9
2.2.2. Aldehídos	10
2.2.3. Metanol	10
2.2.4. Furfural	11
2.2.5. Extracto seco	11
2.3. Importancia del perfil fisicoquímico en la clasificación del mezcal	12

2.4.	Fundamentos del aprendizaje automático y algoritmos de clasificación	13
2.5.	Aplicación de aprendizaje automático en la clasificación de bebidas alcohólicas.....	15
2.6.	Relevancia del desbalance de clases en la modelación del sistema mezcalero.....	17
2.7.	Integración del análisis sensorial descriptivo (CATA)	19
2.8.	Relevancia de la clasificación por especie de maguey.....	20
2.9.	Literatura citada	21
3.	CLASIFICACIÓN DE MEZCALES ARTESANALES POR ESPECIE DE AGAVE EMPLEANDO SU PERFIL FISICOQUÍMICO Y APRENDIZAJE AUTOMÁTICO	27
3.1.	Introducción	27
3.2.	Materiales y métodos.....	29
3.2.1.	Conjunto de datos	29
3.2.2.	Manejo metodológico del desbalance de clases y definición de escenarios de análisis	30
3.2.3.	Preprocesamiento y estandarización.....	32
3.2.4.	Proceso de validación	32
3.2.5.	Selección de algoritmos de clasificación	33
3.2.6.	Ajuste de hiperparámetros	34
3.2.7.	Análisis sensorial (CATA)	34
3.2.8.	Análisis estadísticos	37
3.3.	Resultados y discusión	38
3.3.1.	Tasa de desbalance.....	40
3.3.2.	Selección de hiperparámetros y evaluación de modelos.....	41

3.3.3.	Análisis CATA	51
3.4.	CONCLUSIONES	55
3.5.	Literatura citada	56

LISTA DE CUADROS

Cuadro 1. Hiperparámetros evaluados durante el proceso de ajuste.	35
Cuadro 2. Análisis de parámetros fisicoquímicos del mezcal entre especies de agave mediante la prueba de Kruskal–Wallis.	38
Cuadro 3. Tasas de desbalance del conjunto de datos completo.	40
Cuadro 4. Tasas de desbalance en el escenario A.	40
Cuadro 5. Tasas de desbalance en el escenario B.	41
Cuadro 6. Valores de los hiperparámetros ajustados al aplicar Búsqueda en cuadrícula en la clasificación de mezcal elaborado con <i>A. angustifolia</i>	42
Cuadro 7. Valores de los hiperparámetros ajustados al aplicar Búsqueda en cuadrícula en la clasificación de mezcal elaborado con <i>A. karwinskii</i>	45
Cuadro 8. f1-score (%) en la clasificación de especies minoritarias con modelos binarios.	47
Cuadro 9. f1-score (%) en la clasificación de especies minoritarias con modelos multiclase.	48
Cuadro 10. Frecuencia en que los términos de las preguntas CATA fueron usadas por los consumidores (n = 100) para describir las cuatro muestras de mezcal artesanal y los resultados de la prueba de Q de Cochran para la comparación entre muestras.	53

LISTA DE FIGURAS

Figura 1. Desempeño de los algoritmos de clasificación aplicados a muestras de <i>A. angustifolia</i> con parámetros por defecto y ajustados.....	43
Figura 2. Clasificación para las muestras elaboradas con <i>A. karwinskii</i>	45
Figura 3. Flujo de clasificación propuesto para muestras desconocidas.	49
Figura 4. Representación de los atributos (●) y muestras de mezcal (●) en las primeras dos dimensiones del análisis de correspondencia.	54

DEDICATORIAS

A mi madre, **Aida Juárez Molina**, por ser el motor y amor de mi vida; por todo lo que nos has dado y enseñado a mis hermanos y a mí, TE AMO.

A mis hermanos **Alexa y Enrique**, por ser parte fundamental de mi vida y de lo que soy. Su amor significa más de lo que las palabras pueden expresar.

A **Andrés, Paulina, Blanca, Emanuel y Ozuna**, por su amistad sincera, su apoyo constante y por hacer de este camino una experiencia más amena.

A **Montserrat y Kevin**, por los momentos que hemos compartido y por la relación tan especial que hemos formado. Su apoyo, su compañía y su presencia han sido un regalo en esta etapa de mi vida.

AGRADECIMIENTOS

A la **Secretaría de Ciencia, Humanidades, Tecnología e Innovación (SECIHTI)**, por el apoyo económico otorgado durante la realización de esta investigación y por la beca de maestría que hizo posible la continuidad de mis estudios.

A la **Universidad Autónoma Chapingo Posgrado en Ciencia y al Posgrado en Ciencia y Tecnología Agroalimentaria**, por ser mi casa de estudios y el espacio donde fue posible desarrollar este trabajo.

Al **Dr. Eleazar Aguirre Mandujano**, en su calidad de director de tesis, por su conducción académica, sus aportes y el apoyo brindado para este trabajo.

Al **Dr. Eleazar Aguirre Anaya**, por su dirección, por cada reunión y asesoría, y por el acompañamiento académico que permitió la culminación de esta tesis.

Al **Dr. Artemio Pérez López**, por su asesoría, sus observaciones y el apoyo brindado en el desarrollo del proyecto.

Al **Centro de Investigación en Computación y al Laboratorio de Ciberseguridad**, por permitirme realizar una estancia académica en sus instalaciones, lo cual enriqueció significativamente esta investigación.

A la **Asociación de Maguey y Mezcal Artesanal A.C.**, por facilitar el acceso a los datos que hicieron posible la realización de este trabajo.

A todos los que sin ser mencionados aportaron a mi desarrollo durante la maestría, GRACIAS.

DATOS BIOGRÁFICOS

Datos personales

Nombre	Angel Arturo Juárez Molina
Fecha de nacimiento	06 de junio de 2000
Lugar de nacimiento	Zacapoaxtla, Puebla.
CURP	JUMA000606HPLRLNA2
Profesión	Ingeniero Agroindustrial
Cédula Profesional	13837530



Desarrollo académico

Bachillerato	Preparatoria Agrícola, Universidad Autónoma Chapingo
Licenciatura	Ingeniero Agroindustrial, Universidad Autónoma Chapingo

RESUMEN GENERAL

CLASIFICACIÓN DE MEZCALES MEDIANTE INTELIGENCIA ARTIFICIAL Y ANÁLISIS SENSORIAL DESCRIPTIVO

El mezcal es una bebida destilada mexicana cuya diversidad biológica genera una amplia variabilidad en su composición química. Aunque esta heterogeneidad es parte de su identidad, también complica la caracterización objetiva por especie de *Agave*, aspecto relevante para la certificación del producto. Aunque su composición ha sido ampliamente estudiada, su uso con fines de clasificación ha sido poco explorado, a diferencia de otras bebidas destiladas. En ese contexto, los datos fisicoquímicos de certificación generados de manera rutinaria bajo métodos estandarizados representan una fuente accesible para desarrollar herramientas complementarias basadas en aprendizaje automático. El objetivo de esta investigación fue clasificar mezcales por especie de agave utilizando modelos de aprendizaje automático entrenados con dichos datos, para identificar los modelos idóneos para dicha tarea. Se analizaron 393 muestras certificadas de Oaxaca. La metodología incluyó un análisis estadístico mediante la prueba de Kruskal–Wallis para identificar variables con diferencias significativas entre especies. Posteriormente, se preparó y exploró el conjunto de datos, evaluando su desbalance y generando escenarios analíticos que permitieron comparar el desempeño de los modelos bajo distintas configuraciones. En la etapa de modelado se entrenaron más de mil modelos, resultado de evaluar seis algoritmos supervisados bajo múltiples configuraciones mediante validación cruzada y búsqueda de hiperparámetros. Se utilizó la métrica f1-score para identificar las configuraciones con mejor desempeño. De manera complementaria, se aplicó la metodología CATA para evaluar si las diferencias químicas identificadas por los modelos se reflejaban en la percepción del consumidor. Los resultados mostraron diferencias significativas entre especies en alcoholes superiores y metanol, además de modelos con alta capacidad de clasificación. En conjunto, los hallazgos demuestran que los datos fisicoquímicos de certificación pueden alimentar modelos capaces de distinguir mezcales por especie y ofrecen una base para el desarrollo de futuras herramientas computacionales de autenticación.

Palabras clave: mezcal artesanal, perfil fisicoquímico, clasificación, aprendizaje automático, metodología CATA.

Tesis de Maestría en Ciencia, Programa de Posgrado en Ciencia y Tecnología Agroalimentaria, Universidad Autónoma Chapingo.

Autor: Angel Arturo Juárez Molina

Director de Tesis: Dr. Eleazar Aguirre Mandujano

GENERAL ABSTRACT

CLASSIFICATION OF MEZCAL USING ARTIFICIAL INTELLIGENCE AND DESCRIPTIVE SENSORY ANALYSIS

Mezcal is a Mexican distilled beverage whose biological diversity generates wide variability in its chemical composition. Although this heterogeneity is part of its identity, it also complicates the objective characterization by *Agave* species, an aspect relevant to product certification. Although its composition has been widely studied, its use for classification purposes has been little explored, unlike other distilled beverages. In this context, the physicochemical certification data routinely generated under standardized methods represent an accessible source for developing complementary tools based on machine learning. The objective of this research was to classify mezcales by agave species using machine learning models trained with these data, in order to identify the most suitable models for this task. A total of 393 certified samples from Oaxaca were analyzed. The methodology included a statistical analysis using the Kruskal–Wallis test to identify variables with significant differences between species. Subsequently, the dataset was prepared and explored, evaluating its imbalance and generating analytical scenarios that allowed comparison of model performance under different configurations. In the modeling stage, more than one thousand models were trained as a result of evaluating six supervised algorithms under multiple configurations through cross-validation and hyperparameter tuning. The f1-score metric was used to identify the configurations with the best performance. Complementarily, the CATA methodology was applied to assess whether the chemical differences identified by the models were reflected in consumer perception. The results showed significant differences between species in higher alcohols and methanol, as well as models with high classification capacity. Taken together, the findings demonstrate that physicochemical certification data can be used to train models capable of distinguishing mezcales by species and provide a basis for the development of future computational authentication tools.

Key words: artisanal mezcal, physicochemical profile, classification, machine learning, CATA methodology.

Tesis de Maestría en Ciencia, Programa de Posgrado en Ciencia y Tecnología Agroalimentaria, Universidad Autónoma Chapingo.

Autor: Angel Arturo Juárez Molina

Director de Tesis: Dr. Eleazar Aguirre Mandujano

1. INTRODUCCIÓN GENERAL

El mezcal es una bebida alcohólica destilada mexicana con Denominación de Origen (DO), reconocida por su valor cultural, histórico y económico. Su elaboración se basa en la fermentación y destilación del mosto de diferentes especies del género *Agave*, lo que le confiere una notable diversidad biológica y tecnológica. Esta variabilidad, característica esencial del producto, también representa un desafío científico para su caracterización, ya que factores como la especie de agave, el entorno geográfico y las prácticas artesanales influyen directamente en su composición química y sensorial.

La producción y comercialización del mezcal están reguladas por la Norma Oficial Mexicana NOM-070-SCFI-2016, que define los criterios de calidad y establece los rangos permisibles de parámetros fisicoquímicos, tales como alcoholes superiores, metanol, aldehídos, furfural y extracto seco. Estos valores, determinados rutinariamente mediante cromatografía de gases, garantizan el cumplimiento normativo y la inocuidad del producto. Sin embargo, su aplicación se ha restringido a la verificación de conformidad, sin aprovecharse su potencial como fuente de información que refleje la diversidad biológica y química del mezcal.

Aunque se han realizado estudios que analizan la composición química y volátil del mezcal mediante técnicas instrumentales de alta resolución como cromatografía de gases acoplada a espectrometría de masas (GC-MS), espectroscopía infrarroja (FTIR) y resonancia magnética nuclear (NMR), estos enfoques se han centrado principalmente en la descripción analítica del producto (Sánchez-Fernández et al., 2025; Vera-Guzmán et al., 2018).

Sin embargo, los datos generados por estos análisis, al igual que los obtenidos durante el proceso de certificación, representan un recurso con un potencial aún poco explorado para tareas de clasificación, a diferencia de lo observado en otras bebidas alcohólicas destiladas, donde se han aprovechado exitosamente para predecir el origen geográfico, la materia prima o la calidad sensorial del producto (Swift et al., 2023; Zhou, Yu, Ai, et al., 2025; Zhou, Yu, Zhu, et al., 2025).

En este contexto, los datos fisicoquímicos generados durante el proceso de certificación del mezcal constituyen una fuente de información accesible, estandarizada y disponible a gran escala, que puede aprovecharse para fines científicos más allá del cumplimiento normativo. Su análisis mediante técnicas de aprendizaje automático (machine learning, ML) representa una oportunidad para identificar patrones complejos en matrices químicas multivariadas y distinguir muestras según la especie de agave empleada. Con la ventaja de utilizar datos que ya son generados de manera rutinaria en el proceso de certificación, por lo que su uso no implica un aumento en costos de análisis.

Con fundamento en lo expuesto, en la presente investigación se utilizaron los datos fisicoquímicos obtenidos durante la certificación del mezcal como insumo para desarrollar modelos de clasificación capaces de distinguir la bebida según la especie de *Agave* utilizada. Paralelamente, se incorporó un análisis sensorial mediante la metodología Check All That Apply (CATA), con el objetivo de explorar la correspondencia entre los atributos perceptuales y los patrones químicos observados, fortaleciendo así la comprensión integral del producto desde la perspectiva analítica y sensorial.

Con base en lo anterior, se plantearon las siguientes hipótesis:

1. El perfil fisicoquímico generado durante el proceso de certificación del mezcal contiene información suficiente para clasificar la bebida según la especie de *Agave* utilizada en su producción.

2. El perfil sensorial del mezcal presenta diferencias perceptibles entre especies de agave, lo que permite diferenciar las muestras a partir de sus atributos sensoriales.

El objetivo general de esta tesis fue clasificar mezcales por especie de agave utilizando modelos de aprendizaje automático entrenados con datos fisicoquímicos de certificación, para identificar los modelos idóneos para dicha tarea. De manera específica, se buscó:

- Evaluar el desempeño de diferentes algoritmos de aprendizaje automático aplicados a la clasificación de mezcales por especie de agave, con el fin de identificar los modelos con mejor rendimiento
- Determinar el ambiente de evaluación y analizar el desempeño de los modelos desarrollados mediante la métrica f1-score, con el propósito de identificar los algoritmos con mayor capacidad de clasificación por especie de agave.
- Analizar las diferencias perceptuales entre muestras mediante la metodología CATA.

La tesis se estructura en tres apartados principales. El primer apartado corresponde a la Introducción General, donde se presentan los antecedentes del estudio, la justificación, las hipótesis y los objetivos de la investigación. El segundo apartado desarrolla el marco teórico, en el que se abordan los fundamentos del aprendizaje automático y los antecedentes químicos y sensoriales del mezcal, proporcionando el sustento conceptual que guía el análisis. Finalmente, el tercer apartado expone la metodología, los resultados y la discusión derivados del trabajo experimental, en los que se evalúa la aplicación de modelos de clasificación y el análisis sensorial en muestras de mezcal artesanal.

1.1. Literatura citada

- Andrade, J. M., Ballabio, D., Gómez-Carracedo, M. P., & Pérez-Caballero, G. (2017). Nonlinear classification of commercial Mexican tequilas. *Journal of Chemometrics*, 31(12), e2939. <https://doi.org/10.1002/cem.2939>
- Pérez-Caballero, G., Andrade, J. M., Olmos, P., Molina, Y., Jiménez, I., Durán, J. J., Fernández-Lozano, C., & Miguel-Cruz, F. (2017). *Authentication of tequilas using pattern recognition and supervised classification*. *TrAC Trends in Analytical Chemistry*, 94, 1–9. <https://doi.org/10.1016/j.trac.2017.07.008>
- Sánchez-Fernández, R. E., Pérez-López, A., Morales-Solís, A., Manilla-Téllez, Y., Reyes-Carmona, E. D., & Ávila-Urbe, G. (2025). *Chemistry of mezcal: Volatile profile of artisanal mezcal made from wild agaves of Oaxaca*. *Foods*, 14(7), 1222. <https://doi.org/10.3390/foods14071222>
- Swift, J. R., Turner, M. A., & Reynolds, J. C. (2023). A rapid dynamic headspace method for authentication of whiskies using artificial neural networks. *Food Chemistry Advances*, 3, 100417. <https://doi.org/10.1016/j.focha.2023.100417>
- Vera-Guzmán, A. M., Guzmán-Gerónimo, R. I., López, M. G., & Chávez-Servia, J. L. (2018). *Volatile compound profiles in mezcal spirits as influenced by agave species and production processes*. *Beverages*, 4(1), 9. <https://doi.org/10.3390/beverages4010009>
- Zhou, C., Yu, Y., Ai, J., Song, C., Cui, Z., Zhou, Q., Zhao, S., Huang, R., Ao, Z., Peng, B., Chen, P., Feng, X., Li, D., & Liu, Y. (2025). *Fruit wines classification enabled by combining machine learning with comprehensive volatile profiles of GC-TOF/MS and GC-IMS*. *Food Research International*, 204, 115890. <https://doi.org/10.1016/j.foodres.2025.115890>
- Zhou, C., Yu, Y., Zhu, Y., Chai, J., Liu, R., Cui, Z., Li, D., Feng, X., & Liu, Y. (2025). *Machine learning discrimination of quality grades of base liquor integrating GC-TOF/MS and GC-IMS data analysis: Case study of strong-flavor Chinese*

baijiu. Food Chemistry, 492, 145281.
<https://doi.org/10.1016/j.foodchem.2025.145281>

2. REVISIÓN DE LITERATURA SOBRE LAS BASES FISICOQUÍMICAS, SENSORIALES Y COMPUTACIONALES PARA LA CLASIFICACIÓN DEL MEZCAL POR ESPECIE DE AGAVE

El presente capítulo tiene como propósito establecer las bases conceptuales que sustentan el uso de modelos de clasificación basados en ML para diferenciar mezcales según la especie de agave utilizada en su elaboración. A través de la revisión de literatura, se abordan los aspectos biológicos y fisicoquímicos que explican la variabilidad intrínseca del mezcal, así como los fundamentos analíticos y computacionales que permiten traducir dicha variabilidad en patrones clasificables.

El punto de partida de esta investigación se encuentra en la relación entre la composición fisicoquímica del mezcal y la especie de agave, vínculo que ha sido abordado principalmente desde un enfoque descriptivo, orientado a la identificación de compuestos volátiles (Sánchez-Fernández et al., 2025) y metabolitos secundarios y marcadores específicos asociados a la materia prima (Vera-Guzmán et al., 2018). Dichos trabajos han contribuido de manera fundamental a comprender la base química y biológica del producto; sin embargo, aún persiste un vacío metodológico respecto a la aplicación de esa información para desarrollar modelos de clasificación automatizada. En este sentido, el presente estudio no busca reemplazar los enfoques previos, sino ampliar su alcance al explorar la posibilidad de emplear los datos fisicoquímicos ya disponibles para identificar patrones que permitan diferenciar las muestras según la especie de agave utilizada.

Asimismo, se justifica la selección de los algoritmos de aprendizaje automático empleados, el tratamiento del desbalance de clases, y la integración del análisis sensorial como un análisis complementario. Con ello, se establece la coherencia teórica entre el planteamiento del problema, los objetivos de la investigación y las decisiones metodológicas presentadas en el capítulo 3.

2.1. El mezcal y su variabilidad biológica

El mezcal es una bebida alcohólica destilada de origen mexicano obtenida a partir de la fermentación y destilación del mosto extraído de distintas especies del género *Agave*. México alberga más de 200 especies, de las cuales aproximadamente 159 poseen potencial para la producción de mezcal (Anta-Fonseca, 2016; Vera-Guzmán et al., 2018). En el estado de Oaxaca, la producción se encuentra dominada por la especie cultivada *Agave angustifolia*, que representa alrededor del 85.25 % del volumen total, mientras que especies silvestres como *A. potatorum*, *A. marmorata* y *A. karwinskii* participan con el 1 %, 0.46 % y 0.44 %, respectivamente (COMERCAM, 2025).

Las diferencias metabólicas entre especies de agave se reflejan principalmente en la composición de carbohidratos, compuestos nitrogenados, ácidos orgánicos y metabolitos secundarios, los cuales determinan la disponibilidad de sustratos fermentables y la formación de precursores de compuestos volátiles durante la producción del mezcal (Nolasco-Cancino et al., 2018; Vera-Guzmán et al., 2018). Por ejemplo, especies cultivadas como *Agave angustifolia* Haw. presentan altos contenidos de fructanos de cadena corta, que se hidrolizan con mayor facilidad durante la cocción, generando mostos ricos en azúcares reductores y promoviendo fermentaciones más activas. En contraste, especies silvestres como *A. potatorum* o *A. marmorata* poseen fructanos más complejos y de cadena larga, lo que condiciona una liberación más lenta de azúcares y, por ende, fermentaciones de menor rendimiento, pero con una mayor diversidad de metabolitos secundarios (Sánchez-Fernández et al., 2025).

Además del contenido de carbohidratos, la composición de aminoácidos y compuestos nitrogenados varía significativamente entre especies. Estos

compuestos son esenciales en la formación de alcoholes superiores y aldehídos durante la fermentación, ya que actúan como precursores en la ruta del catabolismo de los aminoácidos de Ehrlich (Hazelwood et al., 2008). En especies como *A. karwinskii*, se ha reportado una mayor proporción de aminoácidos de cadena ramificada, que favorecen la síntesis de 3-metil-1-butanol y 2-metil-1-propanol, alcoholes que son parte de la cuantificación total de alcoholes superiores (Vera-Guzmán et al., 2018). Estas diferencias en la disponibilidad de precursores nitrogenados explican en parte las variaciones observadas en la composición final del mezcal según la especie empleada.

En términos de distribución, las especies de agave utilizadas para mezcal se agrupan en dos grandes categorías: cultivadas y silvestres. Las cultivadas, como *A. angustifolia*, se reproducen predominantemente por clonación vegetativa, lo que promueve una relativa homogeneidad genética y química. En cambio, las especies silvestres, como *A. potatorum*, *A. marmorata* o *A. karwinskii*, se reproducen por semilla y presentan una elevada variabilidad genética intraespecífica, lo que se traduce en una mayor diversidad química entre individuos (Aguirre-Dugua & Eguiarte, 2013). Este mosaico genético contribuye a la diferenciación de los perfiles químicos del mezcal, incluso dentro de una misma especie, y refuerza la importancia de la clasificación biológica.

Asimismo, la edad de madurez de las plantas influye notablemente en su composición química. Durante su desarrollo, las plantas acumulan azúcares y compuestos nitrogenados en la piña como reservas energéticas, pero la proporción de estos metabolitos varía según el estado fisiológico. En *A. angustifolia*, la concentración de fructanos y azúcares reductores alcanza su punto máximo a los 7 u 8 años, mientras que en especies silvestres como *A. marmorata*, este proceso puede tardar hasta 20 años, lo que repercute en la composición inicial del mosto y, por tanto, en el perfil final del mezcal (Nolasco-Cancino et al., 2018; Sánchez-Fernández et al., 2025).

La variabilidad biológica, en conjunto, es la responsable de la singularidad del mezcal frente a otros destilados. Cada especie aporta un conjunto distintivo de

características fisicoquímicas que se manifiestan en la bebida final. Por esta razón, la clasificación del mezcal por especie de agave no solo es viable, sino que posee un fundamento biológico sólido. Reconocer esta diversidad es esencial para comprender el origen de las diferencias químicas y, por ende, para justificar la aplicación de herramientas de clasificación, las cuales buscan aprovechar dicha variabilidad para distinguir objetivamente los mezcales según su especie de origen.

2.2. Perfil fisicoquímico del mezcal

El perfil fisicoquímico del mezcal representa la descripción cuantitativa de los componentes básicos que determinan su calidad, seguridad y autenticidad. Estos parámetros son establecidos por la Norma Oficial Mexicana NOM-070-SCFI-2016, la cual regula la producción, clasificación y comercialización del mezcal, definiendo los valores mínimos y máximos permitidos para compuestos clave como alcoholes superiores, aldehídos, metanol, furfural y extracto seco. Dichos compuestos se determinan mediante cromatografía de gases en laboratorios acreditados por los organismos evaluadores de la conformidad (OEC), garantizando resultados comparables, trazables y confiables (Secretaría de Economía, 2017).

Desde el punto de vista químico, estos parámetros reflejan el equilibrio entre las distintas fases del proceso de producción: cocción, fermentación y destilación. Su análisis no solo permite evaluar el cumplimiento normativo, sino también inferir información sobre la materia prima, la microbiota fermentativa y las condiciones tecnológicas empleadas. De este modo, el perfil fisicoquímico puede considerarse una huella composicional del mezcal, susceptible de emplearse como fuente de información para fines de clasificación.

2.2.1. Alcoholes superiores

Los alcoholes superiores comprenden compuestos como el 1-propanol, 2-metil-1-propanol, 3-metil-1-butanol y 2-metil-1-butanol. Se forman durante la fermentación alcohólica a partir del metabolismo de los aminoácidos a través de

la ruta de Ehrlich, en la cual la descarboxilación y reducción de los α -cetoácidos derivados de leucina, isoleucina y valina producen estos alcoholes (Hazelwood et al., 2008). Estos compuestos tienen una doble relevancia: en bajas concentraciones contribuyen al cuerpo y complejidad aromática de la bebida, pero en exceso pueden generar notas olfativas desagradables y afectar su calidad sensorial (González Seguí et al., 2019). Su proporción depende fuertemente de la especie de agave, del tipo de fermentación y de la cepa microbiana predominante (Pino et al., 2012).

2.2.2. Aldehídos

Los aldehídos son compuestos intermedios del metabolismo fermentativo que ejercen una influencia significativa en el aroma y la percepción sensorial del mezcal, y su presencia puede aportar información relevante para la clasificación de la bebida. El más representativo es el acetaldehído, formado durante la descarboxilación del piruvato por acción de la piruvato descarboxilasa, cuya concentración varía según la especie de agave y las condiciones fermentativas. Este compuesto confiere notas punzantes y frutales, y en niveles elevados puede reflejar procesos de oxidación o una destilación incompleta (Lachenmeier et al., 2005). Asimismo, otros aldehídos como el furfural y el 5-hidroximetilfurfural (HMF) se originan durante la cocción del agave por degradación térmica de azúcares reductores, actuando como indicadores del tipo de materia prima y del tratamiento térmico aplicado (Ávila-Reyes et al., 2025). Dado que sus concentraciones dependen de la especie de agave, estos compuestos pueden considerarse variables discriminantes útiles en modelos de clasificación basados en perfiles fisicoquímicos.

2.2.3. Metanol

El metanol es un compuesto que se genera durante la cocción y fermentación del agave a partir de la desmetoxilación de las pectinas de la pared celular. Su concentración depende directamente de la composición estructural de cada especie, ya que el contenido y tipo de pectinas varía entre agaves. Estas diferencias explican las variaciones observadas en los niveles de metanol en

mezcales producidos con distintas materias primas. Por ello, más allá de su relevancia regulatoria, el metanol constituye una variable con potencial discriminante dentro de los modelos de clasificación, al reflejar características intrínsecas de la especie de agave empleada en la elaboración del destilado (González Seguí et al., 2019).

2.2.4. Furfural

El furfural y sus derivados se originan por la degradación térmica de los carbohidratos durante la cocción de las piñas del agave. Su formación depende de la temperatura, del tiempo de exposición al calor y del tipo de horno empleado. De acuerdo con Pérez et al. (2013), la presencia de furfural es un marcador químico del uso de agave como materia prima, ya que este compuesto no se genera a partir de otras fuentes vegetales en condiciones equivalentes (López-Ramírez et al., 2013). En el contexto de la clasificación, las variaciones en su concentración entre especies pueden reflejar diferencias en la composición de carbohidratos y en la respuesta térmica de cada agave durante la cocción, lo que convierte al furfural en un parámetro útil para discriminar entre muestras elaboradas con distintas especies.

2.2.5. Extracto seco

El extracto seco agrupa los componentes no volátiles que permanecen después de la evaporación del alcohol y el agua en condiciones controladas. Está conformado por azúcares residuales, ácidos orgánicos, minerales, compuestos fenólicos y otros sólidos disueltos. En el mezcal, su valor suele oscilar entre 0.02 y 0.1 g/100 mL, dependiendo de la especie de agave y del método de destilación (López-Ramírez et al., 2013). Estas diferencias reflejan la proporción de sólidos solubles transferidos desde la materia prima, lo que sugiere que el extracto seco puede aportar información útil para la clasificación de mezcales, al estar influido por características intrínsecas de cada especie y por la eficiencia metabólica del proceso fermentativo.

2.3. Importancia del perfil fisicoquímico en la clasificación del mezcal

La combinación de los compuestos que conforman el perfil fisicoquímico forma una huella química característica del mezcal, donde las proporciones relativas entre compuestos funcionan como indicadores del origen botánico y del proceso tecnológico. Diversos estudios han demostrado que los perfiles de alcoholes, aldehídos y furfural difieren significativamente entre especies de *Agave*. Sánchez-Fernández et al. (2025) identificaron variaciones notables en los niveles de alcoholes superiores como 1-propanol, 2-metil-1-propanol y 1-butanol, así como en etil acetato, ácido acético y furfural, al comparar mezcales elaborados con *A. potatorum*, *A. karwinskii*, *A. angustifolia* y *A. marmorata*. Por ejemplo, *A. potatorum* presentó mayores concentraciones de etil acetato, asociadas con notas dulces y afrutadas, mientras que *A. karwinskii* y *A. angustifolia* mostraron contenidos más bajos de este éster y perfiles dominados por alcoholes superiores. Además, el contenido de furfural fue significativamente diferente entre especies, reflejando la influencia de la especie de agave. De manera complementaria, Vera-Guzmán et al. (2018) reportaron diferencias en los niveles de metanol y alcoholes superiores entre especies silvestres y cultivadas, lo que evidencia la influencia de la composición estructural del agave en la formación de estos compuestos durante la fermentación. En conjunto, estos estudios confirman que los perfiles de alcoholes, aldehídos y furfural constituyen indicadores químicos sensibles a la especie de agave y, por tanto, variables clave para su utilización en modelos de clasificación.

El uso de los datos fisicoquímicos generados rutinariamente durante el proceso de certificación representa una ventaja científica y práctica, ya que son obtenidos de manera estandarizada (mediante cromatografía de gases), reproducible, y abarcan prácticamente la totalidad del mercado formal del mezcal, al ser un requisito indispensable para su certificación, lo que los convierte en un recurso de información confiable y accesible para la investigación. Su aprovechamiento en el ámbito del ML permite explorar nuevas posibilidades de análisis basadas

en la detección de relaciones multivariadas entre variables fisicoquímicas, sin incrementar los costos analíticos ni requerir infraestructura adicional.

En este contexto, la clasificación del mezcal por especie de agave adquiere relevancia al establecer una conexión directa entre los patrones fisicoquímicos observables y la diversidad biológica de la materia prima. Esta relación justifica el uso de los datos obtenidos rutinariamente durante la certificación como indicadores del origen botánico.

2.4. Fundamentos del aprendizaje automático y algoritmos de clasificación

La inteligencia artificial (IA) constituye un campo interdisciplinario cuyo objetivo es desarrollar sistemas capaces de realizar tareas que requieren inteligencia humana, como el razonamiento, el reconocimiento de patrones o la toma de decisiones (Russell & Norvig, 2021). Dentro de la IA, el ML representa una rama que permite que los sistemas aprendan a partir de los datos sin ser explícitamente programados para cada tarea, mejorando su desempeño conforme incrementa la experiencia (Aurélien Géron, 2019).

El aprendizaje supervisado se basa en el uso de conjuntos de datos etiquetados, en los cuales se entrena un modelo para aprender la relación entre las variables predictoras y la variable objetivo. Este proceso implica tres etapas fundamentales:

1. Entrenamiento, en el que el modelo aprende los patrones subyacentes en los datos mediante la optimización de una función de pérdida;
2. Validación, donde se ajustan los hiperparámetros para evitar sobreajuste y mejorar la generalización; y
3. Evaluación, en la que se mide el desempeño del modelo sobre un conjunto independiente de datos no vistos, a fin de estimar su capacidad predictiva real (Yang & Shami, 2020).

Entre los algoritmos más utilizados en la literatura se encuentran las Máquinas de Soporte Vectorial (SVM), Naive Bayes (NB), el método de k-Vecinos más Cercanos (KNN), los Árboles de Decisión (DT), los Bosques Aleatorios (RF) y el Perceptrón Multicapa (MLP). Cada uno representa un enfoque matemático distinto para la detección de patrones, la construcción de fronteras de decisión y la generalización de la información. La diversidad de principios que los sustentan permite abarcar diferentes perspectivas del modelado: desde esquemas lineales y probabilísticos hasta arquitecturas neuronales no lineales.

Las Máquinas de Soporte Vectorial (SVM) se basan en la construcción de un hiperplano que separa de forma óptima las clases dentro del espacio de características, maximizando la distancia entre las observaciones de distintos grupos. Gracias al uso de funciones kernel, este método puede resolver problemas no lineales y de alta dimensionalidad, por lo que ha sido empleado en estudios de clasificación de bebidas fermentadas y destiladas donde las variables fisicoquímicas presentan correlaciones complejas (Andrade et al., 2017).

Naive Bayes (NB), por su parte, se fundamenta en el teorema de Bayes y asume independencia condicional entre variables predictoras. Aunque este supuesto rara vez se cumple completamente en datos reales, el modelo se ha mostrado eficiente en conjuntos pequeños, ofreciendo una línea base interpretativa y computacionalmente ligera (Sulzmann et al., 2007).

El algoritmo de k-Vecinos más Cercanos (KNN) clasifica nuevas instancias de acuerdo con la categoría predominante entre los k puntos más próximos en el espacio de características. Su naturaleza no paramétrica y su capacidad para adaptarse a la estructura de los datos lo hacen apropiado para estudios exploratorios, particularmente en la comparación de similitudes químicas entre muestras de distintas regiones o materias primas (Singh & Domijan, 2019).

Por su parte, los Árboles de Decisión (DT) y los Bosques Aleatorios (RF) constituyen una familia de métodos basados en divisiones sucesivas del conjunto de datos según criterios de ganancia de información o reducción de impureza. Los Árboles de Decisión permiten generar reglas interpretables que describen

cómo los valores de las variables determinan una clasificación. Los Bosques Aleatorios, en cambio, integran múltiples árboles mediante técnicas de ensamblado (*bagging*), aumentando la estabilidad del modelo y reduciendo el sobreajuste. Su eficacia ha sido ampliamente demostrada en la clasificación de bebidas alcohólicas, logrando altos niveles de precisión (Fawagreh et al., 2014; Warren-Vega et al., 2023).

Finalmente, el Perceptrón Multicapa (MLP) representa una de las arquitecturas más utilizadas dentro de las redes neuronales artificiales. Está compuesto por capas de neuronas interconectadas que transforman las entradas mediante funciones de activación no lineales, lo que permite capturar relaciones complejas entre variables. En la ciencia de los alimentos, los MLP han sido empleados para modelar fenómenos de fermentación, predecir propiedades sensoriales y clasificar productos a partir de datos fisicoquímicos y cromatográficos (Zhou, Yu, Zhu, et al., 2025). Su flexibilidad y capacidad de aprendizaje profundo los posicionan como una herramienta poderosa para sistemas donde los patrones no son evidentes ni lineales.

La revisión de estos algoritmos evidencia que cada uno aborda la clasificación desde principios matemáticos distintos, lo que permite analizar el sistema desde perspectivas complementarias. Comparar sus desempeños resulta fundamental para determinar qué tipo de modelado (lineal, probabilístico o no lineal) se ajusta mejor a la estructura de los datos y a la complejidad del sistema químico.

2.5. Aplicación de aprendizaje automático en la clasificación de bebidas alcohólicas.

En el ámbito de los alimentos, el aprendizaje automático ha adquirido relevancia por su capacidad para analizar sistemas multivariados (Kamilaris & Prenafeta-Boldú, 2018). A diferencia de los métodos estadísticos convencionales, los algoritmos de ML pueden procesar datos complejos y detectar relaciones no lineales entre variables, lo que los hace especialmente útiles en matrices químicas como las de los destilados, donde múltiples compuestos contribuyen simultáneamente al perfil característico de la bebida.

En este contexto, la aplicación de modelos de ML en la clasificación de bebidas alcohólicas ha surgido como una herramienta poderosa para la diferenciación de productos según su composición química, su origen botánico o su calidad comercial. En este campo, uno de los primeros estudios de referencia fue el de Andrade et al. (2017), quienes aplicaron técnicas de inteligencia artificial para la autenticación de tequilas a partir de sus perfiles cromatográficos. Utilizando datos obtenidos por cromatografía de gases con detector de ionización de flama (GC-FID) y cromatografía de gases acoplada a espectrometría de masas (GC-MS), los autores evaluaron modelos como Máquinas de Soporte Vectorial, Bosques Aleatorios y Redes Neuronales Artificiales, alcanzando en todos los casos precisiones superiores al 94 % en la identificación de tequilas auténticos y adulterados. Este trabajo demostró que los datos fisicoquímicos, pueden contener patrones estructurados capaces de alimentar modelos de clasificación robustos, abriendo una nueva perspectiva para los destilados derivados del agave.

De manera complementaria, Warren-Vega et al. (2023) exploraron la relación entre la composición volátil y la clasificación de tequilas de diferentes zonas de producción. Utilizando perfiles obtenidos por HS-SPME/GC-MS y combinando técnicas multivariadas (PCA, PLS-DA) con el algoritmo Random Forest, con lo que lograron discriminar las muestras con una exactitud superior al 90 %. El estudio demostró que los compuestos volátiles principalmente alcoholes superiores y aldehídos pueden funcionar como marcadores químicos de diferenciación geográfica y tecnológica, y que los métodos de ML superan en desempeño a las aproximaciones puramente descriptivas.

Aunado a lo anterior, se han documentado estudios en otras bebidas alcohólicas destiladas, en el caso del whisky, Swift et al. (2023) combinaron análisis de compuestos volátiles mediante GC-MS y espectrometría de masas de alta resolución con algoritmos de Random Forest y redes neuronales profundas, logrando clasificar whiskies escoceses con una precisión superior al 95 %, en ambos casos. Este estudio demostró que los perfiles químicos multivariados

pueden discriminar diferencias geográficas y tecnológicas relacionadas con la maduración y el proceso de destilación, evidenciando que el ML permite identificar marcadores predictivos incluso cuando las variaciones individuales no son estadísticamente significativas.

En un enfoque aún más complejo, Zhou, Yu, Zhu, et al. (2025) desarrollaron un sistema de clasificación para el baijiu chino, combinando GC-TOF/MS y GC-IMS con algoritmos de Random Forest y regresión logística. En su trabajo, más de 500 compuestos volátiles fueron cuantificados y analizados, logrando precisiones de hasta 91 % en la discriminación de distintos grados de calidad. Este estudio representa un hito metodológico, ya que integró datos de alta dimensionalidad provenientes de diferentes plataformas instrumentales, confirmando que la sinergia entre la química analítica y la inteligencia artificial permite establecer relaciones predictivas entre composición y calidad en matrices alcohólicas complejas.

2.6. Relevancia del desbalance de clases en la modelación del sistema mezcalero

En los problemas de clasificación supervisada, el desbalance de clases se presenta cuando las categorías del conjunto de datos no están representadas de manera equitativa, es decir, cuando una o varias clases contienen significativamente más observaciones que otras. Esta situación es común en contextos reales donde la recolección de datos no puede controlarse experimentalmente, como en los sistemas biológicos o productivos. En el caso del mezcal, esta condición refleja una realidad estructural del sector: la producción está dominada por *Agave angustifolia*, mientras que las especies silvestres como *A. karwinskii*, *A. potatorum* y *A. marmorata* representan fracciones mucho menores (COMERCAM, 2025).

Desde el punto de vista computacional, el desbalance afecta el proceso de aprendizaje de los modelos, pues los algoritmos tienden a favorecer la clase mayoritaria para minimizar el error global, lo que puede producir resultados con alta exactitud aparente, pero con un bajo poder predictivo en las clases

minoritarias (He & Garcia, 2009). En consecuencia, los modelos pueden generar predicciones sesgadas, incapaces de reconocer adecuadamente las observaciones pertenecientes a grupos con menor representación (Johnson & Khoshgoftaar, 2019).

Para abordar el problema del desbalance entre clases, la literatura ha propuesto diversas estrategias que incluyen la modificación del conjunto de datos mediante submuestreo o sobremuestreo, el ajuste de pesos en las clases minoritarias y el uso de métricas diseñadas para penalizar errores en categorías poco representadas (López et al., 2013). Entre estas técnicas, SMOTE (Synthetic Minority Over-Sampling Technique) y el submuestreo aleatorio (RUS) son dos de los métodos más empleados en estudios de clasificación. Aunque SMOTE puede generar instancias sintéticas útiles para equilibrar la distribución, distintos autores han advertido que, en sistemas biológicos o productos con variabilidad natural marcada, puede introducir patrones que no reflejan completamente la complejidad del fenómeno estudiado (Mukherjee & Khushi, 2021). El submuestreo, por su parte, puede simplificar el conjunto de datos a costa de eliminar información relevante. A pesar de estas limitaciones, ambas técnicas resultan útiles para evaluar si el rendimiento de los modelos puede mejorar en contextos donde el desbalance severo impide alcanzar niveles adecuados de clasificación bajo las condiciones originales del conjunto de datos.

En escenarios con desbalance de clases, la evaluación del desempeño no debe basarse únicamente en la exactitud (accuracy), sino en métricas que integren tanto la precisión (precision) como la exhaustividad (recall), como el f1-score. Esta métrica pondera los falsos positivos y falsos negativos, proporcionando una medida más robusta y equitativa del rendimiento de los modelos en contextos desbalanceados (He & Garcia, 2009).

El estudio del desbalance de clases no solo tiene implicaciones técnicas, sino también interpretativas, ya que en muchos sistemas como en la cadena productiva de mezcal, la estructura desigual de los datos refleja dinámicas reales de disponibilidad y producción. Por tanto, el desarrollo de modelos capaces de

mantener un buen desempeño bajo estas condiciones puede considerarse no solo un avance metodológico, sino también una representación más fiel de la complejidad inherente a los sistemas agroindustriales.

2.7. Integración del análisis sensorial descriptivo (CATA)

El análisis sensorial es una herramienta fundamental en la ciencia de los alimentos, cuyo propósito es evaluar las propiedades sensoriales de un producto a través de la percepción humana (ISO 13299, 2016). En el caso de bebidas alcohólicas como el mezcal, esta disciplina adquiere relevancia no solo por su relación con la aceptación del consumidor, sino también por su capacidad para capturar la complejidad sensorial derivada de factores biológicos y tecnológicos asociados a la producción (Varela & Ares, 2012).

Dentro de los distintos enfoques sensoriales disponibles, la metodología Check All That Apply (CATA) se ha consolidado como una de las más eficientes para la descripción rápida de productos complejos. El método consiste en presentar a un grupo de consumidores o jueces entrenados una lista predefinida de atributos (sensoriales o descriptivos), solicitando que marquen todos aquellos que perciban en la muestra (Ares et al., 2015). Este enfoque genera datos categóricos que, al ser procesados mediante análisis estadísticos multivariados como el análisis de correspondencia (AC) permiten visualizar patrones perceptivos y diferenciar las muestras según su perfil sensorial (Varela & Ares, 2012).

La metodología CATA ha demostrado ser una herramienta eficaz en el ámbito de las bebidas alcohólicas, ya que facilita la identificación de atributos sensoriales relevantes directamente desde la perspectiva del consumidor. En el caso del vino, su aplicación por Nougarede et al. (2023) en un panel conformado por estudiantes no entrenados permitió comparar la técnica CATA con el método Adapted-Pivot-Test, encontrando que ambos generaron descripciones sensoriales consistentes, aunque la prueba CATA mostró mayor discriminación entre muestras y una mejor reproducibilidad en los atributos seleccionados. Los participantes lograron identificar diferencias perceptibles en color, intensidad aromática, acidez y astringencia, lo que evidenció la utilidad del método para

obtener perfiles sensoriales fiables incluso con paneles inexpertos. De manera similar, en el estudio de Zhou, Yu, Zhu, et al. (2025) aplicado a muestras de baijiu tipo fuerte, la metodología CATA fue empleada junto con una escala hedónica de nueve puntos para evaluar cinco grados de calidad de licor base. En el caso del mezcal, la integración del análisis sensorial representa una aproximación paralela y complementaria al estudio de su composición fisicoquímica. Aunque los parámetros empleados en el proceso de certificación a excepción de los alcoholes superiores no están directamente asociados con atributos perceptibles, ambos niveles de información reflejan distintos aspectos de la misma realidad: la complejidad del producto derivada de la especie de agave, el proceso artesanal y las condiciones del entorno productivo (Sánchez-Fernández et al., 2025; Vera-Guzmán et al., 2018).

En síntesis, la metodología CATA ofrece un enfoque sensorial accesible, reproducible y coherente con la diversidad del mezcal. Su integración en estudios de clasificación no solo enriquece la comprensión de la variabilidad sensorial entre especies de agave, sino que también sienta las bases para correlacionar las percepciones humanas con los perfiles fisicoquímicos, fortaleciendo la caracterización integral de esta bebida artesanal.

2.8. Relevancia de la clasificación por especie de maguey

La revisión de literatura permitió establecer que, aunque la composición química del mezcal ha sido ampliamente documentada, la mayoría de los estudios existentes han permanecido en un nivel descriptivo, centrado en la identificación de compuestos volátiles, metabolitos secundarios o atributos perceptuales, sin avanzar hacia el desarrollo de modelos predictivos basados en aprendizaje automático. Este vacío metodológico delimita la necesidad de integrar enfoques computacionales capaces de aprovechar los datos fisicoquímicos ya disponibles para la clasificación del producto según la especie de agave empleada.

En este contexto, el presente trabajo se sustenta en la hipótesis de que el perfil fisicoquímico del mezcal, determinado durante el proceso de certificación bajo condiciones estandarizadas, contiene información suficiente para diferenciar la

bebida por especie de maguey. Dicha hipótesis se apoya en la evidencia reportada en la literatura, que demuestra que parámetros como los alcoholes superiores, aldehídos, furfural y metanol reflejan la variabilidad biológica de la especie de agave utilizada en el proceso de producción.

El análisis de los antecedentes también evidenció que la aplicación del aprendizaje automático en bebidas alcohólicas ha logrado resolver con éxito problemas de autenticidad, diferenciación y control de calidad en matrices químicas complejas como el vino, el whisky o el baijiu. Sin embargo, aún no se han reportado trabajos en el contexto del mezcal artesanal, lo que evidencia la pertinencia de esta investigación al adaptar estrategias consolidadas en otros destilados a un sistema productivo con particularidades biológicas, tecnológicas y económicas propias.

En consecuencia, el propósito de esta investigación no se limita a la clasificación por especie de agave, sino que busca sentar las bases para la incorporación de técnicas de inteligencia artificial en el estudio del mezcal. Con ello, se amplía el horizonte de análisis hacia un enfoque de clasificación, donde los datos rutinarios generados durante la certificación se transforman en un recurso de valor analítico y estratégico para el sector mezcalero.

2.9. Literatura citada

- Aguirre-Dugua, X., & Eguiarte, L. E. (2013). Genetic diversity, conservation and sustainable use of wild *Agave cupreata* and *Agave potatorum* extracted for mezcal production in Mexico. *Journal of Arid Environments*, 90, 36–44. <https://doi.org/10.1016/j.jaridenv.2012.10.018>
- Andrade, J. M., Ballabio, D., Gómez-Carracedo, M. P., & Pérez-Caballero, G. (2017). Nonlinear classification of commercial Mexican tequilas. *Journal of Chemometrics*, 31(12), e2939. <https://doi.org/10.1002/cem.2939>
- Anjos, O., Santos, A. J. A., Estevinho, L. M., & Caldeira, I. (2016). FTIR–ATR spectroscopy applied to quality control of grape-derived spirits. *Food Chemistry*, 205, 28–35. <https://doi.org/10.1016/j.foodchem.2016.02.128>

- Anta-Fonseca, S. (2016). *Diagnóstico de la cadena de valor forestal en las regiones de Oaxaca* [Informe técnico].
- Ares, G., Antúnez, L., Bruzzone, F., Vidal, L., Giménez, A., Pineau, B., Beresford, M. K., Jin, D., Paisley, A. G., Chheang, S. L., Roigard, C. M., & Jaeger, S. R. (2015). Comparison of sensory product profiles generated by trained assessors and consumers using CATA questions: Four case studies with complex and/or similar samples. *Food Quality and Preference*, 45, 75–86. <https://doi.org/10.1016/j.foodqual.2015.05.007>
- Ávila-Reyes, S. V., Jiménez-Aparicio, A. R., Melgar-Lalanne, G., Fajardo-Espinoza, F. S., & Hernández-Sánchez, H. (2025). Mezcal: A review of chemistry, processing, and potential health benefits. *Foods*, 14(8), 1408. <https://doi.org/10.3390/foods14081408>
- COMERCAM. (2025). *Informe público 2025*.
- Elgeldawi, E., Sayed, A., Galal, A. R., & Zaki, A. M. (2021). Hyperparameter tuning for machine learning algorithms used for Arabic sentiment analysis. *Informatics*, 8(4), 79. <https://doi.org/10.3390/informatics8040079>
- Fawagreh, K., Gaber, M. M., & Elyan, E. (2014). Random forests: From early developments to recent advancements. *Systems Science and Control Engineering*, 2(1), 602–609. <https://doi.org/10.1080/21642583.2014.956265>
- Frausto-Reyes, C., Medina-Gutiérrez, C., Sato-Berrú, R., & Sahagún, L. R. (2005). Qualitative study of ethanol content in tequilas by Raman spectroscopy and principal component analysis. *Spectrochimica Acta Part A*, 61(11–12), 2657–2662. <https://doi.org/10.1016/j.saa.2004.10.008>
- Gámez-Martínez, M., Alfaro-Cortés, E., Alfaro-Navarro, J., & García-Rubio, N. (2009). Árboles de clasificación para el análisis de gráficos de control multivariantes. *Revista de Matemática: Teoría y Aplicaciones*, 16(1), 30–42.

- González Seguí, H. O., Hernández López, J. D. J., & Giersiepen, J. H. (2019). Metanol: Tolerancias y exigencias en las normas para mezcal y bebidas de agave. *Revista RIVAR*, 7(19), 1–20. <https://doi.org/10.35588/rivar.v7i19.4246>
- Hazelwood, L. A., Daran, J.-M., van Maris, A. J. A., Pronk, J. T., & Dickinson, J. R. (2008). The Ehrlich pathway for fusel alcohol production: A century of research on *Saccharomyces cerevisiae* metabolism. *Applied and Environmental Microbiology*, 74(8), 2259–2266. <https://doi.org/10.1128/aem.02625-07>
- He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263–1284. <https://doi.org/10.1109/tkde.2008.239>
- International Organization for Standardization. (2016). *ISO 13299: Sensory analysis — Methodology — General guidance for establishing a sensory profile*. ISO. <https://www.iso.org/standard/58024.html>
- Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *Journal of Big Data*, 6, 27. <https://doi.org/10.1186/s40537-019-0192-5>
- Kamilaris, A., & Prenafeta-Boldú, F. X. (2018). Deep learning in agriculture: A survey. *Computers and Electronics in Agriculture*, 147, 70–90. <https://doi.org/10.1016/j.compag.2018.02.016>
- Koutsoukas, A., Monaghan, K. J., Li, X., & Huan, J. (2017). Deep-learning: Investigating deep neural networks hyper-parameters and comparison of performance to shallow methods for modeling bioactivity data. *Journal of Cheminformatics*, 9, 42. <https://doi.org/10.1186/s13321-017-0226-y>
- Lachenmeier, D. W., Richling, E., López, M. G., Frank, W., & Schreier, P. (2005). Multivariate analysis of FTIR and ion chromatographic data for the quality control of tequila. *Journal of Agricultural and Food Chemistry*, 53(6), 2151–2157. <https://doi.org/10.1021/jf048637f>

- Lee, K. J., Trowbridge, A. C., Bruce, G. D., Dwapanyin, G. O., Dunning, K. R., Dholakia, K., & Schartner, E. P. (2024). Learning algorithms for identification of whisky using portable Raman spectroscopy. *Current Research in Food Science*, 8, 100729. <https://doi.org/10.1016/j.crfs.2024.100729>
- López, V., Fernández, A., García, S., Palade, V., & Herrera, F. (2013). An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information Sciences*, 250, 113–141. <https://doi.org/10.1016/j.ins.2013.07.007>
- López-Ramírez, J. E., Martín-del-Campo, S. T., Escalona-Buendía, H., García-Fajardo, J. A., & Estarrón-Espinosa, M. (2013). Physicochemical quality of tequila during barrel maturation: A preliminary study. *CYTA – Journal of Food*, 11(3), 189–197. <https://doi.org/10.1080/19476337.2012.727033>
- Mukherjee, M., & Khushi, M. (2021). SMOTE-ENC: A novel SMOTE-based method to generate synthetic data for nominal and continuous features. *Applied System Innovation*, 4(1), 18. <https://doi.org/10.3390/asi4010018>
- Nolasco-Cancino, H., Santiago-Urbina, J. A., Wachter, C., & Ruíz-Terán, F. (2018). Predominant yeasts during artisanal mezcal fermentation and their capacity to ferment maguey juice. *Frontiers in Microbiology*, 9, 2900. <https://doi.org/10.3389/fmicb.2018.02900>
- Nougarède, S., Diot, A., Maza, E., Samson, A., Olivier-Salvagnac, V., Caillé, S., Geffroy, O., & Chervin, C. (2023). Comparison of Check-All-That-Apply and Adapted-Pivot-Test methods for wine sensory characterization with a panel of untrained students. *Journal of Sensory Studies*, 38(5). <https://doi.org/10.1111/joss.12862>
- Pérez, E., González-Hernández, J. C., Chávez-Parga, M. C., & Cortés-Penagos, C. (2013). Caracterización fermentativa de levaduras productoras de etanol a partir de jugo de *Agave cupreata* en la elaboración de mezcal. *Revista Mexicana de Ingeniería Química*, 12(3), 543–556.

- Pérez-Caballero, G., Andrade, J. M., Olmos, P., Molina, Y., Jiménez, I., Durán, J. J., Fernández-Lozano, C., & Miguel-Cruz, F. (2017). Authentication of tequilas using pattern recognition and supervised classification. *TrAC Trends in Analytical Chemistry*, 94, 1–9. <https://doi.org/10.1016/j.trac.2017.07.008>
- Pino, J. A., Tolle, S., Gök, R., & Winterhalter, P. (2012). Characterisation of odour-active compounds in aged rum. *Food Chemistry*, 132(3), 1436–1441. <https://doi.org/10.1016/j.foodchem.2011.11.133>
- Russell, S., & Norvig, P. (2021). *Artificial intelligence: A modern approach* (4th ed.). Pearson.
- Sánchez-Fernández, R. E., Pérez-López, A., Morales-Solís, A., Manilla-Téllez, Y., Reyes-Carmona, E. D., & Ávila-Urbe, G. (2025). Chemistry of mezcal: Volatile profile of artisanal mezcal made from wild agaves of Oaxaca. *Foods*, 14(7), 1222. <https://doi.org/10.3390/foods14071222>
- Secretaría de Economía. (2017). *NOM-070-SCFI-2016: Bebidas alcohólicas — Mezcal — Especificaciones*. Diario Oficial de la Federación. <https://www.dof.gob.mx>
- Singh, M., & Domijan, K. (2019). Comparison of machine learning models in food authentication studies. In *Proceedings of the 30th Irish Signals and Systems Conference (ISSC)* (pp. 1–6). <https://doi.org/10.1109/issc.2019.8904924>
- Sujka, K., & Koczoń, P. (2018). The application of FT-IR spectroscopy in discrimination of differently originated and aged whisky. *European Food Research and Technology*, 244(11), 2035–2043. <https://doi.org/10.1007/s00217-018-3113-5>
- Sulzmann, J. N., Fürnkranz, J., & Hüllermeier, E. (2007). On pairwise naive Bayes classifiers. In J. Kok (Ed.), *Machine Learning: ECML 2007* (pp. 371–382). Springer. https://doi.org/10.1007/978-3-540-74958-5_35
- Swift, J. R., Turner, M. A., & Reynolds, J. C. (2023). A rapid dynamic headspace method for authentication of whiskies using artificial neural networks. *Food*

Chemistry Advances, 3, 100417.
<https://doi.org/10.1016/j.focha.2023.100417>

Varela, P., & Ares, G. (2012). Sensory profiling: The blurred line between sensory and consumer science. A review of novel methods for product characterization. *Food Research International*, 48(2), 893–908.
<https://doi.org/10.1016/j.foodres.2012.06.037>

Vera-Guzmán, A. M., Guzmán-Gerónimo, R. I., López, M. G., & Chávez-Servia, J. L. (2018). Volatile compound profiles in mezcal spirits as influenced by agave species and production processes. *Beverages*, 4(1), 9.
<https://doi.org/10.3390/beverages4010009>

Warren-Vega, W. M., Contreras-Atrisco, Z. A., Ramírez-Quezada, M. F., & Romero-Cano, L. A. (2023). A novel approach of artificial intelligence for the study of the relation of physicochemical profile and color acquired by tequila 100% agave in its maturation process. *Journal of Food Composition and Analysis*, 123, 105533. <https://doi.org/10.1016/j.jfca.2023.105533>

Yang, L., & Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, 415, 295–316.
<https://doi.org/10.1016/j.neucom.2020.07.061>

Zhou, C., Yu, Y., Ai, J., Song, C., Cui, Z., Zhou, Q., Zhao, S., Huang, R., Ao, Z., Peng, B., Chen, P., Feng, X., Li, D., & Liu, Y. (2025). Fruit wines classification enabled by combining machine learning with comprehensive volatile profiles of GC-TOF/MS and GC-IMS. *Food Research International*, 204, 115890.
<https://doi.org/10.1016/j.foodres.2025.115890>

Zhou, C., Yu, Y., Zhu, Y., Chai, J., Liu, R., Cui, Z., Li, D., Feng, X., & Liu, Y. (2025). Machine learning discrimination of quality grades of base liquor integrating GC-TOF/MS and GC-IMS data analysis: Case study of strong-flavor Chinese baijiu. *Food Chemistry*, 492, 145281.
<https://doi.org/10.1016/j.foodchem.2025.145281>

3. CLASIFICACIÓN DE MEZCALES ARTESANALES POR ESPECIE DE AGAVE EMPLEANDO SU PERFIL FISICOQUÍMICO Y APRENDIZAJE AUTOMÁTICO

3.1. Introducción

El mezcal es una bebida alcohólica destilada elaborada a partir de diferentes especies del género *Agave*, cuya producción se encuentra protegida por la Denominación de Origen Mezcal (DOM) y regulada por la NOM-070-SCFI-2016, la cual establece los criterios de calidad, autenticidad y parámetros fisicoquímicos que deben cumplirse durante el proceso de certificación. Dichos parámetros incluyen el contenido de alcoholes superiores, metanol, aldehídos, furfural y extracto seco, determinados mediante técnicas cromatográficas estandarizadas. Si bien estos datos son esenciales para garantizar el cumplimiento normativo, su uso se ha limitado a la verificación de conformidad, sin explorar su potencial como fuente de información para la clasificación de la bebida (COMERCAM, 2025).

En este sentido, la certificación asegura que el destilado cumpla con los valores de referencia, pero no aporta información suficiente para caracterizar la bebida. Esta carencia limita el potencial de los datos generados en la certificación, pues impide aprovecharlos para fines de trazabilidad o autenticación del producto que contribuya a evitar posibles fraudes o adulteraciones, fenómenos que representan una amenaza constante en la integridad de la cadena productiva del mezcal (Quintero Arenas et al., 2020). Frente a estas limitaciones, surge la necesidad de explorar metodologías complementarias que permitan dar un valor añadido a la información ya disponible (Andrade et al., 2017). Una alternativa prometedora es el aprendizaje automático (machine learning, ML), capaz de identificar patrones complejos en datos multivariados y de clasificar productos con base en sus características químicas.

Investigaciones previas en destilados como tequila, whisky y baijiu han demostrado que los algoritmos de ML alcanzan altos niveles de precisión (superiores al 90 %) al discriminar categorías de origen, calidad o materia prima (Swift et al., 2023; Warren-Vega et al., 2023; Zhou, Yu, Zhu, et al., 2025). Sin embargo, hasta ahora no se ha reportado su aplicación al mezcal.

La mayoría de los estudios realizados en mezcal se han enfocado en la caracterización descriptiva de su composición química o volátil, sin explorar el uso de modelos de clasificación que permitan clasificar muestras nuevas o automatizar la identificación de la especie de agave empleada (Sánchez-Fernández et al., 2025; Vera-Guzmán et al., 2018). Esta brecha metodológica evidencia la ausencia de herramientas computacionales aplicadas al contexto artesanal, donde los recursos para análisis avanzados son limitados.

Ante este panorama, el presente trabajo propone un enfoque que aprovecha los datos fisicoquímicos rutinarios de certificación ya disponibles, confiables y estandarizados como insumo para el entrenamiento de modelos de aprendizaje automático orientados a la clasificación de mezcales según la especie de maguey utilizada en su elaboración. Este planteamiento no implica costos analíticos adicionales y permite explorar la capacidad de dichos parámetros para reflejar la variabilidad biológica inherente al producto.

Complementariamente, se integró un análisis sensorial mediante la metodología Check All That Apply (CATA), con el fin de identificar agrupamientos perceptuales entre las muestras (Galmarini & Visalli, 2024; Varela & Ares, 2012).

El estudio plantea que los datos fisicoquímicos generados rutinariamente durante la certificación del mezcal pueden constituir la base de modelos predictivos robustos capaces de distinguir entre especies de agave, aportando así herramientas para la clasificación y el control de calidad en la industria mezcalera.

De este modo, el objetivo del estudio fue clasificar mezcales por especie de agave utilizando modelos de aprendizaje automático entrenados con datos

fisicoquímicos de certificación, para identificar los modelos idóneos para dicha tarea.

3.2. Materiales y métodos

3.2.1. Conjunto de datos

El conjunto de datos empleado en esta investigación estuvo compuesto por 393 instancias provenientes de registros fisicoquímicos generados de manera rutinaria durante el proceso de certificación del mezcal, conforme a la NOM-070-SCFI-2016, por el organismo evaluador de la conformidad (OEC) Asociación de Agave y Mezcal Artesanal A.C. Los análisis se realizaron en laboratorios acreditados bajo procedimientos cromatográficos estandarizados, garantizando la confiabilidad de las mediciones.

Cada registro correspondió a una muestra certificada de mezcal artesanal con su respectiva especie de agave identificada. Se consideraron cinco variables continuas: alcoholes superiores, metanol, furfural, aldehídos y extracto seco, expresadas en miligramos por 100 mL de alcohol anhidro.

Si bien el conjunto de datos es único, para los fines del análisis se generaron cuatro escenarios independientes de clasificación binaria, cada uno correspondiente a una de las especies de mayor interés productivo: *Agave angustifolia*, *A. karwinskii*, *A. marmorata* y *A. potatorum*. En todos los casos, la variable objetivo se redefinió siguiendo la estrategia Uno contra todos, donde la clase positiva estuvo conformada exclusivamente por las muestras pertenecientes a la especie objetivo, mientras que la clase negativa agrupó el resto de las especies presentes en el conjunto de datos. Esta clase negativa incluyó también especies con muy baja representación, como *A. atrovirens*, *A. rodhocanta*, *A. convallis*, *A. cupreata* y *A. salmiana*, las cuales no se consideraron como clase positiva debido a que contaban con un número muy pequeño de instancias, generalmente una sola muestra, reflejando su escasa presencia en el mercado. La inclusión de estas especies dentro de la clase negativa permitió conservar la estructura completa del conjunto de datos sin modificar su

composición original, garantizando así un análisis coherente con la disponibilidad real de materia prima en el sistema mezcalero.

El uso de modelos de clasificación binaria se fundamentó en el interés de evaluar de forma independiente la capacidad de los algoritmos para distinguir cada especie de *Agave*. Este enfoque permitió analizar con mayor precisión el desempeño de clasificación frente a cada clase objetivo, especialmente en contextos con proporciones naturalmente desbalanceadas, como en la producción de mezcal. De esta manera, cada modelo se entrenó para identificar una sola especie frente a las demás, lo que posibilita una evaluación más clara de su capacidad discriminante y una interpretación específica de los patrones asociados a cada variedad (Rožman et al., 2024).

3.2.2. Manejo metodológico del desbalance de clases y definición de escenarios de análisis

El conjunto de datos presenta una marcada asimetría en la representación de especies de agave, que refleja directamente la estructura actual del sistema productivo del mezcal. De acuerdo con cifras del COMERCAM, en el año 2025 *A. angustifolia* constituyó aproximadamente 85.25 % de la producción certificada, mientras que *A. potatorum* cerca del 1 %, *A. marmorata* alrededor de 0.46 % y *A. karwinskii* aproximadamente 0.44 %. De tal forma que la distribución desigual observada en el conjunto de datos no es producto de un sesgo de muestreo, sino de la realidad productiva del sector. Desde el punto de vista del aprendizaje automático, tal desbalance puede influir en el entrenamiento al inducir que los algoritmos favorezcan la clase mayoritaria (Buda et al., 2018). Para cuantificar esta condición, se calculó la tasa de desbalance como la razón entre los tamaños de las clases mayoritaria y minoritaria, siguiendo a Johnson & Khoshgoftaar (2019).

En primer lugar, se entrenaron los modelos con el conjunto de datos, tal y como fue proporcionado por el OEC, preservando su distribución natural entre especies. Este escenario constituyó el punto de partida para evaluar la capacidad real de los algoritmos bajo las condiciones propias del proceso de certificación.

Posteriormente y con el fin de cumplir con el objetivo de este trabajo se incorporaron procedimientos adicionales orientados a examinar distintos enfoques para manejar o modificar el desbalance entre clases. Estos procedimientos incluyeron técnicas de sobremuestreo, submuestreo y la definición de escenarios alternativos con distribuciones más equilibradas, lo que permitió analizar cómo distintas configuraciones del conjunto de datos influyeron en el comportamiento de los algoritmos. Entre ellos, se consideraron técnicas de sobremuestreo y submuestreo, ampliamente utilizadas en clasificación desbalanceada. El sobremuestreo se exploró mediante SMOTE, técnica que genera instancias sintéticas de las clases minoritarias mediante interpolación entre observaciones reales (Johnson & Khoshgoftaar, 2019). El submuestreo, por su parte, se implementó mediante la técnica Random Undersampling (RUS), que consiste en reducir el tamaño de la clase mayoritaria mediante la eliminación aleatoria de observaciones (Hasanin et al., 2019). Estas estrategias se implementaron dentro del proceso metodológico para evaluar su efecto sobre la clasificación; sin embargo, en contextos químicos o biológicos se ha documentado que el sobremuestreo puede generar patrones que no representan adecuadamente la variabilidad natural de los datos, mientras que el submuestreo puede eliminar información valiosa al reducir la diversidad de observaciones reales (Mukherjee & Khushi, 2021).

Además de estas técnicas, se definieron dos escenarios alternativos para analizar el comportamiento de los algoritmos bajo configuraciones del conjunto de datos menos asimétricas, pero sin generar información sintética. En el Escenario A, se excluyeron las observaciones de *A. angustifolia*, con el fin de evaluar la clasificación entre especies que presentaban menor tasa de desbalance. En el Escenario B, se eliminaron tanto *A. angustifolia*, como *A. karwinskii*, de manera que el análisis se enfocó exclusivamente en las especies con menor representación muestral.

Finalmente, y en coherencia con el objetivo de esta investigación, se incorporó una etapa metodológica adicional que amplió el proceso de análisis inicialmente

planteado. Esta segunda fase surgió de manera natural conforme avanzó el estudio, ya que, tras aplicar los modelos binarios en los diferentes escenarios y configuraciones del conjunto de datos, se identificó la necesidad de profundizar en alternativas que permitieran seguir avanzando hacia el cumplimiento de este.

Particularmente, en el Escenario B, la exclusión de *A. angustifolia* y *A. karwinskii* dio lugar a un conjunto de datos compuesto exclusivamente por las especies minoritarias restantes y dado que tras explorar ampliamente el comportamiento de los modelos bajo esquemas binarios y distintas estrategias de manejo del desbalance, resultaba pertinente examinar también la clasificación simultánea de todas las especies presentes en el escenario. Así, la incorporación del enfoque multiclase se constituyó como una etapa complementaria del estudio, orientada a ampliar el análisis y a mantener la coherencia con el objetivo de identificar modelos capaces de distinguir mezcales según su especie de agave.

3.2.3. Preprocesamiento y estandarización

Dado que las variables presentan diferentes unidades y rangos de magnitud, se realizó una estandarización mediante la ecuación 1, calculando para cada variable x_i su valor normalizado:

$$z_i = \frac{x_i - \mu}{\sigma} \quad (1)$$

donde μ es la media y σ la desviación estándar de la variable en el conjunto de entrenamiento.

Esta etapa garantizó que todas las variables contribuyan equitativamente al proceso de aprendizaje, evitando que aquellas con mayor rango numérico dominen el ajuste de los modelos (Jain et al., 2000).

3.2.4. Proceso de validación

Se implementó una validación cruzada estratificada de tipo K-folds con 10 particiones ($k = 10$). Antes de la división, los datos fueron barajados

aleatoriamente para evitar sesgos de orden, y se establecieron semillas aleatorias fijas con el propósito de asegurar la reproducibilidad de los resultados.

Esta estrategia permitió estimar el rendimiento medio y su variabilidad, maximizando el aprovechamiento de las muestras disponibles. Su elección se fundamenta en que la estratificación corrige posibles sesgos de clase en conjuntos desbalanceados, asegurando una evaluación justa del modelo (Elgeldawi et al., 2021).

3.2.5. Selección de algoritmos de clasificación

La selección de los algoritmos constituye un elemento central en esta investigación, ya que cada modelo representa un enfoque matemático distinto para aprender patrones presentes en el perfil fisicoquímico del mezcal. Dado que no existían antecedentes sobre el comportamiento de este tipo de datos en tareas de aprendizaje automático, fue necesario incluir modelos diversos en estructura, complejidad y forma de construir sus fronteras de decisión. Esto permitió explorar diferentes maneras de modelar la relación entre las variables fisicoquímicas y la especie de agave, sin asumir de antemano qué tipo de modelo sería más adecuado.

Para ello, se evaluaron seis algoritmos supervisados ampliamente utilizados en el análisis químico y en estudios de clasificación en bebidas alcohólicas (Warren-Vega et al., 2023; Swift et al., 2023; Zhou, Yu, Ai, et al., 2025): Máquinas de Soporte Vectorial (SVM), Naive Bayes (NB), k-Vecinos más Cercanos (KNN), Árboles de Decisión (DT), Bosques Aleatorios (RF) y Perceptrón Multicapa (MLP). En conjunto, estos modelos cubren una gama amplia de fundamentos matemáticos que incluye fronteras lineales, probabilidad, vecindad, segmentación jerárquica y redes neuronales.

En conjunto, estos algoritmos abarcan enfoques lineales, probabilísticos, basados en distancia, jerárquicos y neuronales. Esta diversidad metodológica permitió realizar una comparación equilibrada y robusta del desempeño de las

distintas aproximaciones matemáticas, sin asumir de manera anticipada cuál de ellas es la más adecuada para caracterizar la variabilidad química del mezcal.

3.2.6. Ajuste de hiperparámetros

El ajuste se realizó mediante búsqueda en cuadrícula (GridSearchCV) con validación cruzada estratificada ($k = 10$) empleando como métrica principal el f1-score de la clase positiva, debido al desbalance de clases.

Los rangos de valores incluidos en la búsqueda fueron propuestos en esta investigación con el propósito de explorar el rendimiento de los modelos bajo diferentes configuraciones, ya que no se identificaron estudios previos que reportaran ajustes específicos de hiperparámetros para datos fisicoquímicos en bebidas destiladas, estos se muestran en el Cuadro 1. El procedimiento de ajuste permitió comparar el desempeño entre modelos con parámetros por defecto y modelos ajustados.

3.2.7. Análisis sensorial (CATA)

Caracterización de muestras para el análisis sensorial

Se emplearon cuatro mezcales artesanales provenientes de la comunidad de Sitio de Xitlapehua, Miahuatlán de Porfirio Díaz, Oaxaca ($16^{\circ}21'45''$ N, $96^{\circ}32'30''$ W), elaborados con las cuatro especies de agave analizadas en el proceso de clasificación: *Agave angustifolia*, *A. potatorum*, *A. marmorata* y *A. karwinskii*. El criterio de selección del palenque fue que siguiera los métodos de producción artesanal conforme a lo establecido en la NOM-070-SCFI-2016, asegurando el cumplimiento de la denominación de origen y de los procesos tradicionales de elaboración. Se optó por un solo productor y zona geográfica para controlar factores externos como técnicas de fermentación y destilación, de manera que las diferencias observadas en los perfiles fisicoquímicos, compuestos volátiles y características sensoriales pudieran atribuirse principalmente a la especie de agave utilizada. Las muestras fueron recolectadas en frascos de vidrio y almacenadas a temperatura ambiente para su análisis posterior.

Cuadro 1. Hiperparámetros evaluados durante el proceso de ajuste.

Algoritmo	Parámetro	Valores	Aporte
SVM	C	0.1,1,10,100	Controla el equilibrio entre el ajuste al dataset y evitar el sobre ajuste.
	Kernel	Linear, rbf	Es una función que transforma los datos a un espacio de mayor dimensión para que sean separables mediante un hiperplano.
	Gamma	Scale, auto, 0.1, 1	Afecta la influencia de un solo ejemplo (solo para kernels no lineales)
KNN	n_neighbors	3,5,7,9,11,15	Número de vecinos considerados para la clasificación.
	priors	None, [0.5, 0.5]	Se refiere a las probabilidades iniciales de cada clase antes de observar los datos.
Naive bayes	var_smoothing	$10^{-9}, 10^{-8}, 10^{-7}, 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^0$	Agrega una pequeña cantidad al cálculo de la varianza para evitar inestabilidad numérica.
	Criterion	Gini, entropy	Métrica para evaluar la calidad de las divisiones.
Árboles de clasificación	Max_depth	3,5, 10, none	Profundidad máxima del árbol para evitar sobreajuste.
	min_samples_split	2, 5, 10	Mínimo de muestras requeridas para dividir un nodo
	min_samples_leaf	1, 2, 5,10,15	Mínimo de muestras requeridas para formar una hoja
	splitter	Best, random	Estrategia para dividir los nodos
	n_estimate	10,50,100,200,500	Número de árboles en el bosque. Más árboles suelen mejorar la estabilidad, pero aumenta el costo computacional.
Bosques de clasificación	Criterion	Gini, entropy	Métrica para evaluar la calidad de las divisiones.
	Max_depth	None, 10,20,30	Profundidad máxima de los árboles. None, permite que crezcan hasta que las hojas sean puras o mínimas.
	Min_samples_split	2,5,10	Número mínimo de muestras requeridas para dividir un nodo interno. Valores más altos reducen el sobreajuste.
	Min_samples_leaf	1,2,4	Número mínimo de muestras requeridas para dividir un nodo interno. Valores altos reducen el sobreajuste.
	Learning_rate	0.001,0.01,0.1	Determina la velocidad con la que el modelo actualiza sus pesos durante el entrenamiento.
Perceptrón multicapa	Hidden_units_1	8,12,16	Determina la cantidad de neuronas en la primera capa oculta
	hidden_units_2	0,4,6,8	Determina la cantidad de neuronas en la segunda capa oculta

Grupo focal

Se realizó un grupo focal en el laboratorio de Evaluación Sensorial de la Universidad Autónoma Chapingo, mismo que estuvo integrado por 12 consumidores de mezcal, a quienes se les presentaron las cuatro muestras del

destilado y se les solicitó que escribieran en una hoja que se les proveyó los descriptores o atributos presentes en cada una de las muestras y que mejor permitieran diferenciarlas, invitándolos a dividirlos en apariencia, aroma y sabor señalando que no había respuestas incorrectas. Los términos generados fueron utilizados en la metodología CATA.

Aplicación de la técnica CATA

Se reclutaron a 100 consumidores de mezcal de ambos géneros cuya edad fluctuó entre los 18 y 34 años, se les sirvieron 20 mL de mezcal en vasos codificados previamente con números aleatorios de tres dígitos, además se les proporcionaron agua y galletas para que se enjuagaran la boca entre muestra y muestra, mismas que fueron balanceadas, aleatorizadas y presentadas de manera monódica a cada uno de los panelistas. Además, se les proporcionó un formato donde se les presentó la pregunta CATA con 26 descriptores que fueron determinados por el grupo focal, y se les pidió que marcaran aquellos atributos que consideraban apropiados para describir a la muestra proporcionada, haciendo hincapié en que no había términos incorrectos y que no había un número máximo ni mínimo de atributos a considerar (Nougarède et al., 2023; Zhou, Yu, Zhu, et al., 2025).

El método CATA se eligió por su simplicidad, reproducibilidad y capacidad para capturar diferencias perceptuales entre productos sin requerir paneles entrenados (Nougarède et al., 2023; Varela & Ares, 2012; Zhou, Yu, Zhu, et al., 2025). Además, su análisis mediante Análisis de Correspondencias (CA) facilita la visualización de relaciones entre atributos y muestras, permitiendo identificar posibles agrupamientos perceptuales que complementen la clasificación obtenida por aprendizaje automático.

Entorno de desarrollo

Todos los modelos de aprendizaje automático se implementaron utilizando la biblioteca Scikit-learn en Python 3.8, ejecutados en el entorno Spyder dentro de Anaconda. La manipulación de datos se realizó con las bibliotecas NumPy y

Pandas. El equipo de cómputo contaba con un procesador Intel Core i7 de 13^a generación, 16 GB de RAM, tarjeta gráfica NVIDIA GeForce RTX 4070 y unidad de almacenamiento sólido de 512 GB, con sistema operativo Windows 11.

Cada experimento, conjunto de hiperparámetros y resultado fue documentado y almacenado de forma estructurada para su posterior análisis comparativo.

Consideraciones éticas

Los datos empleados corresponden a resultados analíticos obtenidos en procesos de certificación sin información personal ni comercial sensible. Su uso con fines académicos y científicos fue autorizado por la Asociación de Agave y Mezcal Artesanal A.C., garantizando la confidencialidad de los productores y lotes involucrados.

3.2.8. Análisis estadísticos

Variables fisicoquímicas

Se aplicó la prueba de Kruskal-Wallis para determinar diferencias significativas en el contenido de las variables analizadas entre las distintas especies de maguey en el mezcal. El análisis estadístico se llevó a cabo mediante Infostat versión 2020 (Universidad Nacional de Córdoba, Argentina).

Datos CATA

Primero, se obtuvo la frecuencia de uso de cada término descriptivo incluido en la pregunta CATA. Luego, se realizó la prueba Q de Cochran con el objetivo de evaluar si existió diferencia significativa ($p \leq 0.05$) en la proporción de cada uno de los términos entre las muestras (McKean & Manoukian, 1988). Luego, se aplicó un Análisis de Correspondencia (AC) de acuerdo con la matriz de frecuencias para cada descriptor que resultó significativo, para ello se empleó la distancia de Hellinger con la finalidad de cuantificar las similitudes entre variables. Dichos análisis se realizaron en el programa XLSTAT versión 2019 (Addinsoft, EE. UU.).

3.3. Resultados y discusión

Antes del entrenamiento de los modelos de clasificación, se aplicó la prueba no paramétrica de Kruskal–Wallis (H) con el fin de determinar si existían diferencias estadísticamente significativas entre las especies de *Agave* en relación con las variables fisicoquímicas analizadas: alcoholes superiores, metanol, aldehídos, furfural y extracto seco. Esta prueba fue seleccionada debido a que los datos no cumplían con los supuestos de normalidad ni homocedasticidad requeridos por un ANOVA, lo que hace de la prueba de Kruskal–Wallis una alternativa robusta para comparar medianas entre grupos independientes (Ostertagová et al., 2014).

Los resultados obtenidos al aplicar la prueba de Kruskal–Wallis se muestran en el Cuadro 2, donde se observan diferencias significativas entre especies en los niveles de alcoholes superiores ($H = 22.44$, $p < 0.05$) y metanol ($H = 44.30$, $p < 0.05$), mientras que extracto seco, furfural y aldehídos no mostraron diferencias estadísticamente significativas ($p > 0.05$). Este comportamiento sugiere que ciertos parámetros fisicoquímicos, particularmente los alcoholes superiores, podrían ser marcadores potenciales de variabilidad química asociada a la especie de maguey.

Cuadro 2. Análisis de parámetros fisicoquímicos del mezcal entre especies de agave mediante la prueba de Kruskal–Wallis.

Compuesto	H ^z	<i>A. angustifolia</i>	<i>A. karwinskii</i>	<i>A. marmorata</i>	<i>A. potatorum</i>
Alcoholes superiores	22.44	263.43±58.240a ^y	278.95±97.82ab	356.11±104.45c	306.69±77.10bc
Extracto seco	2.80	0.11±0.21a	0.28±0.41a	0.10±0.03a	0.13±0.19a
Metanol	44.30	191.95±50.08a	202.96±49.36a	234.49±51.16b	255.60±41.89b
Furfural	3.78	2.71±1.17a	2.90±1.15a	2.53±0.99a	2.36±1.12a
Aldehídos	7.45	12.06±4.76a	11.55±5.80a	11.88±5.57a	10.67±6.60a

^zLas poblaciones pueden considerarse estadísticamente diferentes si el parámetro H calculado es mayor que $\chi^2(3, 0.05) = 7.81$.

^yProporciones con diferente letra dentro del mismo renglón, son estadísticamente diferentes ($H > 7.81$).

En el caso de los alcoholes superiores, *A. marmorata* presentó las concentraciones más elevadas ($356.11 \pm 104.45 \text{ mg } 100 \text{ mL}^{-1}$), seguidas de *A. potatorum* ($306.69 \pm 77.10 \text{ mg } 100 \text{ mL}^{-1}$), en tanto que *A. angustifolia* y *A. karwinskii* mostraron valores significativamente menores. Este comportamiento sugiere una marcada influencia de la materia prima sobre la producción de alcoholes superiores durante la fermentación, dado que su concentración depende del contenido de aminoácidos y azúcares fermentables del mosto (Pino et al., 2012). Estos resultados coinciden con lo reportado por Vera-Guzmán et al. (2018), quienes asociaron las diferencias en el contenido de alcoholes con la composición metabólica intrínseca del agave utilizado.

En cuanto al metanol, las especies se agruparon en dos conjuntos bien diferenciados: *A. angustifolia* y *A. karwinskii* mostraron las menores concentraciones (191.95 ± 50.08 y $202.96 \pm 49.36 \text{ mg } 100 \text{ mL}^{-1}$, respectivamente), mientras que *A. marmorata* y *A. potatorum* registraron los valores más altos (234.49 ± 51.16 y $255.60 \pm 41.89 \text{ mg } 100 \text{ mL}^{-1}$). Este patrón puede relacionarse con la proporción de pectinas presentes en la planta, precursoras del metanol durante la hidrólisis térmica del agave, que varía entre especies (Lopez et al., 2003).

Por el contrario, las diferencias en furfural, aldehídos y extracto seco no alcanzaron significancia estadística. Estos compuestos se asocian más estrechamente con las condiciones de cocción y destilación que con factores biológicos, lo cual podría explicar la homogeneidad observada entre las especies analizadas (Pérez et al., 2013).

En conjunto, el análisis sugiere que los alcoholes superiores y el metanol son los parámetros fisicoquímicos con mayor poder discriminante entre especies de *Agave*, reflejando diferencias metabólicas en la materia prima que se conservan a lo largo del proceso artesanal de producción. Este hallazgo proporciona una base empírica para la aplicación de modelos de aprendizaje automático, los cuales pueden integrar simultáneamente estas variables y revelar relaciones no lineales que escapan al análisis univariado tradicional (Andrade et al., 2017).

3.3.1. Tasa de desbalance

En el Cuadro 3 se presentan las tasas de desbalance del conjunto completo, donde se observa que, mientras *A. angustifolia* muestra un desbalance moderado (1.69), las demás especies alcanzan valores superiores a 6, 11 y hasta 13.9, confirmando la marcada asimetría del conjunto de datos y su correspondencia con la realidad productiva (COMERCAM, 2025).

Cuadro 3. Tasas de desbalance del conjunto de datos completo.

Especie de maguey (clase positiva)	Número de instancias en la clase positiva	Número de instancias en la clase negativa	Tasa de desbalance
<i>A. angustifolia</i>	247	146	1.69
<i>A. karwinskii</i>	39	245	6.28
<i>A. marmorata</i>	25	287	11.48
<i>A. potatorum</i>	26	362	13.92

En el escenario A, se eliminaron todas las instancias correspondientes a *A. angustifolia*, permitiendo observar el comportamiento de los modelos en un sistema donde desaparece la clase dominante. El Cuadro 4 muestra que, aunque persiste cierto grado de desbalance, las tasas disminuyen de manera notable, lo que genera un entorno más equilibrado para evaluar el aprendizaje de las especies con menor representación.

Cuadro 4. Tasas de desbalance en el escenario A.

Especie de maguey en análisis	Número de instancias en la clase positiva	Número de instancias en la clase negativa	Tasa de desbalance
<i>A. karwinskii</i>	39	99	2.53
<i>A. marmorata</i>	25	140	5.60
<i>A. potatorum</i>	26	115	4.42

En el escenario B, además de *A. angustifolia* se excluyeron las instancias de *A. karwinskii*, lo que dio lugar a un conjunto formado únicamente por *A. marmorata* y *A. potatorum*, especies con frecuencias similares. El Cuadro 5 evidencia que las tasas de desbalance en este caso son próximas a 1, configurando un sistema prácticamente equilibrado. Este escenario permitió analizar si la ausencia de las

clases mayoritarias incrementa la capacidad discriminante de los algoritmos o modifica su desempeño en condiciones menos asimétricas.

Cuadro 5. Tasas de desbalance en el escenario B.

Especie de maguey en análisis	Número de instancias en la clase positiva	Número de instancias en la clase negativa	Tasa de desbalance
<i>A. marmorata</i>	25	73	2.92
<i>A. potatorum</i>	26	70	2.69

En cada uno de los tres escenarios (el conjunto completo y los dos conjuntos reducidos) se aplicó el mismo proceso de entrenamiento, validación cruzada estratificada y evaluación con f1-score de la clase positiva.

3.3.2. Selección de hiperparámetros y evaluación de modelos

Entorno real del conjunto de datos

En el Cuadro 6 se presentan los hiperparámetros seleccionados para la clasificación de mezcal elaborado con *A. angustifolia* para cada uno de los algoritmos analizados y en la Figura 1, se presentan los resultados de clasificación para dicha variedad. En este análisis, el modelo SVM alcanzó un f1-score de 84 % empleando sus parámetros por defecto, superando el desempeño obtenido tras el ajuste mediante búsqueda en cuadrícula. Este comportamiento indica que, para este conjunto de datos, SVM ofrece un mejor rendimiento sin necesidad de modificaciones adicionales. En contraste, estudios aplicados en matrices químicas similares han mostrado que el algoritmo SVM puede alcanzar desempeños elevados incluso utilizando configuraciones estándar. Por ejemplo, en la clasificación de vinos de frutas mediante variables químicas derivadas de GC–MS y GC–IMS, Zhou, Yu, Ai, et al. (2025) reportaron un f1-score de 85.97 %, sin requerir ajustes de hiperparámetros. Por otro lado, los algoritmos NB y MLP sí mostraron un desempeño superior tras el ajuste de hiperparámetros. En particular, el Perceptrón Multicapa (MLP) obtuvo la mejor clasificación para *A. angustifolia*, alcanzando un f1-score de 89 %, lo que refleja su capacidad para modelar relaciones no lineales entre las variables fisicoquímicas. Este comportamiento es coherente con lo observado en estudios previos sobre

bebidas alcohólicas, como el de Zhou, Yu, Zhu, et al. (2025), quienes aplicaron redes neuronales para clasificar distintos grados de calidad de *baijiu* a partir de perfiles químicos obtenidos por GC-TOF/MS y GC-IMS. En su trabajo, las redes neuronales lograron un f1-score de 86 %, métricas comparables a las obtenidas en este estudio. La similitud en el rendimiento sugiere que, aun cuando el baijiu y el mezcal presentan matrices químicas y contextos productivos distintos, los modelos neuronales comparten una notable capacidad para capturar patrones estructurales relevantes en sistemas complejos basados en perfiles químicos multivariados. Esto refuerza la pertinencia del MLP dentro del conjunto de modelos evaluados y confirma su utilidad en tareas de clasificación basadas en datos fisicoquímicos de bebidas destiladas.

Cuadro 6. Valores de los hiperparámetros ajustados al aplicar Búsqueda en cuadrícula en la clasificación de mezcal elaborado con *A. angustifolia*.

Algoritmo	Hiperparámetros
SVM	C: 10; kernel: Rbf; Gamma: 0.1
NB	Priors: 0.5,0.5; Var_smoothing: 1
KNN	n_neighbors: 11
DT	Criterion: Gini; Max_depth: 3; min_samples_split: 2; min_samples_leaf: 10; splitter: best
RF	N_estimate: 50; Criterion: Entropy; Max_depth: 10; Min_samples_split: 10; Min_samples_leaf: 4
MLP	Learnig_rate: 0.001; Hidden_units_1: 12; Hidden_units_2: 6

Dado que *A. angustifolia* representa cerca del 85 % de la producción total de mezcal en Oaxaca (COMERCAM, 2025), la correcta clasificación de esta especie adquiere especial relevancia. Su predominio en la producción implica que los modelos desarrollados podrían aplicarse de manera efectiva a una proporción significativa del mezcal elaborado en la región, contribuyendo a la identificación objetiva de una de las variedades más representativas del sector y que además, al tratarse de la especie con menor costo comercial, su correcta clasificación permitiría detectar muestras que, al declarar pertenecer a especies de mayor valor, podrían estar incurriendo en prácticas de etiquetado incorrecto y en

declaraciones que no corresponden con la especie realmente utilizada. Por tanto, el desempeño obtenido demuestra que es posible reconocer patrones químicos consistentes asociados a *A. angustifolia*, lo que resalta el potencial del aprendizaje automático como herramienta para la clasificación sistemática del mezcal artesanal.

En esta etapa no se reportan los modelos entrenados para *A. karwinskii*, *A. marmorata* y *A. potatorum* utilizando el conjunto completo, ya que en los tres casos el f1-score obtenido fue inferior al 50 %, valor que en una tarea de clasificación binaria equivale a un rendimiento comparable al obtenido por azar (Van De Bijl et al., 2024). Siguiendo dicho criterio, para las configuraciones del conjunto de datos siguientes, se presentan únicamente los modelos que superaron de manera consistente dicho umbral, garantizando así que los resultados analizados reflejen una capacidad real de clasificación y no un desempeño atribuible al desequilibrio extremo entre clases.

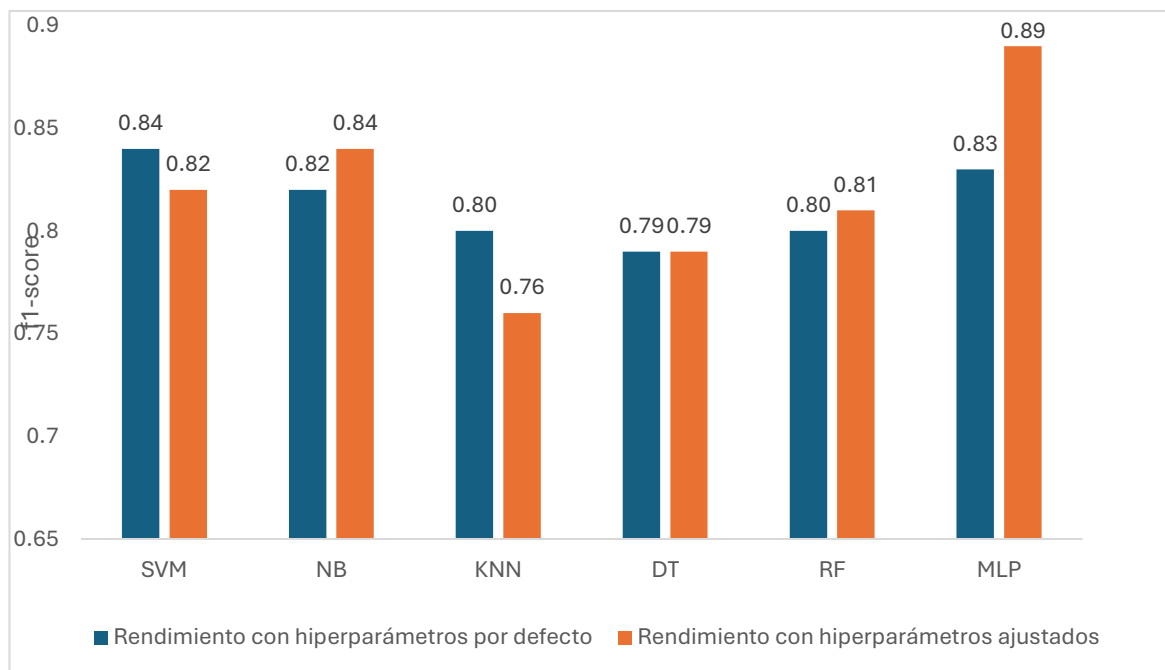


Figura 1. Desempeño de los algoritmos de clasificación aplicados a muestras de *A. angustifolia* con parámetros por defecto y ajustados.

Empleo de SMOTE y RUS

Las pruebas realizadas con SMOTE y RUS mostraron que los algoritmos presentaron un rendimiento limitado en las especies con menor representación. En la mayoría de los casos, las métricas de precisión, recall y f1-score fueron insuficientes para sustentar una evaluación sólida del desempeño. Debido a ello, y dado que estos valores no aportan evidencia concluyente para la selección de modelos, los resultados detallados de estas pruebas no se incluyeron.

Para este análisis también se consideraron las métricas de precisión y recall, dado que las técnicas de sobremuestreo y submuestreo generan contextos artificialmente más balanceados en los que estas métricas suelen ofrecer una interpretación más clara del comportamiento del modelo (Bagui & Li, 2021). Sin embargo, incluso bajo estas condiciones más equilibradas, el desempeño obtenido no fue relevante.

En conjunto, estos resultados indican que las técnicas de sobre y submuestreo no mejoraron la capacidad discriminante de los modelos en el contexto de este estudio.

Eliminación de la clase mayoritaria *A. angustifolia*

Al eliminar las instancias de *A. angustifolia*, se evaluó la capacidad de los modelos para clasificar mezcales producidos con especies no dominantes. En este contexto reducido, el Cuadro 7 presenta los hiperparámetros seleccionados para la clasificación de las muestras elaboradas con *A. karwinskii*. Los resultados del desempeño de los modelos se muestran en la Figura 2, donde se incluyen únicamente aquellos algoritmos que alcanzaron un f1-score superior al 50 %, siguiendo el mismo criterio adoptado previamente para asegurar que los modelos reportados presenten un rendimiento mayor al obtenido por azar en una clasificación binaria. Entre los modelos evaluados, el algoritmo SVM con ajuste presentó el mejor desempeño, con un f1-score de 77 %, a pesar de presentar una tasa de desbalance en los datos de 2.53 (véase Cuadro 4). Este resultado es relevante, ya que el desbalance de clases suele afectar negativamente el

rendimiento de los clasificadores al sesgar la frontera de decisión hacia la clase mayoritaria (Cervantes et al., 2017; He & Garcia, 2009). Sin embargo, mantener la distribución original de los datos puede resultar más apropiado en estudios exploratorios, pues preserva la representatividad y evita el riesgo de introducir ruido o sobreajuste asociado a técnicas de remuestreo (Batista et al., 2004).

Cuadro 7. Valores de los hiperparámetros ajustados al aplicar Búsqueda en cuadrícula en la clasificación de mezcal elaborado con *A. karwinskii*.

Algoritmo	Hiperparámetros
SVM	C: 100; kernel: Rbf; Gamma: scale
NB	Priors:0.5,0.5; Var_smoothing: 10^{-9}
DT	Criterion: Gini; Max_depth: 3; min_samples_split: 2; min_samples_leaf: 10; splitter: best
RF	N_estimate: 10; Criterion: Gini; Max_depth: None; Min_samples_split: 5; Min_samples_leaf: 1

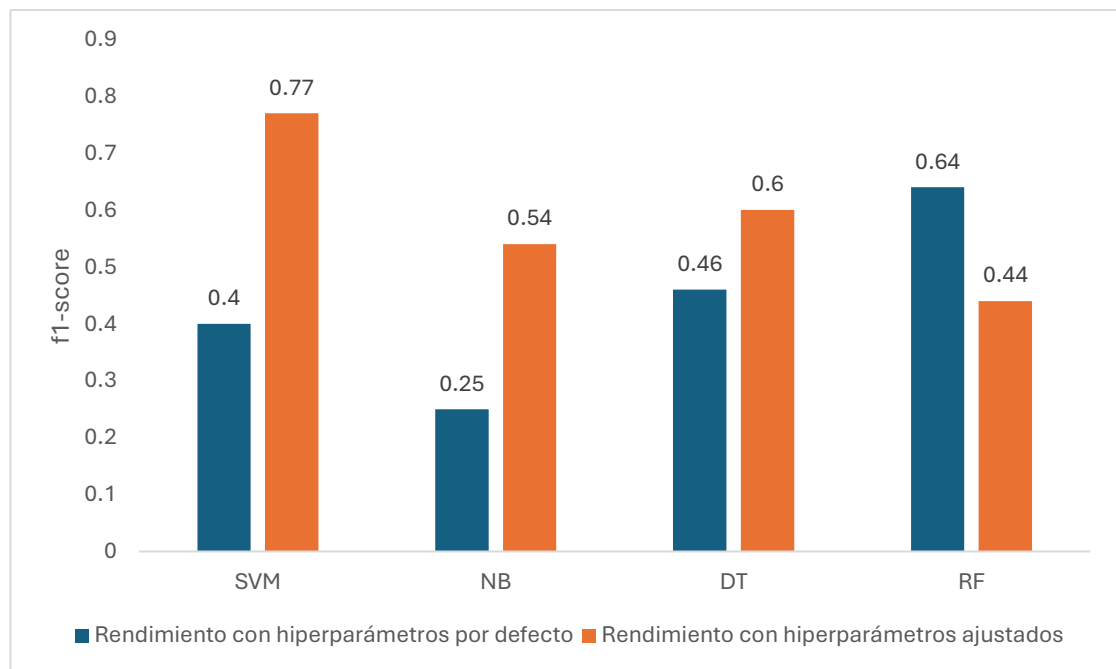


Figura 2. Clasificación para las muestras elaboradas con *A. karwinskii*.

En este sentido, el desempeño obtenido por SVM puede considerarse satisfactorio, dado que refleja la capacidad del modelo para identificar patrones reales bajo condiciones de desbalance moderado, sin recurrir a manipulación

artificial de los datos. Este enfoque contribuye a una evaluación más realista de la aplicabilidad del modelo en contextos de certificación, donde el desbalance entre especies de agave es un fenómeno intrínseco al mercado del mezcal.

En contraste, estudios como el de Andrade et al. (2017) reportaron precisiones cercanas al 99 % en la clasificación de tequila utilizando SVM sobre conjuntos de datos completamente balanceados, lo cual evidencia el efecto positivo que puede tener un balance adecuado en el rendimiento de los modelos. En el caso de *A. karwinskii*, aun después de aplicar técnicas de balanceo como el sobremuestreo sintético y el submuestreo aleatorio, no se observaron mejoras sustanciales en el desempeño. Esto sugiere que, más allá del desbalance, la propia variabilidad química de la especie y la limitada disponibilidad de muestras influyen en la dificultad de su clasificación. Bajo estas condiciones reales, un f1-score de 77 % sigue siendo un resultado consistente y relevante, que refleja la capacidad del modelo para identificar patrones distintivos.

Clasificación de las clases minoritarias *A. marmorata* y *A. potatorum*.

En el Escenario A, correspondiente a la eliminación de *A. angustifolia*, las especies *A. marmorata* y *A. potatorum* no alcanzaron f1-scores superiores a 50 %, lo que indica que la remoción de la clase mayoritaria no fue suficiente para mejorar su capacidad de clasificación. Este desempeño se relaciona con las elevadas tasas de desbalance que caracterizan a estas especies (5.60 y 4.42, respectivamente; véase Cuadro 4), condiciones que limitan la habilidad de los algoritmos para identificar patrones químicos distintivos aun cuando la clase dominante ha sido excluida.

Con el propósito de explorar si un entorno aún más equilibrado podía favorecer su clasificación, en el Escenario B se excluyeron también las muestras de *A. karwinskii*, segunda especie con mayor representación y previamente identificada como clasificable. Este ajuste permitió centrar el análisis en las especies minoritarias restantes y evaluar su separabilidad en ausencia de clases influyentes. Sin embargo, aún bajo estas condiciones más equilibradas, el comportamiento de los modelos continuó siendo limitado, en el Cuadro 8 se

muestran los resultados de clasificación obtenidos. Este resultado sugiere que, más allá del desbalance, las señales químicas de *A. marmorata* y *A. potatorum* son menos marcadas o más complejas de capturar mediante los algoritmos empleados, lo que dificulta su discriminación incluso en escenarios controlados.

Cuadro 8. f1-score (%) en la clasificación de especies minoritarias con modelos binarios.

Algoritmo	<i>A. Marmorata</i>		<i>A. potatorum</i>	
	NM ^x	GS ^y	NM	GS
KNN	0	40	28.57	0.29
NB	42.86	0	25	28.57
SVM	0	33	0	33
DT	28.57	29	22.22	50
RF	33.33	57	44	33
MLP	0	33	0	33

^xAlgoritmos sin ajuste (hiperparámetros por defecto)

^yAlgoritmos con ajuste por medio de Búsqueda en Cuadrícula.

Si bien los modelos binarios permiten analizar el comportamiento predictivo de cada especie de manera individual y aportan información valiosa sobre su separabilidad química, su desempeño se ve limitado cuando algunas clases presentan una representación muy reducida. Siendo el caso de *A. marmorata* y *A. potatorum*, para las cuales los modelos binarios no alcanzaron valores de f1-scores superiores a 50 %, incluso después de modificar la composición del conjunto de datos. Ante esta limitación, se decidió ampliar el análisis incorporando modelos multiclase, con el propósito de evaluar de manera simultánea la capacidad de los algoritmos para discriminar entre especies minoritarias y un grupo de muestras que no pertenecen a ellas, y así explorar patrones de separación que no fueron captados bajo el esquema estrictamente binario (Gantert et al., 2024). Para ello, se construyó un escenario multiclase en el que *A. marmorata* y *A. potatorum* se consideraron clases explícitas, y se añadió una categoría adicional denominada “No pertenece”, integrada por las muestras que no correspondían a *A. angustifolia* ni *A. karwinskii* (ya modeladas en las

etapas previas) ni a las especies minoritarias en evaluación. Esta clase incluyó mezcales elaborados con especies como *A. atrovirens*, *A. rodhocanta*, *A. convallis*, *A. cupreata* y *A. salmiana*, que en los modelos binarios habían sido agrupadas dentro de la clase negativa debido a su escaso número de instancias. De este modo, el enfoque multiclase no sustituyó al esquema binario, sino que lo complementó, al ofrecer una perspectiva conjunta sobre la estructura del espacio de características y permitir evaluar si, bajo una formulación distinta del problema, era posible mejorar la clasificación de las especies minoritarias. Los resultados obtenidos tras este cambio de enfoque se presentan en el Cuadro 9.

Con lo anterior, fue posible clasificar las muestras de mezcal elaboradas con *A. marmorata* mediante los algoritmos DT sin ajuste y SVM con ajuste, alcanzando un f1-score de 75 %. Por otro lado, para las muestras elaboradas con *A. potatorum*, el mejor rendimiento alcanzado fue solo un f1-score de 50 %, lo cual se consideró insuficiente para una clasificación confiable.

Cuadro 9. f1-score (%) en la clasificación de especies minoritarias con modelos multiclase.

	NM ^x	NM	NM	GS ^y	GS	GS
	Tepeztate	Potatorum	No es	Tepeztate	Potatorum	No es
SVM	40	33.33	63.16	75	44.44	46.15
DT	75	50	57.14	50	50	20

^xAlgoritmos sin ajuste (hiperparámetros por defecto)

^yAlgoritmos con ajuste por medio de Búsqueda en Cuadrícula.

La clasificación obtenida para los mezcales elaborados con *A. potatorum* sugieren que la señal discriminativa de esta especie es tenue dentro del espacio de características utilizado, lo que dificulta su separación respecto a otras clases, en ese sentido, se ha documentado que en escenarios de desbalance o con alta variabilidad intraclase, algunas categorías tienden a presentar bajos valores de precisión, recall y, en consecuencia, de f1-score, incluso utilizando modelos robustos de clasificación (Johnson & Khoshgoftaar, 2019).

Finalmente, la figura 3 sintetiza el flujo de clasificación propuesto a partir de los modelos con mejor desempeño obtenidos en este estudio. Este diagrama

representa la culminación de todo el proceso metodológico aplicado al conjunto de datos y refleja, de manera operativa, cómo podrían utilizarse los modelos entrenados para clasificar una muestra desconocida de mezcal en función de la especie de agave empleada en su elaboración. Dicho flujo es un sistema secuencial donde cada especie es evaluada únicamente por los modelos que demostraron la mayor capacidad de clasificación. Este enfoque evita la aplicación indiscriminada de todos los modelos a todas las muestras y favorece un proceso más eficiente.

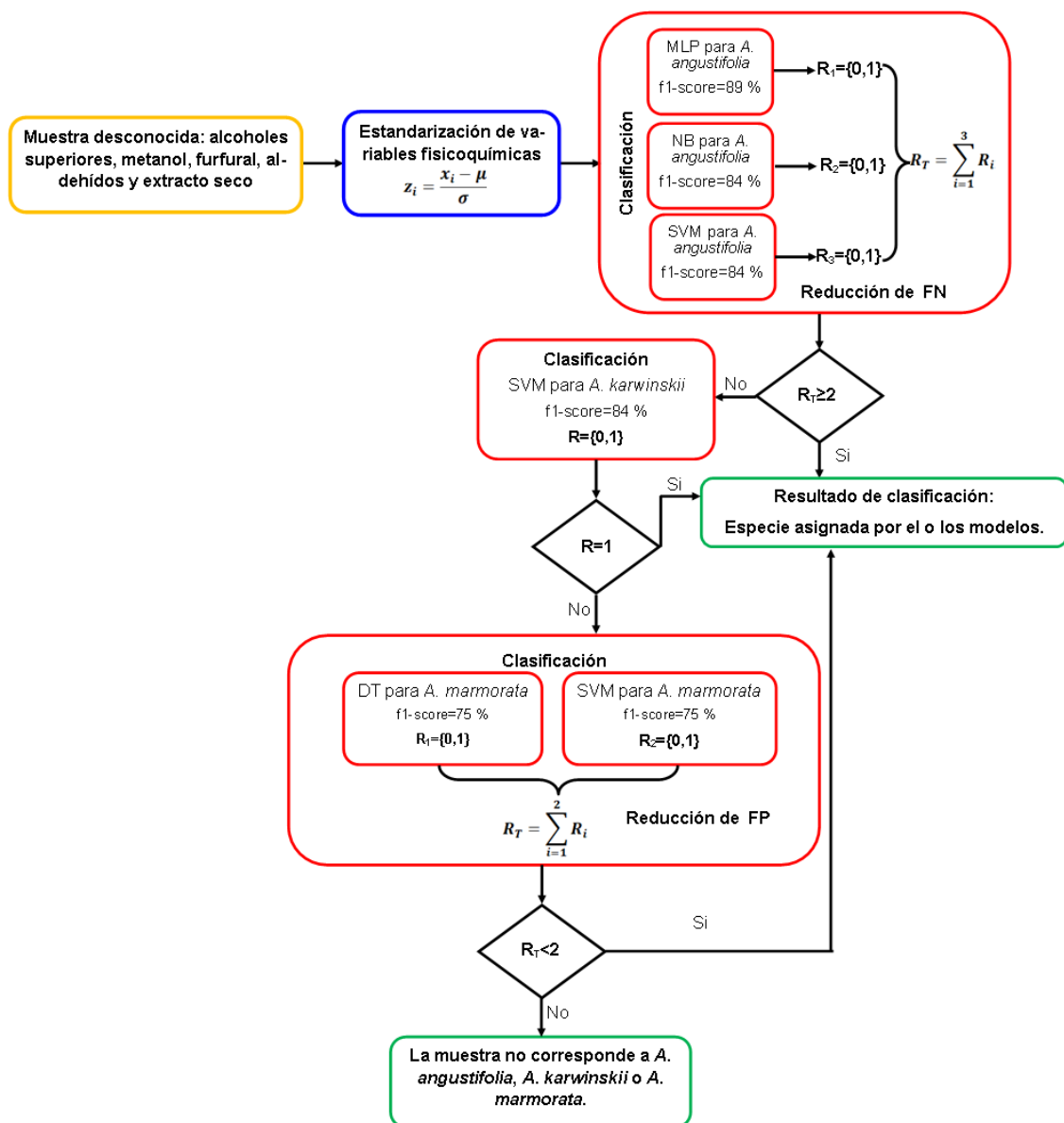


Figura 3. Flujo de clasificación propuesto para muestras desconocidas.

Para llegar a este esquema final fue necesario recorrer varias etapas analíticas. En primera instancia, se evaluaron seis algoritmos de clasificación bajo un enfoque binario Uno contra todos, explorando su comportamiento en el conjunto de datos original y, posteriormente, bajo escenarios de desbalance moderado, severo y balanceado. Esto implicó entrenar y comparar más de un millar de modelos resultantes de la combinación entre algoritmos, hiperparámetros y configuraciones del conjunto de datos. Adicionalmente, se probaron técnicas de sobremuestreo (SMOTE) y submuestreo (RUS), así como dos escenarios alternativos basados en la eliminación de clases mayoritarias. Solo después de verificar que estas estrategias no aportaban mejoras consistentes, se optó por mantener la distribución real de los datos para las especies con señales distinguibles, y crear un escenario multiclase para explorar la separabilidad de las especies minoritarias.

Los modelos seleccionados en cada etapa del flujo se definieron a partir de un análisis exhaustivo orientado a identificar cuáles algoritmos ofrecían un desempeño suficiente y estable para su aplicación operativa. En el caso de *A. angustifolia*, que presentó los niveles más altos de discriminación, se estableció un esquema de votación mayoritaria entre los modelos SVM, NB y el MLP. Este diseño prioriza la reducción de falsos negativos, ya que se trata de la especie más común y de menor valor económico; por lo tanto, evita que un mezcal de *A. angustifolia* sea clasificado como una especie silvestre previniendo interpretaciones erróneas o sobrevaloraciones indebidas. Para *A. karwinskii* se seleccionó el modelo SVM por su consistencia una vez eliminadas las instancias de *A. angustifolia* del conjunto de datos, lo que permitió mejorar la estabilidad de la clasificación. Finalmente, para *A. marmorata*, especie silvestre de mayor valor y con un patrón químico más sutil, se eligieron los modelos SVM y RF. En este caso el diseño priorizó la reducción de falsos positivos, ya que clasificar por error un mezcal económico como si fuera de *A. marmorata* tendría repercusiones directas en la autenticidad y el valor comercial; por ello se integraron ambos modelos para reforzar la confiabilidad en la decisión final.

Desde una perspectiva práctica, este flujo de clasificación puede considerarse como una herramienta complementaria para los OEC. Que si bien no busca reemplazar al proceso de certificación establecido por la NOM-070-SCFI-2016, tiene el potencial de funcionar como un sistema adicional de verificación que ayude a identificar inconsistencias. Esto resulta especialmente relevante para diferenciar mezcales elaborados con especies cultivadas como *A. angustifolia*, cuyo costo es menor, de los producidos con especies de mayor valor comercial como *A. karwinskii* y *A. marmorata*.

Su utilidad potencial reside en que no requiere análisis adicionales, no modifica los procedimientos actuales y puede implementarse como un filtro automatizado adicional al análisis normativo, proporcionando así un nivel adicional de certeza en la identificación de la especie utilizada, y abriendo la posibilidad de que, en un futuro, estas técnicas puedan integrarse de manera formal como herramientas complementarias dentro del marco normativo vigente.

3.3.3. Análisis CATA

Como complemento de la clasificación, se aplicó la metodología CATA y se agruparon a las muestras por medio de los atributos sensoriales que los caracterizaron. En el cuadro 10 se muestran las frecuencias de uso para cada uno de los atributos y/o términos utilizados en la pregunta CATA, donde 17 de los 26 atributos empleados fueron significativos ($p \leq 0.05$) para describir las diferencias entre las muestras de mezcal con base en la prueba Q de Cochran. Las mayores frecuencias relativas de mención la tuvieron los términos “olor a acetona”, “olor a plástico”, “olor a esmalte de uñas”, “aroma a alcohol”, “aroma dulce”, “aroma a agave cocido”, “sabor herbal”, “sabor a solvente”, “sabor dulce”, “astringente”, “desagrado” y “felicidad”, siendo todos ellos significativos ($p \leq 0.05$) para describir las diferencias entre las muestras, junto con los siguientes atributos “aroma a hierba”, “aroma frutal”, “aroma a madera”, “aroma a pan”, “sabor a plástico”, “sabor a maguey cocido” y “sabor frutal”.

Los mezcales pudieron dividirse en dos grupos de acuerdo con los atributos sensoriales. Por un lado, el mezcal elaborado con *A. marmorata* se diferenció

significativamente del resto por presentar mayores frecuencias de mención en los atributos “olor a acetona”, “olor a plástico”, “olor a esmalte”, “sabor a solvente”, “sabor a plástico” y “desagradable”.

Esta tendencia puede vincularse con la presencia de alcoholes superiores, compuestos cuya variabilidad entre especies resultó estadísticamente significativa mediante la prueba de Kruskal-Wallis y que, según lo reportado en la literatura, contribuyen directamente a la percepción de notas alcohólicas, punzantes y de solvente cuando se encuentran en concentraciones elevadas, efecto que se observó en dichas muestras con respecto a las demás, pero sin exceder los límites marcados por la NOM-070-SCFI-2016 (Pino et al., 2012). Por lo tanto, conocer la composición específica y las proporciones relativas de estos compuestos podría aportar información valiosa sobre la expresión sensorial del mezcal y su relación con la especie de agave utilizada en la producción.

Por otro lado, los mezcales elaborados con *A. angustifolia*, *A. karwinskii* y *A. potatorum* se distinguieron significativamente de *A. marmorata* por presentar mayores frecuencias de mención en los atributos “aroma frutal”, “aroma dulce” y “sabor dulce”. Estas características sensoriales pueden vincularse con la presencia de alcoholes superiores en concentraciones moderadas, los cuales, a diferencia de niveles elevados, contribuyen a una percepción más equilibrada y agradable de la bebida. Diversos estudios han señalado que estos compuestos participan activamente en la generación de notas frutales, dulces y herbales propias de los destilados (Cameleyre et al., 2015; Cordente et al., 2021; Huang & Cadwallader, 2023; Tang et al., 2024).

En la Figura 4, se muestran los términos significativos y las muestras, mediante el Análisis de Correspondencia (AC), mismo que explicó el 97.38 % de la variación de los datos en las primeras dos dimensiones, 93.41 % y 3.97 %, respectivamente. En el trabajo de Veríssimo et al. (2021) al caracterizar vinos tintos tropicales por medio de la metodología CATA se obtuvieron resultados inferiores a los reportados en este trabajo, cuyas primeras dos dimensiones del AC explicaron el 47.52 % y 33.29 %, respectivamente. El mezcal elaborado con

A. marmorata estuvo caracterizado por los atributos “olor a esmalte de uñas”, “olor a plástico”, “olor a acetona”, “sabor a plástico” y “sabor a solvente”. El mezcal elaborado con *A. karwinskii* estuvo caracterizado por los atributos “astringente”, “sabor herbal” y “aroma a alcohol”. Finalmente, los mezcales elaborados con *A. angustifolia* y *A. potatorum* estuvieron caracterizados por los atributos “aroma a alcohol”, “sabor a maguey cocido”, “aroma a agave cocido”, “aroma dulce”, “sabor herbal” y “sabor dulce”.

Cuadro 10. Frecuencia en que los términos de las preguntas CATA fueron usadas por los consumidores (n = 100) para describir las cuatro muestras de mezcal artesanal y los resultados de la prueba de Q de Cochran para la comparación entre muestras.

Atributos	valores-p	<i>A. marmorata</i>	<i>A. karwinskii</i>	<i>A. potatorum</i>	<i>A. angustifolia</i>
Olor a acetona	0.000	0.750 (b ²)	0.370 (a)	0.290 (a)	0.350 (a)
Olor a plástico	0.000	0.660 (c)	0.370 (b)	0.230 (ab)	0.180 (a)
Olor a esmalte de uñas	0.000	0.750 (c)	0.530 (b)	0.180 (a)	0.250 (a)
Aroma a hierba	0.001	0.220 (a)	0.460 (b)	0.420 (b)	0.370 (ab)
Aroma a alcohol	0.046	0.720 (a)	0.740 (a)	0.850 (a)	0.830 (a)
Aroma frutal	0.000	0.110 (a)	0.340 (b)	0.380 (b)	0.400 (b)
Aroma dulce	0.000	0.280 (a)	0.520 (b)	0.640 (b)	0.590 (b)
Aroma a madera	0.023	0.290 (a)	0.480 (a)	0.330 (a)	0.350 (a)
Aroma a agave cocido	0.001	0.260 (a)	0.480 (b)	0.520 (b)	0.430 (ab)
Aroma ahumado	0.609	0.330 (a)	0.410 (a)	0.370 (a)	0.390 (a)
Aroma a maguey verde	0.176	0.230 (a)	0.360 (a)	0.310 (a)	0.320 (a)
Aroma cítrico	0.523	0.280 (a)	0.300 (a)	0.350 (a)	0.340 (a)
Aroma a vainilla	0.056	0.040 (a)	0.130 (a)	0.150 (a)	0.110 (a)
Aroma a pan	0.022	0.010 (a)	0.070 (a)	0.090 (a)	0.100 (a)
Sabor ahumado	0.129	0.390 (a)	0.490 (a)	0.510 (a)	0.530 (a)
Sabor herbal	0.001	0.360 (a)	0.610 (b)	0.460 (ab)	0.520 (ab)
Sabor a solvente	0.000	0.610 (b)	0.430 (a)	0.300 (a)	0.290 (a)
Sabor dulce	0.000	0.210 (a)	0.480 (b)	0.480 (b)	0.530 (b)
Sabor ahumado	0.184	0.400 (a)	0.490 (a)	0.530 (a)	0.510 (a)
Sabor a plástico	0.000	0.630 (c)	0.360 (b)	0.220 (ab)	0.120 (a)
sabor ácido	0.280	0.520 (a)	0.410 (a)	0.460 (a)	0.450 (a)
Sabor a maguey cocido	0.020	0.280 (a)	0.410 (a)	0.460 (a)	0.440 (a)
Sabor frutal	0.000	0.110 (a)	0.300 (b)	0.270 (ab)	0.370 (b)
Sabor a madera	0.056	0.370 (a)	0.510 (a)	0.400 (a)	0.350 (a)
Picante	0.760	0.460 (a)	0.490 (a)	0.500 (a)	0.440 (a)
Astringente	0.017	0.620 (a)	0.440 (a)	0.460 (a)	0.440 (a)

²Proporciones con distinta letra en la misma fila, son estadísticamente diferentes ($p \leq 0.05$).

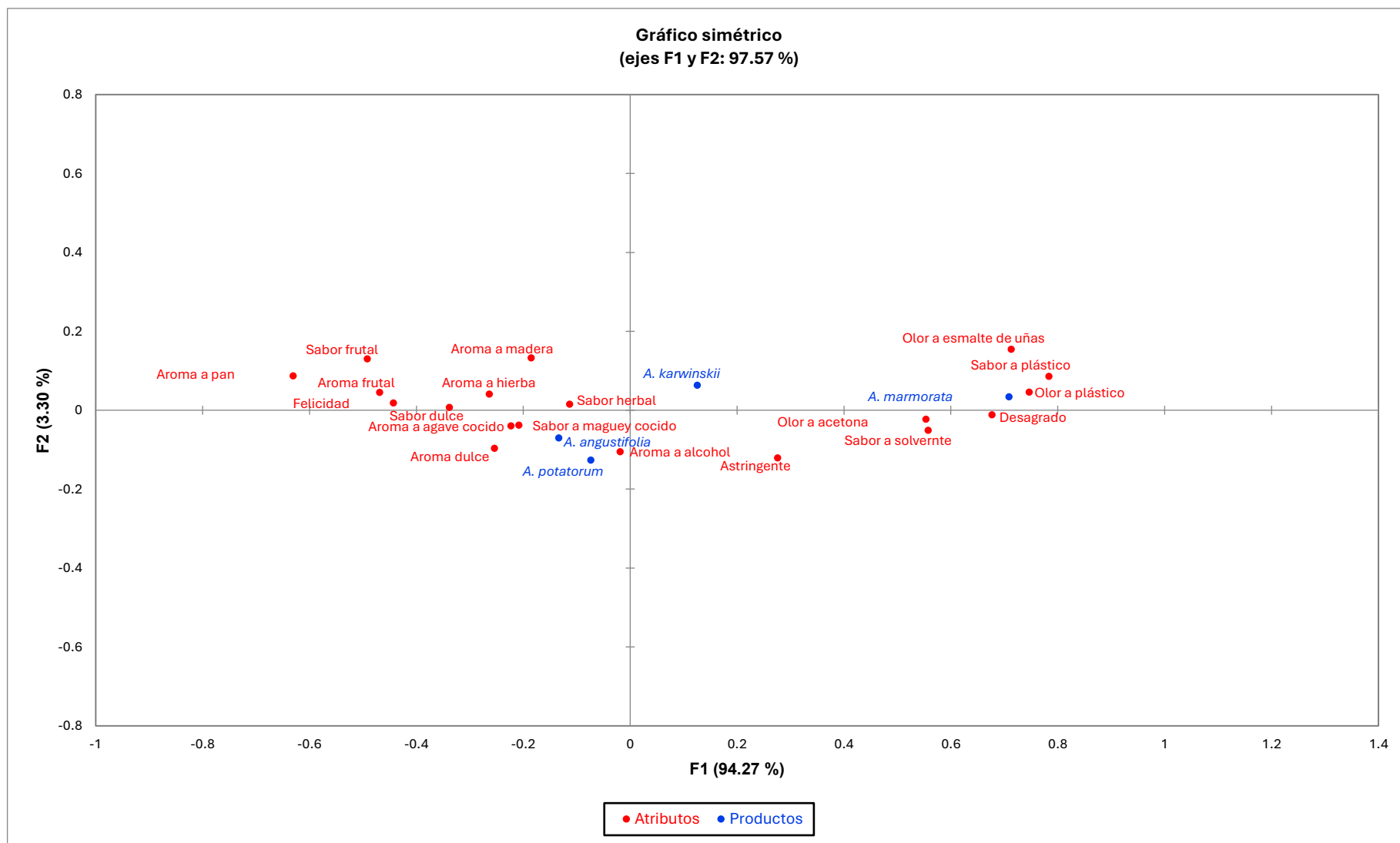


Figura 4. Representación de los atributos (●) y muestras de mezcal (●) en las primeras dos dimensiones del análisis de correspondencia.

3.4. CONCLUSIONES

Los resultados obtenidos demostraron que el perfil fisicoquímico del mezcal puede ser utilizado de manera efectiva para alimentar modelos de clasificación capaces de distinguir entre las especies de agave empleadas en su elaboración. Entre las variables analizadas, los alcoholes superiores y el metanol presentaron diferencias estadísticamente significativas entre especies, evidenciando que la composición química del producto está directamente influenciada por la materia prima y las condiciones fermentativas particulares de cada variedad de *Agave*. Estos hallazgos respaldan el uso de dichos compuestos como indicadores de variabilidad botánica en el proceso de elaboración de mezcal.

Los modelos de aprendizaje automático mostraron que, pese al tamaño limitado del conjunto de datos, los algoritmos lograron identificar relaciones consistentes entre las variables fisicoquímicas que distinguen a cada especie de agave. En particular, el MLP alcanzó un f1-score de 89 % al clasificar el mezcal elaborado con *A. angustifolia*, mientras que los modelos SVM y NB 84 %. Este resultado reviste especial importancia, ya que *A. angustifolia* representa alrededor del 85 % de la producción total de mezcal en Oaxaca. Además, al tratarse de una especie cultivada y de menor costo comercial en comparación con especies silvestres como *A. karwinskii* o *A. marmorata*, su adecuada clasificación resulta fundamental para identificar casos en los que un mezcal elaborado con *A. angustifolia* pudiera declararse como proveniente de una especie silvestre de mayor valor. De esta manera, los modelos contribuyen a prevenir prácticas de etiquetado incorrecto y a fortalecer la autenticidad del producto que llega al consumidor final.

Para el mezcal producido con *A. karwinskii*, el modelo SVM presentó un f1-score de 77 %, y para las muestras de *A. marmorata*, tanto SVM como RF alcanzaron un 75 %. Estos resultados evidencian el potencial de estos algoritmos para clasificar mezcales artesanales por especie de maguey utilizando únicamente datos fisicoquímicos rutinarios.

De manera complementaria, la metodología CATA corroboró que las diferencias químicas identificadas por los modelos se reflejan en la percepción del consumidor, al asociarse atributos sensoriales específicos con cada especie de maguey. Esta coincidencia entre los patrones fisicoquímicos y perceptuales refuerza la idea de que la diversidad biológica del agave se expresa tanto en la composición química del destilado como en su perfil sensorial.

En conjunto, los resultados muestran que los algoritmos de aprendizaje automático pueden aprovechar de manera efectiva los datos fisicoquímicos generados durante la certificación para distinguir entre especies de agave empleadas en la elaboración del mezcal. Este enfoque contribuye al fortalecimiento del conocimiento en torno a la relación entre la diversidad biológica del maguey y las características del mezcal, y sienta las bases para el desarrollo de metodologías de análisis más integrales que consideren, de forma conjunta, aspectos químicos, tecnológicos y sensoriales del producto.

3.5. Literatura citada

- Andrade, J. M., Ballabio, D., Gómez-Carracedo, M. P., & Pérez-Caballero, G. (2017). Nonlinear classification of commercial Mexican tequilas. *Journal of Chemometrics*, 31(12), e2939. <https://doi.org/10.1002/cem.2939>
- Bagui, S., & Li, K. (2021). Resampling imbalanced data for network intrusion detection datasets. *Journal of Big Data*, 8(1), 6. <https://doi.org/10.1186/s40537-020-00390-x>
- Batista, G. E. A. P. A., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, 6(1), 20–29. <https://doi.org/10.1145/1007730.1007735>
- Buda, M., Maki, A., & Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106, 249–259. <https://doi.org/10.1016/j.neunet.2018.07.011>

- Cameleyre, M., Lytra, G., Tempere, S., & Barbe, J. C. (2015). Olfactory impact of higher alcohols on red wine fruity ester aroma expression in model solution. *Journal of Agricultural and Food Chemistry*, 63(44), 9768–9778. <https://doi.org/10.1021/acs.jafc.5b03489>
- Cervantes, J., Garcia-Lamont, F., Rodriguez, L., López, A., Castilla, J. R., & Trueba, A. (2017). PSO-based method for SVM classification on skewed data sets. *Neurocomputing*, 228, 187–197. <https://doi.org/10.1016/j.neucom.2016.10.041>
- COMERCAM. (2025). *Informe público 2025*.
- Cordente, A. G., Nandorfy, D. E., Solomon, M., Schulkin, A., Kolouchova, R., Francis, I. L., & Schmidt, S. A. (2021). Aromatic higher alcohols in wine: Implication on aroma and palate attributes during Chardonnay aging. *Molecules*, 26(16), 4979. <https://doi.org/10.3390/molecules26164979>
- Elgeldawi, E., Sayed, A., Galal, A. R., & Zaki, A. M. (2021). Hyperparameter tuning for machine learning algorithms used for Arabic sentiment analysis. *Informatics*, 8(4), 79. <https://doi.org/10.3390/informatics8040079>
- Galmarini, M. V., & Visalli, M. (2024). Sensory research and temporal descriptive methods: Where is this relationship going? *Science Talks*, 11, 100375. <https://doi.org/10.1016/j.sctalk.2024.100375>
- Gantert, L., Zeffiro, T., Sammarco, M., & Campista, M. E. M. (2024). Multiclass classification of faulty industrial machinery using sound samples. *Engineering Applications of Artificial Intelligence*, 136, 108943. <https://doi.org/10.1016/j.engappai.2024.108943>
- Hasanin, T., Khoshgoftaar, T. M., Leevy, J. L., & Bauder, R. A. (2019). Severely imbalanced Big Data challenges: Investigating data sampling approaches. *Journal of Big Data*, 6(1), 85. <https://doi.org/10.1186/s40537-019-0274-4>

- He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263–1284. <https://doi.org/10.1109/tkde.2008.239>
- Huang, X., & Cadwallader, K. R. (2023). A critical review of the flavor chemistry of tequila. In *ACS Symposium Series* (Vol. 1455, pp. 1–36). American Chemical Society. <https://doi.org/10.1021/bk-2023-1455.ch001>
- Jain, A. K., Duin, R. P. W., & Mao, J. (2000). Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1), 4–37. <https://doi.org/10.1109/34.824819>
- Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *Journal of Big Data*, 6, 27. <https://doi.org/10.1186/s40537-019-0192-5>
- Lopez, M. G., Mancilla-Margalli, N. A., & Mendoza-Díaz, G. (2003). Molecular structures of fructans from *Agave tequilana* Weber var. azul. *Journal of Agricultural and Food Chemistry*, 51(27), 7835–7841. <https://doi.org/10.1021/jf030383v>
- McKean, J. W., & Manoukian, E. B. (1988). Mathematical nonparametric statistics. *Journal of the American Statistical Association*, 83(404), 934–935. <https://doi.org/10.2307/2290166>
- Mukherjee, M., & Khushi, M. (2021). SMOTE-ENC: A novel SMOTE-based method to generate synthetic data for nominal and continuous features. *Applied System Innovation*, 4(1), 18. <https://doi.org/10.3390/asi4010018>
- Nougarède, S., Diot, A., Maza, E., Samson, A., Olivier-Salvagnac, V., Caillé, S., Geffroy, O., & Chervin, C. (2023). Comparison of Check-All-That-Apply and Adapted-Pivot-Test methods for wine sensory characterization with a panel of untrained students. *Journal of Sensory Studies*, 38(5), e12862. <https://doi.org/10.1111/joss.12862>

- Ostertagová, E., Ostertag, O., & Kováč, J. (2014). Methodology and application of the Kruskal–Wallis test. *Applied Mechanics and Materials*, 611, 115–120. <https://doi.org/10.4028/www.scientific.net/amm.611.115>
- Pérez, E., González-Hernández, J. C., Chávez-Parga, M. C., & Cortés-Penagos, C. (2013). Caracterización fermentativa de levaduras productoras de etanol a partir de jugo de *Agave cupreata* en la elaboración de mezcal. *Revista Mexicana de Ingeniería Química*, 12(3), 543–556.
- Pino, J. A., Tolle, S., Gök, R., & Winterhalter, P. (2012). Characterisation of odour-active compounds in aged rum. *Food Chemistry*, 132(3), 1436–1441. <https://doi.org/10.1016/j.foodchem.2011.11.133>
- Quintero-Arenas, M. A., Meza-Márquez, O. G., Velázquez-Hernández, J. L., Gallardo-Velázquez, T., & Osorio-Revilla, G. (2020). Quantification of adulterants in mezcal by means of FT-MIR and FT-NIR spectroscopy coupled to multivariate analysis. *CYTA – Journal of Food*, 18(1), 157–164. <https://doi.org/10.1080/19476337.2020.1740327>
- Rožman, M., Kišić, A., & Oreški, D. (2024). Comparative analysis of nonlinear models developed using machine learning algorithms. *WSEAS Transactions on Information Science and Applications*, 21, 303–307. <https://doi.org/10.37394/23209.2024.21.29>
- Sánchez-Fernández, R. E., Pérez-López, A., Morales-Solís, A., Manilla-Téllez, Y., Reyes-Carmona, E. D., & Ávila-Uribe, G. (2025). Chemistry of mezcal: Volatile profile of artisanal mezcal made from wild agaves of Oaxaca. *Foods*, 14(7), 1222. <https://doi.org/10.3390/foods14071222>
- Swift, J. R., Turner, M. A., & Reynolds, J. C. (2023). A rapid dynamic headspace method for authentication of whiskies using artificial neural networks. *Food Chemistry Advances*, 3, 100417. <https://doi.org/10.1016/j.focha.2023.100417>
- Tang, P., Wang, L., Zhao, Q., Lu, J., Qiao, M., Li, C., Xiao, D., & Guo, X. (2024). Characterization of key aroma compounds and relationship between aroma

- compounds and sensory attributes in different quality of high-temperature Daqu. *LWT*, 194, 115801. <https://doi.org/10.1016/j.lwt.2024.115801>
- Van de Bijl, E., Klein, J., Pries, J., Bhulai, S., Hoogendoorn, M., & van der Mei, R. (2024). The Dutch Draw: Constructing a universal baseline for binary classification problems. *Journal of Applied Probability*, 1–21. <https://doi.org/10.1017/jpr.2024.52>
- Varela, P., & Ares, G. (2012). Sensory profiling: The blurred line between sensory and consumer science. A review of novel methods for product characterization. *Food Research International*, 48(2), 893–908. <https://doi.org/10.1016/j.foodres.2012.06.037>
- Vera-Guzmán, A. M., Guzmán-Gerónimo, R. I., López, M. G., & Chávez-Servia, J. L. (2018). Volatile compound profiles in mezcal spirits as influenced by agave species and production processes. *Beverages*, 4(1), 9. <https://doi.org/10.3390/beverages4010009>
- Veríssimo, C. M., de Macêdo Moraes, S., de Andrade Lima, L. L., Pereira, G. E., & Maciel, M. I. S. (2021). A short training as an enhancer of sensory ability: The case of red wine consumers. *Journal of Sensory Studies*, 36(2), e12629. <https://doi.org/10.1111/joss.12629>
- Warren-Vega, W. M., Contreras-Atrisco, Z. A., Ramírez-Quezada, M. F., & Romero-Cano, L. A. (2023). A novel approach of artificial intelligence for the study of the relation of physicochemical profile and color acquired by tequila 100% agave in its maturation process. *Journal of Food Composition and Analysis*, 123, 105533. <https://doi.org/10.1016/j.jfca.2023.105533>
- Zhou, C., Yu, Y., Ai, J., Song, C., Cui, Z., Zhou, Q., Zhao, S., Huang, R., Ao, Z., Peng, B., Chen, P., Feng, X., Li, D., & Liu, Y. (2025). Fruit wines classification enabled by combining machine learning with comprehensive volatile profiles of GC-TOF/MS and GC-IMS. *Food Research International*, 204, 115890. <https://doi.org/10.1016/j.foodres.2025.115890>

Zhou, C., Yu, Y., Zhu, Y., Chai, J., Liu, R., Cui, Z., Li, D., Feng, X., & Liu, Y. (2025). Machine learning discrimination of quality grades of base liquor integrating GC-TOF/MS and GC-IMS data analysis: Case study of strong-flavor Chinese baijiu. *Food Chemistry*, 492, 145281. <https://doi.org/10.1016/j.foodchem.2025.145281>