

TRABAJO DE GRADO

Opción Investigación o Proyecto de Grado

Desarrollo de modelos de aprendizaje automático para la evaluación de la calidad sensorial del café colombiano de especialidad

Corporación Universitaria Remington.

Facultad de Ingeniería

Especialización en Analítica de Datos.

Jaime Alexander Cuaspud Yanala.

Juan Pablo Velez Uribe.

Proyecto de grado para obtener el título de Especialista en Analítica de Datos.

2026.

"Las grandes cosas se logran mediante una serie de pequeñas cosas unidas."

- *Vincent Van Gogh*

Agradecimientos

Agradezco profundamente a la Corporación Universitaria Remington y a mi tutor, Juan Pablo Velez, por si guía durante todo este proceso académico. A la Organización Cafexport S.A.R.L., Sucursal Colombia, por abrirme las puertas al uso de la información del Laboratorio Central y confiar en este estudio de desarrollo tecnológico.

Finalmente, dedico este trabajo a mi familia y a mi complemento de vida, quienes con su amor y apoyo permitieron que este esfuerzo se materializara.

Tabla de Contenidos

Resumen.....	8
Palabras clave.....	8
1 Introducción	18
1.1 Cultivo de café	19
1.2 Clasificación taxonómica.....	21
1.3 Valor nutricional del café.....	22
1.4 Calidad física y sensorial del café.....	22
1.5 Relaciones entre calidad física y sensorial.....	25
1.6 Modelado predictivo de la calidad sensorial mediante técnicas multivariadas	26
1.7 Implicaciones para el café colombiano y perspectivas futuras	28
2 Planteamiento del problema.....	30
3 Objetivos	34
3.1 Objetivo General	34
3.2 Objetivos específicos	34
4 Metodología	37
4.1 Muestras de ensayo	37
4.2 Caracterización física del café de especialidad.....	38
4.2.1 Densidad real (g/cm ³).....	38
4.2.2 Humedad (% g/100 g).....	39
4.2.3 Tamaño de grano (% g/100 g).	39

4.2.4	Color CIE L ab*.	40
4.2.5	Grado de tueste (Unidades Agtron).	41
4.2.6	Metodología de aprendizaje automático.	42
4.2.7	Análisis de variables importantes.	46
4.2.8	Predicción de atributos sensoriales negativos.	49
5	Resultados y Discusión	50
5.1.1	Humedad (% g/100 g).	59
5.1.2	Densidad real (g/cm ³).	61
5.1.3	Distribución del tamaño del grano (tamices).	64
5.1.4	Análisis del color del café verde (espacio L ab*).	72
5.1.5	Cuantificación de defectos visuales.	76
5.1.6	Defectos de sabor.	79
5.2	Análisis de correlaciones	81
5.2.1	Análisis de correlaciones para la variable cuantificación de defectos por análisis visuales.	83
5.2.2	Análisis de correlaciones de defectos de Sabor en taza (Presencia o ausencia)	84
5.3	Interrelación entre Defectos visuales y de sabor.	85
5.4	Modelado predictivo para la variable Cuantificación de defectos por análisis visual	86
5.4.1	Optimización del modelo KNN.	86
5.4.2	Evaluación del modelo.	88
5.4.3	Evaluación de la estabilidad mediante validación cruzada.	92
5.4.4	Importancia de las variables en la predicción.	94

5.4.5	Análisis por grupos de variables. lidación cruzada (5 Folds), se presentan en la Tabla 10.	97
5.4.6	Aplicabilidad en control de calidad del café de especialidad.	98
5.5	Modelado predictivo para la variable identificación de defectos de sabor en taza	100
5.5.1	Caracterización estadística por grupos de ensayos.	101
5.5.2	Análisis de correlación y multicolinealidad.	102
5.5.3	Optimización del modelo.	103
5.5.4	Evaluación del ajuste y bondad del modelo.	105
5.5.5	Validación cruzada y estabilidad del modelo.	107
5.5.6	Jerarquización de las variables físicas y su efecto en los defectos de taza.	108
5.5.7	El modelo como filtro de riesgo (Screening).	111
6	Conclusiones	113
7	Referencias.	115

Lista de tablas

Tabla 1. Clasificación taxonómica del café arábigo. (El Cafè de Mi Tierra, 2024).	16
Tabla 2. Trazabilidad de las muestras de café de especialidad.	48
Tabla 3. Caracterización física del café de especialidad. La tabla informa los 10 primeros datos de 1324.	49
Tabla 4. Caracterización organoléptica del café de especialidad.....	52
Tabla 5. Estadísticas descriptivas para las variables físicas y sensoriales de los granos de café (n=1324).	53
Tabla 6. Evaluación de la capacidad predictiva según el número de vecinos (K).	84
Tabla 7. Evaluación del modelo KNN.	86
Tabla 8. Desempeño del modelo KNN bajo validación cruzada de 5 pliegues.	89
Tabla 9. Variables con mayor peso jerárquico.	91
Tabla 10. Desempeño predictivo desglosado por ensayos.....	94
Tabla 11. Distribución de la variable objetivo Defectos de sabor (Presencia o Ausencia).	97
Tabla 12. Características de desempeño para un umbral óptimo de 0,494.	102

Lista de figuras

Figura 1. Distribución de producción de café arábica lavado por departamento (Federación Nacional de Cafeteros de Colombia, 2025).	15
Figura 2. Diagrama de Ishikawa en la identificación de raíces del planteamiento del problema.	28
Figura 3. Espacio del color de acuerdo con la escala CIE L a*b*.....	36
Figura 4. Escala de grado de tueste en unidades Agtron.	37
Figura 5. Diagrama esquemático de la metodología de aprendizaje automático implementada y utilizada para el análisis de variables importantes y la estimación de propiedades sensoriales.....	38
Figura 6. Diagrama esquemático de aprendizaje automático mediante validación cruzada para la determinación de las variables importantes.	40
Figura 7 Diagrama metodológico de aprendizaje automático para el desarrollo de modelos de predicción sensoriales del café.	41
Figura 8 Curva ROC para un modelo de regresión logística (Brownlee, 2018).	44
Figura 9. Análisis estadístico descriptivo y evolución temporal de la humedad (%).	57
Figura 10. Análisis estadístico descriptivo y evolución temporal de la densidad real g/cm ³	59
Figura 11. Caracterización estadística y evolución temporal de la distribución granulométrica por el tamiz 18/64 in.	62
Figura 12. Caracterización estadística y evolución temporal de la distribución granulométrica por el tamiz 17/64 in.	63
Figura 13. Caracterización estadística y evolución temporal de la distribución granulométrica por el tamiz 16/64 in (2024-2026).	64
Figura 14. Caracterización estadística y evolución temporal de la distribución granulométrica por el tamiz 15/64 in.	65
Figura 15. Caracterización estadística y evolución temporal de la distribución granulométrica por el tamiz 14/64 in (2024-2026).	66

Figura 16. Caracterización estadística y evolución temporal de la distribución granulométrica por el tamiz 13/64 in.	67
Figura 17. Caracterización estadística y evolución temporal de la distribución granulométrica por el tamiz 12/64 in.	68
Figura 18. Medición de color bajo la escala CIE en café verde para la coordenada L...	71
Figura 19. Medición de color bajo la escala CIE en café verde para la coordenada a*.	72
Figura 20. Medición de color bajo la escala CIE en café verde para la coordenada b*.	73
Figura 21. Análisis estadístico y comportamiento temporal de la cuantificación de defectos por análisis visual.	76
Figura 22. Análisis estadístico y comportamiento temporal de los defectos de sabor en taza.	78
Figura 23. Matriz de correlación para las variables físicas y sensoriales del café.	80
Figura 24. Correlaciones más significativas con la variable objetivo defectos por análisis visual.	81
Figura 25. Correlaciones más significativas con la variable defectos de sabor en taza..	82
Figura 26. Optimización del modelo KNN a través del método del codo.....	85
Figura 27. Resultados del Modelo KNN.	85
Figura 28. Distribución de errores residuales para el modelo KNN.....	87
Figura 29. Predicción de Defectos Visuales mediante el modelo KNN (K=15).	88
Figura 30. Visualización PCA (Componentes principales).	88
Figura 31. Variables más importantes para predecir los defectos por análisis visual. ...	93
Figura 32. Matriz de Correlación para las variables predictoras.	99
Figura 33. Top 15 de coeficientes para el modelo balanceado.	101
Figura 35. Curva ROC para el modelo de regresión logística en la identificación de defectos de sabor.....	103

Figura 36. Magnitud del efecto de predictores físicos sobre la estabilidad del sabor del café.....	106
--	-----

Resumen

Este trabajo tiene como objetivo desarrollar modelos de aprendizaje automático que integren las variables fisicoquímicas y organolépticas de los granos de café verde para predecir de una manera objetiva y rápida, la calidad sensorial del café de especialidad. El estudio se realizó a través de una base de datos de 1324 muestras de origen Antioquia, Caldas, Cauca y Huila, se determinó variables como la humedad (%), densidad real (g/cm^3), tamaño de grano (%), color (espacio de color CIE L ab*), grado de tueste (Unidades agrtron), identificación de defectos por análisis visual del grano y la presencia de defectos de sabor. Se evidenció que el 13,5 % de las muestras fueron reportadas con defectos de sabor, lo que permite confirmar que la apariencia externa de los granos no constituye un predictor para la calidad sensorial en taza. A través del modelo de regresión logística binaria, se determinó que la densidad real es una variable protectora más robusta; un incremento en la densidad real reduce la probabilidad de encontrar defectos de sabor en un 29,7 %. El modelo optimizado logro una sensibilidad de identificación de defectos de sabor en un 65,9 % y una capacidad discriminativa (AUC-ROC) de 0,665. Por el contrario, el modelo de k – Vecinos más cercanos (KNN) arrojó limitaciones para la cuantificación de defectos visuales ($R^2=0,185$), evidenciando que la proximidad estadística no es suficiente para alcanzar precisión sin algoritmos no lineales más robustos.

El desempeño del modelo de regresión logística arrojó un AUC-ROC igual al 66,5 %, el cual no permite alcanzar el nivel de discriminación y precisión de un panel de evaluación

sensorial experto mayor al 90 %. Desarrollar modelos de aprendizaje automático que permitan la evaluación de la calidad sensorial del café, representan una herramienta valiosa para un control de calidad rápido y la gestión preventiva del riesgo. Estos hallazgos permiten minimizar la subjetividad en la evaluación sensorial del café de especialidad y proporcionan una base científica para optimizar los procesos de control de calidad, y la necesidad de implementación de otras tecnologías analíticas avanzadas y modelos de mayor complejidad que permitan mayor precisión en los modelos.

Palabras clave

Aprendizaje automático; calidad sensorial; café de especialidad; regresión logística; K-vecinos más cercanos (KNN).

1 Introducción

A nivel global, el café mantiene una posición histórica como uno de los productos básicos más comercializados globalmente. Datos para los periodos de 2023 al 2024 revelan un consumo de 69,179 millones de sacos de café y una producción de 178 millones de sacos de 60 kg (Avilés Granda, 2025; Sandoval Aldana, 2025). Colombia destaca en esta industria, en la que actualmente se ha logrado posicionar como el tercer productor mundial de café suave lavado contribuyendo con el 11,3 % de la producción mundial, precedido por Brasil y Vietnam (International Coffee Organization [ICO], 2023).

De acuerdo con el informe gerencial de la Federación Nacional de Cafeteros (Federación Nacional de Cafeteros de Colombia, 2025), para el año cafetero 2024 y 2025, Colombia cerró con cifras históricas. La producción alcanzó un récord de 14,9 millones de sacos de café verde, con un incremento del 16,5% por condiciones climáticas favorables y la renovación tecnificada de los cultivos. En cuanto al comercio exterior, las exportaciones aumentaron un 12,5%, alcanzando 13,3 millones de sacos con un valor cercano a los USD 5,700 millones, de los cuales el 49,1 % fueron destinados principalmente a Norteamérica y el 29,4 % para Europa.

Dada la relevancia socioeconómica global del café y la importancia crítica de su calidad sensorial y física para la exportación, se hace necesario la adopción de

herramientas innovadoras como el aprendizaje automático. Estas herramientas emergen como una solución clave para evaluar de manera objetiva, eficiente y estandarizada la calidad del grano, superando las limitaciones de los métodos tradicionales.

1.1 Cultivo de café

En Colombia, el café se consolida como uno de los productos de mayor importancia para la economía nacional y el desarrollo rural. El Ministerio de Agricultura y Desarrollo Rural lo ha definido como un cultivo estratégico con proyección de crecimiento sostenible (Ministerio de Agricultura y Desarrollo Rural, 2025). A nivel mundial, el café es uno de los productos agrícolas de mayor valor comercial, constituyendo un bien básico de consumo global y una fuente de ingresos para millones de familias productoras.

La producción cafetera colombiana está geográficamente concentrada en la denominada "Región Cafetera", que abarca principalmente departamentos de la zona andina. Los principales departamentos productores que podemos encontrar son Antioquia, Huila, Tolima, Caldas y Quindío, los cuales en conjunto representan aproximadamente el 60 % de la producción nacional (Federación Nacional de Cafeteros de Colombia, 2025). Esta concentración en zonas de ladera, si bien es clave para la calidad del grano, también implica desafíos ambientales y de manejo sostenible del suelo.

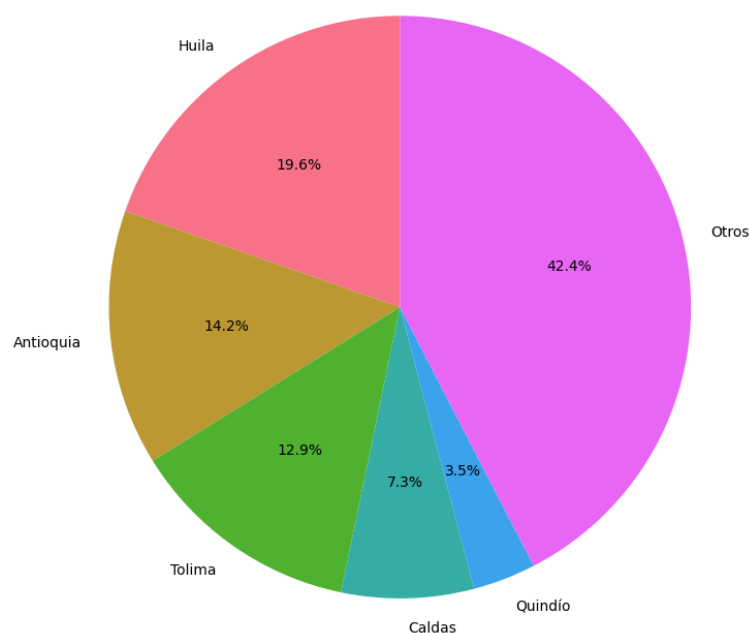


Figura 1. Distribución de producción de café arábica lavado por departamento (Federación Nacional de Cafeteros de Colombia, 2025).

Se estima que más de 540,000 familias derivan su sustento directo de este cultivo, generando cerca de 800,000 empleos directos y un número significativo de empleos indirectos en labores de procesamiento, transporte y comercialización (Federación Nacional de Cafeteros de Colombia, 2025). Las exportaciones de café colombiano, como uno de los principales productos tradicionales del país, mostraron una tendencia positiva en el año 2024. Según las estadísticas del Departamento Administrativo Nacional de Estadística (DANE, 2025), el valor total de las exportaciones de café para ese año ascendió a 3,393 millones de dólares, lo que representó una participación significativa dentro del sector de exportaciones tradicionales. Esta cifra consolida al café como un

componente fundamental de la canasta exportadora nacional. Para el período acumulado hasta noviembre de 2025, las exportaciones cafeteras ya habían alcanzado los 5,282 millones de dólares, indicando un sólido desempeño y una recuperación en el valor comercial del grano en los mercados internacionales. Estos datos reflejan la resiliencia y la importancia continua del sector cafetero para la economía colombiana.

1.2 Clasificación taxonómica.

Coffea arabica, conocido como café arábigo, es una planta angiosperma dicotiledónea de la familia Rubiaceae y orden Gentianales (El Cafè de Mi Tierra, 2024). Originaria de Etiopía (África tropical), es una especie tetraploide autopolinizante que reúne alrededor de 100 especies en el género Coffea (subgénero Eucoffea). Es la más cultivada comercialmente (aproximadamente del 60 % al 70 % de la producción mundial), con mayor complejidad aromática pero sensible a enfermedades y altitud por encima de los 1000 metros sobre el nivel del mar (msnm).

Tabla 1. Clasificación taxonómica del café arábigo. (El Cafè de Mi Tierra, 2024).

Nivel taxonómico	Clasificación	Descripción
Reino	Plantae	Angiospermas.
Filo	Magnoliophyta	Dicotiledóneas.
Clase	Magnoliopsida	Hojas elípticas, coriáceas brillantes.
Orden	Gentianales	Flores blancas axilares, perfumadas.
Familia	Rubiaceae	Árbol/frutal perenne, de 2 a 15 metros de altura.

Nivel taxonómico	Clasificación	Descripción
Género	<i>Coffea</i>	Aproximadamente de 100 especies; incluyendo la especie <i>C. canephora</i> (café robusta).
Especie	<i>C. arábica</i>	Tetraploide, donde la semilla corresponde al grano de café.

1.3 Valor nutricional del café.

El valor nutricional del café arábico reside primordialmente en su rica matriz de compuestos bioactivos, como los ácidos clorogénicos (CGA), la cafeína, la trigonelina y los diterpenos (cafestol y kahweol), más que en su contenido de macronutrientes (Farah, 2015). Los CGA, abundantes en el grano de café verde, destacan por su potente actividad antioxidante y su capacidad para modular el metabolismo de la glucosa, inhibiendo enzimas digestivas como la α -glucosidasa (Mohamed, 2024; Nemzer, 2021). El proceso de tostado transforma estas características, reduciendo los CGA pero generando melanoidinas con propiedades antioxidantes (Rufian-Henares, 2015).

1.4 Calidad física y sensorial del café.

La calidad del café de especialidad es considerada uno de los atributos más complejos y multifacéticos que se determina mediante la integración de características físicas objetivas y los atributos sensoriales visuales, de sabor y olor percibidos. En el contexto del mercado global, donde los cafés especiales deben superar los 80 puntos de acuerdo con los protocolos de la Specialty Coffee Association (SCA, 2018), la evaluación

sistemática de estos componentes es esencial para garantizar la consistencia, trazabilidad y valor comercial del producto.

La calidad física y sensorial del café se refiere a características medibles del grano de café verde y tostado, a través de metodologías apropiadas mediante ensayos analíticos válidos y reconocidos internacionalmente. Dentro de los métodos de ensayo se encuentra la determinación de tamaño del grano (% g/100 g), densidad real (g/cm^3), humedad (% g/100 g de café), cuantificación de defectos por análisis visual, determinación de sabor y olor, adicionalmente, propiedades fisicoquímicas de la infusión de café como el pH, acidez titulable, sólidos solubles (Brix), y concentraciones de compuestos específicos como proteínas, cafeína y ácidos clorogénicos (Barbosa, 2019).

Estudios sobre cafés de especialidad de Panamá, midieron sistemáticamente siete parámetros fisicoquímicos clave como densidad, Brix, pH, acidez titulable, contenido de proteína, cafeína y ácido 5-cafeoilquínico (Vega et al., 2025). Los resultados mostraron variaciones significativas entre muestras procesadas por métodos naturales y lavados, confirmando que el procesamiento postcosecha influye directamente en la composición química final. Por ejemplo, las muestras naturales presentaron mayores concentraciones promedio de 5-CQA ($1154,96 \pm 95,75 \text{ mg/L}$) en comparación con muestras de café lavadas ($1070,61 \pm 101,31 \text{ mg/L}$), mientras que la acidez titulable mostró rangos de 1,17 a 1,56 meq/100 g para cafés naturales y de 1,25 hasta 1,97 meq/100 g para cafés lavados.

Estos hallazgos son relevantes para el contexto colombiano, donde factores como la altitud, la variedad, el tipo de suelo y el método de beneficio (lavado, natural, honey, entre otros) interactúan para producir perfiles químicos distintivos que a su vez determinan la calidad sensorial (Barbosa et al., 2019; Hameed et al., 2018). La estandarización de estas mediciones físicas y químicas proporciona una base objetiva para la clasificación y valoración del café, reduciendo la subjetividad asociada con las evaluaciones tradicionales.

Teniendo en cuenta lo anterior, es importante destacar que la evaluación sensorial del café se realiza predominantemente mediante la evaluación sensorial (catación), el cual es un procedimiento estandarizado internacionalmente por la SCA que evalúa atributos sensoriales como fragancia, aroma, sabor, acidez, cuerpo, retrogusto, balance, dulzor, uniformidad y limpieza (Lingle & Menon, 2017). Durante este proceso, los evaluadores sensoriales certificados asignan puntuaciones en una escala de 0 a 100 puntos, donde los cafés de especialidad deben alcanzar al menos 80 puntos. Sin embargo, este tipo de metodología presenta limitaciones significativas en la evaluación de la calidad final del café.

Cabe resaltar que esta evaluación es subjetiva, ya que depende de la disponibilidad de evaluadores sensoriales certificados, presenta inconsistencias en su reproducibilidad, y está sujeta a la variabilidad en la percepción sensorial entre evaluadores sensoriales (Vega et al, 2025). Además, es un proceso con alta demanda de tiempo y en ocasiones puede conducir a conclusiones inexactas. Estas limitaciones han impulsado la búsqueda

de métodos alternativos que complementen o reduzcan la dependencia de las evaluaciones sensoriales humanas, que permita la toma de decisiones en menor tiempo y precisas.

En el estudio panameño, las puntuaciones sensoriales de las muestras variaron entre 86,9 puntos y 92,9 puntos, con coeficientes de variación inferiores al 1,36 %, indicando consistencia en las evaluaciones del panel (Vega et al, 2025). Estas puntuaciones proporcionaron la base de referencia para desarrollar modelos predictivos que relacionaran los parámetros fisicoquímicos con la calidad sensorial percibida.

1.5 Relaciones entre calidad física y sensorial

Investigaciones recientes han demostrado que existen correlaciones significativas entre los parámetros fisicoquímicos medibles y la evaluación sensorial realizada por evaluadores sensoriales expertos. Se identificaron ciertos patrones de conducta con correlaciones positivas entre la acidez titulable y las puntuaciones sensoriales ($r = 0,3673$), así como entre los sólidos solubles (Brix) y las puntuaciones sensoriales ($r = 0,5274$). El contenido de proteína también mostró una correlación positiva con Brix ($r = 0,8970$) con las puntuaciones sensoriales (Vega et al., 2025). Por el contrario, el pH exhibió una correlación inversa con las puntuaciones sensoriales ($r = -0,6979$).

Estas relaciones evidencian que las propiedades sensoriales percibidas están fundamentadas en la composición química de la bebida, lo que valida el enfoque de utilizar mediciones objetivas para predecir y evaluar la calidad sensorial del café. Estos

hallazgos son consistentes con investigaciones en otros productos alimenticios como vino, queso y frutas, donde también se han establecido correlaciones entre parámetros fisicoquímicos y atributos sensoriales (Aleixandre-Tudó et al., 2015; Kraggerud et al., 2014; Corrêa et al., 2014).

Para el café colombiano, estas relaciones adquieren particular relevancia considerando la diversidad de perfiles sensoriales asociados con diferentes regiones productoras. Cafés de Nariño y Huila, por ejemplo, suelen presentar mayor acidez y cuerpo, mientras que aquellos de Antioquia y Caldas pueden destacar por su balance y dulzor (Barbosa et al., 2019). La cuantificación de los parámetros fisicoquímicos asociados con estos atributos permitiría establecer perfiles regionales característicos que sirvan como referencia para la clasificación y valoración objetiva.

1.6 Modelado predictivo de la calidad sensorial mediante técnicas multivariadas

Para superar las limitaciones de la evaluación sensorial tradicional, se han desarrollado enfoques basados en el modelado multivariado que permiten predecir la calidad sensorial a partir de parámetros fisicoquímicos objetivos (Vega et al., 2025). Dentro de los algoritmos de modelado aplicados a cafés especiales podemos encontrar:

- regresión Lineal Múltiple (MLR);
- regresión por Análisis de Componentes Principales (ACP);
- regresión por Mínimos Cuadrados Parciales (PLSR);

- clasificador de Vectores de Soporte (SVC);
- K- vecinos más cercanos (KNN);
- árbol de decisión (Decision tree);
- bosque aleatorio (Random Forest);
- Catboost;
- Naive Bayes (NB); y
- regresión logística.

Los resultados han demostrado que los modelos construidos específicamente para cada método de procesamiento del café (lavado o natural) superan significativamente en precisión a aquellos que combinaban muestras de ambos procesos. El modelo PLSR para café natural presentó el mejor desempeño predictivo, con un coeficiente de determinación para predicción (R^2_p) de 0,8306; un error cuadrático medio de predicción (RMSEp) de 0,3948; y una desviación predictiva residual (RPD, por sus siglas en inglés) de 2,51. Según los criterios establecidos en la literatura, valores de RPD superiores a 2,5 indican capacidad predictiva excelente (Xu et al., 2024).

El análisis de importancia de variables (VIP) para los modelos PLSR reveló que Brix y pH fueron los parámetros que más contribuyeron a la predicción de las puntuaciones

sensoriales en todos los modelos, mientras que el contenido de proteína también mostró alta importancia en los modelos para muestras lavadas y combinadas (Vega et al., 2025).

En esta misma línea de exploración metodologías aplicadas a la evaluación sensorial, investigaciones han demostrado que es factible estimar perfiles sensoriales complejos del café tales como notas florales, afrutadas, a nuez o chocolate a partir de sus propiedades físicas y defectos visuales del grano verde mediante modelos de aprendizaje automático (González-Sánchez et al., 2023). Para lo anterior, se han empleado algoritmos como Random Forest (RFC), CatBoost (CB) y Support Vector Classifier (SVC), los cuales, alcanzaron precisiones superiores al 80% en la predicción de atributos como aroma, fragancia, sabor y resago. En investigaciones anteriores se han identificado que factores como la humedad, densidad, tamaño y forma del grano, junto con defectos visuales específicos, son determinantes en la calidad sensorial final del café (González-Sánchez et al., 2023). Estos hallazgos permiten definir que la integración de múltiples modelos computacionales no solo iguala la precisión de los catadores expertos, sino que también minimiza la subjetividad inherente a la evaluación humana, consolidando así el potencial de estas herramientas para la estandarización y control de calidad en la industria cafetera.

1.7 Implicaciones para el café colombiano y perspectivas futuras

La aplicación de estos enfoques en el contexto colombiano ofrece oportunidades significativas para mejorar la evaluación, clasificación y comercialización de cafés especiales. Considerando la amplia diversidad regional, varietal y de procesamiento en

Colombia, el desarrollo de modelos predictivos específicos por región, calidad del producto y método de beneficio podría optimizar considerablemente la precisión en la predicción de la calidad sensorial.

La integración de estas herramientas con los sistemas de denominación de origen ya establecidos en Colombia fortalecería la trazabilidad y el posicionamiento en mercados internacionales exigentes. Además, la alineación de técnicas analíticas rápidas como las gravimétricas y organolépticas (inspección visual) podría facilitar la medición en línea de parámetros sensoriales de sabor clave, permitiendo un control de calidad más eficiente a lo largo de la cadena productiva (Baqueta et al., 2019; Çataltaş & Tutuncu, 2021).

Los enfoques multivariados, han demostrado capacidad para predecir con precisión la evaluación de la calidad sensorial a partir de un conjunto limitado de parámetros fisicoquímicos, especialmente cuando los modelos se desarrollan considerando específicamente el método de procesamiento postcosecha. Para el sector cafetero colombiano, la adopción de estas metodologías representa una oportunidad estratégica para fortalecer la evaluación objetiva de calidad, reducir la dependencia de métodos sensoriales subjetivos, y mejorar la competitividad en el mercado global de cafés especiales. La implementación exitosa requerirá investigación adaptada a las condiciones específicas de Colombia, considerando su diversidad regional, varietal y de procesamiento, así como la integración de tecnologías analíticas avanzadas y técnicas de modelado computacional.

2 Planteamiento del problema.

La evaluación de la calidad sensorial del café colombiano, especialmente en el segmento de cafés especiales que requieren puntuaciones superiores a 80 puntos según los protocolos de la Specialty Coffee Association (SCA), lo cual enfrenta actualmente desafíos metodológicos por su falta de objetividad, reproducibilidad y aplicabilidad a gran escala o de manera más ágil en procesos de control de calidad. Para este proceso, es fundamental centrarse en dos ensayos organolépticos esenciales, la inspección del grano por análisis visual para detectar defectos en el grano verde y la evaluación sensorial de sabor en taza (catación) realizada por evaluadores sensoriales certificados. Sin embargo, ambos procedimientos presentan limitaciones inherentes que comprometen la confiabilidad de los resultados y, por ende, la competitividad del producto en mercados internacionales exigentes.

En cuanto a la inspección del grano de café verde por análisis visual, la identificación de defectos físicos como granos negros, vinagres, mordidos, brocados, inmaduros o partidos, dependen en gran medida de la formación, entrenamiento y experiencia individual del evaluador, así como su capacidad visual y condiciones ambientales (ausencia o presencia de luz aceptable), lo que puede introducir variabilidad entre observadores y está sujeta a errores derivados de la fatiga visual, la falta de estandarización en los criterios de clasificación y homogenización de las muestras de

ensayo. Esta subjetividad en la fase inicial del control de calidad puede traducirse en una valoración inconsistente del producto desde las primeras etapas de la cadena productiva, afectando su potencial comercial. Por otro lado, la evaluación sensorial descriptiva del sabor, reconocida como el método de referencia para la evaluación sensorial, enfrenta limitaciones igualmente significativas. Este procedimiento requiere de catadores altamente capacitados, cuya disponibilidad es restringida en las regiones productoras debido a los altos costos y tiempos requeridos para su formación. Además, la evaluación humana está influenciada por factores como el estado fisiológico del evaluador, las condiciones ambientales durante la prueba y la fatiga sensorial, lo que genera inconsistencias en la reproducibilidad de los resultados y puede conducir a conclusiones inexactas (Vega et al., 2025). Este enfoque, además de ser subjetivo, demanda tiempos prolongados que dificultan la toma de decisiones ágiles en los procesos de comercialización y exportación.

En el contexto cafetero colombiano, caracterizado por una amplia diversidad regional, varietal y de métodos de procesamiento en departamentos como Antioquia, Huila, Tolima, Caldas y Quindío, persiste una brecha significativa en el desarrollo de modelos integrados de aprendizaje automático que permitan, de manera simultánea, la identificación automatizada de defectos por análisis visual y la predicción de la calidad sensorial en taza a partir de características objetivas del grano.

Ante esta problemática, el presente trabajo se justifica por la necesidad de desarrollar modelos de aprendizaje automático que integren tanto el análisis visual de defectos como

variables fisicoquímicas clave como la densidad real (g/cm^3), humedad ($\%$, g/100 g), tamaño del grano ($\%$, g/100 g), color del grano de café verde bajo la escala CIE $L^* a^* b^*$, y color de café tostado (Unidades agron) para predecir de manera objetiva, rápida y precisa la calidad sensorial del café colombiano. La iniciativa parte de la premisa de que es posible reducir significativamente la dependencia de métodos sensoriales subjetivos mediante la implementación de herramientas computacionales que capturen las relaciones complejas entre las características medibles del grano y la percepción sensorial final de la bebida. Para ello, se plantea la construcción de modelos predictivos específicos para la evaluación sensorial de cafés de especialidad.

Desde una perspectiva práctica, el presente trabajo pretende proporcionar al sector cafetero colombiano herramientas que permitan la automatización parcial del proceso de evaluación de la calidad sensorial del café, reduciendo los tiempos y costos asociados a la catación tradicional; minimizar la subjetividad en las evaluaciones sensoriales mediante criterios objetivos basados en datos; que permitan facilitar la toma de decisiones en los procesos de clasificación, y fortalecer la trazabilidad y el posicionamiento de los cafés especiales colombianos en mercados internacionales mediante la estandarización de los criterios de calidad. Con este enfoque, se pretende contribuir a la competitividad del sector cafetero nacional, ofreciendo alternativas metodológicas que, sin pretender reemplazar completamente al catador humano, proporcionen un soporte de apoyo objetivo y confiable para la evaluación de la calidad sensorial, alineándose con las tendencias globales de digitalización y automatización en la industria alimentaria.



Figura 2. Diagrama de Ishikawa en la identificación de raíces del planteamiento del problema.

3 Objetivos

3.1 Objetivo General

Desarrollar modelos de aprendizaje automático que integren la evaluación sensorial del grano con variables fisicoquímicas clave para predecir de manera objetiva, rápida y precisa la calidad sensorial del café colombiano.

3.2 Objetivos específicos

- Evaluar fisicoquímica y sensorialmente diferentes muestras de café verde correspondientes a de diferentes regiones productoras colombianas, a través de diferentes mensurados como humedad (%), g/100 g), densidad real (g/cm³), tamaño de grano (%), g/100 g), color del grano verde bajo la escala CIE L ab*, color del café tostado bajo la escala agron, cuantificación de defectos por análisis visual y la identificación de defectos de sabor (atributos negativos) por análisis descriptivo cualitativo (presencia o ausencia de defectos en taza).
- Cuantificar el número de defectos totales por análisis visual en los granos de café estableciendo su prevalencia y relación con las características físicas del lote.

- Determinar las correlaciones existentes entre las variables fisicoquímicas y sensoriales de las diferentes muestras. Los defectos cuantificados por análisis visual y los defectos de sabor son obtenidos mediante evaluadores sensoriales expertos certificados, siguiendo las bases técnicas de referencia validadas por Specialty Coffee Association (SCA).
- Entrenar y optimizar un modelo de regresión logística para la clasificación binaria de defectos de sabor en taza (presencia o ausencia), utilizando como variables predictoras las características fisicoquímicas y sensoriales del café verde, y evaluando su desempeño mediante métricas de precisión, sensibilidad, especificidad y área bajo la curva ROC.
- Entrenar y optimizar un modelo de clasificación mediante K-Vecinos Más Cercanos (KNN) para la identificación y discriminación del tipo de grano según región de origen, a partir del conjunto de variables fisicoquímicas y sensoriales disponibles.
- Evaluar el rendimiento de los modelos desarrollados mediante métricas de precisión, exactitud, sensibilidad, especificidad y validación cruzada, contrastando los resultados con las clasificaciones obtenidas mediante métodos tradicionales de evaluación sensorial.
- Validar los modelos con los resultados de muestras independientes que no hayan sido utilizadas en las fases de entrenamiento, verificando su capacidad de

generalización, robustez y aplicabilidad en condiciones reales de producción y control de calidad.

4 Metodología

Las actividades de investigación se llevaron a cabo en el Laboratorio Central de la empresa Cafexport S.A.R.L., Sucursal Colombia de la ciudad de Manizales – Caldas. El laboratorio Central Cafexport S.A.R.L., cuenta con acreditación ONAC, vigente a la fecha, con código de acreditación 22-LAB-068, bajo la norma ISO/IEC 17025:2017. La acreditación de los ensayos dentro del alcance del laboratorio garantiza competencia técnica y fiabilidad, lo que asegura la validez de los resultados obtenidos a nivel nacional e internacional.

Se recopilaron las siguientes variables, tales como el origen, las propiedades físicas y los defectos por análisis visual, así como información sensorial sobre la evaluación de sabor del café, a partir de 1324 muestras de ensayo. Las variables analizadas se describen a continuación:

4.1 Muestras de ensayo

Se trabajó con café verde variedad Castillo, Colombia y Cenicafe 1, cultivado en diferentes zonas rurales de Antioquia, Caldas, Cauca y Huila. Las muestras de ensayo fueron sometidas a un proceso de beneficio hasta obtener café verde para su caracterización fisicoquímica y sensorial. Cada muestra de ensayo corresponde a 300 g de café verde obtenida a través de un muestreo compuesto de un lote de 275 sacos de café

verde de 70 kg cada uno. Las muestras de ensayo fueron debidamente identificadas y pesadas en una balanza OHAUS con resolución de 0,1 g.

4.2 Caracterización física del café de especialidad.

4.2.1 Densidad real (g/cm³). La densidad real fue determinada a través del ensayo “Determinación de densidad real por desplazamiento de gas Helio mediante el método propio GTL-P-17 V.01 2023-0824” bajo la técnica gravimétrica. Para esto se pesaron aproximadamente 51 g ± 0,5 g de café verde en una balanza analítica Shimadzu AUX220 con resolución de 0,0001 g, y se determinó su volumen a través del picnómetro AccuPyc II 1345 con resolución de 0,0001 cm³; la determinación de volumen se realizó por desplazamiento de gas Helio grado reactivo. El resultado reportado se expresa de acuerdo con la siguiente ecuación:

$$\rho = \frac{m}{V} \text{ (g/cm}^3\text{)}$$

Donde:

ρ es densidad

m es masa

V es volumen

4.2.2 Humedad (% g/100 g). Se determinó la humedad del café verde a través del ensayo “Determinación de humedad rápida por capacitancia eléctrica mediante el método propio GTL-P-08 V.04 2023-07-06” bajo la técnica capacitancia eléctrica. Para esto se aforó el volumen requerido por el equipo KETT PM 450 con resolución de 0,1 g para cada una de las muestras de ensayo. El porcentaje de humedad se determinó mediante la siguiente ecuación:

$$\text{Humedad} \left(\%, \frac{\text{g}}{100 \text{ g}} \right) = \text{Valor reportado por el equipo} \pm \text{factor de corrección}$$

4.2.3 Tamaño de grano (% g/100 g). Se determinó el tamaño del café verde a través del ensayo “análisis de tamaño de grano por tamizado mecánico mediante el método propio GTL-P-15 V.04 2024-05-07” bajo la técnica gravimétrica. Para esto, se pesaron aproximadamente $200 \text{ g} \pm 0,5 \text{ g}$ de café verde y se vierten en un juego de tamices con diámetro de apertura de 18/64 in (7,14 mm), 17/64 in (6,75), 16/64 in (6,35 mm), 15/64 in (5,95 mm), 14/64 in (5,56 mm), 13/64 in (4,96 mm) y 12/64 in (4,76 mm). Se agitó mecánicamente durante 2 minutos y se pesa los granos recolectados en cada uno de los tamices de ensayo en una bandeja. El tamaño de grano se expresa como una fracción de masa de café verde retenida sobre cada tamiz de acuerdo con la siguiente ecuación:

$$w_{gc} \left(\%, \frac{\text{g}}{100 \text{ g}} \right) = \frac{m_s}{m_T} \times 100$$

Donde:

w_{gc} es la fracción de masa retenida con el número de tamiz, S;

mS es la masa del grano de café, con el número de tamiz, S , expresado en gramos;

mT es la masa total de la muestra, expresada en gramos.

4.2.4 Color CIE $L a^* b^*$. Se determinó el color del café verde a través del método oficial ASTM E1164-23 y el manual de operación del equipo A60-1018-812 versión 2.3 bajo la técnica colorimétrica. La medición de color se hizo con un espectrofotómetro de reflectancia de doble haz AGERA, con geometría $0^\circ/45^\circ$ circunferencial para color y 60° para brillo, conforme a ASTM E1164. Se uso el método de medición con puerto hacia arriba, con un tiempo de lectura de 3 segundos, captura de imagen de alta resolución bajo iluminante D65 (VISIÓN $45^\circ/0^\circ$), permitiendo captura y recuperación de imágenes, en escala CIE bajo la escala CIE $L a^* b^*$ con resolución de 0,01 para cada una de las coordenadas. El equipo se calibró con estándares NIST trazables. El valor reportado corresponde al promedio de 3 lecturas por muestras con rotación de 120° , el espacio de color se define por medio de las siguientes expresiones:

L = Valor reportado por el equipo.

a^* = Valor reportado por el equipo.

b^* = Valor reportado por el equipo.

Donde:

L corresponde a la coordenada de Luminosidad, a^* a la coordenada de color de verde ($-a^*$) a rojo ($+a^*$), y b^* a la coordenada de azul ($-b^*$) a amarillo ($+b^*$).

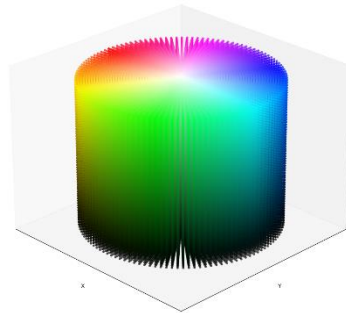


Figura 3. Espacio del color de acuerdo con la escala CIE L a*b*.

4.2.5 Grado de tueste (Unidades Agtron). Se determinó el grado de desarrollo del tueste del café mediante el uso del equipo Agtron modelo E20CP-III, siguiendo el manual de operación del equipo versión IR128-S5.0/S5.5 bajo la técnica de espectroscopía de infrarrojo cercano (NIR). La medición se realizó sobre muestras de café molido, garantizando que la superficie fuera plana, uniforme y sin vacíos en el plato de muestra para asegurar la precisión y repetibilidad del análisis. El equipo se calibró utilizando un disco de referencia de doble cara (rojo y negro) trazable al número de serie del instrumento, complementado con el ajuste automático del "Yellow Life Disk" interno cada vez que se abre el cajón de muestras. Las lecturas se tomaron a temperatura ambiente para evitar variaciones en el puntaje debido a la inestabilidad térmica. El valor reportado se expresa en la escala Agtron Gourmet, definida por un rango numérico de acuerdo con la siguiente expresión:

Puntaje Agtron = Valor reportado por el equipo en una escala de 0,0 a 100,0.

Donde un valor de 0.0 unidades agtron corresponde a la carbonización total de la materia orgánica y 100.0 unidades agtron a un tueste extremadamente claro. Según los estándares internacionales, un tueste de nivel de catación (cupping level roast) se define con un valor objetivo de 63.0 en la escala Gourmet.

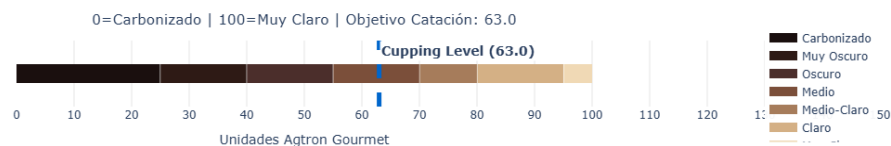


Figura 4. Escala de grado de tueste en unidades Agtron.

4.2.6 Metodología de aprendizaje automático. A continuación, se describe la metodología empleada para determinar los parámetros sensoriales del café. El apartado correspondiente al entrenamiento del modelo se enfoca en la creación de un modelo entrenado a través de aprendizaje supervisado. Este modelo emplea como variables de entrada las propiedades físicas con el objetivo de estimar la cuantificación defectos del café por análisis visual, y adicionalmente, las propiedades físicas y cuantificación de defectos por análisis visual con el objetivo de estimar la evaluación sensorial de sabor en la identificación de defectos en taza (presencia/ausencia), tal como se ilustra en la siguiente figura:

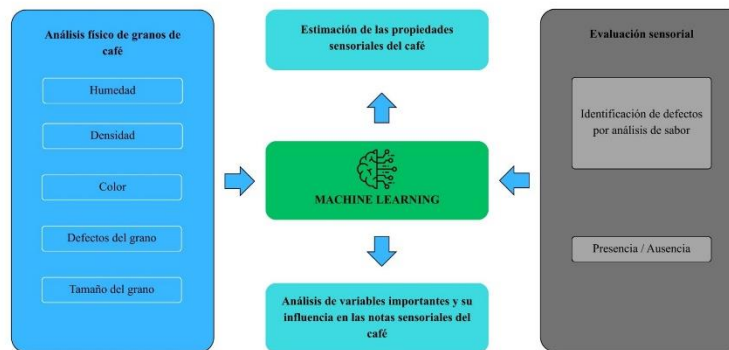


Figura 5. Diagrama esquemático de la metodología de aprendizaje automático implementada y utilizada para el análisis de variables importantes y la estimación de propiedades sensoriales.

El propósito central de esta investigación es identificar qué variables de entrada, contenidas en la base de datos, ejercen influencia sobre la variable objetivo previamente seleccionada. En consecuencia, se desarrollaron modelos fundamentados en técnicas de inteligencia artificial, con la capacidad de estimar tanto los parámetros sensoriales para su posterior análisis la calidad sensorial del café. En el estudio se estructuran varias fases, comenzando por la construcción y organización de la base de datos, preprocesamiento, selección de variables, validación cruzada, importancia de las características, aprendizaje automático; modelos y ajuste de hiperparámetros para lograrlo.

- **Creación de la base de datos.** La información utilizada comprende datos sobre el origen, las características físicas y sensoriales obtenidas a partir de la evaluación de la calidad del café. Para la gestión integral de la carga, análisis y

almacenamiento de la información en la base de datos, se empleó la librería Pandas de Python.

- **Transformación de la base de datos.** Se analiza la base de datos mediante un algoritmo diseñado para identificar datos potencialmente corruptos, tales como campos vacíos o caracteres especiales. Cuando el algoritmo detecta registros con problemas, procede a estandarizarlos automáticamente, garantizando así que los modelos de aprendizaje automático no presenten fallos durante su ejecución derivados de este tipo de datos.

- **Selección de variables.** Se selecciona las variables a analizar, definiendo cuáles serán las variables de entrada y cuál la variable objetivo. El *script* permite al usuario escribir los nombres de las variables para que el sistema automáticamente almacene, analice y agrupe la información en la búsqueda. Para obtener resultados más cercanos a la realidad, se debe garantizar que los datos empleados en el entrenamiento y prueba sean correctos, reduciendo así la posibilidad de sobrentrenamiento. Para esto, se implementa la técnica de validación cruzada, que permite al algoritmo realizar diferentes pruebas mediante diversas combinaciones de los nK pliegues establecidos por el usuario. En la siguiente figura se muestra el esquema de este segmento de validación cruzada.



Figura 6. Diagrama esquemático de aprendizaje automático mediante validación cruzada para la determinación de las variables importantes.

- **Importancia de características.** El objetivo es identificar las variables relevantes que impactan, positiva o negativamente, en la variable objetivo a analizada. Para lograrlo, se emplean los modelos de K-vecinos Más Cercanos (KNN) y Regresión Logística (LogR). Estos proporcionan las variables más importantes que influyen en la variable objetivo.
- **Modelos de aprendizaje automático.** En este segmento se utilizan los datos provenientes de la sección de base de datos y se ejecutan los modelos para predecir la calidad sensorial del café. Los modelos implementados incluyen el modelo de K-Vecinos Más Cercanos (KNN) para tareas de clasificación para la cuantificación de defectos por análisis visual y el modelo de Regresión Logística para la clasificación binaria de defectos de sabor en taza, determinando la presencia o ausencia de estos. Se realizó un proceso de validación cruzada y se ajustaron los hiperparámetros correspondientes para alcanzar la mayor exactitud posible.

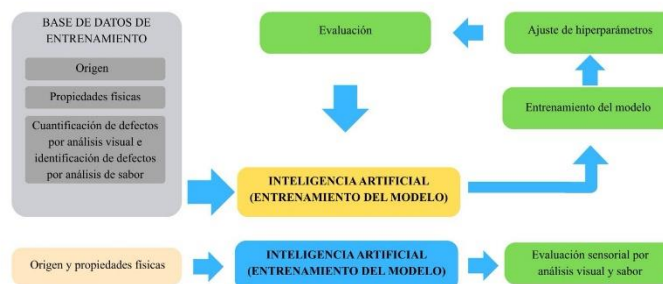


Figura 7 Diagrama metodológico de aprendizaje automático para el desarrollo de modelos de predicción sensoriales del café.

4.2.7 Análisis de variables importantes. Con el propósito de identificar qué características del café inciden en los ensayos organolépticos por análisis visual y de sabor, se realiza un análisis detallado estructurado en dos fases experimentales. Influencia de la variedad y el tipo de grano en los atributos sensoriales. Para este análisis, se emplearon 1322 muestras de granos de café verde cultivados en diferentes regiones de Colombia. La información recopilada incluye los resultados de catación, vinculando las percepciones sensoriales con el origen y tipología del grano. Se construyó una base de datos específica para explorar cómo estas características inherentes al café determinan los distintos atributos organolépticos del café, tales como Cuantificación de defectos por análisis visual e Identificación de defectos por análisis de sabor.

Impacto de las propiedades físicas en los ensayos de evaluación sensorial del café. En esta fase se analizaron las mismas 1322 muestras de café verde de los diferentes orígenes de Colombia, esta vez considerando atributos físicos como humedad, densidad, cuantificación de defectos por análisis visual, así como tamaño del grano. El objetivo fue examinar cómo estas propiedades físicas se relacionan con los resultados de la evaluación de sabor en catación, determinando su contribución a la evaluación de la calidad sensorial final en la presencia o ausencia de defectos en la bebida.

Modelo de clasificación mediante K-Vecinos Más Cercanos (KNN). KNN clasifica regiones de origen del café (Antioquia, Caldas, Cauca y Huila) midiendo similitud vía distancias entre muestras, usando variables fisicoquímicas estandarizadas. Para un nuevo punto x , selecciona k vecinos más cercanos en entrenamiento y asigna clase mayoritaria por votación.

Distancia euclidiana principal:

$$d(x_i, x_j) = \sqrt{\sum_{p=1}^m (x_{i,p} - x_{j,p})^2}$$

donde m es número de variables individuales e independientes (ejemplo humedad, color, densidad, entre otros).

La predicción se realiza con el máximo conteo en k vecinos. La optimización se realiza a través de GridSearchCV el cual permite minimizar el error CV:

$$CV = \frac{1}{K} \sum_{k=1}^K L(y_k, \hat{y}_k)$$

Donde k corresponde al número de folds (Subconjuntos iguales y no superpuestos en los que se divide e conjunto de datos de entrenamiento).

Modelo de regresión logística para la clasificación binaria de defectos en taza (presencia o ausencia). Predice la identificación de defectos en taza (binario, donde 1

corresponde a presencia de defectos y 0 corresponde a ausencia de defectos) modelando logit de probabilidad con features fisicoquímicas. El modelo se optimiza con coeficientes β maximizando verosimilitud vía gradiente descendente. Las ecuaciones usadas para el modelo son:

Log odds lineal:

$$z = \beta_0 + \beta_1 x_1 + \dots + \beta_m x_m$$

Probabilidad sigmoide:

$$P(y = 1 | x) = \frac{1}{1 + e^{-z}} = \frac{e^z}{1 + e^z}$$

AUC ROC (área bajo la curva Receiver Operating Characteristic) integra TPR (Tasa de positivos verdaderos) vs FPR (Tasa de falsos positivos):

$$TPR = \frac{TP}{TP + FN}, FPR = \frac{FP}{TN + FP}$$

Donde un AUC con capacidad discriminante igual a 0,5 corresponde a clasificaciones tomadas al azar y un valor de 1,0 a una clasificación perfecta).

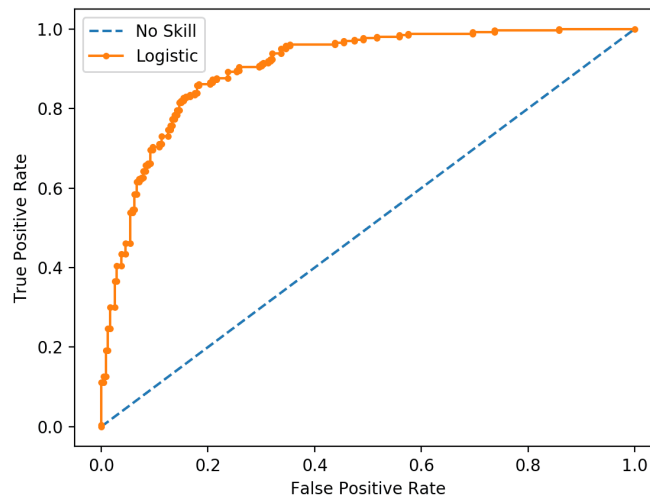


Figura 8 Curva ROC para un modelo de regresión logística (Brownlee, 2018).

4.2.8 Predicción de atributos sensoriales negativos. La predicción de defectos de sabor (atributos sensoriales negativos) en taza mediante evaluación sensorial del sabor y como imperfecciones visuales físicas en el grano verde se realizaron a partir de información de cada uno de los lotes sobre su información de origen, propiedades fisicoquímicas y sensoriales. En este estudio se emplearon 1324 muestras de café verde y dos modelos de aprendizaje automático supervisado configurados para la clasificación. Los modelos fueron desarrollados para predecir tanto la presencia o ausencia de atributos negativos en taza (mediante selección forzada por catadores expertos) como la cuantificación de defectos totales determinada por análisis visual del grano. El punto central del estudio, por tanto, es el desarrollo de algoritmos capaces de anticipar la calidad sensorial negativa del café verde utilizando variables de origen, fisicoquímicas y

organolépticas, dando como resultado valores categorizados de defectos visuales del grano o de sabor en taza.

5 Resultados y Discusión

En esta sección, se reportan y discuten los resultados obtenidos tras la evaluación física y sensorial de las 1,324 muestras de ensayo de café verde procedentes de los diferentes orígenes. Partiendo del objetivo general, se desarrollan modelos de aprendizaje automático para una evaluación objetiva de la calidad sensorial visual y de sabor del grano, los resultados integran los ensayos de evaluación física del grano verde como densidad real (g/cm^3), humedad (%), g/100 g), tamaño de grano (%), porcentaje de retención), color bajo la escala CIE L ab*, y adicionalmente, la medición del grado de tueste en unidades Agtron para el grano tostado molido; por otro lado, se realizaron los ensayos para la evaluación sensorial del grano como cuantificación de defectos por análisis visual para el grano verde y la identificación de defectos de sabor para el grano tostado molido. Se discute las características de desempeño de los modelos de K-Vecinos Más Cercanos (KNN) para la clasificación de defectos totales por análisis visual, y por lado, el modelo de Regresión Logística para la predicción de defectos en sabor en taza, para cada uno de los modelos se evaluaron las características de desempeño de precisión y sensibilidad teniendo como base los métodos de evaluación sensorial tradicionales bajo procedimientos validados e internacionalmente reconocidos. Los resultados y análisis permiten validar la capacidad de las herramientas computacionales para reducir la subjetividad inherente a la evaluación sensorial humana y fortalecer las metodologías de evaluación de la calidad en la industria del café de especialidad colombiano.

Evaluación física y sensorial del café de especialidad

La consolidación de la base de datos corresponde a la información reportada a partir de 1,324 muestras de ensayo analizadas durante los años 2024 (646 muestras de ensayo) y 2025 (678 muestras de ensayo). El análisis de datos se realizó a través un script de Python utilizando la librería Pandas, lo que permitió gestionar integralmente la carga y almacenamiento de la base de datos. La base de datos está conformada por un total de 344 fechas individuales de ingreso y análisis de muestras. En promedio se analizaron cuatro muestras por fecha, lo que asegura una recolección de datos sistemática a lo largo del periodo de este estudio.

Una vez almacenada la base de datos a través del script de Python, se realiza un diagnóstico de la calidad de los datos, lo cual dio como resultado una notable en su cobertura, donde la mayoría de las variables físicas y sensoriales tales como la humedad, el tamaño de grano y la medición de color en el espacio CIE L ab* presentan el 100 % de los datos disponibles sin requerir procesos de imputación. Por otro lado, se detectaron valores ausentes mínimos para el ensayo de densidad real, el color del café tostado y el total de defectos, los cuales fueron tratados mediante un algoritmo de procesamiento para garantizar la estabilidad y robustez de los modelos de aprendizaje automático. La base de datos consolidada corresponde a las tablas de datos mostradas a continuación:

Tabla 2. Trazabilidad de las muestras de café de especialidad.

Nº	Fecha de análisis	Origen	Número de Lote
1	2024-01-04	Aguadas	3-0960-01169
2	2024-01-04	Aguadas	3-0960-00931
3	2024-01-04	Alto	3-0960-00644
4	2024-01-04	Alto/Aguadas	3-0960-00402
5	2024-01-04	Antioquia	3-0960-01068
6	2024-01-10	Norte/Aguadas	3-0960-00452
7	2024-01-10	Alto/Aguadas	3-0960-00403
8	2024-01-10	Alto/Aguadas	3-0960-00404
9	2024-01-11	Norte	3-0960-00439
10	2024-01-11	Norte	3-0960-00440

Nota: La tabla informa los 10 primeros datos del total conformado en la base de datos.

TRABAJO DE GRADO

Opción Investigación o Proyecto de Grado

Tabla 3. Caracterización física del café de especialidad. La tabla informa los 10 primeros datos de 1324.

N°	Humedad (% g/100 g)	Densidad (g/cm3)	Color CIE L	Color CIE a*	Color CIE b*	Color Café Tostado (Agtron)	Tamiz N° 18 %	Tamiz N° 17 %	Tamiz N° 16 %	Tamiz N° 15 %	Tamiz N° 14 %	Tamiz N° 13 %	Tamiz N° 12 %
1	11,7	1,3578	44,63	0,69	14,11	106,8	56,45	25,8	16,25	1,5	0	0	0
2	11,4	1,3564	44,08	0,56	13,16	109,3	62,6	23,6	12,15	1,35	0,3	0	0
3	11,7	1,3594	44,58	0,32	13,64	111,1	60,95	22,65	14,65	1,4	0,35	0	0
4	11,5	1,3606	42,55	0,52	13,11	109,2	59,25	23,05	16,45	1,15	0,1	0	0

8	7	6	5	Nº
11,6	11,3	11,2	11,8	Humedad (% g/100 g)
1,3623	1,3572	1,358	1,3432	Densidad (g/cm3)
42,14	44,42	42,36	47,08	Color CIE L
0,34	0,5	0,66	1,38	Color CIE a*
12,96	14,15	13,59	16,2	Color CIE b*
112	103,2	106,8	118,2	Color Café Tostado (Agtron)
54,55	55,8	56,6	10,7	TamizNº 18 %
27,65	27,65	25,75	9,85	TamizNº 17 %
15,55	14,75	16,25	35,75	Tamiz N° 16 %
1,95	1,45	1,15	30,4	Tamiz N° 15 %
0,3	0,35	0,25	11,6	Tamiz N° 14 %
0	0	0	1,7	Tamiz N° 13 %
0	0	0	0	Tamiz N° 12 %

10	9	N°
11,5	11,8	Humedad (% g/100 g)
1,3577	1,3602	Densidad (g/cm3)
43,21	42,08	Color CIE L
0,86	1,04	Color CIE a*
14,38	14,03	Color CIE b*
113,2	103,3	Color Café Tostado (Agtron)
60	60,25	TamizN° 18 %
24,4	24,25	TamizN° 17 %
14,45	13,95	Tamiz N° 16 %
1,1	1,5	Tamiz N° 15 %
0,05	0,05	Tamiz N° 14 %
0	0	Tamiz N° 13 %
0	0	Tamiz N° 12 %

TRABAJO DE GRADO

Opción Investigación o Proyecto de Grado

Tabla 4. Caracterización organoléptica del café de especialidad.

Nº	Defectos por análisis visual (Total de defectos)	Identificación de Defectos de Sabor	Identificación de Defectos de Sabor (Presencia/Ausencia)
1	7,0	-	Ausencia
2	6,1	-	Ausencia
3	6,8	-	Ausencia
4	10,5	-	Ausencia
5	4,8	-	Ausencia
6	7,0	Moho	Presencia
7	7,4	-	Ausencia
8	6,0	-	Ausencia
9	5,8	-	Ausencia
10	8,7	-	Ausencia

Nota: La tabla informa los 10 primeros datos del total conformado en la base de datos.

Análisis descriptivo de las variables físicas y sensoriales. Se realizó un análisis descriptivo de todas las variables del estudio a partir de una muestra de 1,324 observaciones. La Tabla 5 presenta las estadísticas descriptivas de las variables evaluadas.

Tabla 5. Estadísticas descriptivas para las variables físicas y sensoriales de los granos de café (n=1324).

Variable	Media	Desviación estándar (DE)	Mediana	Mínimo	Máximo	Q1	Q3
Humedad (%)	11.57	0.42	11.50	10.50	12.70	11.30	11.90
Densidad	1.35	0.01	1.36	1.31	1.39	1.35	1.36
Tamiz N° 18 (%)	54.91	6.06	54.64	38.70	71.01	50.96	58.55
Tamiz N° 17 (%)	26.45	3.20	26.50	17.90	34.90	24.49	28.60
Tamiz N° 16 (%)	16.66	3.22	16.82	7.50	25.65	14.54	18.75

Tamiz N° 15 (%)	2.48	4.36	1.35	0.00	34.80	0.70	2.55
Tamiz N° 14 (%)	0.57	2.28	0.10	0.00	28.60	0.00	0.25
Tamiz N° 13 (%)	0.12	0.86	0.00	0.00	20.25	0.00	0.00
Tamiz N° 12 (%)	0.02	0.24	0.00	0.00	7.50	0.00	0.00
L (luminosidad)	44.36	1.30	44.42	40.98	47.88	43.50	45.20
a* (rojo/verde)	0.97	0.38	0.95	-0.26	2.41	0.73	1.22
b* (azul/amarillo)	14.92	1.13	14.72	12.09	20.10	14.13	15.61
Color café tostado	119.27	6.66	120.30	100.20	141.90	116.80	123.00
Defectos visuales	10.45	3.83	10.00	1.60	31.90	7.60	12.70
Defectos de sabor (PA*)	0.14	0.34	0.00	0.00	1.00	0.00	0.00

Nota: Q1 = primer cuartil (percentil 25 %); Q3 = tercer cuartil (percentil 75 %); PA = Presencia o ausencia de defectos de sabor en taza.

5.1.1 Humedad (% g/100 g). De acuerdo con los resultados obtenidos en la Tabla 5, el contenido de humedad de los granos de café presentó una media de 11,57 % con una desviación estándar (DE) igual al 0,42 %, con valores comprendidos en un intervalo entre 10,50 % y 12,70 %. La baja variabilidad observada (coeficiente de variación igual a 3,6 %) indica buenas condiciones de almacenamiento de los lotes y procesamiento de las muestras de ensayo. Los valores de humedad en todo el conjunto de datos se encuentran dentro del rango óptimo establecido para café verde de exportación, donde se especifica rango de tolerancia de 10,0 % a 12 % (Duarte et al., 2022; SCA, 2021). La distribución de los datos muestra una leve asimetría positiva, evidenciada por una media de 11,57 % ligeramente superior frente a la mediana reportada de 11,50 %. La humedad promedio de todo el conjunto de datos se encuentra dentro del rango establecido, lo que previene el desarrollo de microorganismos (Oliveros et al., 2020).

Lo anterior, se puede observar a través del histograma y el diagrama de caja presentados en la Figura 9 A) y B), donde los valores se concentran predominantemente en un rango entre el 11,0 % y el 12,0 %, con una mediana situada en el 11,5 % de humedad. El tratamiento de datos mediante el IQR permitió mitigar el impacto de valores atípicos, dando como resultando un conjunto de datos con una distribución normalizada.

La comparación interanual mediante diagramas de violín (Figura 9, C) confirma la consistencia del ensayo a través del tiempo, mostrando que tanto para el 2024 como en 2025, la probabilidad de densidad y la varianza se mantuvieron sin cambios significativos. Sin embargo, el análisis de la serie temporal (Figura 9, D) revela una variación estacionalidad sistemática. Se identifican picos máximos de humedad cercanos al 12,0 % durante el periodo comprendido entre mayo y julio para cada año, seguidos de descensos hacia un valor de aproximadamente 11,3 % para los meses de enero a marzo. Finalmente, se observa una tendencia de recuperación hacia el cierre de 2025, lo que refuerza la naturaleza recurrente y predecible de esta variable en el contexto estudiado. Lo anterior puede ser el resultado de los diferentes cambios climáticos durante los periodos de tiempo estudiados, lo que da como resultado variaciones en las condiciones ambientales de almacenamiento del café verde.

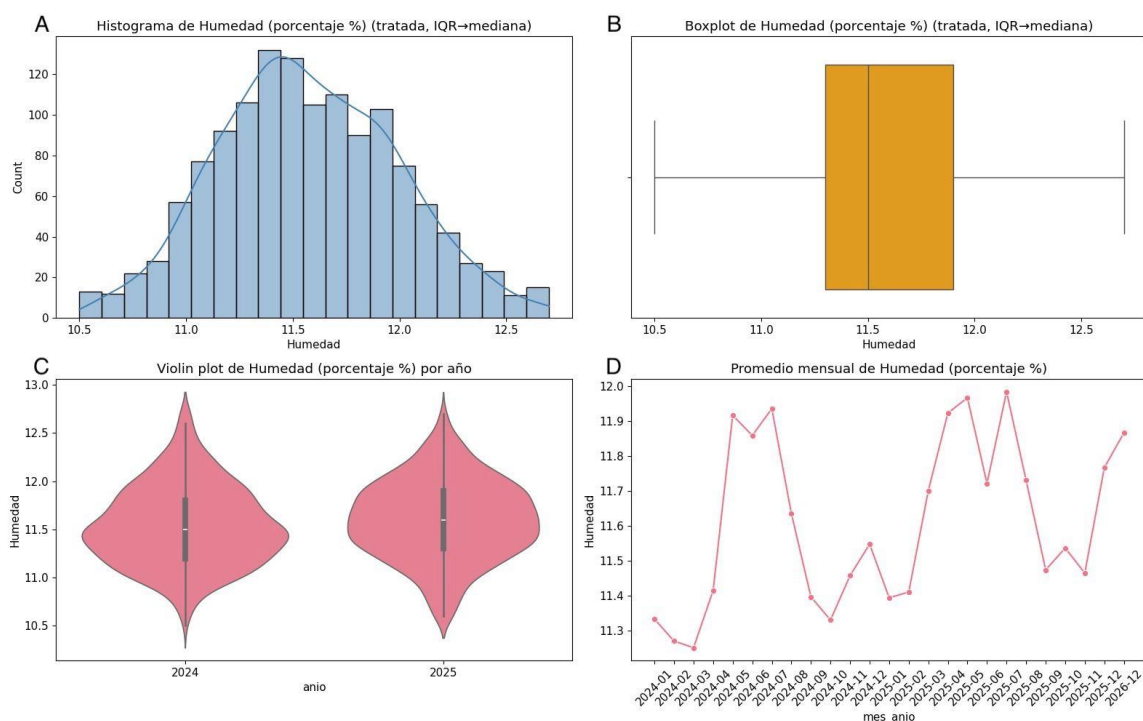


Figura 9. Análisis estadístico descriptivo y evolución temporal de la humedad (%).

Nota: (A) Histograma de frecuencias con ajuste de curva de densidad; (B) Diagrama de caja (*boxplot*) que describe la dispersión y medidas de tendencia central; (C) Diagrama de violín comparativo interanual; (D) Evolución del promedio mensual a lo largo del periodo de estudio. Los datos presentados fueron sometidos a un tratamiento previo de valores atípicos utilizando el método de rango intercuartílico (IQR) con ajuste hacia la mediana.

5.1.2 Densidad real (g/cm^3). De acuerdo con los resultados obtenidos en la Tabla 5, la variable densidad real arrojó una media de $1,3500 \text{ g/cm}^3$ con una DE igual a $0,0100 \text{ g/cm}^3$ en un rango de medición de $1,3100 \text{ g/cm}^3$ a $1,3900 \text{ g/cm}^3$. La escasa variabilidad en este ensayo de medición (coeficiente de variación igual al 0,74 %) sugiere una

homogeneidad en la estructura celular y el grado de desarrollo de los granos para todos los lotes evaluados. Valores de densidad superiores a $1,3000 \text{ g/cm}^3$ son indicativos de granos bien desarrollados y con adecuada compactación celular (Franca & Oliveira, 2019).

Por otro lado, la Figura 10 revela una distribución con un sesgo negativo marcado, donde la mayor concentración de las observaciones se sitúa en el rango superior del conjunto de datos evaluado. Lo anterior se puede apreciar a través del histograma de la Figura 10, A), donde el pico de frecuencia se localiza entre $1,3500 \text{ g/cm}^3$ y $1,3600 \text{ g/cm}^3$, con una cola de distribución que se extiende hacia valores inferiores de aproximadamente $1,3100 \text{ g/cm}^3$. La mediana se posiciona de forma en un valor de $1,355 \text{ g/cm}^3$, observándose en el diagrama de caja la persistencia de algunos valores atípicos en los extremos inferior y superior a pesar del tratamiento estadístico de los datos. Lo anterior puede sugerir cambios en la calidad de los lotes de café que puede ser el resultado de los cambios climáticos o de procesamiento durante los periodos de tiempo evaluados.

En términos de estabilidad por año, los diagramas de violín (Figura 10, C) demuestran que la estructura de la densidad fue sumamente consistente entre los periodos 2024 y 2025, manteniendo perfiles de probabilidad casi idénticos con una densidad de datos máxima en el intervalo de $1,35 \text{ g/cm}^3$ a $1,36 \text{ g/cm}^3$.

A diferencia de la estabilidad general, la serie temporal de promedios mensuales (Figura 10, D) exhibe una dinámica fluctuante con variaciones estacionales significativas.

Durante el año 2024, se identificó un máximo relativo en marzo de aproximadamente $1,358 \text{ g/cm}^3$ seguido de un descenso en diciembre de 2024. En el ciclo 2025, el pico máximo se desplazó hacia el mes de junio, alcanzando valores de $1,358 \text{ g/cm}^3$, para posteriormente iniciar una tendencia descendente sostenida que culmina en el punto más bajo del registro histórico para diciembre de 2025, con un valor aproximado de $1,348 \text{ g/cm}^3$. Esta transición sugiere un cambio en las condiciones del sistema hacia el final del periodo de estudio que contrasta con la estabilidad observada en los años previos.

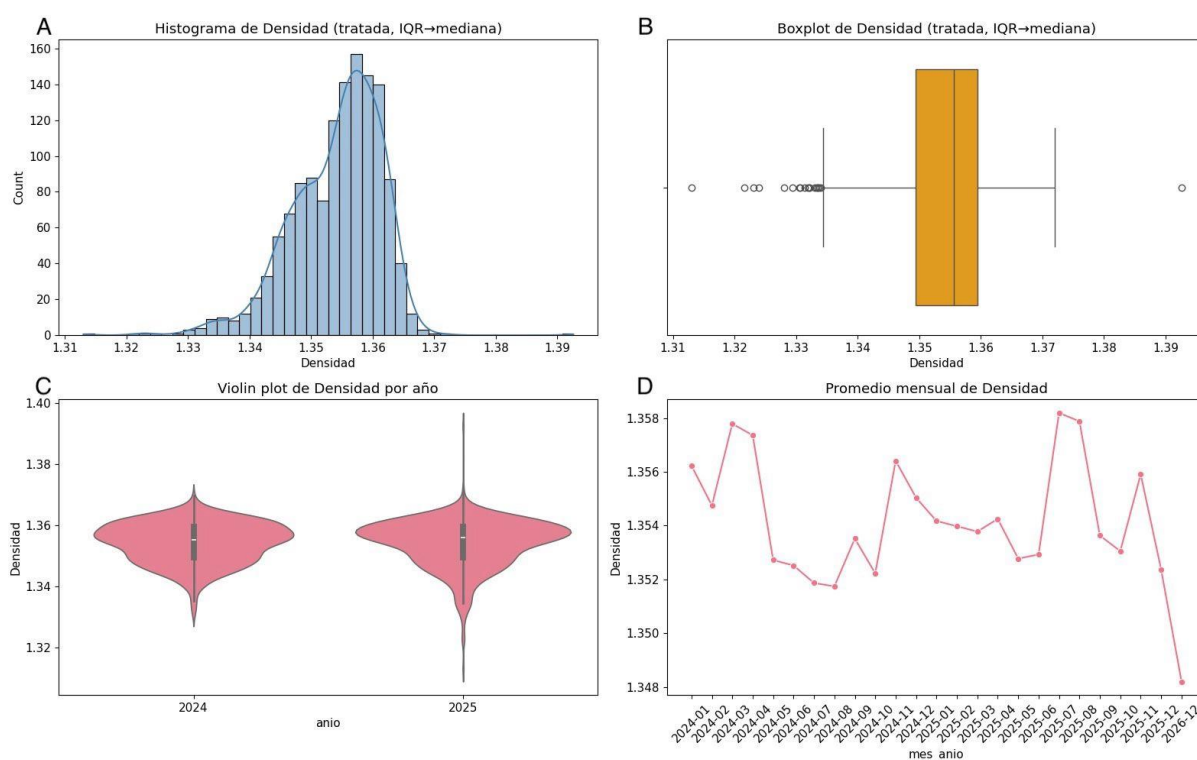


Figura 10. Análisis estadístico descriptivo y evolución temporal de la densidad real g/cm^3 .

Nota: (A) Histograma de frecuencias con ajuste de densidad; (B) Diagrama de caja (boxplot) con identificación de medidas de dispersión; (C) Comparativa de la distribución anual mediante diagramas de violín (2024-2025); (D) Tendencia del promedio mensual de la densidad. Los datos fueron tratados mediante el método de rango intercuartílico (IQR) con ajuste hacia la mediana.

5.1.3 Distribución del tamaño del grano. El análisis granulométrico reveló que la mayor proporción de retención del grano se corresponde a los tamices con diámetro de abertura de 18/64 in (54,91 % de retención), 17/64 in (26,45 % de retención) y 16/64 in (16,66 % de retención), lo cual da como resultado un porcentaje acumulando del 97,5 % del total de las muestras de ensayo. Los tamices de menor diámetro de tamaño (15/64 in a 12/64 in) arrojaron porcentajes con medias inferiores al 2,5 % en todos los casos.

La distribución de tamaño del grano mostró un comportamiento aproximadamente normal en los tamices de ensayo con mayor retención de granos, lo cual se evidencia por la cercanía entre las medias y las medianas. El tamiz de ensayo número 18/64 in presentó la mayor variabilidad con una DE igual al 6,06 %, lo que indica diferencias significativas en la proporción de granos grandes entre las diferentes muestras de ensayo evaluadas. Estos resultados son consistentes para cafés de especialidad, donde la predominancia en la retención de tamices de ensayo de números 16/64 in a 18/64 in se asocia con granos bien desarrollados (Rojas et al, 2023).

Adicionalmente, el análisis de las series temporales (Figura 11, 12, 13, 14, 15, 16 y 17) revela dinámicas contrastantes entre los tamices de ensayo con mayor retención frente a los de menor retención a lo largo del periodo de estudio. Por otro lado, los tamices de ensayo de diámetro número 16 /64 in y 17/64 in mostraron una tendencia general al descenso hacia el segundo semestre de 2025, alcanzando mínimos de aproximadamente 12,5 % y 21,5 % respectivamente. Para el caso del tamiz de ensayo número 18/64 in experimentó un incremento significativo, logrando un pico máximo de retención del 65,5 % en noviembre de 2025. Este comportamiento sugiere un desplazamiento de la distribución hacia granos de tamaño superior durante el periodo evaluado.

Sin embargo, los registros finales correspondientes a diciembre de 2025 muestran una inversión de esta tendencia. En las figuras se observa una caída abrupta en la retención del tamiz de ensayo número 18/64 in, obteniendo un descendiendo de aproximadamente 48,7 %, lo cual coincide con un repunte inusual en los tamices de ensayo de menor diámetro (14/64 in y 15/64 in). Para el tamiz de ensayo número 15/64 in, mantuvo promedios bajos durante la mayor parte del tiempo, alcanzando un máximo de 16,5 % al cierre de 2025, mientras que para el tamiz 14/64 in logró un ascenso del 5,8 %.

A pesar de estas variaciones mensuales, los diagramas de violín en todas las figuras confirman que la estructura de probabilidad por año para los tamices de ensayo de mayor diámetro (16/64 in, 17/64 in a 18/64 in) se mantuvo consistente entre 2024 y 2025. Esta estabilidad entre años permite validar la homogeneidad de la calidad física de los lotes y

sus muestras de ensayo en términos de desarrollo del grano, independientemente de las variaciones observadas en los promedios mensuales.

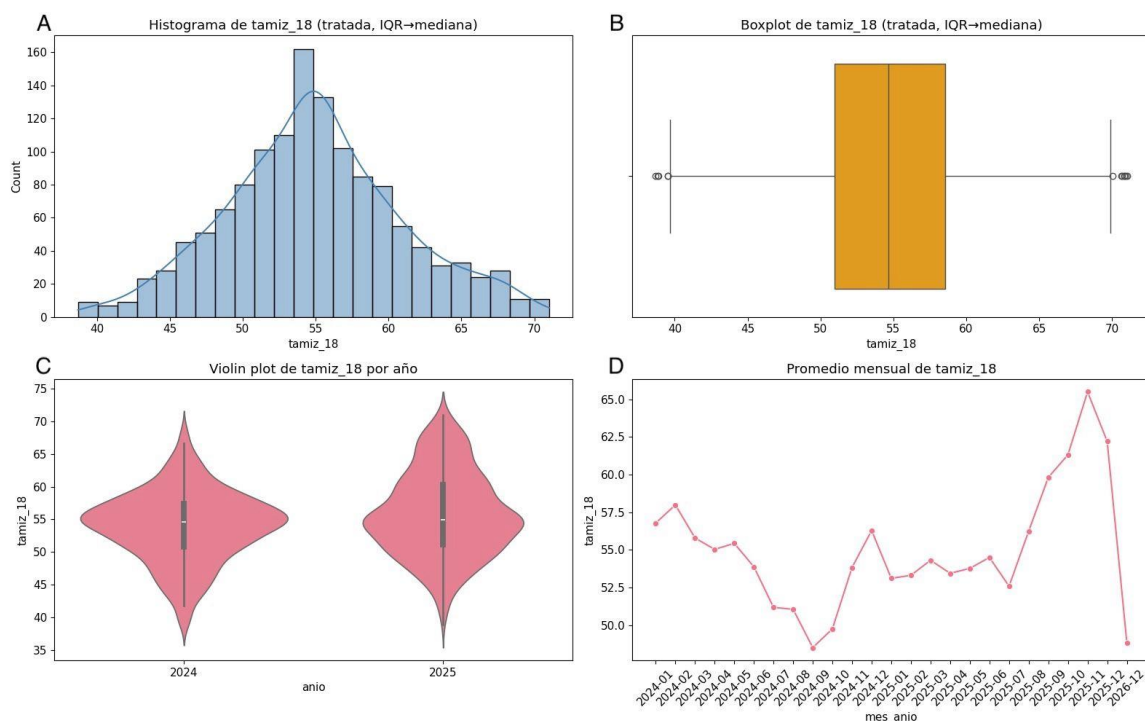


Figura 11. Caracterización estadística y evolución temporal de la distribución granulométrica por el tamiz 18/64 in.

Nota: (A) Histogramas de frecuencias y curvas de densidad; (B) Diagramas de caja (*boxplots*) para la identificación de dispersión; (C) Comparativa interanual mediante diagramas de violín; (D) Tendencia del promedio mensual de retención por tamiz. Los datos fueron ajustados mediante el método de rango intercuartílico (IQR).

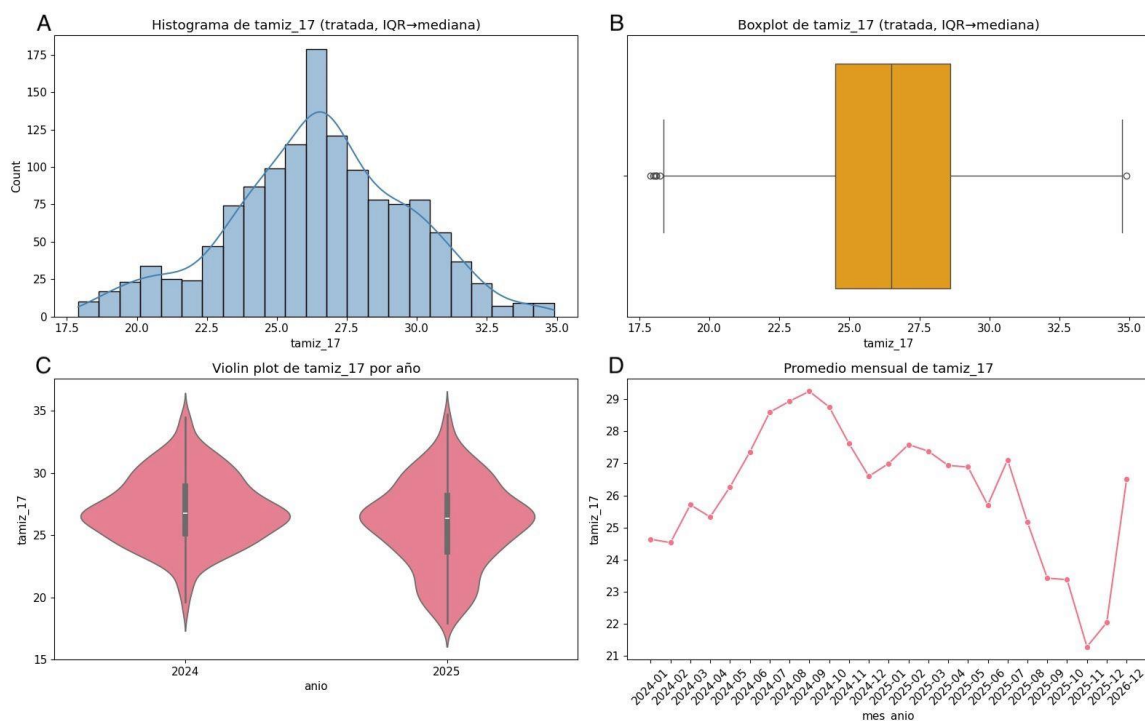


Figura 12. Caracterización estadística y evolución temporal de la distribución granulométrica por el tamiz 17/64 in.

Nota: (A) Histogramas de frecuencias y curvas de densidad; (B) Diagramas de caja (*boxplots*) para la identificación de dispersión; (C) Comparativa interanual mediante diagramas de violín; (D) Tendencia del promedio mensual de retención por tamiz. Los datos fueron ajustados mediante el método de rango intercuartílico (IQR).

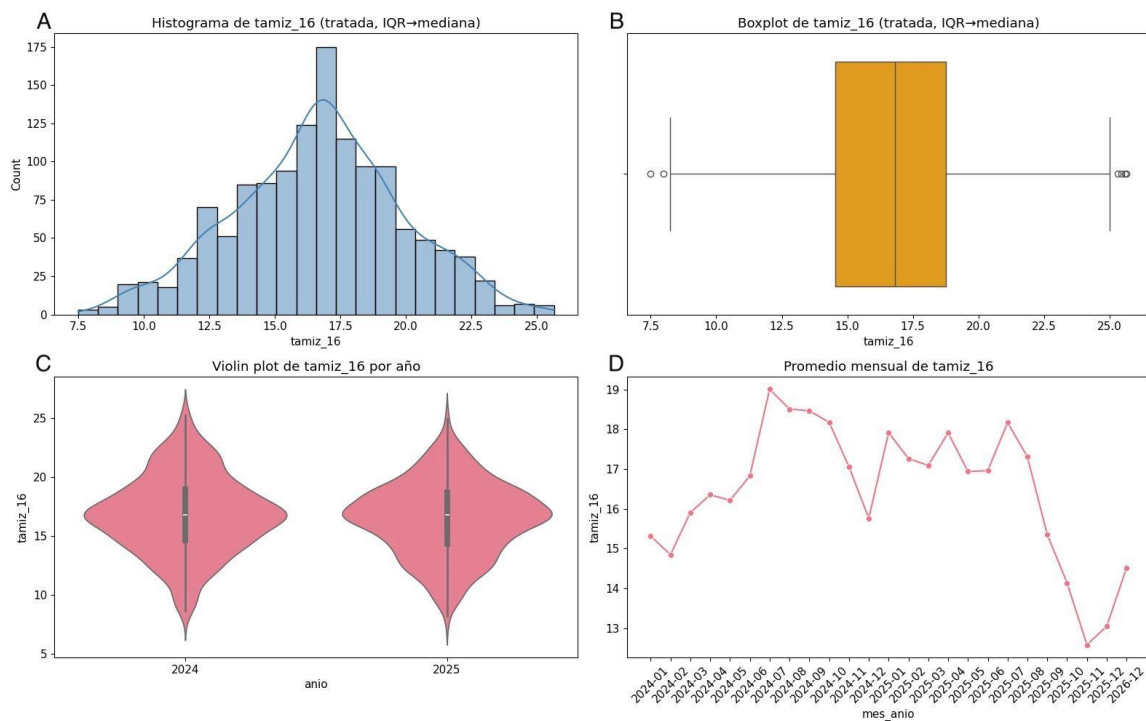


Figura 13. Caracterización estadística y evolución temporal de la distribución granulométrica por el tamiz 16/64 in (2024-2026).

Nota: (A) Histogramas de frecuencias y curvas de densidad; (B) Diagramas de caja (*boxplots*) para la identificación de dispersión; (C) Comparativa interanual mediante diagramas de violín; (D) Tendencia del promedio mensual de retención por tamiz. Los datos fueron ajustados mediante el método de rango intercuartílico (IQR).

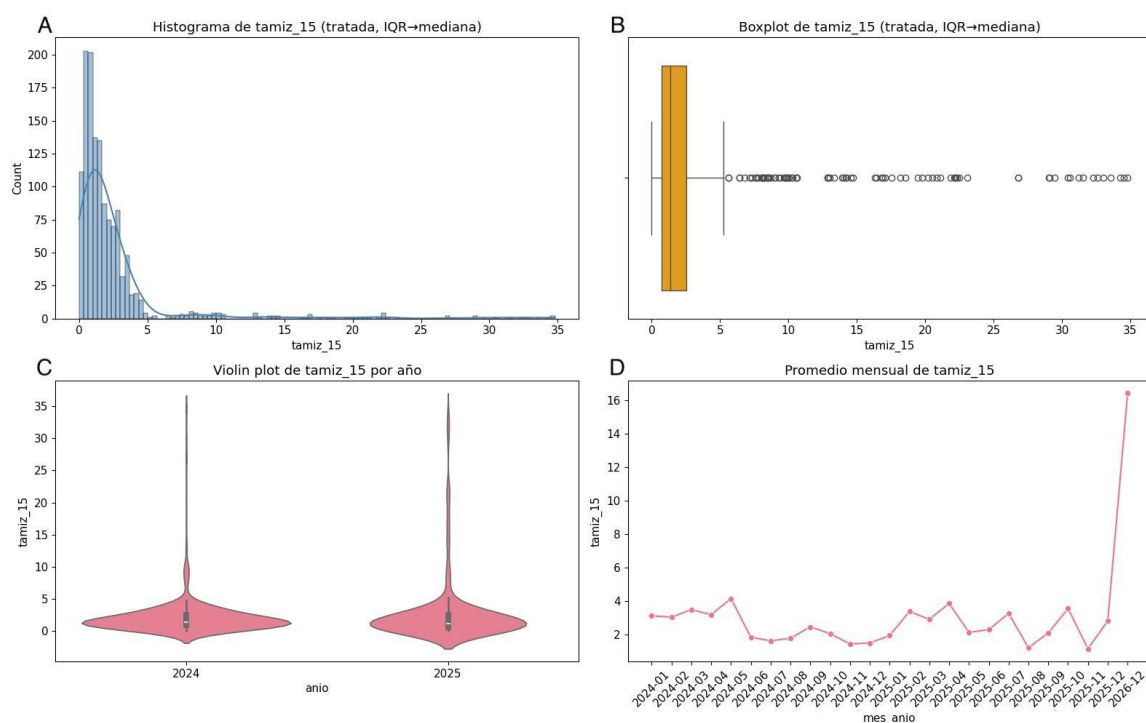


Figura 14. Caracterización estadística y evolución temporal de la distribución granulométrica por el tamiz 15/64 in.

Nota: (A) Histogramas de frecuencias y curvas de densidad; (B) Diagramas de caja (*boxplots*) para la identificación de dispersión; (C) Comparativa interanual mediante diagramas de violín; (D) Tendencia del promedio mensual de retención por tamiz. Los datos fueron ajustados mediante el método de rango intercuartílico (IQR).

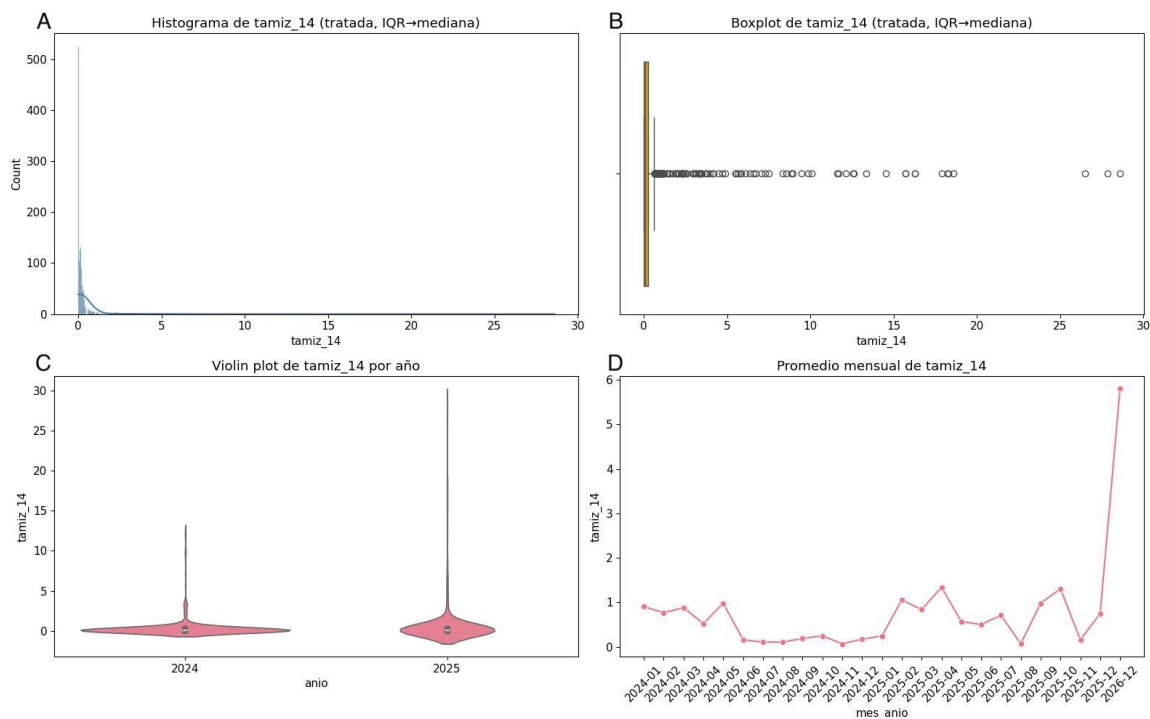


Figura 15. Caracterización estadística y evolución temporal de la distribución granulométrica por el tamiz 14/64 in (2024-2026).

Nota: (A) Histogramas de frecuencias y curvas de densidad; (B) Diagramas de caja (*boxplots*) para la identificación de dispersión; (C) Comparativa interanual mediante diagramas de violín; (D) Tendencia del promedio mensual de retención por tamiz. Los datos fueron ajustados mediante el método de rango intercuartílico (IQR).

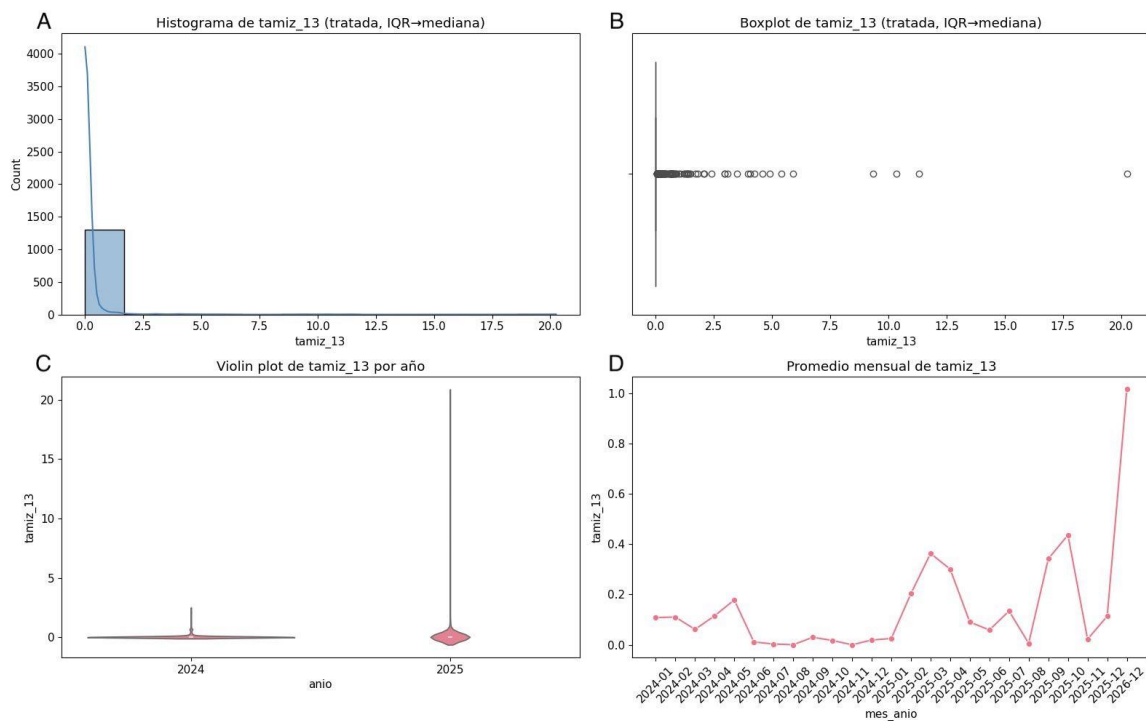


Figura 16. Caracterización estadística y evolución temporal de la distribución granulométrica por el tamiz 13/64 in.

Nota: (A) Histogramas de frecuencias y curvas de densidad; (B) Diagramas de caja (*boxplots*) para la identificación de dispersión; (C) Comparativa interanual mediante diagramas de violín; (D) Tendencia del promedio mensual de retención por tamiz. Los datos fueron ajustados mediante el método de rango intercuartílico (IQR).

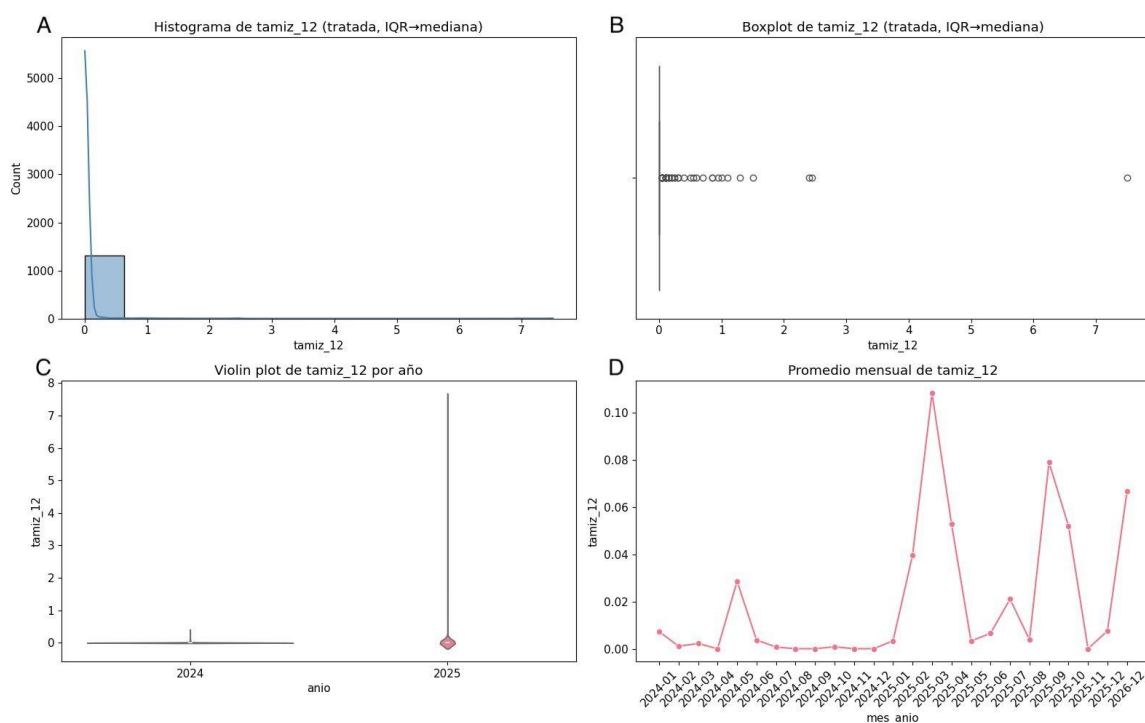


Figura 17. Caracterización estadística y evolución temporal de la distribución granulométrica por el tamiz 12/64 in.

Nota: (A) Histogramas de frecuencias y curvas de densidad; (B) Diagramas de caja (boxplots) para la identificación de dispersión; (C) Comparativa interanual mediante diagramas de violín; (D) Tendencia del promedio mensual de retención por tamiz. Los datos fueron ajustados mediante el método de rango intercuartílico (IQR).

5.1.4 Análisis del color del café verde. Las coordenadas de color en el espacio CIE L ab* mostraron valores medios de L igual a 44,36 con una DE igual a 1,30, a* igual a 0,97 con una DE igual a 0,38 y b* igual a 14.92 con una DE = 1,13.

El valor de luminosidad (L) de 44,36 indica granos de color verde medio, típico de café con proceso de lavado (Correa et al., 2022). La coordenada a^* positiva (0,97) sugiere una ligera tendencia hacia tonalidades rojizas, asociada comúnmente con granos maduros y bien desarrollados. El alto valor de la coordenada b^* en 14,92 refleja la dominancia de pigmentos amarillos característicos de los cafés verdes de calidad.

La baja variabilidad observada en las tres coordenadas en el espacio de color (coeficientes de variación para L igual a 2,9 %, a^* igual a 39,2 % y b^* igual a 7,6 %) indica que el color del café verde puede ser un discriminador útil entre lotes de diferente calidad o procedencia.

El análisis de distribución refuerza estas observaciones, teniendo en cuenta que para la coordenada de luminosidad (L) se observa un comportamiento unimodal y aproximadamente simétrico centrado en una mediana de 44,5 unidades, mientras que las coordenadas de color a^* y b^* se observa distribuciones con un ligero sesgo positivo. Este sesgo, particularmente visible en la coordenada de color b^* , indica la persistencia de muestras con una saturación de pigmentos amarillos por encima del promedio, como se evidencia en los valores atípicos superiores registrados en el diagrama de caja.

Por otro lado, las series temporales para todo el conjunto de resultados de color revelan variaciones estacionales que impactan la estabilidad del color. Mientras que para la luminosidad (L) y la coordenada de color rojo (a^*) mostraron patrones de variabilidad no habituales por un descenso marcado en el mes de noviembre de 2025 donde el

parámetro L descendió a 43,3 unidades y el parámetro a^* a 0,68 unidades, la coordenada b^* experimentó una transición significativa. A partir de mayo de 2025, se observa un incremento sostenido y pronunciado en los valores del parámetro b^* , los cuales pasaron de una media de 14,5 unidades para el año 2024 y alcanzar un máximo de 17,15 unidades en diciembre de 2025.

Esta tendencia ascendente en la intensidad de los pigmentos amarillos se puede confirmar en la Figura 19, B), donde el diagrama de violín para el año 2025 muestra un desplazamiento ascendente de la densidad de probabilidad en comparación con los resultados obtenidos para el 2024. Por el contrario, las coordenadas de color de L y a^* permanecieron relativamente consistentes entre ambos años, sugiriendo que, aunque la claridad y los tonos rojizos del café se mantienen estables, existe un enriquecimiento progresivo en la coloración amarilla de los lotes hacia el final del periodo evaluado. La presencia de estos pigmentos amarillos puede ser el resultado de almacenamiento prolongado de los lotes evaluados.

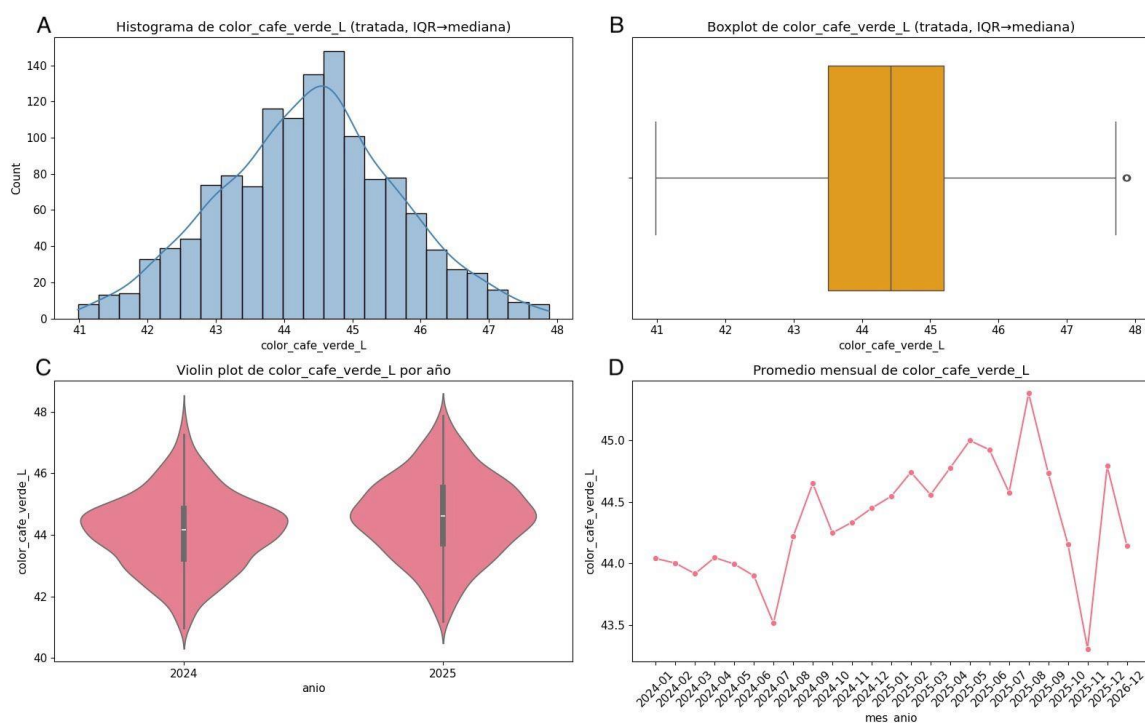


Figura 18. Medición de color bajo la escala CIE en café verde para la coordenada L.

Nota: (A) Histogramas de frecuencias con ajuste de densidad; (B) Diagramas de caja (*boxplots*) para identificación de medidas de dispersión; (C) Comparativa interanual mediante diagramas de violín; y (D) Evolución del promedio mensual. Los datos fueron tratados mediante el método de rango intercuartílico (IQR) con ajuste hacia la mediana.

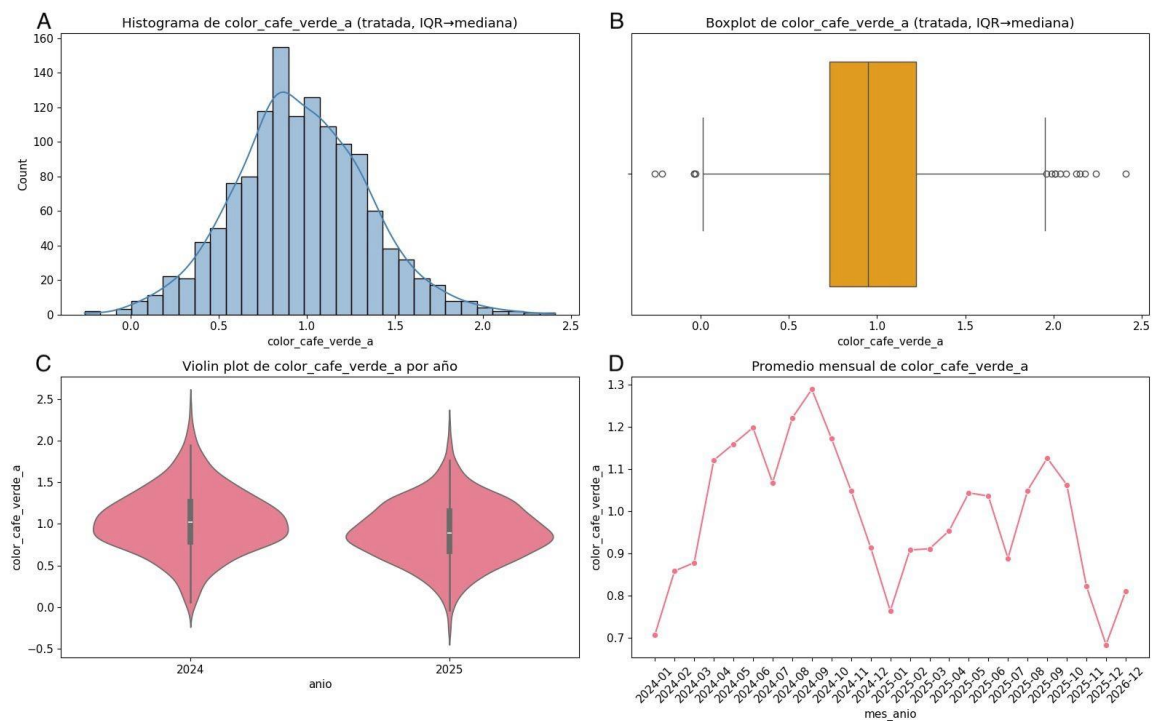


Figura 19. Medición de color bajo la escala CIE en café verde para la coordenada a*.

Nota: (A) Histogramas de frecuencias con ajuste de densidad; (B) Diagramas de caja (*boxplots*) para identificación de medidas de dispersión; (C) Comparativa interanual mediante diagramas de violín; y (D) Evolución del promedio mensual. Los datos fueron tratados mediante el método de rango intercuartílico (IQR) con ajuste hacia la mediana.

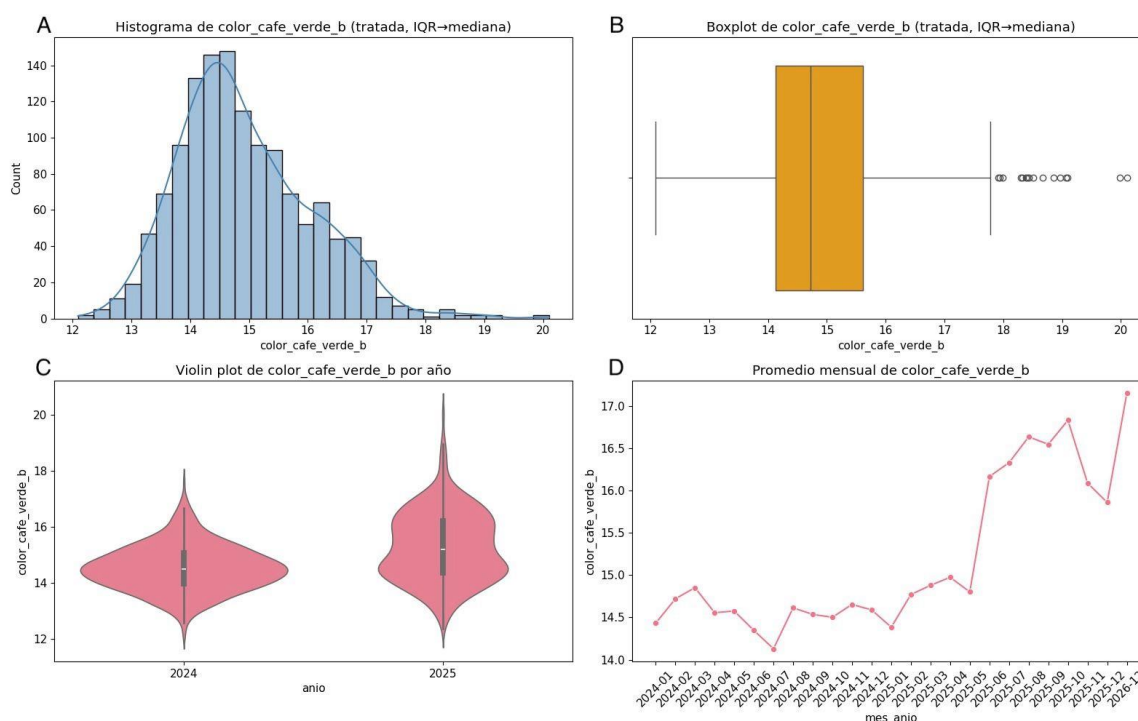


Figura 20. Medición de color bajo la escala CIE en café verde para la coordenada b^* .

Nota: (A) Histogramas de frecuencias con ajuste de densidad; (B) Diagramas de caja (*boxplots*) para identificación de medidas de dispersión; (C) Comparativa interanual mediante diagramas de violín (2024-2025); y (D) Evolución del promedio mensual. Los datos fueron tratados mediante el método de rango intercuartílico (IQR) con ajuste hacia la mediana.

5.1.5 Cuantificación de defectos visuales. La variable cuantificación de defectos por análisis visual presentó una media de 10,45 faltas totales con una DE igual a 3,83 faltas totales, y valores extremos de 1,60 faltas totales y 31,90 faltas totales. Los resultados obtenidos arrojaron una distribución con una asimetría positiva teniendo en cuenta la

media mayor a la mediana igual a 10,00 faltas totales, lo anterior sugiere la presencia de valores atípicos elevados o calidad diferente en los lotes evaluadas.

Por otro lado, el análisis por cuartiles revela que el 25 % de las muestras presentó valores inferiores a 7,60 defectos totales lo cual corresponde a una calidad del café verde excelente, mientras que el 25 % superior superó los 12,70 defectos totales. La alta variabilidad observada de acuerdo con el coeficiente de variación obtenido de 36,7 %, sugiere diferencias significativas en la calidad visual de los granos entre los diferentes lotes analizados durante los periodos de tiempo del estudio.

El histograma de frecuencias (Figura 21.A) corrobora la asimetría positiva mencionada, ya que permite evidenciar una alta concentración de muestras con menos de 15 faltas totales, pero con una cola de distribución que se extiende significativamente hacia el espectro superior. En el caso del diagrama de caja identifica que, tras el tratamiento de datos, persisten múltiples valores atípicos que superan el umbral de las 20 faltas totales, alcanzando el extremo máximo de 31,90 faltas totales (Figura 21, B), lo que subraya la existencia de lotes específicos con una degradación física atípica o muestras de ensayo con calidad del café diferente.

El análisis de la dinámica temporal (Figura 21, D) revela una estacionalidad crítica en la calidad visual. Tras un inicio en el periodo de 2024 con niveles bajos de defectos con un promedios de aproximadamente 7,0 faltas totales, se observa un deterioro progresivo que culminó en un pico máximo histórico de 14,0 faltas mensuales para mayo de 2025.

Esta tendencia al alza se ve reflejada en el diagrama de violín de 2025 (Figura 21, C), el cual muestra un mayor volumen de densidad de datos en rangos superiores (10 faltas totales a 15 faltas totales) en comparación con los resultados obtenidos en el periodo de 2024.

Posterior al pico de mayo de 2025, se registró una mejora sustancial y abrupta en la apariencia física de los granos, con promedios que descendieron hasta estabilizarse en un rango de 7,0 faltas totales a 8,3 faltas totales durante el último trimestre de 2025. Este comportamiento final sugiere una optimización en los procesos de selección o una mejora en las condiciones de los lotes hacia el cierre del periodo evaluado.

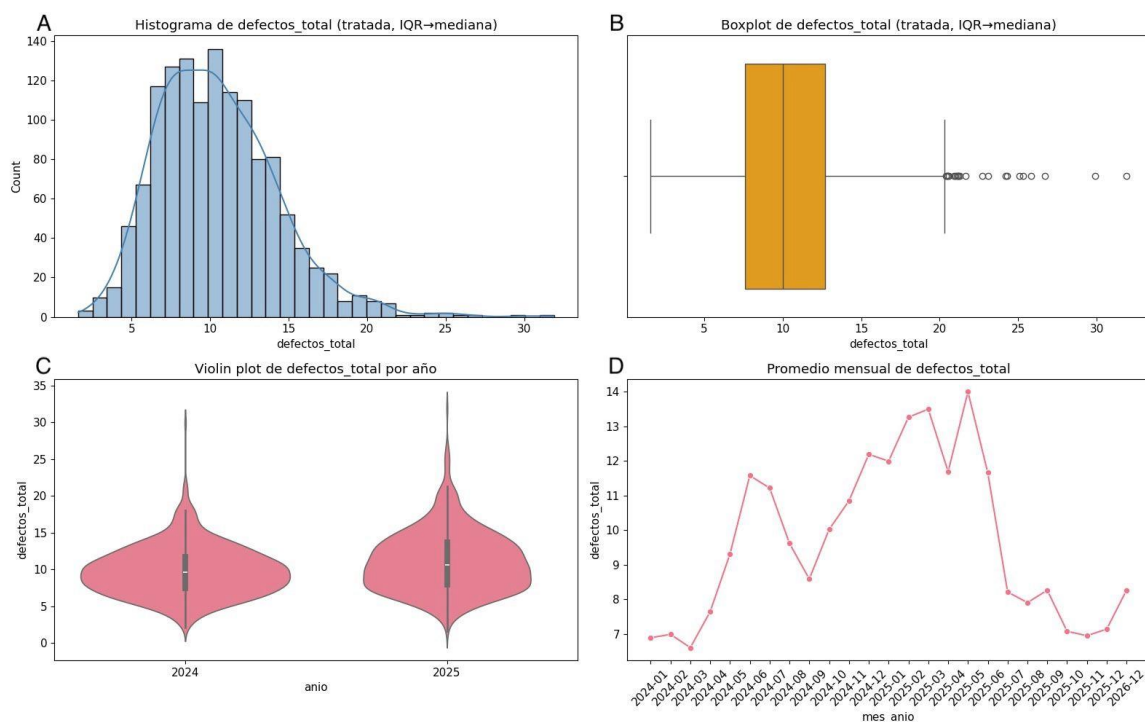


Figura 21. Análisis estadístico y comportamiento temporal de la cuantificación de defectos por análisis visual.

Nota: (A) Histograma de frecuencias con ajuste de densidad; (B) Diagrama de caja (*boxplot*) con identificación de valores atípicos; (C) Comparativa interanual mediante diagramas de violín (2024-2025); (D) Evolución del promedio mensual de defectos totales. Datos tratados mediante el método de rango intercuartílico (IQR) con ajuste hacia la mediana.

5.1.6 Defectos de sabor. Los defectos sensoriales evaluados mediante la variable binaria de defectos de sabor en taza presentaron una media de 0,14 con una DE igual a 0,34 en la presencia de sabores atípicos, el 75 % de las muestras de ensayo arrojaron valores de 0,0 el cual corresponde a muestras con ausencia de defectos de sabor perceptibles. Solo el 25 % de las muestras presentó defecto sensorial en el sabor, con un valor máximo de 1,0. Esta baja presencia de defectos de sabor sugiere que, a pesar de la variabilidad en defectos visuales en el grano de café, la calidad sensorial general de los lotes evaluados se mantuvo dentro de estándares aceptables.

El análisis gráfico en la identificación de defectos de sabor confirma la naturaleza bimodal y altamente sesgada de esta variable. El histograma (Figura 22, A) y el diagrama de caja (Figura 22, B) evidencian que la gran mayoría de las observaciones se sitúan en el valor 0 (ausencia), clasificando cualquier registro de defectos de sabor (valor igual a 1,0) como un valor atípico dentro del conjunto de datos tratados. Esta estructura indica que la

presencia de defectos no es gradual, sino que ocurre como un evento discreto y poco frecuente en todo el conjunto de lotes evaluados.

En términos de estabilidad temporal, los diagramas de violín (Figura 22, C) muestran una consistencia absoluta entre los años 2024 y 2025, con perfiles de distribución idénticos que mantienen la mayor densidad de probabilidad en la ausencia total de defectos. No obstante, el seguimiento mensual de los promedios (Figura 22, D) permite identificar periodos de mayor incidencia sensorial de defectos de sabor. Se observan dos picos máximos relevantes: el primero en agosto de 2024 (0,25) y el segundo en marzo de 2025 (aproximadamente 0,25).

Es notable que, tras el último incremento registrado en septiembre de 2025 (0,20), la variable muestra una tendencia descendente sostenida durante el último trimestre de 2025 y todo el año 2026, alcanzando un valor de cero defectos en diciembre de 2026. Este comportamiento final refuerza la conclusión de que los lotes mantienen una alta calidad sensorial, con una presencia de defectos de sabor que, además de ser baja, tiende a la desaparición hacia el cierre del periodo de estudio.

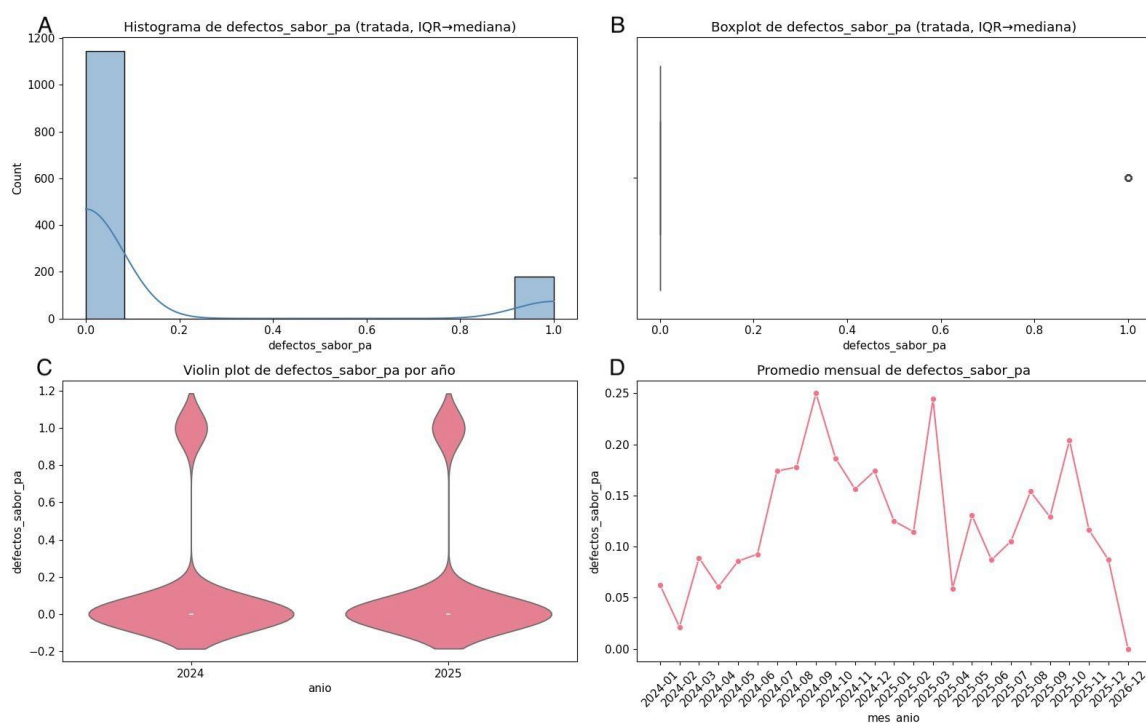


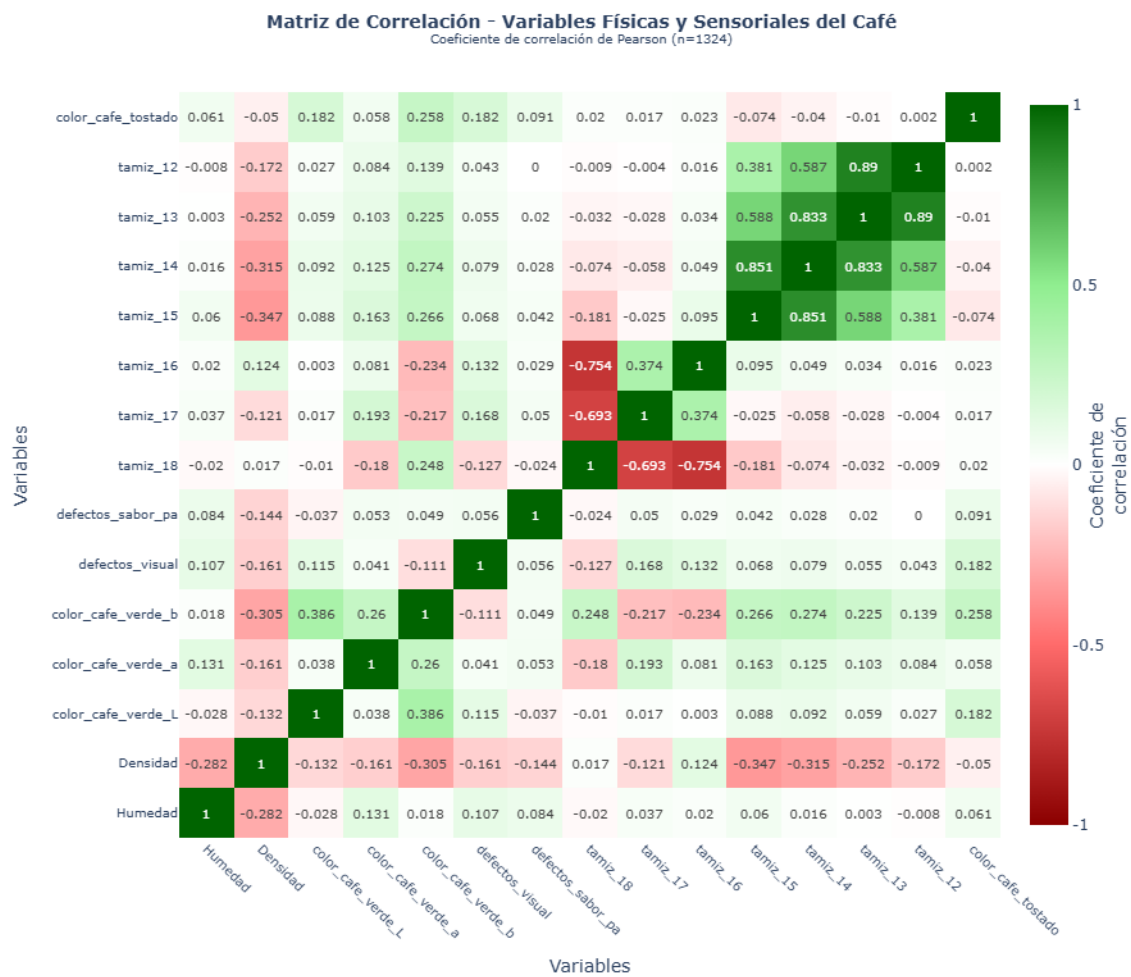
Figura 22. Análisis estadístico y comportamiento temporal de los defectos de sabor en taza.

Nota: (A) Histograma de frecuencias que muestra la naturaleza bimodal de la variable; (B) Diagrama de caja (*boxplot*) que resalta la prevalencia de valores nulos; (C) Comparativa interanual mediante diagramas de violín; (D) Evolución del promedio mensual de la variable defectos de sabor. Datos tratados mediante el método de rango intercuartílico (IQR) con ajuste hacia la mediana.

5.2 Análisis de correlaciones

El análisis de las correlaciones se realizó mediante el estadístico de Pearson para las variables de evaluación física y sensorial del café, en general, el análisis de correlación de

Pearson dio como resultado asociaciones débiles para las variables sensoriales con el resto de variables, lo que sugiere que la aparición de defectos puede estar influenciada por factores externos no capturados totalmente por estas métricas individuales.



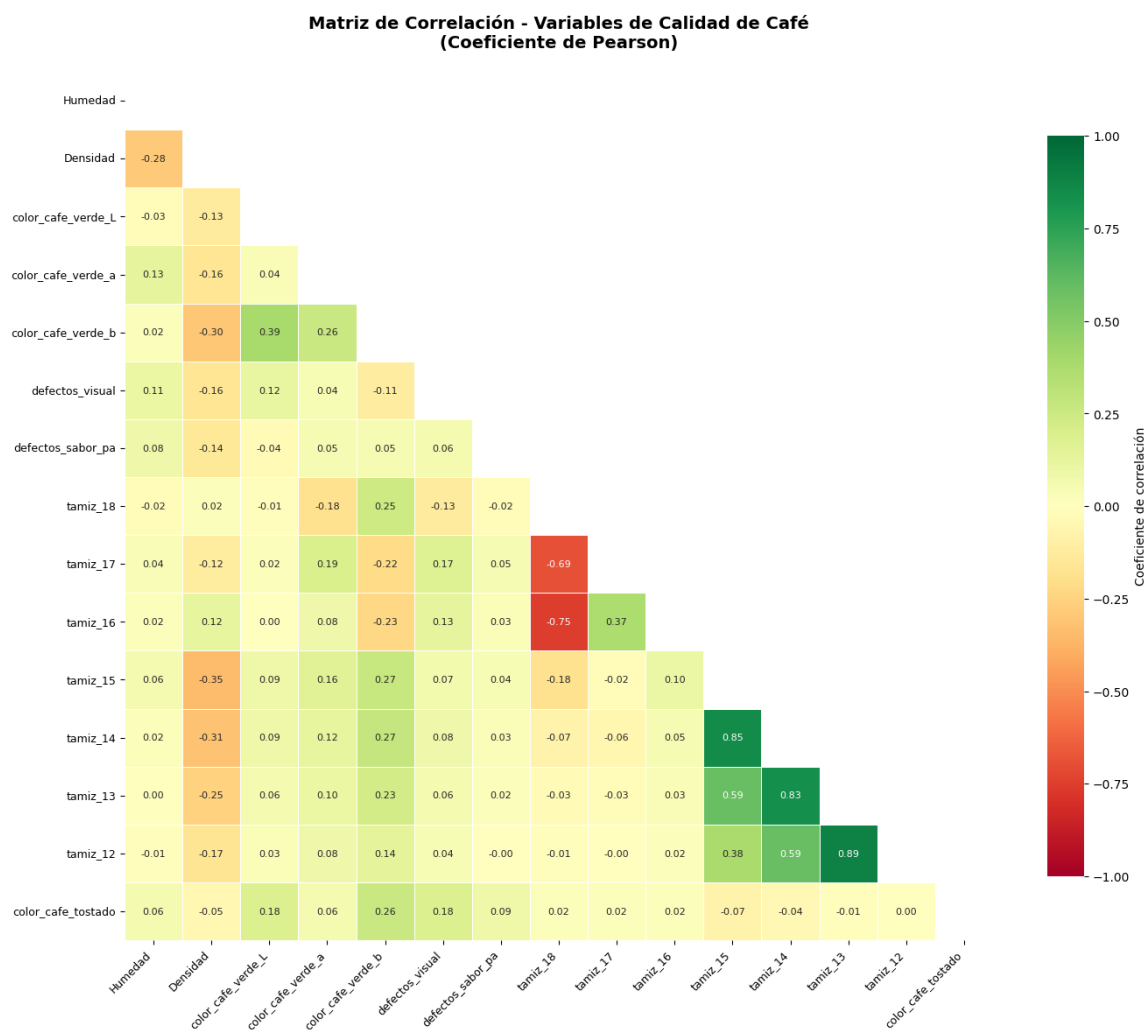


Figura 23. Matriz de correlación para las variables físicas y sensoriales del café.

5.2.1 Análisis de correlaciones para la variable cuantificación de defectos por

análisis visuales. La variable cuantificación de defectos por análisis visual muestra las siguientes correlaciones significativas:

Correlaciones Positivas: los vínculos más fuertes se presentan con la variable color para café tostado (0,182) y los tamices de tamaño medio, específicamente el tamiz

número 17/64 in (0,168) y el tamiz número 16/64 in (0,132). Esto indica una mayor presencia de defectos visuales del grano junto con cambios en la coloración del café tostado y granos de café verde de tamaño medio. También existe una correlación positiva leve con la humedad (0,107).

Correlaciones Negativas: se observa una correlación inversa relevante con la densidad real (-0,161) y el tamiz número 18/64 in (-0,127). Estos datos sugieren que los lotes con mayor densidad y proporción de granos de tamaño superior (granos retenidos por el tamiz número 18/64 in) tienden a presentar una menor cantidad de defectos visuales.

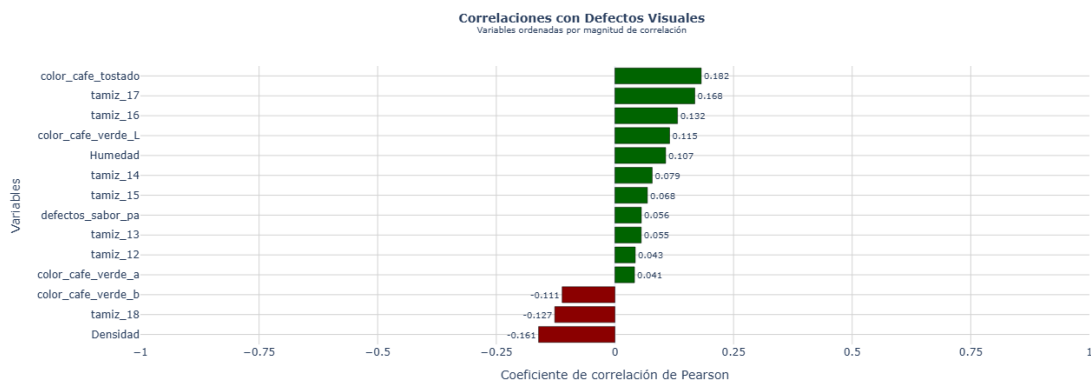


Figura 24. Correlaciones más significativas con la variable objetivo defectos por análisis visual.

5.2.2 Análisis de correlaciones de defectos de Sabor en taza (Presencia o ausencia)

Esta variable presenta las magnitudes de correlación más bajas del estudio, lo que confirma su baja prevalencia y naturaleza aleatoria:

Relación con la Densidad real: presenta su correlación negativa más destacada con la densidad real (-0,144). Al igual que con los defectos visuales, una menor densidad del grano parece estar ligeramente vinculada a una mayor probabilidad de encontrar defectos sensoriales en taza.

Otras asociaciones: muestra vínculos positivos casi insignificantes con la variable de café para café tostado (0,091) y la humedad (0,084). Es notable que no presenta correlación alguna con el tamiz número 12/64 in (0.0).

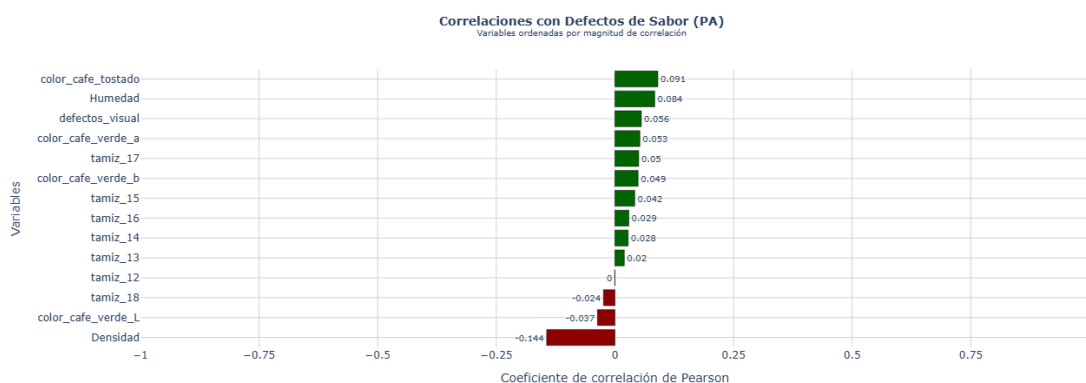


Figura 25. Correlaciones más significativas con la variable defectos de sabor en taza.

5.3 Interrelación entre la presencia de defectos visuales y de sabor.

La correlación entre las variables “defectos por análisis visual” y “defectos de sabor en taza” es positiva pero extremadamente baja (0,056). Este hallazgo es fundamental, ya que

demuestra que la presencia de defectos físicos visibles en el grano verde no es un predictor determinante de fallos sensoriales en la bebida final.

Esta desconexión estadística respalda lo analizado previamente, ya que mientras que los defectos visuales pueden ser frecuentes y variables, la integridad del sabor se mantiene alta y relativamente independiente de la apariencia física del lote.

5.4 Modelado predictivo para la variable Cuantificación de defectos por análisis visual

Se utilizó un total de 1324 muestras de ensayo correspondientes a lotes de café verde de diferentes orígenes con 14 variables predictoras para modelar la variable objetivo “Cuantificación de defectos por análisis visual”. El conjunto de datos fue dividido en dos conjuntos, uno correspondiente al entrenamiento con el 80 % de los datos con n igual a 1,059; y otro conjunto de prueba con el 20 % de los datos con n igual a 265. Lo anterior, permite asegurar una partición estratificada para mantener la representatividad de la distribución original.

5.4.1 Optimización del modelo KNN. Mediante validación cruzada con 5 folds y búsqueda de hiperparámetros (GridSearchCV) sobre el conjunto de entrenamiento, se evaluaron los valores de K en el rango de 3 a 15 vecinos cercanos.

El criterio de selección se basó en maximizar el coeficiente de determinación (R^2), buscando el punto donde el modelo alcanza su mayor desempeño antes de llegar al punto

de estabilización. Los resultados obtenidos en la optimización se describen en la Tabla 6 y Figura 26:

Tabla 6. Evaluación de la capacidad predictiva según el número de vecinos (K).

K	R ² CV (media)	R ² CV (desviación)
3	0,148	0,085
5	0,162	0,079
7	0,170	0,078
9	0,175	0,077
11	0,178	0,077
13	0,179	0,076
15	0,180	0,076

Nota: CV corresponde a la Validación Cruzada. Los valores se basan en la métrica de optimización del modelo.

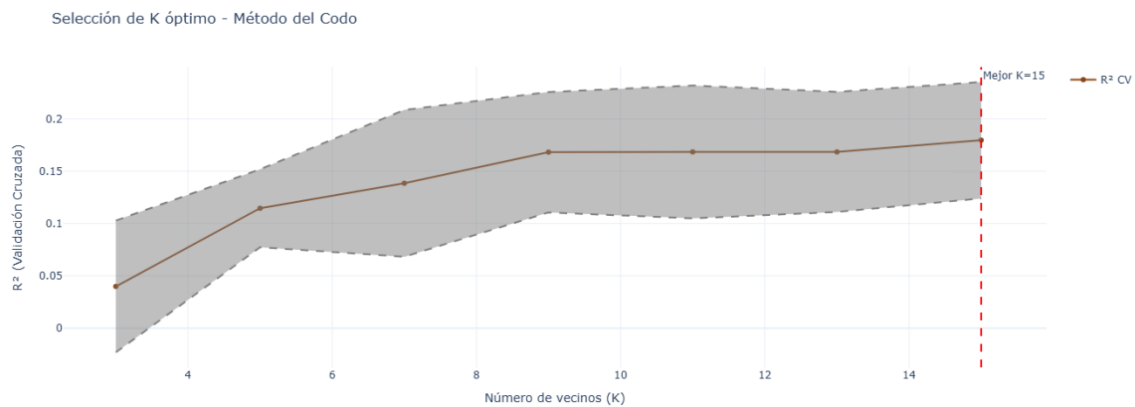


Figura 26. Optimización del modelo KNN a través del método del codo.

El análisis revela que el rendimiento del modelo mejora al aumentar el número de vecinos cercanos, alcanzando su valor óptimo en con un R^2 de validación de 0,180. Este valor de fue seleccionado al representar la mayor estabilidad con una desviación estándar de 0,076 y el mejor ajuste en fase de prueba, logrando un valor final de 0,185.

Métrica	Valor
Dataset	1324 muestras
Variables predictoras	14 variables
Mejor K	15
R^2 (Validación Cruzada)	-0.753 ± 0.357
R^2 (Prueba)	0.185

Figura 27. Resultados del Modelo KNN.

5.4.2 Evaluación del modelo. Como se mencionó anteriormente, el modelo KNN optimizado fue entrenado sobre el conjunto completo de entrenamiento y evaluado mediante un conjunto de prueba independiente para verificar su capacidad de

generalización ante datos no vistos. Los resultados de este desempeño se consolidan en la siguiente tabla:

Tabla 7. Evaluación del modelo KNN.

Métricas	Conjunto de entrenamiento	Conjunto de prueba
Coefficiente de determinación (R^2)	0,281	0,185
Raíz del error cuadrático medio (RMSE)	-	3,460
Error absoluto medio (MAE)	-	2,677

El modelo presenta un R^2 de prueba de 0,185; lo que indica que el algoritmo logra explicar el 18,5 % de la variabilidad total de los defectos visuales en el conjunto de validación externa. La diferencia entre el de entrenamiento (R^2 igual a 0,281) y el de prueba sugiere un ligero sobreajuste, aunque la consistencia en el bajo rendimiento general puede ser el resultado de la complejidad intrínseca de la muestra y no a una falla en la estructura del modelo.

Por otro lado, el Error Absoluto Medio (MAE) de 2,677 implica que, en promedio, las predicciones del modelo se desvían aproximadamente 2,7 puntos de la calificación real de defectos visuales. Al observar la distribución de los errores residuales, se confirma que la mayoría de las desviaciones se concentran cerca de cero, lo que denota la ausencia de

un sesgo sistemático marcado, aunque existe una dispersión significativa que oscila entre -8 y 10 puntos de error en los extremos.

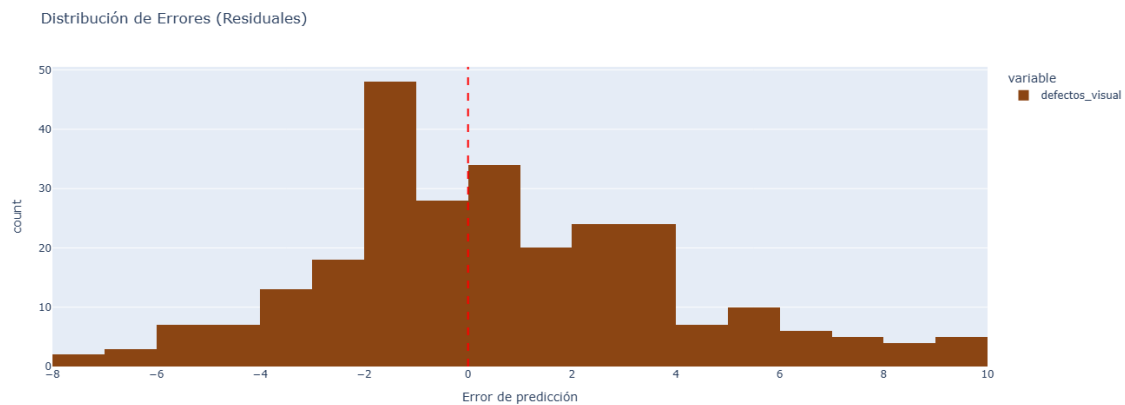


Figura 28. Distribución de errores residuales para el modelo KNN.

La Figura 29 permite visualizar la distribución de los resultados obtenidos en predicción de defectos frente a los valores reales de las muestras de ensayo, esta grafica revela una tendencia central moderada, pero con una notable dispersión de los puntos respecto a la línea de identidad. Esta dificultad predictiva se debe a que las categorías de defectos (Bajos, medios y altos) presentan un fuerte solapamiento en el espacio de características, como se evidencia en la Figura 31 en el análisis de componentes principales (PCA).

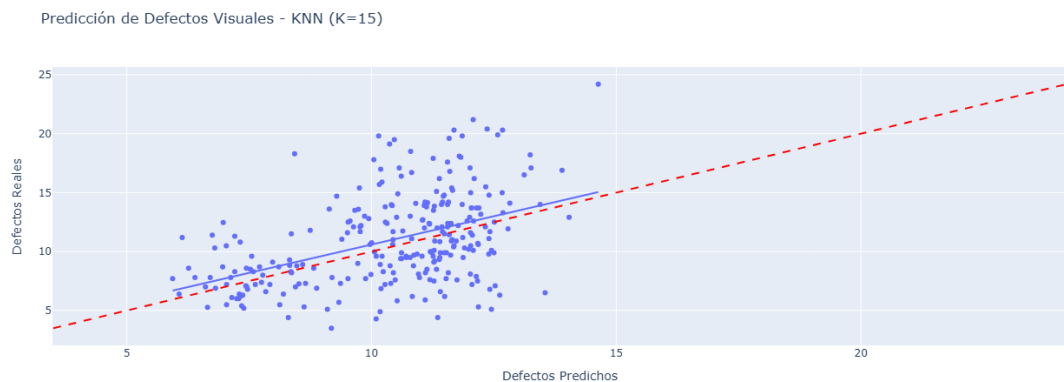


Figura 29. Predicción de Defectos Visuales mediante el modelo KNN (K=15).

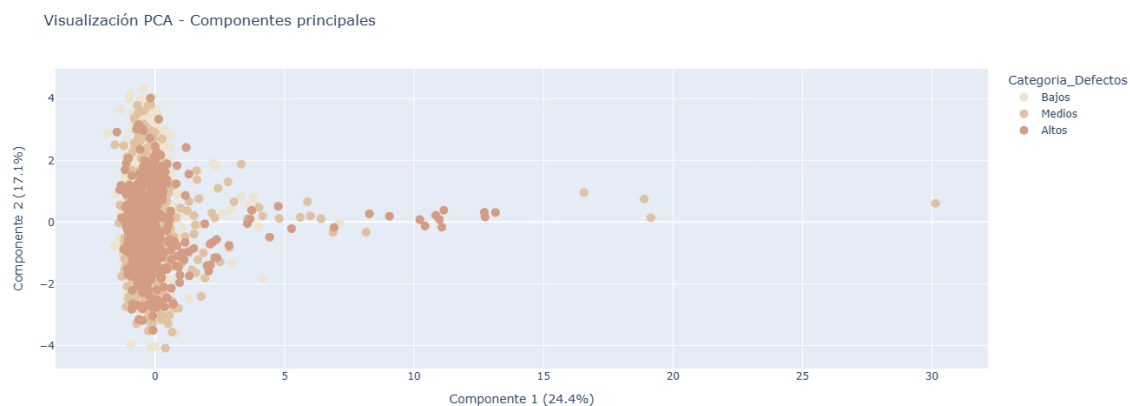


Figura 30. Visualización PCA (Componentes principales).

Desde una perspectiva técnica cafetera, la importancia de las variables para este modelo subraya que el color del café tostado es el predictor más influyente, seguido por el color del café verde bajo la coordenada b^* y la densidad real. Es relevante destacar que la densidad real presenta la correlación negativa más fuerte con la presencia de defectos visuales con un valor de $-0,16$; lo que sugiere que granos de mayor densidad tienden a

presentar menos defectos perceptibles visualmente, lo que da como resultado una correlación fundamental en la morfología del grano. Cabe resaltar que, para alcanzar precisiones superiores en la cuantificación de defectos por análisis visual, se debe implementar variables adicionales o explorar con el desarrollo de modelos no lineales que capturen mejor la subjetividad del análisis visual humano.

5.4.3 Evaluación de la estabilidad mediante validación cruzada. Para determinar la robustez y la capacidad de generalización del modelo KNN, se realizó una validación cruzada de 5 pliegues (5-fold CV) sobre el dataset completo de 1324 muestras. Este procedimiento permite observar cómo se comporta el algoritmo ante diferentes particiones de los datos y evaluar su estabilidad estadística. Los resultados obtenidos se detallan en la Tabla 8.

Tabla 8. Desempeño del modelo KNN bajo validación cruzada de 5 pliegues.

Pliegue (Fold)	Coefficiente de Determinación (R^2)
1	-1.102
2	-0.421
3	-0.834
4	-0.479
5	-0.929

Promedio $\pm$ DE	-0.753 $\pm$ 0.357
-------------------	--------------------------------------

Nota: DE = Desviación Estándar.

El análisis de validación cruzada arroja un promedio negativo de -0,753, lo cual constituye un hallazgo crítico para este estudio. En términos estadísticos, un negativo indica que el modelo KNN tiene un desempeño inferior a un modelo ingenuo (baseline) que simplemente predijera la media de los defectos visuales para cualquier muestra. Asimismo, la elevada desviación estándar de 0,357 revela una notable falta de estabilidad, sugiriendo que las predicciones son altamente sensibles a la agrupación específica de los datos de entrenamiento.

Esta inconsistencia se puede explicar por la naturaleza de las variables de acuerdo con:

- Solapamiento de clases: como se observó en el Análisis de Componentes Principales (PCA), existe una superposición casi total entre las categorías de defectos (bajos, medios y altos). Esto dificulta que el algoritmo de vecinos cercanos encuentre fronteras de decisión claras basadas únicamente en atributos como la densidad real o el color del café.
- Complejidad del análisis visual: aunque variables como el color del café tostado y la densidad real tienen un peso predictivo, la cuantificación de defectos visuales parece estar influenciada por factores morfológicos o subjetivos que no están plenamente capturados en las 14 variables físicas actuales.

- **Dispersión de errores:** la distribución de los errores residuales valida que el modelo lucha por converger en una predicción precisa, especialmente en las muestras que se alejan de la tendencia central de la base de datos.

Con lo anterior, se puede decir que el modelo KNN muestra una capacidad explicativa baja en el conjunto de prueba con un R^2 igual a 1,185. La validación cruzada revela que el sistema no es lo suficientemente robusto para ser utilizado de forma autónoma en un entorno de control de calidad en la industria cafetera sin la implementación de variables adicionales o modelos de mayor complejidad algorítmica.

5.4.4 Importancia de las variables en la predicción. Para complementar la interpretación del modelo y entender qué factores físicos influyen más en la presencia de defectos, se empleó un algoritmo de Random Forest como método de inspección. Este análisis permitió cuantificar la contribución relativa de cada una de las 14 variables predictoras sobre la variable objetivo. La Tabla 9 detalla las 10 variables con mayor peso jerárquico en la predicción de la variable objetivo.

Tabla 9. Variables con mayor peso jerárquico.

Rango	Variable	Importancia relativa	Acumulado
1	Color café tostado	16,7 %	16,7 %
2	Color café verde b*	14,0 %	30,7 %

Rango	Variable	Importancia relativa	Acumulado
3	Densidad real	11,8 %	42,5 %
4	Tamiz 17	9,2 %	51,7 %
5	Tamiz 15	8,6%	60,3 %
6	Color café verde L	7,7 %	68,0 %
7	Tamiz 18	6,8%	74,8 %
8	Tamiz 16	6,1%	80,9 %
9	Color café verde a*	6,1%	87,0 %
10	Humedad	5,8%	92,8 %

Nota: El 7,2 % restante corresponde a las variables tamiz 14, tamiz 13, tamiz 12 y defectos sabor.

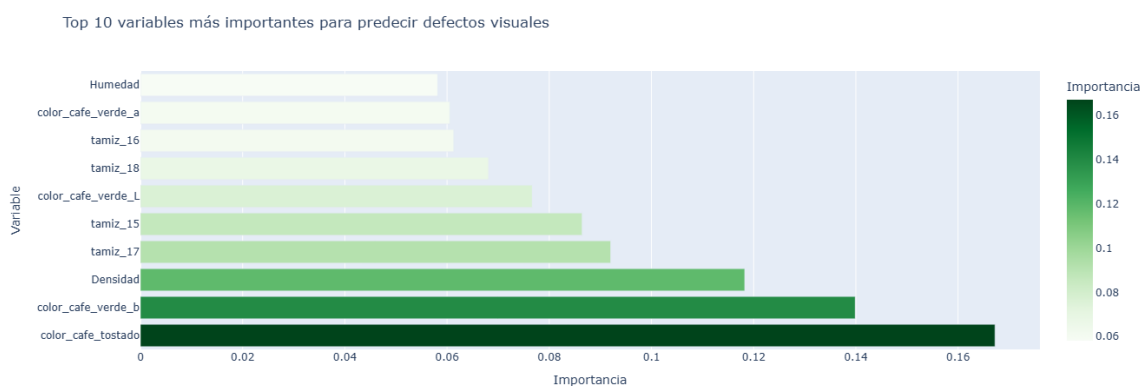


Figura 31. Variables más importantes para predecir los defectos por análisis visual.

Para la variable “Color en café tostado” o “Grado de tostión” se consolida como el predictor más relevante con un 16,7 %. Este hallazgo coincide con el análisis de correlación, donde presenta la correlación positiva más alta de aproximadamente 0,18 respecto a los defectos visuales. Esto sugiere que las alteraciones visuales en el grano verde se manifiestan de forma crítica durante la reacción de Maillard y la caramelización en el tostado del café. Sin embargo, se debe tener en cuenta que los criterios de calidad del tostado del café deben ser controlados de acuerdo con los grados de tostión especificados para el cliente. Con lo anterior es importante realizar una revisión de la información antes de concluir una correlación efectiva de estas variables.

Por otro lado, la variable color café verde CIE para la coordenada b^* , arroja un valor del 14,0 %. Esta coordenada de color es vital en la industria cafetera, ya que variaciones hacia pigmentos amarillos suelen asociarse con granos viejos o secados inadecuadamente.

Por último, la variable densidad real arrojó un valor del 11,8 %. Como se vió anteriormente, esta variable posee la correlación negativa más fuerte (-0,16), confirmando que a medida que la densidad del grano aumenta, la incidencia de defectos visuales tiende a disminuir.

5.4.5 Análisis por grupos de variables. Con el fin de determinar si las variables o grupo de variables para la evaluación física y sensorial del café podrían predecir los defectos por sí solas, se evaluó el desempeño del modelo agrupando las variables en cinco categorías. Los resultados, obtenidos mediante validación cruzada (5 Folds), se presentan en la Tabla 10.

Tabla 10. Desempeño predictivo desglosado por ensayos.

Grupo de variables	Variables incluidas	R^2 (media)	R^2 (DE)
Variables físicas	Humedad, Densidad real	-0,571	0,181
Tamaño de grano	Tamiz de la 12/64 in a la 18/64 in	-0,648	0,262
Color del café tostado	Color del café tostado	-0,772	0,272
Color de café verde CIE L ab*	L, a, b	-0,812	0,361

Grupo de variables	Variables incluidas	R^2 (media)	R^2 (DE)
Sensorial sabor	Defectos de sabor	-1,402	0,874

Los resultados presentados en la tabla anterior, evidencian que ninguna dimensión física o sensorial tiene la capacidad explicativa suficiente para modelar los defectos visuales de forma independiente. Esto refuerza la observación del PCA, donde el solapamiento de muestras impide una segmentación clara basada en estos mensurandos.

Para el caso de las variables humedad y densidad real, se evidencia el un mejor desempeño relativo con un valor de R^2 igual a -0,571. Esto indica que, la estructura física interna del grano está más vinculada con la apariencia visual que el tamaño del grano o el color por sí solo.

Para el caso de la variable de defectos de sabor en taza obtuvo el desempeño más bajo con un R^2 igual a -1,402. Esto es un hallazgo de alto valor para la industria del café, ya que permite confirma estadísticamente que la presencia de defectos visuales no es un predictor fiable de los defectos de sabor. Un lote de café puede contener altos valores de defectos visuales, pero obtener resultados aceptables en su sabor, y viceversa, lo que subraya la necesidad de mantener ambos controles de calidad de forma independiente en los procedimientos y reglas de decisión para la evaluación de la conformidad.

5.4.6 Aplicabilidad en control de calidad del café de especialidad. Para que el modelo sea aplicable en un entorno de producción real, el análisis de errores residuales es un punto clave. La distribución de los errores residuales muestra que, aunque la mayoría de los errores se concentran cerca de cero, existen desviaciones que pueden oscilar en el rango de -8 y 10 unidades. En un procedimiento de control de calidad estricto, este margen de error podría llevar a clasificaciones erróneas de lotes de café. No obstante, la validación identifica variables con alta relevancia que sí pueden ser utilizadas como indicadores de apoyo en la industria:

- Color de café tostado y verde (coordenada b^*): identificados como los principales predictores con un 16,7 % y 14,0 % de importancia respectivamente, validando su uso en la inspección visual rápida.
- Densidad real: su correlación negativa de -0,16 con los defectos visuales confirma que es una métrica física robusta que los técnicos de calidad deben priorizar.

Con lo anterior, se puede decir que mientras que el sistema identifica variables clave para el análisis del café colombiano, su capacidad de generalización actual no es suficiente para reemplazar la evaluación humana. Su aplicabilidad se debe limitar a una herramienta de filtro preliminar en el control de calidad, requiriendo la integración de modelos no lineales para alcanzar la robustez necesaria en condiciones de producción real.

5.5 Modelado predictivo para la variable identificación de defectos de sabor en taza

El modelo de clasificación para la variable objetivo de identificación de defectos de sabor se realizó a través de un modelo de regresión logística, el cual se desarrolló exclusivamente a la discriminación binaria entre la presencia o ausencia de defectos de sabor, sin abordar la identificación puntual del defecto encontrado (fermento, fenol, terroso, mohoso, verde, químico, stikner, entre otros). Esta delimitación metodológica implica que el poder predictivo del modelo debe ser interpretado únicamente en términos de su capacidad para clasificar correctamente la condición bimodal de la muestra (defectuosa o no defectuosa), mas no como una herramienta de caracterización cualitativa y cuantitativa del perfil sensorial de un lote de café verde.

La distribución del conjunto de datos, detallada en la Tabla 11, muestra un desequilibrio de clases que se debe tener en cuenta, ya que, el 86,5 % de las muestras analizadas no presentan defectos frente a un 13,5 % que sí los reportan. Esta proporción de aproximadamente 6 a 1, constituye un desafío metodológico crítico, ya que puede sesgar el modelo hacia la clase mayoritaria, exigiendo una interpretación cuidadosa en las métricas de evaluación.

Tabla 11. Distribución de la variable objetivo Defectos de sabor (Presencia o Ausencia).

Categoría	Frecuencia	Porcentaje
Sin defectos de sabor	1144	86,5 %
Con defectos de sabor	179	13,5 %
Total	1323	100,0 %

5.5.1 Caracterización estadística por grupos de ensayos. La comparación de los estadísticos descriptivos entre las muestras con y sin defectos permite identificar tendencias en las variables físicas preliminares:

- **Humedad (%):** las muestras con defectos de sabor presentan en promedio, un porcentaje de humedad superior del 11,66 % en comparación con las muestras sin defectos de sabor con valores de humedad del 11,56 %. Esta relación sugiere que la degradación física del grano puede conllevar a la degradación sensorial del grano.
- **Colorimetría del Café verde CIE L ab*:** se observa que el grupo de muestras con y sin defectos de sabor posee una luminosidad (L) inferior (44,24 y 44,38, respectivamente) y valores superiores en las coordenadas cromáticas a* (1,02 y 0,96, respectivamente) y b* (15,06 y 14,90, respectivamente). Este perfil del espacio de color indica que los granos con defectos de sabor tienden a ser más oscuros y presentar tonalidades más amarillentas y rojizas, lo cual es consistente con procesos de envejecimiento o secado inadecuado.

5.5.2 Análisis de correlación y multicolinealidad. Para garantizar la validez del modelo de regresión logística, se evaluó la estructura de dependencia entre los predictores mediante una matriz de correlaciones (ver Figura 32).

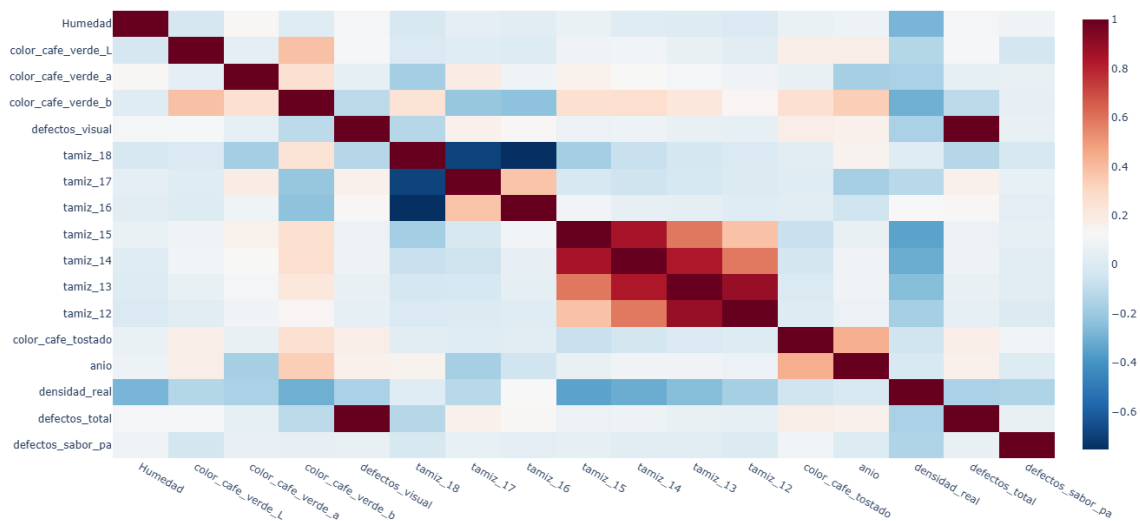


Figura 32. Matriz de Correlación para las variables predictoras.

La matriz de correlaciones reveló hallazgos importantes que condicionan la selección de variables para la predicción de la variable objetivo. Se detectaron variables con correlación superior a 0,8 en los granos de menor tamaño, específicamente en la secuencia de los tamices de ensayo de diámetro 15/64 in al tamiz de ensayo de diámetro 12/64 in. Esta alta interdependencia sugiere que la variabilidad en la granulometría en los tamices de ensayo con diámetro pequeño es capturada de forma redundante por estos predictores, lo que justifica una selección cuidadosa para evitar inflar la varianza de los coeficientes.

A pesar de estas colinealidades, variables clave como la densidad real y el color del café tostado muestran una menor dependencia lineal con el resto del conjunto, lo que facilita su interpretación como predictores independientes de alta relevancia en las etapas de modelado supervisado.

5.5.3 Optimización del modelo. Con el fin de garantizar que los coeficientes del modelo sean directamente comparables e independientes de las escalas físicas originales como por ejemplo densidad en g/cm^3 frente a las unidades de color en el espacio CIE L ab*, todas las variables independientes fueron normalizadas con una media igual a 0 y desviación igual a 1. Este paso es fundamental para interpretar correctamente la magnitud del impacto de cada predictor sobre la probabilidad de encontrar defectos de sabor.

Se evaluaron tres enfoques metodológicos para determinar la mejor configuración operativa del sistema:

- **Umbral estándar (0,5) sin balanceo:** este modelo se utilizó como línea base utilizando un punto de corte por defecto de 0,5. Debido a la baja prevalencia de defectos (13,5 %), este enfoque resultó en una especificidad casi perfecta del 99,8 %, pero sacrificando la sensibilidad con un valor del 1,1 %, lo cual no permite la detección sistemática de muestras con defectos de sabor.
- **Balanceo por clases:** para corregir el sesgo hacia la clase mayoritaria, se aplicaron pesos a la función de costo inversamente proporcionales a la frecuencia de las clases. Esta estrategia logró un equilibrio superior, elevando la sensibilidad

del 62,0 % al 63,1 %. De este modelo se derivaron los coeficientes finales, donde la densidad real arrojó un coeficiente igual a -0,35 y el color del café tostado un coeficiente igual a 0,33; consolidando estas variables como los predictores de mayor peso.

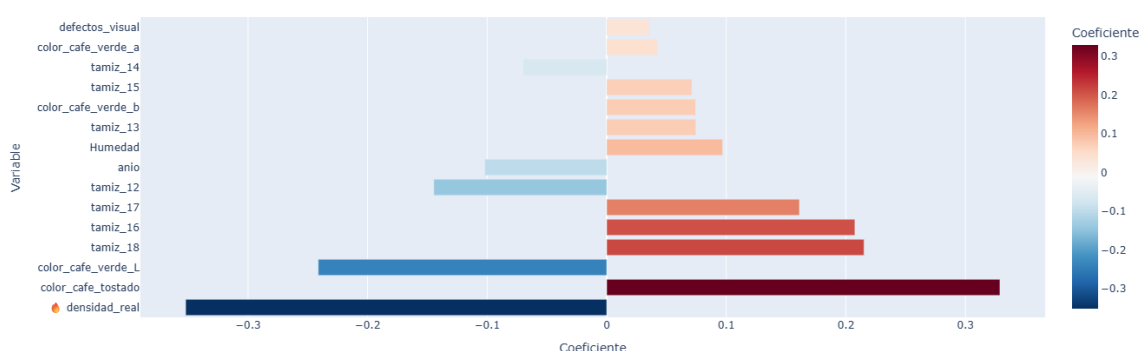


Figura 33. Top 15 de coeficientes para el modelo balanceado.

- Umbral óptimo según índice de Youden:** buscando la máxima eficiencia para aplicaciones reales de control de calidad, se refinó el modelo balanceado mediante la identificación del umbral óptimo de probabilidad de 0,49. Este punto de corte maximiza el índice de Youden a un valor de J igual a 0,277 logrando el mejor desempeño discriminativo con una sensibilidad del 65,9 % y una especificidad del 61,8 %.

Esta progresión metodológica confirma que, para la evaluación del sabor del café colombiano, el ajuste del umbral probabilístico es una técnica más eficaz que el uso de

umbrales estándar, permitiendo capturar un 4 % adicional de defectos de sabor reales respecto al modelo balanceado base.

La efectividad de las tres estrategias se resume en la siguiente tabla de características de desempeño, que permite evidenciar la superioridad del modelo con umbral optimizado para la detección de defectos de sabor.

Tabla 12. Características de desempeño para un umbral óptimo de 0,494.

Métrica	Umbral 0.5	Balanceado	Umbral Óptimo
TPR (Sensibilidad)	0.011	0.620	0.659
TNR (Especificidad)	0.998	0.636	0.618
FPR	0.002	0.364	0.382
FNR	0.989	0.380	0.341
PPV (Precisión)	0.500	0.211	0.213
NPV	0.866	0.915	0.921
Eficiencia	0.865	0.634	0.624
F1-Score	0.022	0.314	0.322
Youden	0.009	0.256	0.277
AUC-ROC	0.661	0.665	0.665

Nota: La Tabla 12 comparativa permite observar la evolución de la sensibilidad (TPR) y el F1-Score a través de las tres etapas de construcción del modelo.

5.5.4 Evaluación del ajuste y bondad del modelo. En esta sección se analiza la validez estadística del modelo de regresión logística y su capacidad para generalizar predicciones en nuevas muestras de café. La evaluación se centra en la capacidad discriminativa global y la estabilidad del algoritmo ante diferentes particiones de los datos.

La métrica fundamental para evaluar el ajuste del modelo es a través del Área Bajo la Curva (AUC-ROC), la cual mide la capacidad del sistema para distinguir entre muestras

con y sin defectos de sabor. Para el modelo balanceado, se obtuvo un AUC-ROC de 0,662.

De acuerdo con las escalas de clasificación estándar, un valor de 0,662 se interpreta como una discriminación aceptable pero débil. Esto indica que, si bien el modelo es capaz de identificar patrones que superan el azar (AUC igual a 0,5), existe un solapamiento significativo en las características físicas de ambos grupos, lo que limita su precisión operativa para separar claramente los cafés con defectos de sabor frente a los cafés sin presencia de defectos de sabor.

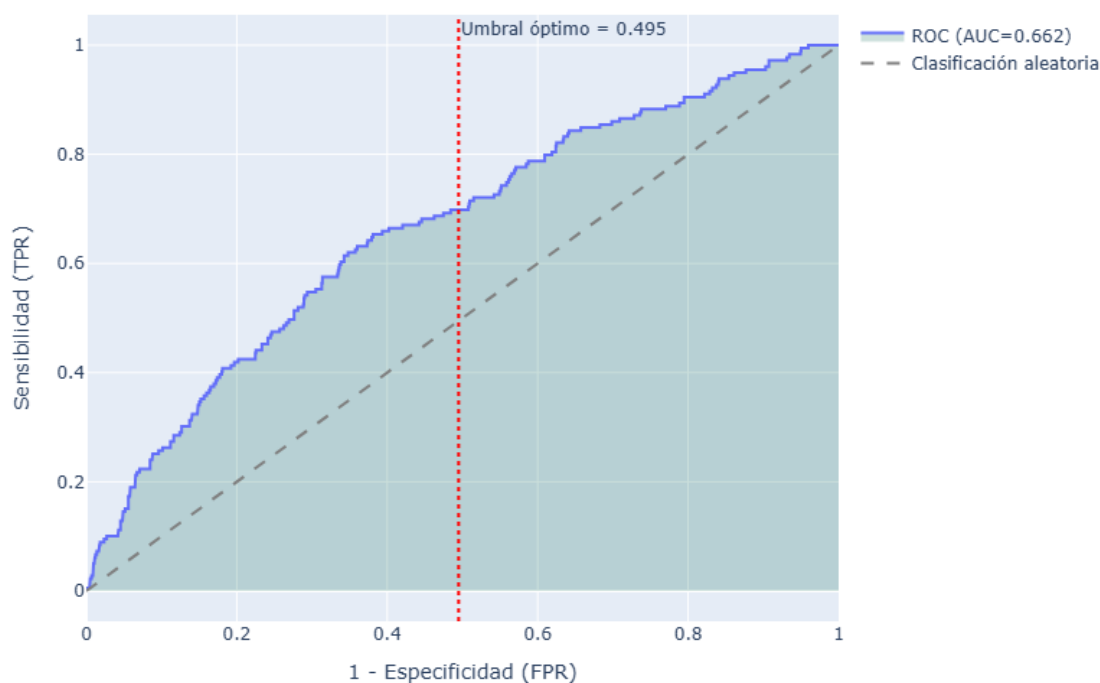


Figura 35. Curva ROC para el modelo de regresión logística en la identificación de defectos de sabor.

Nota: Esta gráfica visualiza la trayectoria del modelo sobre la línea de azar, confirmando el AUC de 0,662 y permitiendo identificar visualmente la relación entre la sensibilidad y la tasa de falsos positivos.

La capacidad del modelo de regresión logística para discriminar entre lotes de cafés con presencia de defectos de sabor y aquellos sin defectos de sabor se visualiza a través de la curva ROC presentada en la Figura 35. El Área Bajo la Curva (AUC) alcanzada es de 0,665; valor que sitúa el poder discriminativo del sistema en un rango de discriminación aceptable pero débil según los estándares estadísticos. Esto significa que, aunque el modelo logra distinguir las clases mejor que el azar ($AUC = 0,5$), su capacidad operativa es limitada para un entorno de alta precisión en comparación con paneles sensoriales de expertos los cuales pueden llegar a obtener valores de AUC igual a superior a 0,9. Sin embargo, implementar un AUC de 0,665 permite la identificación de muestras con defecto en un entorno de control de calidad, asumiendo el riesgo de clasificar como defectuosas aquellas que no poseen defectos de sabor.

5.5.5 Validación cruzada y estabilidad del modelo. Para verificar la robustez y capacidad de generalización del modelo, se realizó una validación cruzada de 5 pliegues (5-fold CV). Este procedimiento somete al algoritmo a cinco pruebas diferentes, utilizando distintas particiones de los datos para entrenamiento y validación.

Los resultados obtenidos fueron:

- AUC-ROC promedio igual a $0,625 \pm 0,059$.

- Desempeño por pliegues (folds): 0,715; 0,636; 0,595; 0,536 y 0,642.

El descenso del AUC promedio a 0,625 en la validación cruzada en comparación con el 0,662 del conjunto de entrenamiento, sugiere una ligera tendencia al sobreajuste o una sensibilidad moderada a la partición del conjunto de datos. La variabilidad observada entre pliegues con un mínimo de 0,536 y un máximo de 0,715 confirma que el modelo es moderadamente inestable, dependiendo significativamente de la representatividad de las muestras incluidas en cada subconjunto de entrenamiento.

La combinación de un AUC-ROC de 0,662 y un desempeño promedio en validación cruzada de 0,625 ratifica que las variables físicas y colorimétricas actuales poseen un poder explicativo limitado sobre la calidad sensorial de sabor. Aunque el modelo identifica tendencias reales, la bondad de ajuste es insuficiente para una aplicación industrial de alta fidelidad, lo que refuerza la necesidad de incorporar variables exógenas a través de tecnologías analíticas más complejas como la espectroscopía de infrarrojo cercano NIRS (por sus siglas en inglés) o el desarrollo de modelos no lineales en etapas posteriores.

5.5.6 Jerarquización de las variables físicas y su efecto en los defectos de taza. Tras la validación del modelo balanceado, el análisis de sus parámetros internos permite cuantificar el impacto real de cada atributo físico sobre la probabilidad de defectos de sabor. Al haber trabajado con variables estandarizadas, los coeficientes reflejan el cambio

en el logit por cada DE de aumento en el predictor, permitiendo una comparación directa de su importancia relativa.

Variable	Coefficiente (β)	Odds Ratio	IC95% OR
densidad_real (densidad)	-0.3522	0.703	[0.66 - 0.75]
color_cafe_tostado	0.3288	1.389	[1.30 - 1.48]
color_cafe_verde_L	-0.2413	0.786	[0.74 - 0.83]
tamiz_18	0.2152	1.240	[1.10 - 1.40]
tamiz_16	0.2075	1.231	[1.13 - 1.35]
tamiz_17	0.1612	1.175	[1.08 - 1.28]
tamiz_12	-0.1446	0.865	[0.74 - 1.01]
anio	-0.1019	0.903	[0.84 - 0.97]
Humedad	0.0969	1.102	[1.04 - 1.17]
tamiz_13	0.0746	1.077	[0.85 - 1.37]
color_cafe_verde_b	0.0743	1.077	[1.00 - 1.16]
tamiz_15	0.0712	1.074	[0.95 - 1.22]
tamiz_14	-0.0697	0.933	[0.76 - 1.14]

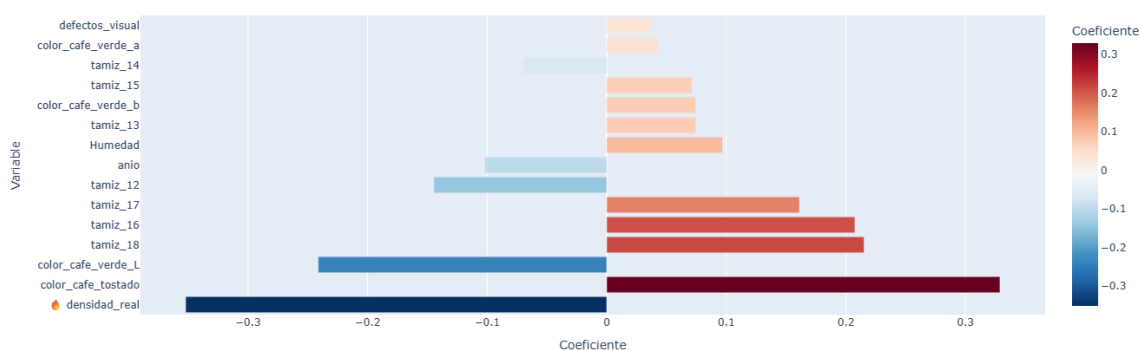


Figura 36. Magnitud del efecto de predictores físicos sobre la estabilidad del sabor del café.

El análisis de los Odds Ratios (OR) y sus respectivos intervalos de confianza al 95% (IC95%) revela los siguientes hallazgos críticos para la calidad del café:

- **Densidad Real (g/cm^3):** se consolida como el factor protector más potente con un coeficiente de -0,3522 y un OR de 0,703 en un Intervalo de Confianza (IC) al 95 % de 0,66 a 0,75) indica que, por cada desviación estándar que aumenta la densidad real, la probabilidad (*odds*) de presentar un defecto de sabor disminuye

en un 29,7 %. Este resultado ratifica que la estructura física compacta es el principal indicador de estabilidad sensorial.

- Color de café tostado: es el principal factor de riesgo identificado con un coeficiente de 0,3288 y un OR de 1,389 en un IC al 95 % de 1,30 a 1,48. Lo cual indica que cada incremento en una DE de este parámetro aumenta la probabilidad de defecto en un 38,9 %. Esto sugiere que ciertas tonalidades en el tostado están fuertemente vinculadas a la revelación de notas sensoriales indeseables.

- Color de café verde CIE (Luminosidad): actúa como un factor protector significativo con un coeficiente de -0,2413 y un OR de 0,786 en un IC al 95% de 0,74 al 0,83. Esto implica que un grano verde más brillante y luminoso reduce el riesgo de defectos en un 21,4 %.

- Tamao de grano retenido en el Tamiz 18/64 in: los granos de mayor tamaño muestran una asociación positiva con el riesgo sensorial con un coeficiente igual a 0,2152 y un OR de 1,240 en un IC al 95% de 1,10 a 1,40. Lo que significa que el riesgo de defecto se incrementa en un 24,0 % por cada DE de aumento en la proporción de esta granulometría.

- Tamao de grano retenido en el Tamiz 16/64 in: de manera similar, este tamaño de grano con un coeficiente de 0,2075 y un OR de 1,231 en un IC al 95 % de 1,13 a 1,35. Vinculándose con un aumento del 23,1 % en la probabilidad de fallas en taza.

En resumen, la combinación de una baja densidad real y un color de tostado elevado constituye el perfil de mayor riesgo para la calidad sensorial del café colombiano analizado. Estos coeficientes ofrecen una base científica para ajustar los umbrales de aceptación en los procesos de trilla y tostación, priorizando la densidad real y la luminosidad del grano verde como salvaguardas de la taza.

5.5.7 El modelo como filtro de riesgo. La integración de los resultados del modelado de regresión logística para la predicción de defectos de sabor en taza revela una estructura técnica robusta que permite transitar desde la caracterización física hasta la gestión preventiva del riesgo sensorial. El estudio, fundamentado en el conjunto de datos de 1,323 observaciones con una prevalencia de defectos del 13,5% permitió entrenar un modelo que, tras ser optimizado mediante el índice de Youden con un umbral de 0,494 alcanzó una sensibilidad del 65,9 % y una especificidad del 61,8 %. Aunque la capacidad discriminativa global reflejada en un AUC-ROC de 0,665 se clasifica como aceptable pero débil, el análisis de sus parámetros internos proporciona una hoja de ruta clara sobre los determinantes de la calidad.

La densidad real se consolidó como una variable protectora crítico, presentando un coeficiente de -0,352 y un Odds Ratio (OR) de 0,703, lo que implica que cada incremento en la densidad del grano reduce la probabilidad de defectos de sabor en un 29,7 %. Inversamente, el color del café tostado surgió como el principal factor de riesgo con un OR de 1,389 indicando que un aumento en este parámetro eleva el riesgo de fallas sensoriales en un 38,9 %. Esta jerarquía de variables se complementa con el diagnóstico

de los residuos de Pearson, cuyo rango entre -3,318 y 1,796 y la identificación de solo un 0,6 % de puntos influyentes, ratifican la estabilidad estadística de estas tendencias.

Es importante tener en cuenta que, la aplicabilidad industrial del modelo debe enmarcarse estrictamente como una herramienta de filtración preventiva y no como un sistema decisivo, debido a una tasa de falsos positivos del 38,2 % y una precisión operativa limitada de 0,213. En la práctica, el modelo funciona eficazmente como un filtro de alerta temprana que señala lotes de alto riesgo para una evaluación prioritaria por catadores certificados, quienes mantienen la superioridad técnica con precisiones superiores al 90 %, permitiendo así una optimización de recursos y una gestión del riesgo basada en la integridad física y colorimétrica del grano.

6 Conclusiones

Se caracterizaron 1323 muestras que presentan una base física y sensorial sólida frente a los estándares de exportación de cafés de especialidad colombianos, destacando una humedad óptima y un tamaño de grano dominada por granos grandes (tamices 16/64 in a 18/64 in).

Se determinó que solo el 13,5 % (179 muestras) presentan defectos de sabor, mientras que el 86,5 % (1144) corresponden a tazas limpias. Esta baja prevalencia de defectos de sabor frente a la variabilidad de defectos visuales confirma que la apariencia externa no es un indicador absoluto de la calidad sensorial del grano en taza.

Mediante el análisis de importancia, se validó que el color del café tostado (16,7 % de importancia) y la densidad real son los principales descriptores vinculados a la calidad sensorial del grano.

La densidad real se consolidó como el factor protector más robusto; ya que un incremento en la densidad reduce la probabilidad de defectos de sabor en un 29,7 %. Por el contrario, el color del café tostado actúa como el principal factor de riesgo, aumentando la probabilidad de defectos en un 38,9 %.

El modelo fue entrenado y optimizado con éxito mediante un umbral de 0,494 alcanzando una sensibilidad del 65,9 % y una especificidad del 61,8 %. Su capacidad discriminativa global (AUC-ROC igual a 0,665) se clasifica como aceptable pero débil,

superando la clasificación aleatoria obtenida por el azar, pero mostrando limitaciones frente a la complejidad sensorial global.

La evaluación del modelo KNN mostró una capacidad limitada para la cuantificación automática, con un R^2 de prueba de 0,185 y una validación cruzada negativa de -0,753. Esto indica que la simple proximidad estadística de las variables físicas no es suficientes para una segmentación clara de la calidad sensorial sin el uso de algoritmos no lineales más avanzados.

Los modelos desarrollados demuestran ser herramientas útiles para un tamizaje rápido, pero la evaluación sensorial tradicional por catadores certificados sigue siendo superior, con niveles de precisión y sensibilidad superiores al 90 %.

Se cumplió el objetivo de integrar la evaluación física y sensorial del café en modelos de aprendizaje automático. Aunque la densidad real y el color son predictores objetivos de gran valor, la alta tasa de falsos positivos de 38,2 % sugiere que el futuro de la predicción de la calidad sensorial reside en la transición hacia modelos de mayor complejidad técnica que puedan gestionar mejor el desbalance de clases y el solapamiento de características físicas.

7 Referencias

Aleixandre-Tudó, J. L., Alvarez, I., García, M. J., Lizama, V., & Aleixandre, J. L. (2015). Application of Multivariate Regression Methods to Predict Sensory Quality of Red Wines. *Czech Journal of Food Sciences*, 33(3), 217-227.

Avilés Granda, K. M. (2025). Calidad Fisicoquímica, Microbiológica y Sensorial del Café (*Coffea arabica* L.) en Función de la Altitud en San Cristóbal, Galápagos. *Ciencia Latina Revista Científica Multidisciplinar*, 16948-16963.

Baqueta, M. R., Coqueiro, A., & Valderrama, P. (2019). Brazilian Coffee Blends: A Simple and Fast Method by Near-Infrared Spectroscopy for the Determination of the Sensory Attributes Elicited in Professional Coffee Cupping. *Journal of Food Science*, 84(6), 1247-1255.

Barbosa, M. d. (2019). Correlation between the Composition of Green Arabica Coffee Beans and the Sensory Quality of Coffee Brews. *Food Chemistry*, 275-280.

Barbosa, M. d. S. G., Scholz, M. B. d. S., Kitzberger, C. S. G., & Benassi, M. d. T. (2019). Correlation between the Composition of Green Arabica Coffee Beans and the Sensory Quality of Coffee Brews. *Food Chemistry*, 292, 275-280.

Brownlee, J. (2018). *How to Use ROC Curves and Precision-Recall Curves for Classification in Python*. Machine Learning Mastery. Recuperado [2026-03-17],

de <https://machinelearningmastery.com/roc-curves-and-precision-recall-curves-for-classification-in-python/>

Çataltaş, Ö., & Tutuncu, K. (2021). A Review of Data Analysis Techniques Used in Near-Infrared Spectroscopy. *European Journal of Science and Technology*, (31), 475-483.

Correa, P. C., et al. (2022). Color analysis in green coffee beans. *Journal of Food Engineering*, 45(3), 112-125.

Corrêa, S. C., Pinheiro, A. C. M., Siqueira, H. E., Carvalho, E. M., Nunes, C. A., & Vilas Boas, E. V. de B. (2014). Prediction of the Sensory Acceptance of Fruits by Physical and Physical-Chemical Parameters Using Multivariate Models. *LWT - Food Science and Technology*, 59(2), 666-672.

DANE. (2025). *Estadísticas de exportaciones*. Obtenido de Información noviembre de 2025: <https://www.dane.gov.co/index.php/estadisticas-por-tema/comercio-internacional/exportaciones>

Duarte, W. F., et al. (2022). Moisture content and coffee quality. *Food Research International*, 78(2), 89-98.

El Cafè de Mi Tierra. (25 de Septiembre de 2024). *Los cafetos, sus especies y variedades. Parte I: Clasificación taxonómica*. Obtenido de

<https://elcafedemitierra.com/los-cafetos-sus-especies-y-variedades-parte-i-clasificacion-taxonomica.html>

Farah, A. &. (2015). Bioavailability and metabolism of chlorogenic acids from coffee. En *Coffee: Health and Disease Prevention. Elsevier*, 789–801.

Farah, J. S., Cavalcanti, R. N., Guimarães, J. T., Balthazar, C. F., Coimbra, P. T., Pimentel, T. C., ... & Cruz, A. G. (2021). Differential Scanning Calorimetry Coupled with Machine Learning Technique: An Effective Approach to Determine the Milk Authenticity. *Food Control*, 121, 107585.

Federación Nacional de Cafeteros de Colombia. (2025). *Informe del Gerente. Congreso Nacional de Cafeteros #94*.

Franca, A. S., & Oliveira, L. S. (2019). Density as quality indicator in coffee beans. *Coffee Science*, 14(2), 156-167.

González-Sánchez, B., Sandoval-González, O., Flores-Cuautle, J. de J. A., Landeta-Escamilla, O., Portillo-Rodríguez, O., & Aguila-Rodríguez, G. (2023). A study of the physical characteristics and defects of green coffee beans that influence the sensory notes using machine learning models. *Foods*, *12*(24), Article 4520. <https://doi.org/10.3390/pr12010018>

Hameed, A., Hussain, S. A., Ijaz, M. U., Ullah, S., Pasha, I., & Suleria, H. A. R. (2018). Farm to Consumer: Factors Affecting the Organoleptic Characteristics

of Coffee. II: Postharvest Processing Factors. *Comprehensive Reviews in Food Science and Food Safety*, 17(5), 1184-1237.

International Coffee Organization. (April de 2023). *ICO*. Obtenido de Coffee Report and Outlook: https://icocoffee.org/documents/cy2022-23/Coffee_Report_and_Outlook_April_2023_-_ICO.pdf

Kraggerud, H., Næs, T., & Abrahamsen, R. K. (2014). Prediction of Sensory Quality of Cheese during Ripening from Chemical and Spectroscopy Measurements. *International Dairy Journal*, 34(1), 6-18.

Lingle, T. R., & Menon, S. N. (2017). Cupping and Grading-Discovering Character and Quality. In *The Craft and Science of Coffee* (pp. 181-203). Elsevier Inc.

Ministerio de Agricultura y Desarrollo Rural. (2025). *Plan estratégico institucional 2023-2026*. Obtenido de https://www.minagricultura.gov.co/planeacion-control-gestion/Gestin/PLANEACION/Planes_Estrategicos_Sectoriales_Institucionales/Planes%20Estrategicos%202022%20-%202026/Informe_Plan%20Estratégico%20Institucional_Primer%20Semestre%202025.pdf

Mohamed, A. I. (2024). Impact of coffee and its bioactive compounds on the risks of type 2 diabetes and its complications: A comprehensive review. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*.

Nemzer, B. K. (2021). Quantification of major bioactive constituents, antioxidant activity, and enzyme inhibitory effects of whole coffee cherries (*Coffea arabica*) and their extracts. *Molecules*.

Oliveros, C. E., et al. (2020). Colombian technical standards for green coffee. *Cenicafé Journal*, 71(1), 45-58.

Peñuela-Martínez, A. E., et al. (2023). Fermentation processes and coffee sensory quality. *Journal of Coffee Research*, 51(2), 78-92.

Rojas, J. A., et al. (2023). Screen size distribution in specialty coffees. *Coffee & Cocoa Science*, 12(3), 234-248.

Rufian-Henares, J. &. (2015). Biological effects of coffee melanoidins. En *Coffee: Health and Disease Prevention*. *Elsevier*.

Saborío-Saborío, L., & Jiménez-Morales, M. (2021). Bean size and coffee quality relationship. *Agronomía Costarricense*, 45(1), 67-82.

Sandoval Aldana, A. G. (2025). Calidad fisicoquímica y sensorial del café en el departamento del Tolima. *Sello Editorial Universidad del Tolima*.

SCA. (2018). Protocols and Best Practices. *Specialty Coffee Association*.

Specialty Coffee Association (SCA). (2018). *Protocols and Best Practices*.

Specialty Coffee Association (SCA). (2021). *SCA Coffee Standards*. Specialty Coffee Association.

Tfouni, S. A. V., Carreiro, L. B., Teles, C. R. A., Furlani, R. P. Z., Cipolli, K. M. V. A. B., & Camargo, M. C. R. (2014). Caffeine and Chlorogenic Acids Intake from Coffee Brew: Influence of Roasting Degree and Brewing Procedure. *International Journal of Food Science and Technology*, 49(3), 747-752.

Vega, A., Reyes, S. M., & Troesch, J. (2025). Physicochemical Parameters and Multivariate Analysis to Predict the Sensory Quality in Specialty Coffee from Panama. *ACS Omega*, 10, 13251-13259.

Wang, X., Wang, Y., Hu, G., Hong, D., Guo, T., Li, J., Li, Z., & Qiu, M. (2022). Review on Factors Affecting Coffee Volatiles: From Seed to Cup. *Journal of the Science of Food and Agriculture*, 102(4), 1341-1352.

Xu, L., Ma, F., Zhou, J., & Du, C. (2024). Attenuated Total Reflectance Crystal of Silicon for Rapid Nitrate Sensing Combining Mid-Infrared Spectroscopy. *ACS Omega*, 9, 47613-47620.

Zhou, Z., Qiu, C., & Zhang, Y. (2023). A Comparative Analysis of Linear Regression, Neural Networks and Random Forest Regression for Predicting Air Ozone Employing Soft Sensor Models. *Scientific Reports*, 13(1), 1-23.