UNIVERSIDAD NACIONAL AUTÓNOMA DE CHOTA

FACULTAD DE CIENCIAS AGRARIAS

Escuela Profesional de Ingeniería Agroindustrial

INFLUENCIA DE LAS VARIABLES PRODUCTIVAS Y DE PROCESAMIENTO SOBRE LA CALIDAD EN TAZA DE CAFÉ DE 40 LOCALIDADES DE LA REGIÓN CAJAMARCA MEDIANTE ALGORITMOS DE MACHINE LEARNING CON INTELIGENCIA ARTIFICIAL

Tesis para optar el título profesional de ingeniero agroindustrial

Autor (es):

Bach. Luis Alberto Idrogo Bustamante

Bach. Ronald Yoel Núñez Benavides

Asesor:

Dr. Ricardo Abel Del Castillo Torres



Chota – Perú

2026

CONSTANCIA DE ORIGINALIDAD Y USO DE INTELIGENCIA ARTIFICIAL

CO-N°003 -2026-FCA-UNACH- RADCT

El que suscribe, Asesor del Informe Final de Tesis de la Escuela Profesional de Ingeniería Agroindustrial de la Facultad de Ciencias Agrarias de la Universidad Nacional Autónoma de Chota, hace constar que el Informe Final de Tesis Titulado "Influencia de las Variables Productivas y de Procesamiento sobre la Calidad en Taza de Café de 40 Localidades de la Región Cajamarca Mediante Algoritmos de Machine Learning con Inteligencia Artificial", perteneciente a los tesistas Luis Alberto Idrogo Bustamante y Ronald Yoel Núñez Benavides; presenta un GRADO DE SIMILITUD del 5 % y USO DE INTELIGENCIA ARTIFICIAL DE 0 %, en base al análisis realizado por el Software Antiplagio Turnitin (04/10/2026) y este corresponde a una similitud menor al 20 %, por lo que se considera producción intelectual propia (Art. 27e del Reglamento de Grados y Títulos 2025 aprobado mediante RESOLUCIÓN DE COMISIÓN ORGANIZADORA N° 770-2025-UNACH).

La presente se emite a los interesados para los trámites que estimen conveniente.

Chota, 04 de octubre de 2026.

Dr. Ricardo Abel Del Castillo Torres
Asesor de Informe Final de Tesis
Escuela Profesional de Ingeniería Agroindustrial
Facultad de Ciencias Agrarias

Archivo
RADCT/ ra

ACTA SE SUSTENTACIÓN DE TESIS

UNIVERSIDAD NACIONAL AUTÓNOMA DE CHOTA
VICEPRESIDENCIA ACADÉMICA
Reglamento de Grados y Títulos

ACTA DE SUSTENTACIÓN DE INFORME FINAL DE TESIS

REG. N° 047 -2026 -FCA

El jurado evaluador designado con RESOLUCIÓN DE COORDINACIÓN DE FACULTAD N.° 240-2026-FCA/UNACH:

Nombres y apellidos	Cargo
Dr. Augusto Antonio Mechato Anastasio	Presidente
M.Sc. Rubén Iván Marchena Chanduvi	Secretario
Dr. Tony Steven Chuquizuta Trigoso	Vocal

De la tesis titulada:

"INFLUENCIA DE LAS VARIABLES PRODUCTIVAS Y DE PROCESAMIENTO SOBRE LA CALIDAD EN TAZA DE CAFÉ DE 30 LOCALIDADES DE LA REGIÓN CAJAMARCA MEDIANTE ALGORITMOS DE MACHINE LEARNING CON INTELIGENCIA ARTIFICIAL"

Que ha sustentado el(los) Bachiller (es):

Nombres y apellidos	DNI
Luis Alberto Idrogo Bustamante	73419390
Ronald Yoel Núñez Benavides	71584034

Para obtener el título profesional de:

Ingeniero Agroindustrial

Acuerdan por:

X	Unanimidad		Mayoría

X	Aprobar		Desaprobar

Otorgando la calificación de:

15	Aprobado
	Excelente
X	Bueno
	Regular

	Desaprobado

Colpa Huacaris, 15 de julio del 2026

(Dr. Augusto Antonio Mechato Anastasio)
Presidente

(M.Sc. Rubén Iván Marchena Chanduvi)
Secretario

(Dr. Tony Steven Chuquizuta Trigoso)
Vocal

(Dr. Ricardo Abel Del Castillo Torres)
Asesor

DEDICATORIA

A Dios, por la vida, la sabiduría y la fortaleza para culminar este camino de formación profesional.

A nuestros padres y madres, por su amor incondicional, su sacrificio constante y por ser la base inquebrantable que ha sostenido cada esfuerzo y cada meta alcanzada. Su ejemplo de trabajo es la inspiración de esta investigación.

A los productores de café de la provincia de Cutervo, Cajamarca, quienes con su conocimiento ancestral y su labor diaria nos abrieron las puertas de sus parcelas. A ellos, con la esperanza de que este estudio, se convierta en una herramienta tangible para mejorar la calidad y la sostenibilidad de su café.

AGRADECIMIENTO

Nuestro más profundo agradecimiento se dirige, en primer lugar, a la Universidad Nacional Autónoma de Chota (UNACH) a la Escuela Profesional de Ingeniería Agroindustrial, por brindarnos la plataforma académica y el conocimiento esencial que hizo posible el desarrollo de esta tesis.

Extendemos nuestro sincero reconocimiento a nuestro asesor, el Dr. Ricardo Abel Del Castillo Torres, y a nuestro coasesor, el Mg. Johonathan Baltazar Salazar Campos, por su guía experta, su paciencia invaluable y su apoyo constante en la aplicación de las metodologías avanzadas de Machine Learning y en la interpretación de los resultados.

Agradecemos especialmente al equipo del Laboratorio Ushún Café y a la Asociación de Productores Agropecuarios Faenas de Cutervo por su apoyo logístico fundamental y la facilitación de las muestras. De manera particular, reconocemos la labor Ever Bances Diaz, Catador Profesional Certificado por la Cámara de Café de Cacao del Perú, cuya experiencia y rigor en la aplicación del Protocolo *Specialty Coffee Association* (SCA) garantizaron la calidad técnica de los datos sensoriales.

Finalmente, y con profunda gratitud, agradecemos a cada uno de los productores cafetaleros de las 40 localidades de la provincia de Cutervo, quienes colaboraron activamente con este estudio. Su generosidad al compartir sus conocimientos y sus cosechas fue fundamental y es el verdadero valor social de esta investigación

Índice

RESUMEN **11**
ABSTRACT **12**
CAPÍTULO I. INTRODUCCIÓN **13**
Planteamiento del problema 13
Justificación 15
Formulación del problema 16
Objetivos de investigación 16
Objetivo general 16
Objetivos específicos 16
CAPÍTULO II. MARCO TEÓRICO **18**
2.1. Antecedentes 18
2.2. Bases teóricas 20
2.2.1. Calidad en taza del café *20*
2.2.2. Métodos de evaluación de calidad en taza *21*
2.2.3. Factores que influyen en la calidad en taza *22*
2.3. Definiciones conceptuales 25
2.3.1. Café de especialidad *25*
2.3.2. Fermentación *25*
2.3.3. Café Pergamino *25*
2.3.4. Pilado *26*
2.3.5. Tostado *26*
2.3.6. Catación de café *26*
2.3.7. Inteligencia Artificial *27*
2.3.8. Machine Learning *27*
2.3.9. Algoritmos de ML aplicados al café *27*
CAPÍTULO III. METODOLOGÍA **28**
3.1. Tipo, diseño, nivel y enfoque de investigación 25
3.1.1. Tipo de investigación *28*
3.1.2. Diseño de investigación *28*
3.1.3. Nivel de investigación *28*
3.1.4. Enfoque de investigación *28*
3.2. Población, muestra, muestreo y unidad de análisis 25

3.2.1. Ubicación *29*

3.2.2. Población *29*

3.2.3. Muestra *29*

3.2.4. Muestreo *29*

3.2.5. Unidad de análisis *30*

3.3. Técnicas e instrumentos de recolección de datos 25

3.4. Hipotesis……25

3.5. Operacionalización de variables 25

3.6. Procedimientos de recolección de datos 31

3.6.1. Recolección de datos *33*

3.6.2. Preparación de muestras *33*

3.6.3. Catación *34*

3.6.4. Preparación de la muestra *34*

3.6.5. Preparación de la infusión *34*

3.6.6. Evaluación sensorial *34*

3.6.7. Puntaje y clasificación *35*

3.6.8. Acondicionamiento de datos *35*

3.7. Procedimientos de análsis de datos 34

3.8. Materiales y equipos 36

3.9. Aspectos éticos 36

CAPÍTULO IV. RESULTADOS Y DISCUSIÓN 41

4.1. Resultados del OE1: Generación de la matriz de datos 37

4.1.1. Análisis de las variables de cultivo de café 44

4.1.1.1. Análisis de las variables de cultivo de café 49

4.1.1.2. Análisis de las variables de procesamiento del café 54

4.1.1.3. Análisis de las variables sensoriales de cata de café 55

4.1.2. Preprocesamiento y creación de variables Dummy *68*

4.1.3. Diseño y entrenamiento de modelos de Machine Learning *69*

4.1.4. Validación de modelos y selección del algoritmo óptimo *70*

4.1.5. Evaluación de factores clave mediante análisis SHAP *73*

CAPÍTULO V. CONCLUSIONES Y RECOMENDACIONES 99

5.1. Conclusiones 99

5.2. Recomendaciones 100

REFERENCIAS 103

ANEXOS 110

Índice de tablas

Tabla 1 Técnicas e instrumentos utilizados en la recolección de datos de campo y laboratorio..... 30

Tabla 2 Operacionalización de variables: dimensiones, indicadores y métodos de medición..... 32

Tabla 3 Estructura de la matriz de datos para el modelado (codificación de variables) 35

Tabla 4 Estadísticos descriptivos de los atributos sensoriales de cata (protocolo SCA, n=40)..... 55

Tabla 5 Estadísticos descriptivos agregados del puntaje final en taza (SCA) 67

Tabla 6 Desempeño predictivo agregado de los modelos de ensemble learning (R^2 y RMSE)..... 71

Índice de figuras

Figura 1 Café pilado (oro) libre de defectos, considerado de Especialidad 23
Figura 2 Diferentes tipos de cafés de especialidad 24
Figura 3 Esquema experimental de la investigación 36
Figura 4 Flujograma metodológico del procesamiento computacional, modelado predictivo e interpretabilidad de datos 38
Figura 5 Distritos de la provincia de Cutervo (Cajamarca) de los cuales se extrajeron las muestras del estudio 44
Figura 6 Frecuencia de variedades de café presentes en las muestras del estudio (n=40) 45
Figura 7 Distribución de la altitud de las parcelas cafetaleras (m.s.n.m.) 47
Figura 8 Distribución de la edad de los cultivos de café (años) 48
Figura 9 Distribución del número de cosechas por muestra 49
Figura 10 Distribución y frecuencia de los métodos de fermentación postcosecha (n=40) 50
Figura 11 Distribución del tiempo de fermentación (en horas) por Intervalos de Muestra 52
Figura 12 Distribución y frecuencia de los métodos de secado postcosecha 53
Figura 13 Distribución del tipo de beneficio aplicado (lavado vs. natural) en las muestras 54
Figura 14 Distribución de los puntajes SCA para el aroma 56
Figura 15 Distribución de los puntajes SCA para el sabor 57
Figura 16 Distribución de los puntajes SCA para el sabor residual 58
Figura 17 Distribución de los puntajes SCA para la acidez 58
Figura 18 Distribución de los puntajes SCA para el cuerpo 60
Figura 19 Distribución de los puntajes SCA para el balance 61
Figura 20 Distribución de los puntajes SCA para la taza limpia 62
Figura 21 Distribución de los puntajes SCA para la uniformidad 63
Figura 22 Distribución de los puntajes SCA para el puntaje catador 64
Figura 23 Distribución general del puntaje final de taza de las 40 muestras de cafés evaluados 67
Figura 24 Importancia global de factores sobre la predicción del aroma (SHAP mean absolute) 74
Figura 25 Direccionalidad e impacto de factores sobre la predicción del aroma (SHAP Summary Plot) 74
Figura 26 Importancia global de factores sobre la predicción del sabor (SHAP mean absolute) 76
Figura 27 Direccionalidad e impacto de factores sobre la predicción del sabor (SHAP Summary Plot) 77
Figura 28 Importancia global de factores sobre la predicción del sabor residual (SHAP mean absolute) 79
Figura 29 Direccionalidad e impacto de factores sobre la predicción del sabor residual (SHAP Summary Plot) 80

Figura 30 Importancia global de factores sobre la predicción de la acidez (SHAP mean absolute)..... 82
Figura 31 Direccionalidad e impacto de factores sobre la predicción de la acidez (SHAP Summary Plot)..... 83
Figura 32 Importancia global de factores sobre la predicción del cuerpo (SHAP mean absolute)..... 85
Figura 33 Direccionalidad e impacto de factores sobre la predicción del cuerpo (SHAP Summary Plot) 86
Figura 34 Importancia global de factores sobre la predicción del balance (SHAP mean absolute)..... 88
Figura 35 Direccionalidad e impacto de factores sobre la predicción del cuerpo (SHAP Summary Plot) 88
Figura 36 Importancia global de factores sobre la predicción de la puntuación otorgada por el catador (SHAP mean absolute)..... 90
Figura 37 Direccionalidad e impacto de factores sobre la puntuación otorgada por el catador (SHAP Summary Plot) 91
Figura 38 Importancia global de factores agronómicos y de procesamiento sobre la puntuación final de taza (SHAP) 93
Figura 39 Direccionalidad e impacto global de factores agronómicos y de procesamiento sobre el puntaje final en taza (SHAP Summary Plot) 94
Figura 40 Descomposición de la predicción individual para la puntuación final de taza (SHAP Waterfall Plot)..... 94

Índice de abreviaturas

AFCA: Association of Fine Coffee

IA: Inteligencia Artificial

ML: Machine Learning (Aprendizaje de Máquina)

SCA: Specialty Coffee Association

SHAP: Shapley Additive Explanations

RMSE: Root Mean Square Error (Raíz del Error Cuadrático Medio)

R^2: Coeficiente de determinación

UNACH: Universidad Nacional Autónoma de Chota

Q Grader: Catador certificado de café de especialidad

ICO: International Coffee Organization

RESUMEN

La calidad sensorial del café de especialidad es un factor determinante en la competitividad de los productores cafetaleros de la región Cajamarca, particularmente en la provincia de Cutervo; sin embargo, la alta variabilidad en la calidad de los lotes y la limitada aplicación de herramientas analíticas basadas en datos dificultan la optimización de las prácticas productivas y de procesamiento postcosecha. El objetivo de esta investigación fue analizar y predecir la influencia de las variables productivas y de procesamiento sobre la calidad en taza del café mediante modelos de Machine Learning interpretables. La matriz de datos se amplió de 30 a 40 muestras de café evaluadas bajo el protocolo de catación de la Specialty Coffee Association (SCA), tras identificar en campo comunidades cafetaleras relevantes previamente no registradas, con la finalidad de fortalecer la estabilidad del modelo predictivo y mejorar la capacidad de generalización de los resultados. Se recolectaron variables de cultivo (variedad, altitud, edad del cultivo y número de cosechas) y de procesamiento (método y tiempo de fermentación, método de secado y tipo de beneficio). La metodología se basó en modelos de regresión supervisada, comparando los algoritmos Random Forest, CatBoost y XGBoost mediante validación cruzada y los indicadores R^2 y RMSE. El modelo XGBoost presentó el mejor desempeño predictivo, alcanzando un coeficiente de determinación cercano a 0.99. Los resultados evidenciaron que el tiempo de fermentación controlado, la altitud y la selección varietal, principalmente de la variedad Geisha y Catimor con 48 h de fermentación, son los factores con mayor influencia sobre el puntaje de taza. Se concluye que el modelo desarrollado permite predecir con alta precisión la calidad sensorial del café y es una herramienta objetiva para la toma de decisiones técnicas orientadas a la optimización de los procesos productivos y al fortalecimiento de la competitividad regional.

Palabras clave: Calidad sensorial de café, *Machine Learning* (ML*), Análisis de datos poscosecha, XGBoost, Interpretabilidad (SHAP), Regresión predictiva.*

ABSTRACT

The sensory quality of specialty coffee constitutes a determining factor in the competitiveness of coffee producers in the Cajamarca region, particularly in the province of Cutervo; however, the high variability in batch quality and the limited application of data-driven analytical tools hinder the optimisation of productive and post-harvest processing practices. The objective of this research was to analyse and predict the influence of productive and processing variables on coffee cup quality using interpretable Machine Learning models. The dataset was expanded from 30 to 40 coffee samples evaluated under the Specialty Coffee Association (SCA) cupping protocol; this increase was implemented after identifying previously unregistered but relevant coffee-growing communities in the field, with the purpose of strengthening the stability of the predictive model and improving the generalisation capacity of the results. Cultivation variables (variety, altitude, crop age and number of harvests) and processing variables (fermentation method and duration, drying method and type of processing) were collected. The methodology was based on supervised regression models, comparing the Random Forest, CatBoost and XGBoost algorithms using cross-validation and the performance indicators R^2 and RMSE. The XGBoost model exhibited the highest predictive performance, achieving a coefficient of determination close to 0.99. Interpretation of the results using the SHAP (Shapley Additive exPlanations) method revealed that controlled fermentation time, altitude and varietal selection—particularly the Geisha variety—were the variables with the greatest influence on cup score. It is concluded that the developed model predicts sensory quality with high accuracy and constitutes an objective tool for technical decision-making aimed at optimising productive processes and strengthening regional competitiveness.

Keywords: Sensory quality of coffee, Machine Learning (ML), Post-harvest computational processing, XGBoost, Interpretability (SHAP), Predictive regression.

CAPÍTULO I. INTRODUCCIÓN

Planteamiento del problema

El café mueve uno de los mercados agrícolas más grandes del mundo, y en el segmento de especialidad el precio lo fija la taza. Aroma, sabor, acidez y cuerpo deciden a qué mercado entra un lote. Cuando esos atributos varían entre cosechas, el comprador lo nota. Y lo descuenta. En regiones con poca tecnificación esa variabilidad es la norma, no la excepción (ICO, 2023).

Cajamarca reúne condiciones agroecológicas favorables para café de altura. En Cutervo, sin embargo, la producción descansa en agricultura familiar con equipamiento escaso (Jezeer et al., 2018). Los productores no controlan de forma sistemática el tiempo que fermentan ni la manera en que secan. Tampoco llevan un registro que relacione esas decisiones con el puntaje que obtienen después. El resultado son lotes desiguales, y un comprador que no puede anticipar qué recibirá en la siguiente entrega.

Brasil y Colombia llevan años incorporando analítica de datos a la gestión de calidad del café. En Perú ese uso sigue siendo escaso a nivel productivo. La evaluación descansa en la cata de expertos, un método estandarizado y fiable, pero costoso y que no deja constancia de qué prácticas de campo produjeron cada resultado (Rashid et al., 2021).

A nivel internacional, diversas investigaciones han demostrado que el uso de algoritmos de *Machine Learning* (ML) en la industria cafetalera ha permitido mejorar la predicción de calidad en taza y optimizar procesos postcosecha. Estudios recientes han mostrado que modelos como *XGBoost*, *CatBoost* y *Random Forest* pueden predecir con alta precisión la calidad sensorial del café a partir de datos sobre su producción y procesamiento (Putra et al., 2023). Otros trabajos han probado ensamblados de árboles sobre datos de producción y procesamiento. En el Perú estas aplicaciones apenas empiezan. Ninguna publicada trabaja con datos de Cutervo.

Sin registro propio, el caficultor de Cutervo no puede demostrar consistencia. Sin consistencia demostrable, no accede a los precios del segmento de especialidad (ICO, 2023).

La brecha no es de calidad potencial. Es de información.

Frente a esta limitación, nuestra investigación propone utilizar el aprendizaje automático para analizar la calidad del café de forma objetiva. Al aplicar estas herramientas a los lotes de Cutervo, buscamos entender qué factores productivos impactan la taza final. Nuestro objetivo es reducir la dependencia de evaluaciones tradicionales y dar a los caficultores de la zona datos concretos para competir en el mercado de especialidad.

Justificación

El café peruano, especialmente el de la región de Cajamarca, es ampliamente reconocido por su calidad sensorial y potencial en los mercados de especialidad (SCA, 2024). La provincia de Cutervo, dentro de esta región, cuenta con condiciones agroecológicas favorables para la producción de cafés con atributos sensoriales distintivos, como notas afrutadas, florales y dulces. Sin embargo, la falta de tecnología y conocimiento técnico en las etapas de producción y procesamiento impide que estos atributos se expresen de manera consistente en los lotes comercializados.

Esa constancia depende de decisiones que se toman en finca: cuándo cosechar, cuánto fermentar, cómo secar. Hoy se toman por criterio heredado, sin medir el efecto que tienen sobre la taza. Ningún trabajo publicado ha relacionado esas decisiones con el puntaje SCA en la provincia.

Esta investigación construye ese registro. Cuarenta lotes de nueve distritos, catados bajo protocolo SCA por un catador certificado, con sus variables de cultivo y poscosecha anotadas en campo. Sobre esa matriz se aplican modelos de regresión basados en ensamblados de árboles.

Diversos estudios han demostrado la efectividad de estos modelos en la industria cafetalera. Putra et al. (2023) reportaron que el uso de *LightGBM* permitió alcanzar una precisión del 72% en la predicción de calidad del café, mientras que Huber et al. (2022) compararon XGBoost con arquitecturas de aprendizaje profundo para estimar rendimiento agrícola a partir de datos de teledetección, y obtuvieron mejor desempeño con el ensamblado de árboles. En este estudio, el uso de ML permitirá parametrizar variables críticas, como la altitud del cultivo, los métodos de fermentación y el tiempo de secado, con el fin de proporcionar recomendaciones precisas para mejorar la calidad en taza.

El resultado directo es una base de datos que antes no existía. A partir de ella, una cooperativa de la zona puede comparar sus propias prácticas con las de los 40 lotes evaluados aquí. Esto se traduce directamente en la capacidad de comercializar lotes consistentes en nichos de alto valor, incrementando sus márgenes de ganancia y reduciendo la vulnerabilidad económica de la agricultura familiar rural.

Formulación del problema

¿Cuál es la influencia de las variables productivas y de procesamiento en la calidad en taza del café en 40 localidades de la región Cajamarca mediante algoritmos de Machine Learning?

Objetivos de investigación

Objetivo general

Determinar la influencia de las variables productivas y de procesamiento en la calidad en taza del café en 40 localidades de la región Cajamarca mediante algoritmos de Machine Learning.

Objetivos específicos

- Generar una matriz de datos estructurada a partir de la recopilación de información

sobre variables productivas, de procesamiento y atributos sensoriales del café, garantizando la calidad y consistencia de los datos.

- Implementar técnicas de preprocesamiento, imputación de valores faltantes y creación de variables Dummy (One-Hot Encoding) para asegurar la estandarización de la matriz de datos antes del modelado.
- Diseñar y entrenar un modelo de aprendizaje de máquina basado en algoritmos de *gradient boostings* (*XGBoost*, *CatBoost* y *Random Forest*) que permita correlacionar las variables productivas y de procesamiento con la calidad en taza del café.
- Validar el modelo desarrollado mediante técnicas de validación cruzada y métricas como el coeficiente de determinación (R^2) y la raíz cuadrada del error medio (RMSE), seleccionando el algoritmo con mejor desempeño.
- Evaluar los factores clave que determinan la calidad sensorial del café utilizando diagramas de cascada (*waterfall)* basados en valores SHAP para garantizar una interpretación adecuada de los resultados.

CAPÍTULO II. MARCO TEÓRICO

2.1. Antecedentes

En el ámbito del modelado predictivo agrícola, Huber et al. (2022) compararon Extreme Gradient Boosting con arquitecturas de aprendizaje profundo para estimar rendimiento a partir de datos de teledetección. El ensamblado de árboles superó a las redes neuronales en ese conjunto. Su aportación a esta investigación es metodológica: respalda el uso de modelos basados en árboles sobre matrices tabulares de tamaño moderado, que es la situación de este estudio.

Así mismo dentro de la predicción de la calidad de café, Putra et al. (2023) aplicaron *LightGBM* (*Light Gradient Boosting Machine*) para clasificar muestras de café en diferentes categorías de calidad sensorial. Usaron un conjunto de datos con información sobre composición química y atributos físicos del grano, logrando una precisión del 72% en la clasificación de café de especialidad. La investigación destacó que el modelo podía ajustarse con técnicas de optimización de hiperparámetros para mejorar su rendimiento.

Rashid et al. (2021) implementaron una combinación de *Random Forest* y redes neuronales artificiales (ANNs) para evaluar la calidad del café con base en variables productivas y climáticas. Su estudio indicó que las redes neuronales lograron una mayor capacidad de generalización, alcanzando una precisión del 90% en la predicción de perfiles sensoriales. Concluyeron que la inteligencia artificial tiene el potencial de sustituir parcialmente las pruebas de cata sensorial, reduciendo costos y tiempos en la evaluación de calidad.

Silveira et al. (2016) realizaron un análisis de la calidad sensorial del café en Brasil utilizando técnicas estadísticas avanzadas. Aplicaron modelos de análisis discriminante y regresión logística para evaluar el impacto de la altitud y el tipo de secado en la calidad del

café. Concluyeron que los cafés cultivados por encima de los 1,200 m.s.n.m. y sometidos a procesos de secado controlados obtuvieron puntuaciones significativamente más altas en cata, lo que respalda la hipótesis de la presente investigación.

Avances recientes en IA aplicada a la evaluación del café

Para ampliar este punto de vista sobre la aplicación de ML en la industria cafetalera, es importante considerar estudios recientes que han utilizado técnicas innovadoras para optimizar la calidad del producto.

Un estudio desarrollado por Altube et al. (2018) exploró el uso de imágenes hiperespectrales para analizar las características físicas del café tostado y clasificarlas según su grado de tueste. Utilizando métodos quimiométricos avanzados, los autores lograron distinguir con alta precisión entre diferentes niveles de tostado, lo que evidencia el potencial de las técnicas de visión espectral para complementar o incluso sustituir parcialmente las evaluaciones sensoriales tradicionales.

Diversos estudios han demostrado la utilidad de modelos de aprendizaje automático para predecir el rendimiento agrícola a partir de variables ambientales. Por ejemplo, Jeong et al. (2016) emplearon *Random Forest* y *Support* Vector *Machines* (SVM) para estimar el rendimiento de maíz utilizando como predictores factores climáticos como temperatura,

precipitación y radiación solar. En su estudio, *Random Forest* mostró el mejor desempeño, alcanzando un R^2 superior a 0.80. Estos resultados evidencian el potencial de los modelos de aprendizaje automático para mejorar la toma de decisiones agronómicas y optimizar la planificación productiva.

Ceballos et al. (2023) implementaron algoritmos de optimización basada en ML para ajustar los parámetros del beneficio húmedo del café, reduciendo el consumo de agua y optimizando el tiempo de fermentación. Utilizando modelos de optimización multiobjetivo, lograron mejorar la eficiencia del proceso en un 25%, manteniendo la calidad sensorial del café. Este estudio refuerza la importancia de aplicar IA en la mejora de procesos postcosecha.

Brecha de conocimiento y relevancia de la presente investigación

Si bien los estudios previos han demostrado la efectividad del ML en la predicción de calidad en productos agroindustriales, la mayoría se han realizado en contextos con infraestructura tecnológica avanzada. En Cutervo (Perú), los pequeños productores carecen de acceso a herramientas analíticas y no se han implementado modelos de inteligencia artificial para la optimización de la calidad del café.

La presente investigación busca llenar esta brecha mediante el desarrollo de un modelo de ML accesible y eficiente que permita correlacionar variables productivas y de procesamiento con la calidad sensorial del café, optimizando los procesos postcosecha y mejorando la competitividad de los productores en mercados de alto valor.

2.2. Bases teóricas

A continuación, se indican las bases teóricas-científicas que forman parte del presente estudio.

2.2.1. Calidad en taza del café

La calidad en taza del café es el principal criterio de diferenciación en la

industria de cafés de especialidad y se encuentra determinada por atributos sensoriales que definen su perfil organoléptico. La Specialty Coffee Association (SCA) ha desarrollado un sistema de puntuación en una escala de 0 a 100, estableciendo que los cafés con 80 puntos o más son considerados de especialidad, lo que les confiere un valor comercial superior en mercados especializados (SCA, 2024).

Los principales atributos que definen la calidad en taza son:

- **Fragancia/aroma**: Intensidad y complejidad de los compuestos volátiles que se perciben antes y después de la infusión.
- **Sabor**: Perfil de sabor general, que incluye dulzura, notas afrutadas o florales y otros matices.
- **Acidez**: Sensación brillante y vibrante en el paladar, asociada a compuestos orgánicos presentes en el café.
- **Cuerpo**: Percepción de densidad y textura del café en la boca.
- **Uniformidad**: Consistencia de la muestra en todas las tazas evaluadas.
- **Balance**: Armonía entre acidez, dulzura, cuerpo y retrogusto.
- **Retrogusto**: Persistencia del sabor tras la degustación.

La evaluación de estos atributos sigue protocolos estandarizados de catación que garantizan reproducibilidad y objetividad, minimizando la subjetividad en la apreciación sensorial (Ting et al., 2024).

2.2.2. Métodos de evaluación de calidad en taza

Para garantizar la calidad del café, se han desarrollado diversas metodologías de evaluación que combinan técnicas sensoriales tradicionales con enfoques científicos avanzados.

2.2.2.1. Evaluación sensorial (Catación del café)

Es el método más utilizado tanto en la industria cafetalera como en la investigación

científica. La catación sigue los estándares de la SCA y la Asociación de Cafés Finos de América (AFCA), en los cuales catadores certificados (Q Graders) asignan puntuaciones a muestras de café bajo condiciones controladas (ICO, 2023).

2.2.2.2. Análisis fisicoquímico

La evaluación instrumental de la calidad del café permite cuantificar compuestos clave mediante:

- **Ácidos orgánicos y cafeína:** Determinantes de la acidez y el amargor.
- **Compuestos volátiles:** Principales responsables del perfil aromático.
- **Humedad y densidad del grano:** Factores que afectan la frescura y el desarrollo del tostado (Cao et al., 2023).

2.2.2.3. Visión por computadora y espectroscopía

- **Análisis de imágenes hiperespectrales:** Permite detectar defectos físicos en el grano.
- **Espectroscopía de infrarrojo cercano (NIR):** Técnica de evaluación rápida y no destructiva utilizada para predecir la calidad del café en taza con alta precisión.

2.2.2.4. ML aplicado a la calidad en taza

- Modelos como *Random Forest, XGBoost* y redes neuronales han sido empleados para correlacionar variables productivas con la calidad en taza.
- Estudios recientes han demostrado que estos modelos pueden predecir con una precisión superior al 85% la puntuación de calidad en taza basándose en datos de producción y procesamiento (Putra et al., 2023).

2.2.3. Factores que influyen en la calidad en taza

Uno de los factores determinantes en la calidad en taza del café es la ausencia de defectos físicos y químicos en el grano, ya que cualquier imperfección puede afectar negativamente el perfil sensorial. Los cafés de especialidad deben presentar granos

homogéneos, con adecuada densidad y sin defectos visibles, asegurando una mayor consistencia en la taza final (SCA, 2024).

Figura 1

Café pilado (oro) libre de defectos, considerado de Especialidad



Nota. Granos de café especiales de la Asociación de Productores Agropecuarios Naranjo Yacu del distrito de La Capilla, provincia de Cutervo.

2.2.3.1. Condiciones de cultivo

Las características agroclimáticas influyen significativamente en la calidad del café. La altitud, temperatura, humedad y composición del suelo afectan la biosíntesis de compuestos aromáticos en el grano. Se ha documentado que cultivos a más de 1,200 m.s.n.m. producen cafés con mayor complejidad aromática y acidez debido a una maduración más lenta del fruto (Silveira et al., 2016).

2.2.3.2. Procesamiento postcosecha

El método de procesamiento impacta directamente la percepción sensorial del café. Se han identificado tres técnicas principales:

- **Lavado:** Destaca la acidez y claridad del café.
- **Natural:** Promueve sabores dulces y un cuerpo más pronunciado.

- **Honey:** Genera un balance entre dulzura y acidez.

Figura 2

Diferentes tipos de cafés de especialidad.



Nota. Laboratorio de Ushún Café de la campaña 2024 en provincia de Cutervo.

2.2.3.3. Almacenamiento y tostado

- El almacenamiento en condiciones de baja humedad y temperatura controlada permite preservar las propiedades sensoriales del café verde.
- El proceso de tostado modifica los compuestos químicos del grano, desarrollando los perfiles aromáticos y potenciando atributos como el cuerpo y la acidez. No obstante, un tostado excesivo puede degradar la calidad sensorial, resultando en sabores quemados o amargos (Cao et al., 2023).

2.2.3.4. Método de preparación

- Parámetros como temperatura del agua (90-96 °C), tiempo de extracción y molienda pueden modificar significativamente la percepción sensorial del café.
- Diferentes métodos de preparación, como V60, prensa francesa o espresso, afectan la intensidad y el cuerpo de la bebida final (Ting et al., 2024).

2.3. Definiciones conceptuales

2.3.1. Café de especialidad

El café de especialidad es aquel que presenta atributos sensoriales distintivos y obtiene una puntuación superior a 80 puntos en una escala de 0 a 100, según los estándares de la Specialty Coffee Association (SCA, 2024). La calificación debe ser otorgada por un Q Grader certificado, quien evalúa las características sensoriales del café en taza. Para ser considerado de especialidad, el café no solo debe destacar en calidad en taza, sino que también debe presentar granos libres de defectos físicos, ya que una alta presencia de estos puede descalificarlo, independientemente de su puntaje en cata (SCA, 2024).

2.3.2. Fermentación

La fermentación en café es un proceso bioquímico en el cual microorganismos como bacterias y levaduras descomponen compuestos orgánicos del mucílago del grano, generando subproductos que influyen en su perfil sensorial y calidad en taza. Dependiendo del tiempo y las condiciones del proceso, la fermentación puede intensificar la acidez, dulzura y complejidad aromática del café. Los métodos de procesamiento como natural, honey y lavado varían en el nivel de fermentación aplicada, afectando significativamente el sabor del producto final (Shen et al., 2025).

2.3.3. Café Pergamino

El café pergamino es el estado intermedio del grano tras el proceso de beneficio húmedo. Después de la remoción de la cáscara y el mucílago, el grano es secado hasta alcanzar un contenido de humedad del 10-12%, conservando una capa protectora similar a un pergamino, de donde proviene su nombre. Este es el estado en el que los caficultores comercializan el café antes de su procesamiento final, como el pilado o tostado, que lo preparan para la exportación o el consumo (Martínez et al., 2023).

2.3.4. Pilado

El pilado es el proceso de remoción de la capa de pergamino que cubre el grano seco de café, obteniendo así el café verde u oro. Esta etapa es fundamental en la preparación del café para su comercialización internacional.

2.3.5. Tostado

El tostado es el proceso térmico mediante el cual el café verde es sometido a temperaturas entre 180°C y 230°C en tiempos controlados, generando una serie de reacciones químicas que desarrollan sus atributos sensoriales. El tipo de tostado afecta directamente el cuerpo, acidez y dulzura del café, siendo determinante en la experiencia del consumidor (Cao et al., 2023).

2.3.6. Catación de café

La catación es el método estandarizado para evaluar la calidad sensorial del café mediante la percepción de sus atributos. Según Ting et al. (2024), los principales factores analizados en la catación son:

- **Fragancia y aroma:** Evaluación de los compuestos volátiles del café molido en seco y después de la infusión.
- **Sabor:** Percepción integrada del aroma, acidez y cuerpo.
- **Retrogusto:** Persistencia del sabor tras la degustación.
- **Acidez:** Sensación de frescura y brillo en boca, predominante en cafés lavados.
- **Cuerpo:** Percepción de densidad y textura del café en boca.
- **Balance:** Armonía entre dulzura, acidez, cuerpo y retrogusto.
- **Dulzor:** Presencia de notas dulces y agradables en el café.
- **Taza limpia:** Ausencia de defectos sensoriales.
- **Uniformidad:** Consistencia de las características sensoriales en múltiples tazas.
- **Puntuación final:** Evaluación global del café según estándares de la SCA (2024).

2.3.7. Inteligencia Artificial

La inteligencia artificial (IA) es un campo interdisciplinario que busca desarrollar sistemas capaces de realizar actividades que requieren inteligencia humana, como identificación de patrones y la toma de decisiones. Su aplicación en la industria cafetalera ajusta la producción y mejora la calidad del café en taza por medio de modelos predictivos y análisis avanzados de datos. Un ejemplo de su utilidad reside en la evaluación no destructiva del grano, donde los algoritmos de aprendizaje profundo se combinan con la visión sistematica para identificar defectos físicos y químicos. Este enfoque mejora la trazabilidad y aporta a la uniformidad del lote, factores relevantes para la obtención de cafés de especialidad (Gonzalez-Sanchez et al., 2024).

2.3.8. Machine Learning

El aprendizaje de máquina es una área correspondiente de la inteligencia artificial que se enfoca en algoritmos que pueden aprender patrones en los datos y mejorar sus predicciones sin intervención humana especifica. En el análisis del café, permite modelar la relación entre variables productivas y sensoriales para optimizar la calidad en taza (Mahesh, 2020).

2.3.9. Algoritmos de ML aplicados al café

- **XGBoost:** Técnica de gradient boosting basada en árboles de decisión, destacada por su eficiencia y precisión en modelos predictivos (Chen et al., 2020).
- **CatBoost:** Algoritmo optimizado para manejar variables categóricas, reduciendo el overfitting y tiempos de entrenamiento (Dorogush et al., 2018).
- **Random Forest:** Ensamblado de múltiples árboles de decisión que reduce el sobreajuste y mejora la estabilidad del modelo (Breiman, 2001).

CAPÍTULO III. METODOLOGÍA

3.1. Tipo, diseño, nivel y enfoque de investigación

3.1.1. Tipo de investigación

El presente estudio es una investigación de tipo aplicada, ya que su objetivo es desarrollar soluciones tecnológicas prácticas que puedan ser implementadas por productores y empresas cafetaleras. Los resultados de esta investigación permitirán optimizar la toma de decisiones en el manejo agrícola y mejorar la calidad del café, contribuyendo así a un desarrollo económico sostenible en la región.

3.1.2. Diseño de investigación

El diseño de esta investigación es de tipo no experimental y transversal. Se recopilaron datos en un momento específico de la cadena de producción del café, centrando el análisis en las etapas de cosecha y postcosecha, sin manipulación de las variables productivas. Los datos obtenidos fueron utilizados para desarrollar modelos de regresión mediante algoritmos de aprendizaje automático que permitirán establecer relaciones causales entre las variables productivas y la calidad sensorial del café.

3.1.3. Nivel de investigación

El nivel de esta investigación es explicativo; nuestra finalidad principal es demostrar la causalidad empírica que existe entre las prácticas agronómicas adoptadas en campo y las variaciones del perfil de taza final. Permitiendo explicar cómo estas variables interactúan entre sí y afectan los atributos sensoriales, mediante el uso de modelos predictivos basados en Machine Learning.

3.1.4. Enfoque de investigación

El enfoque de la investigación es cuantitativo, debido a que se basa en la recolección y análisis de datos numéricos relacionados con variables productivas, de procesamiento y atributos sensoriales del café. Estos datos han sido procesados mediante técnicas estadísticas y

algoritmos de aprendizaje automático, permitiendo medir, modelar y predecir la calidad en taza de manera objetiva y reproducible.

3.2. Población, muestra, muestreo y unidad de análisis

3.2.1. Ubicación

La investigación se desarrolló en la provincia de Cutervo, región Cajamarca, Perú, abarcando nueve distritos cafetaleros donde se recolectaron las muestras de estudio.

El análisis de laboratorio se realizó en el Laboratorio Ushún Café, ubicado en la ciudad de Cutervo, donde se llevaron a cabo las pruebas de calidad física y sensorial bajo estándares de la Specialty Coffee Association (SCA).

3.2.2. Población

La población del estudio está conformada por las unidades productoras de café en 40 localidades de la región Cajamarca, que incluyen datos de producción, cosecha, postcosecha y transformación, fundamentales para evaluar la calidad en taza.

3.2.3. Muestra

Para este estudio, se utilizaron 300 gramos de café pergamino de cada una de las 40 muestras recolectadas. Estas muestras fueron obtenidas de 24 comunidades, distribuidas en 8 distritos de la provincia de Cutervo. El proceso de selección buscó representar la diversidad de condiciones productivas de la región.

3.2.4. Muestreo

El muestreo utilizado en esta investigación fue de tipo no probabilístico por conveniencia, debido a las limitaciones logísticas y la disponibilidad de productores en la zona. Se seleccionaron 40 productores cafetaleros, quienes fueron previamente identificados en la base de datos del proyecto Avanzar Rural, gestionado por la oficina Zonal de Cutervo. Las muestras fueron recolectadas de manera aleatoria en 8 distritos de la provincia de Cutervo,

región Cajamarca. Posteriormente, las muestras fueron acondicionadas en un lugar fresco y sin humedad para asegurar su calidad y luego trasladadas al Laboratorio Ushún Café, en la ciudad de Cutervo, para su análisis correspondiente.

3.2.5. Unidad de análisis

La unidad de análisis estuvo conformada por cada muestra de café pergamino (300 g) recolectada de los productores cafetaleros de las diferentes localidades de la provincia de Cutervo.

Cada muestra representa una combinación específica de condiciones productivas y de procesamiento, sobre la cual se evaluaron los atributos sensoriales de calidad en taza mediante el protocolo SCA.

3.3. Técnicas e instrumentos de recolección de datos

En la Tabla 1 a continuación, se detallan las técnicas e instrumentos empleados en la recolección de datos durante las diferentes etapas del desarrollo del estudio:

Tabla 1

Técnicas e instrumentos utilizados en la recolección de datos de campo y laboratorio

Técnicas	Instrumentos	Recolección de datos
Análisis inmediato	Selección y clasificación	Calidad visual de granos sin desperfectos.
Mediciones	Balanza analítica	Pesos de 300 g en pergamino
Descripción de datos de producción	Ficha de catación	Detección de variedad, altura de cultivo, edad del cultivo y n° de cosecha de la planta.
Descripción de datos de procesamiento	Ficha de catación	Detección de horas de fermentación, métodos de fermentación, método de secado y tipo de beneficios

Catación	Escala hedónica	Aceptabilidad general de fragancia, sabor, acidez, cuerpo, balance, postgusto, uniformidad, dulzura, defectos, puntaje global.
Machine Learning	Con algoritmos gradiente boosting para regresión: XGBoost, CatBoost y Random forest en lenguaje Python 3.11.2 con la librería SHAP	Diseño del modelo computacional que atribuya mejores cualidades sensoriales a partir de sus condiciones de producción y procesamiento de café.

3.4. Hipótesis

Se espera que los algoritmos de ML identifiquen y modelen las relaciones clave entre las variables productivas y de procesamiento, permitiendo predecir la calidad sensorial del café en taza en 40 localidades de la región Cajamarca, de acuerdo con los estándares internacionales.

3.5. Operacionalización de variables

El presente estudio considera tanto variables productivas como variables de procesamiento, debido a su rol complementario en la determinación de la calidad en taza del café. A diferencia de los cafés convencionales, que emplean principalmente el proceso lavado, los cafés de especialidad abarcan una diversidad de metodologías postcosecha, como el honey y el natural, además del lavado. Estas técnicas de procesamiento influyen significativamente en el perfil sensorial del café, afectando atributos clave como la acidez, dulzura, cuerpo y complejidad aromática (Silveira et al., 2016).

La calidad final del café en taza es el resultado correlacional entre las variables productivas y de procesamiento, en donde su análisis conjunto es crucial para la precisión del modelo de ML. Factores como la altitud, variedad del cultivo, edad de la planta y condiciones climáticas tienen impacto en la composición química del grano, mientras que las

condiciones de fermentación y secado determinan el desarrollo de compuestos aromáticos esenciales (Ceballos et al., 2023). Subestimar parte de estos factores limitaría la capacidad del modelo para predecir de manera precisa la calidad sensorial del café, lo que restaría idoneidad a los resultados.

Además, en el proceso de beneficio y aprovechamiento, el café sigue una cadena de valor interdependiente, en donde las decisiones tomadas en cada etapa afectan directamente las características sensoriales del producto final. Por ello, la combinación de variables productivas y de procesamiento permite obtener un análisis más completo, proporcionando herramientas para la toma de decisiones en la mejora de la calidad del café de especialidad y optimizando su competitividad en mercados internacionales.

En este contexto, este estudio no se limita a un único método de procesamiento, sino que evalúa el impacto del honey, natural y lavado, con el fin de obtener una modelización más robusta de la calidad sensorial del café. Esta amplitud metodológica no solo permite predecir la calidad con mayor precisión, sino que también ofrece un espectro más representativo de la producción de cafés de especialidad, diferenciándose de los estudios previos que han considerado únicamente el método de procesamiento lavado.

La Tabla 2, muestra las variables independientes, dependientes, indicadores y el método para cada análisis que se realizó.

Tabla 2

Operacionalización de variables: dimensiones, indicadores y métodos de medición

Variables independientes	Dimensiones	Indicadores	Método
Condiciones de producción	**1.** Variedad **2.** Altura de cultivo **3.** Edad del cultivo **4.** N° de cosecha de la planta	**1.** Caturra, Borbón, geisha. **2.** 1500 – 2200 m.s.n.m. **3.** 1 – 3 años. **4.** 1 – 9 cosechas.	Registros de campo mediante ficha de cata.

Condiciones de procesamiento	**1.** Horas de fermentación **2.** Métodos de fermentación **3.** Método de secado **4.** Tipo de beneficios	**1.** 12 – 72h **2.** Tanque tina, bidón hermético, saco de plástico, etc. **3.** En suelo, secador solar, etc. **4.** Lavado, honey, natural.	Registros de campo mediante ficha de cata.

Variables dependientes	**Dimensiones**	**Indicadores**	**Método**
Calidad sensorial del café	**1.** Fragancia **2.** Sabor **3.** Post gusto **4.** Acidez **5.** Cuerpo **6.** Uniformidad **7.** Balance **8.** Dulzura **9.** Tazas limpias **10.** Taza general **11.** Defectos	Puntuación en la escala SCA (0-100) sumado por todas las cualidades.	Evaluación por panel de catadores, procesamiento de datos con análisis de datos con ML para correlacionar atributos sensoriales.

3.6. Procedimientos de recolección de datos

3.6.1. Recolección de datos

En colaboración con el proyecto Avanzar Rural de la OZ Cutervo, se seleccionaron al menos 40 productores de café de la provincia de Cutervo, con el fin de obtener un total de 300 gramos de café pergamino por muestra, con un contenido de humedad entre el 10% y el 12%. Además, se completó una ficha de cata (Anexo A1) para registrar las características de producción y procesamiento del café. Se recopilaron datos sobre variables productivas y de procesamiento de 40 localidades diferentes en la provincia de Cutervo, región Cajamarca.

3.6.2. Preparación de muestras

Una vez recolectadas las muestras, se asigna un código a cada una, conforme al formulario de registro de cata. Las muestras fueron verificadas en el Laboratorio Ushún Café

en la ciudad de Cutervo, donde se evaluó que cumplan con el rango de humedad establecido. Este paso es crucial, ya que una humedad fuera de los parámetros puede alterar los resultados de la evaluación sensorial.

3.6.3. Catación

La cata fue realizada por un Q Grader certificado, especialista en evaluación sensorial del café (el certificado que avala la competencia técnica del evaluador se adjunta en el Anexo A5), siguiendo el protocolo estandarizado de la Specialty Coffee Association (SCA, 2024). Este protocolo garantiza una evaluación objetiva y consistente de los atributos sensoriales del café, como fragancia, acidez, cuerpo, balance y retrogusto. La calificación de las muestras se registrará utilizando la hoja de cata oficial de la SCA, asegurando rigurosidad en la recopilación y análisis de los datos sensoriales.

3.6.4. Preparación de la muestra

El proceso de cata comienza con la selección y tostado de las muestras (ver Anexo A7, Figura A7.1). Los granos deben tostarse entre 8 y 12 minutos, alcanzando un nivel de tostado ligero a medio, según la escala Agtron. Posteriormente, se dejan reposar durante al menos 8 horas antes de proceder con la cata. La molienda debe realizarse justo antes de la cata para preservar la frescura del aroma, utilizando una molienda ligeramente más gruesa que la empleada para filtros de papel.

3.6.5. Preparación de la infusión

El café molido se colocó en vasos de cata estándar (5-6 onzas) a una proporción de 8.25 gramos por cada 150 ml de agua. El agua, a una temperatura aproximada de 93 °C, se vertió directamente sobre los granos y se dejó reposar durante 3 a 5 minutos antes de romper la costra superficial.

3.6.6. Evaluación sensorial

La cata sigue una secuencia en la evaluación de atributos sensoriales, los cuales se

registran en una hoja de cata estandarizada. Los atributos evaluados incluyen fragancia/aroma, sabor, acidez, cuerpo, balance, uniformidad, dulzura, claridad y postgusto. Además, se identifican posibles defectos en la muestra, que se clasifican como "Taints" (de menor intensidad) o "Faults" (de mayor gravedad), los cuales afectan la puntuación final.

3.6.7. Puntaje y clasificación

Cada atributo se puntúa en una escala de 6 a 9, con incrementos de cuartos de punto, lo que permite un análisis detallado de la calidad del café. Las muestras que obtienen puntajes superiores a 80 son consideradas como cafés especiales, mientras que aquellas que reciben puntuaciones inferiores se clasifican en categorías de calidad comercial o defectuosa. El registro detallado de las evaluaciones sensoriales para cada una de las muestras evaluadas se presenta en el Anexo A6.

3.6.8. Acondicionamiento de datos

Una vez obtenidos los puntajes de las 40 muestras, los datos fueron registrados en un archivo de Microsoft Excel para su posterior análisis. Se incluyeron tanto las características productivas como las cualidades sensoriales del café, de acuerdo con el siguiente orden: código de muestra, variables productivas, variables de procesamiento y puntaje en taza, en coherencia con la Tabla 3.

Tabla 3

Estructura de la matriz de datos para el modelado (codificación de variables)

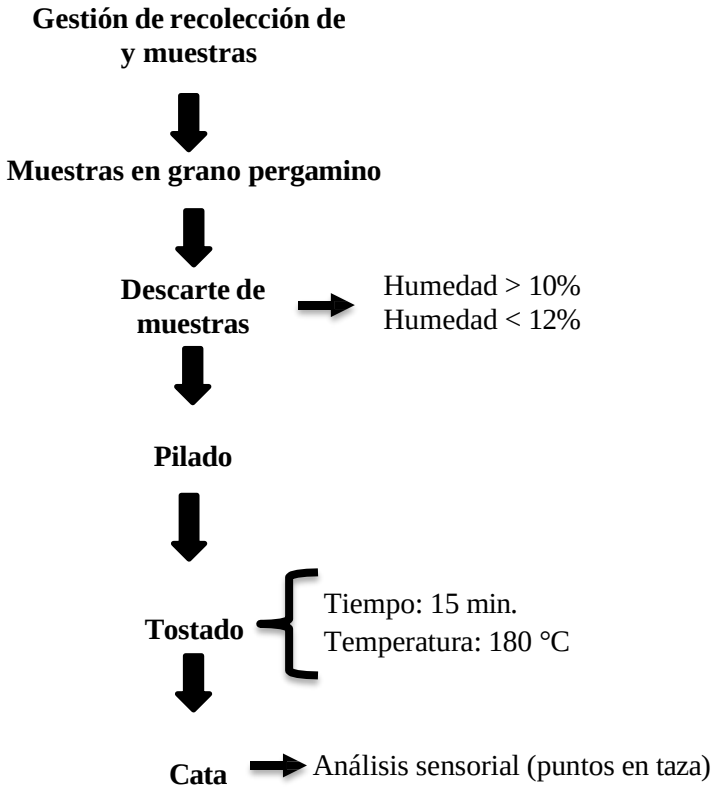
Muestra	**Código de ficha de cata**	**4 variables de Condiciones de producción**	**4 variables de Condiciones de procesamiento**	**10 variables de Calidad sensorial**
1	*501-1*	*Valores de indicador*	*Valores de indicador*	*Valores de indicador*
2	*502-2*	*Valores de indicador*	*Valores de indicador*	*Valores de indicador*
3	*503-3*	*Valores de indicador*	*Valores de indicador*	*Valores de indicador*

.	.	.	.	.
.	.	.	.	.
.	.	.	.	.

Bajo este fichero es que se puede quedar lista la información necesaria para pasar a la etapa de procesamiento de información mediante las técnicas de Inteligencia Artificial propuestas en el estudio.

Figura 3

Esquema experimental de la investigación



3.7. Procedimientos de análisis de datos

El análisis de datos se desarrolló utilizando modelado con algoritmos de ML basados en gradient boosting para regresión: XGBoost, CatBoost y Random Forest. Las variables de entrada (features) fueron las características productivas y de procesamiento, mientras que la variable de salida (response) fue la puntuación de taza contemplando todas sus dimensiones.

a. **Manejo de datos faltantes:** Dado que existen valores faltantes en la base de datos, se aplicó la técnica de imputación utilizando la función *KNNImputer* de la librería sklearn.impute para

completar los datos.

b. **Preprocesamiento de datos:** Para garantizar la correcta interpretación de las características agronómicas y tecnológicas por parte de los algoritmos, las variables categóricas de entrada (variedad, método de fermentación, método de secado y tipo de beneficio) fueron transformadas numéricamente mediante la creación de variables Dummy (One-Hot Encoding). Este proceso se ejecutó utilizando la función *pd.get_dummies* de la librería Pandas en Python, lo que permitió convertir estas categorías en una matriz binaria sin imponer jerarquías artificiales, asegurando así la compatibilidad estructural de los datos con los modelos de Ensemble Learning (DataScientest, 2020).

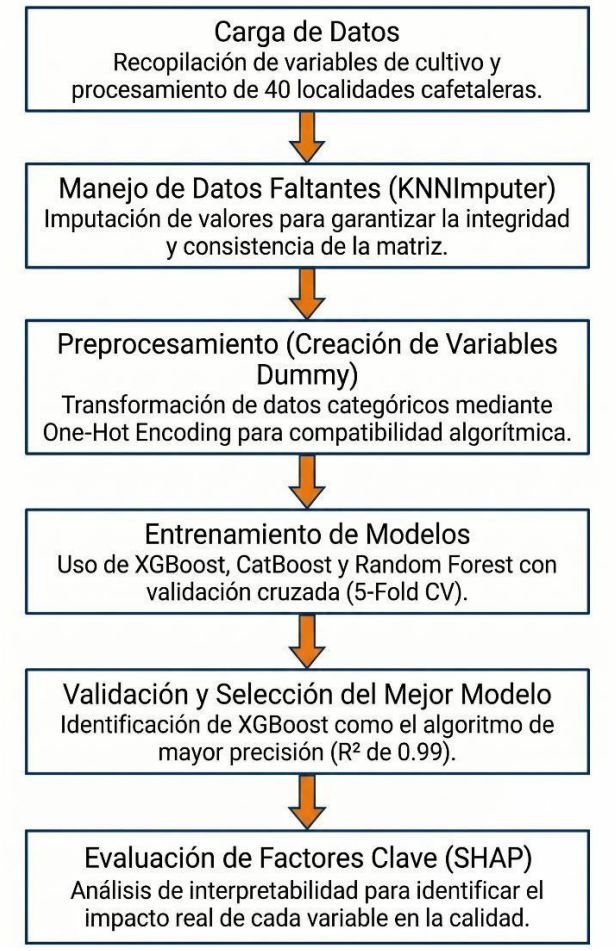
c. **Entrenamiento y validación de modelos:** Los algoritmos *XGBoost, CatBoost* y *Random Forest* fueron entrenados utilizando la técnica de validación cruzada k-fold, lo que permite evaluar el desempeño del modelo en diferentes subconjuntos de datos sin necesidad de una partición manual previa. Los indicadores principales de evaluación serán el coeficiente de determinación (R^2) y la raíz cuadrada del error medio (RMSE). La capacidad de generalización de los modelos fue evaluada bajo la misma metodología, considerando un R^2 superior al 85% y el menor RMSE posible como criterios para seleccionar el modelo más adecuado (DataScientest, 2020).

d. **Evaluación de factores clave:** Para determinar los factores de mayor impacto (variables productivas y de procesamiento) sobre la variable respuesta, se utilizó el diagrama de cascada (*waterfall)* basado en valores SHAP (*Shapley Additive Explanations*), empleando la librería SHAP.

Finalmente, todo el proceso se llevó a cabo utilizando el software Python 3.11.2, garantizando un nivel de confiabilidad del 95%. Para una mejor comprensión del procesamiento, se presenta en la Figura 4 un diagrama de flujo del preprocesamiento, modelado predictivo e interpretabilidad de los datos utilizado en este estudio. Este esquema resume las etapas principales desde la carga de los datos hasta la selección del mejor modelo.

Figura 4

Flujograma metodológico del procesamiento computacional, modelado predictivo e interpretabilidad de datos.



3.8. Materiales y equipos

3.8.1. Materiales

- **Muestras de café pergamino (300 g por unidad):** utilizadas como materia prima para el análisis físico y sensorial del estudio.
- **Fichas de catación SCA:** empleadas para registrar los atributos sensoriales del café (aroma, sabor, acidez, cuerpo, etc.).
- **Vasos de catación (5–6 oz):** utilizados para preparar las infusiones de café durante el proceso de catación.

- **Agua potable a temperatura controlada:** utilizada para la preparación de la infusión en condiciones estandarizadas (93 °C).
- **Cucharas de catación:** empleadas por los catadores para evaluar sensorialmente las muestras.
- **Etiquetas y bolsas herméticas:** utilizadas para la codificación, almacenamiento y conservación de las muestras evitando contaminación.
- **Cuadernos de registro:** utilizados para el registro manual de datos de campo y laboratorio.

3.8.2. Equipos

- **Balanza analítica:** utilizada para pesar con precisión las muestras de café (300 g en pergamino y proporciones de cata).
- **Tostadora de café:** empleada para realizar el tostado controlado de las muestras antes de la evaluación sensorial.
- **Molino de café:** utilizado para obtener la molienda adecuada previa a la preparación de la infusión.
- **Termómetro:** empleado para controlar la temperatura del agua durante la preparación de la bebida.
- **Cronómetro:** utilizado para medir tiempos de tostado, fermentación y preparación de la infusión.
- **Secadores (carpas solares o secador solar):** utilizados en el proceso de secado postcosecha de las muestras.
- **Computadora:** empleada para el procesamiento de datos, análisis estadístico y modelamiento con Machine Learning.
- **Software Python 3.11.2:** utilizado para desarrollar los modelos predictivos.
- **Librerías de Machine Learning (XGBoost, CatBoost, Random Forest, SHAP):**

utilizadas para el entrenamiento, validación e interpretación de los modelos.

3.9. Aspectos éticos

El presente estudio cumple con los principios éticos fundamentales relacionados con la investigación científica, asegurando el respeto a los derechos de los participantes y la integridad de los datos recolectados. Los productores han participado de manera voluntaria tras recibir información clara sobre los objetivos, alcance y métodos del estudio.

Además, la recolección de datos se llevó a cabo de manera confidencial y respetando la privacidad de los productores, evitando la divulgación de información personal o sensible. Los datos se manejaron exclusivamente con fines académicos y de investigación y con el respaldo de catador, cumpliendo con las normativas éticas y legales aplicables, como las establecidas en el Anexo A1.

CAPÍTULO IV. RESULTADOS Y DISCUSIÓN

4.1. Descripción de resultados e inferencias estadísticas

El presente estudio presenta un análisis descriptivo y comparativo de las variables utilizadas durante el desarrollo de la presente investigación, cuyo objetivo determinar la influencia de las variables productivas y de procesamiento en la calidad en taza del café en 40 localidades de la región Cajamarca mediante algoritmos de aprendizaje automático. Esta etapa inicial tiene por finalidad caracterizar la información contenida en la base de datos, identificar patrones de distribución, explorar relaciones potenciales y contextualizar la calidad obtenida en taza, como paso previo a la modelización predictiva.

El análisis se basa en una matriz de datos construida a partir de 40 muestras de café recolectadas en distintas localidades de la región Cajamarca, Perú, específicamente en nueve distritos de la provincia de Cutervo (ver Figura 5). El trabajo de campo se realizó en el marco del proyecto Avanzar Rural, en colaboración con productores cafetaleros locales. Para cada muestra se recolectaron 300 gramos de café pergamino, bajo condiciones controladas, incluyendo entrevistas a los productores para sistematizar la información sobre las prácticas de manejo agronómico y postcosecha.

Una vez recolectadas, las muestras fueron trasladadas al laboratorio Ushún Café, donde se verificaron criterios de calidad inicial como el porcentaje de humedad y se estandarizó su preparación mediante protocolos técnicos de tostado, reposo y molienda. Posteriormente, el equipo y los *Q Graders* certificados procedieron con la catación profesional (el registro visual de estas etapas de preparación y cata se detalla en el Anexo A7), evaluando los atributos organolépticos por medio de el protocolo oficial. Entre los atributos evaluados destacan: aroma, acidez, cuerpo, sabor, dulzura, balance, uniformidad, postgusto y puntaje global, que implican el componente sensorial de la base de datos.

Simultáneamente, se registraron detalladamente variables relacionadas con el entorno

productivo (como altitud, variedad genética, edad del cultivo, número de cosechas) y con el procesamiento postcosecha (método de fermentación, duración, tipo de secado y beneficio).

La matriz integral se construyó con los datos que vincula directamente las condiciones agronómicas y tecnológicas de producción resultados sensoriales obtenidos en laboratorio.

4.1.1. Generación de la matriz de datos.

Para dar cumplimiento al primer objetivo específico, se consolidó una matriz de datos con las 40 muestras evaluadas, integrando las 4 variables de producción, las 4 de procesamiento y los 10 atributos sensoriales. La representación completa de los resultados consolidados en esta matriz de datos se encuentra detallada en el Anexo A2. Asimismo, la base de datos depurada puede ser visualizada e inspeccionada en su formato digital a través del siguiente repositorio:

[https://docs.google.com/spreadsheets/d/18lnvhHQP5AQRHKDaQ3PWfKRI6ZdnNegd/edit?usp=drivesdk&ouid=114113850858660863688&rtpof=true&sd=true]

La matriz final fue objeto de un preprocesamiento para su uso en modelos de aprendizaje automático. Esto incluyó técnicas de imputación de valores faltantes, codificación de variables categóricas y normalización de escalas. Posteriormente, se aplicaron algoritmos de regresión supervisada basados en técnicas de gradient boosting (como XGBoost, CatBoost y Random Forest), con el propósito de evaluar la capacidad predictiva de las variables independientes respecto a la calidad sensorial del café, medida a través de un puntaje promedio de cata.

En función de su naturaleza y función dentro del modelo predictivo, las variables han sido organizadas en tres bloques principales:

- **Variables productivas:** incluyen los factores asociados al entorno agronómico y al manejo del cultivo, como altitud, variedad genética, edad del cafeto y número de

cosechas. Son relevantes para entender cómo influyen las condiciones edafoclimáticas y prácticas de manejo sobre el potencial sensorial del café.

- **Variables de procesamiento:** comprenden las prácticas postcosecha aplicadas a cada muestra, incluyendo el método de fermentación, el tiempo de fermentación, el método de secado y el tipo de beneficio. Estas variables tienen un impacto directo en la transformación fisicoquímica del grano y, por tanto, en su expresión sensorial.
- **Variables sensoriales:** abarcan los atributos evaluados (tales como Fragancia, aroma, sabor, sabor residual, acidez, cuerpo, uniformidad, balance, taza limpia, dulzor, puntaje del catado) por los catadores certificados, los cuales conforman la variable dependiente del modelo. Estos atributos son clave para describir la calidad en taza, y permiten asociar el manejo productivo con la percepción sensorial final.

Cada grupo de variables será analizado mediante herramientas gráficas (gráficos de barras, diagramas de caja, gráficos circulares o de radar) y cuadros de resumen, que permitirá visualizar su distribución. En este contexto a lo largo del capítulo se discutirán los procedimientos desde un enfoque técnico y contextual, lo cual facilitará una mejor interpretación del comportamiento con las variables y de su papel en la construcción de un modelo robusto.

Este enfoque integral no solo permite sentar las bases del análisis predictivo, sino que también proporciona información valiosa para la toma de decisiones en la cadena de valor del café. Actores dedicados a la caficultura podrán identificar prácticas que favorecen una mayor calidad sensorial, en la medida que se fortalece el conocimiento científico sobre la correspondencia entre manejo, procesamiento y calidad del grano.

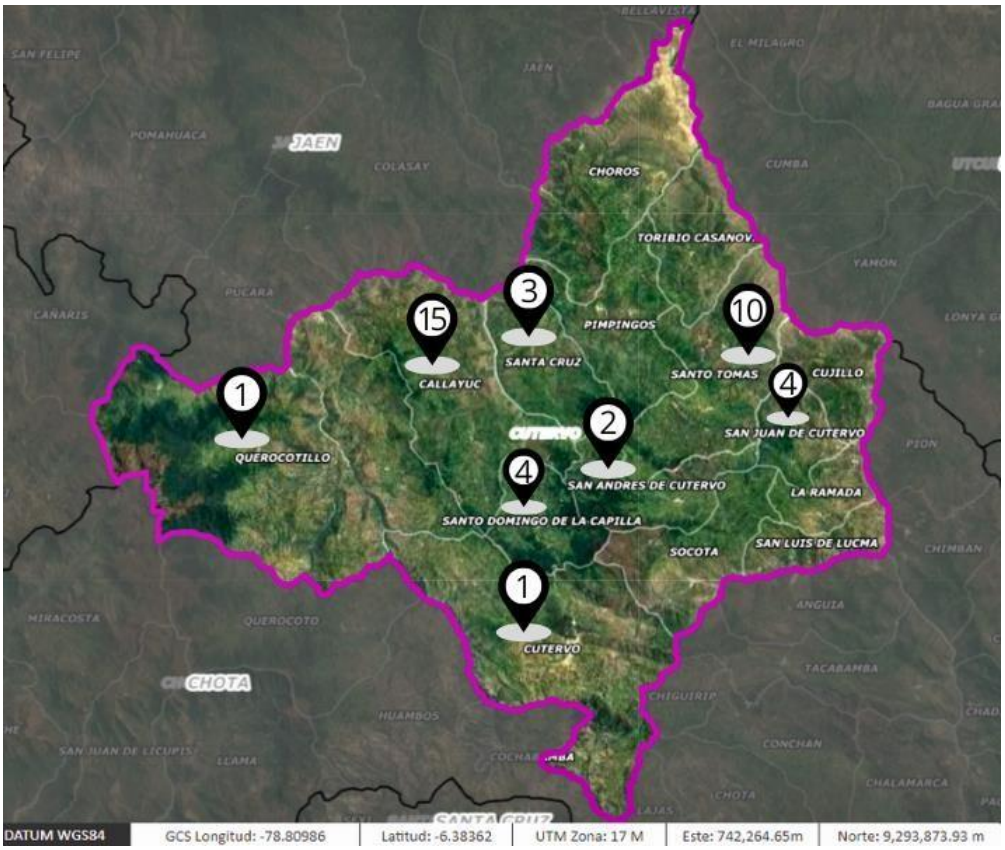
Este enfoque no solo fortalece el análisis cuantitativo del comportamiento de las variables, sino que también sienta las bases para una interpretación contextualizada y práctica de los factores que más inciden en la calidad final del café evaluado. Durante la revisión inicial

de los datos sensoriales se identificó que tres de los diez atributos evaluados presentaron valores constantes en todas las muestras (varianza cero). Este comportamiento impide su uso en análisis estadísticos inferenciales o modelos computacionales, dado que las variables sin variabilidad no aportan información para discriminar o predecir.

Por ello, dichos atributos fueron incluidos únicamente como información descriptiva, pero excluidos de los análisis correlacionales y los modelos de ML, siguiendo buenas prácticas en análisis multivariante.

Figura 5

Distritos de la provincia de Cutervo (Cajamarca) de los cuales se extrajeron las muestras del estudio



Nota. Datos de la Plataforma Nacional de Datos Georreferenciados Geo Perú (IDEP, 2026)

4.1.1.1. Análisis de las variables de cultivo de café

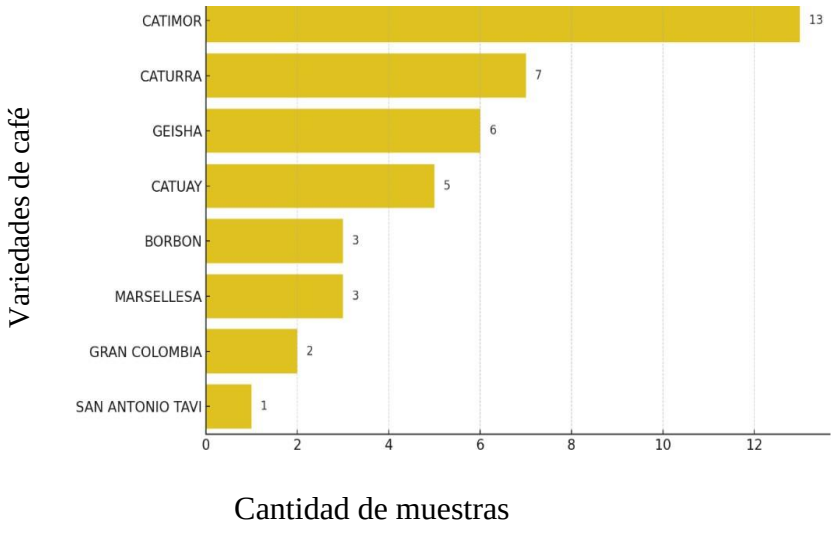
Las muestras proceden de nueve distritos de Cutervo. Los datos de cultivo se tomaron en campo, por entrevista al productor y verificación en parcela durante el acopio. Se describen cuatro variables: variedad, altitud, edad del cafeto y número de cosechas.

4.1.1.1.1. **Variedad genética de cafés**

Las 40 muestras se reparten en ocho variedades (Figura 6). Catimor domina con 13 casos, el 32.5%, seguida de Caturra con 7, Geisha con 6 y Catuay con 5. Las cuatro restantes, Borbón, Marsellesa, Gran Colombia y San Antonio Taví, suman siete muestras entre todas. Ordenadas por puntaje medio en taza, el cuadro cambia. Borbón encabeza con 87.18, por delante de Geisha con 86.29. Catimor queda última, con 82.66. Esa lectura pide una advertencia. Borbón tiene solo tres muestras y Gran Colombia dos, de modo que sus medias no sostienen ninguna conclusión. Únicamente Catimor, Caturra y Geisha reúnen casos suficientes para compararse entre sí. Leroy et al. (2023) sitúan la variedad entre los principales determinantes de la variabilidad del puntaje en arábica, junto con el procesamiento poscosecha.

Figura 6

Frecuencia de variedades de café presentes en las muestras del estudio (n=40)

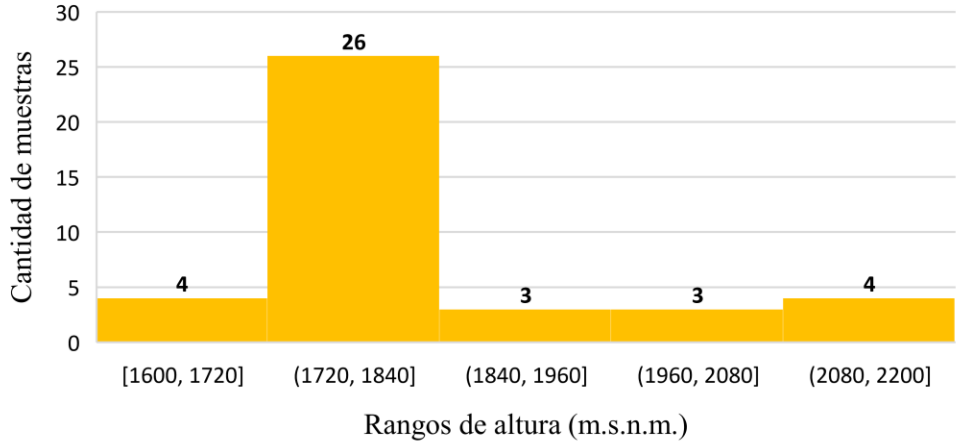


4.1.1.1.2. Altura del cultivo

El recorrido altitudinal de la muestra es estrecho: de 1600 a 2100 msnm (Figura 7). Solo siete valores distintos aparecen en los registros, y 1800 msnm concentra por sí solo veinte parcelas. Veinticuatro de las 40 muestras caen entre 1720 y 1840 msnm. Diez superan los 1900: M2, M3, M4, M12, M13, M25, M27, M32, M33 y M37. Toda la muestra está por encima del umbral que la literatura asocia a cafés de altura. Eso tiene una consecuencia para el análisis posterior. Comparar 1600 con 2100 msnm no es comparar altura con llano, y el rango disponible deja poco margen para que la altitud discrimine entre estas parcelas. Esta franja coincide con lo que la literatura especializada identifica como zona óptima para café de alta calidad, especialmente en variedades arábicas (ICO, 2021).

Figura 7

Distribución de la altitud de las parcelas cafetaleras (m.s.n.m.)

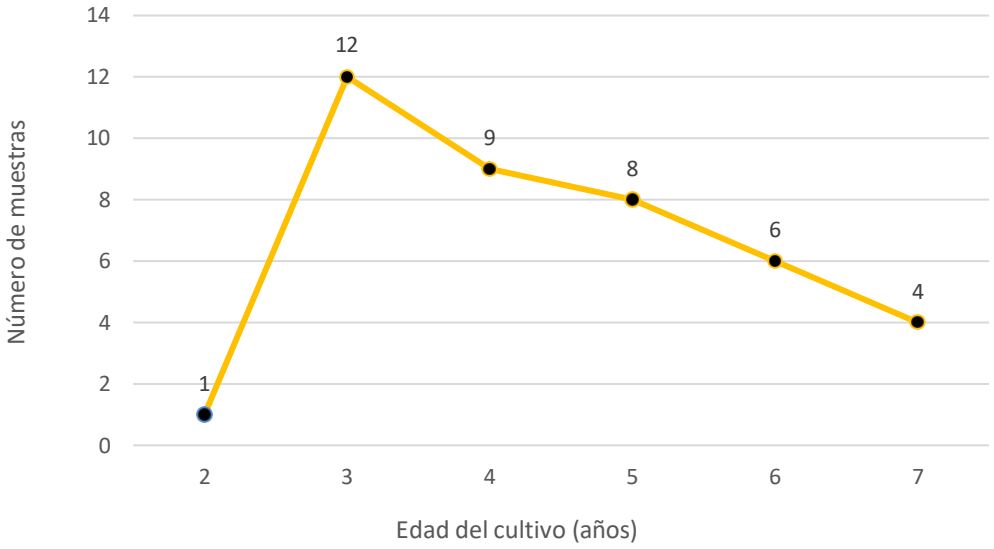


4.1.1.1.3. Edad del Cultivo

El grueso de las plantaciones es joven. Doce muestras proceden de cafetos de tres años, nueve de cuatro y ocho de cinco (Figura 8). Una sola tiene dos años, y las cuatro más antiguas llegan a siete. La media es de 4.45 años. La distribución se inclina hacia las plantaciones jóvenes. No preguntamos por qué. Una renovación reciente de cafetales explicaría este perfil, pero no lo verificamos en campo y no lo afirmamos aquí.

Figura 8

Distribución de la edad de los cultivos de café (años)

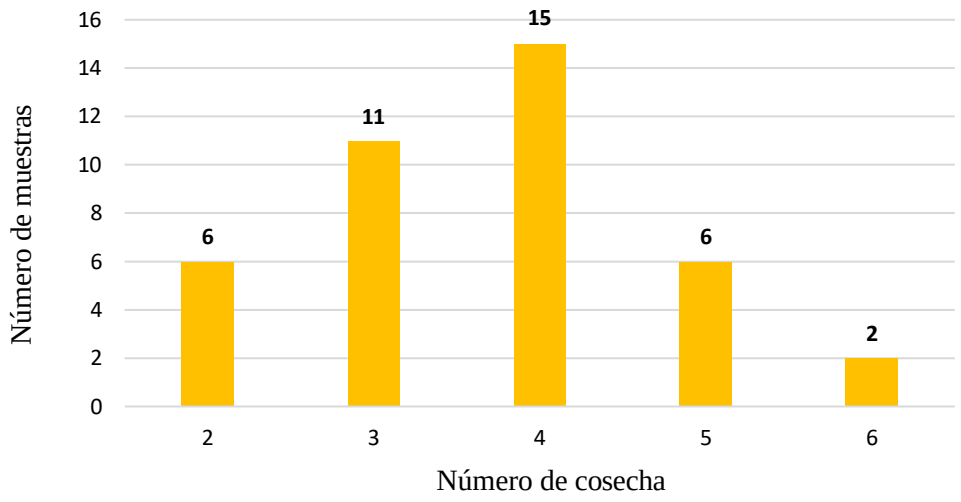


4.1.1.1.4. Número de Cosecha

El número de cosechas acompaña a la edad, como cabía esperar. Quince parcelas registran cuatro cosechas y once registran tres (Figura 9). Seis registran dos, otras seis cinco, y solo dos llegan a seis. La media es de 3.67 cosechas por tanto la calidad se estabiliza a partir de la tercera o cuarta cosecha, y la mayoría de estas parcelas está en ese tramo. Habrá que ver si eso se traduce en algo al cruzarlo con el puntaje. El número de cosechas es también un indicador de desgaste, y los dos efectos apuntan en direcciones opuestas.

Figura 9

Distribución del número de cosechas por muestra



4.1.1.2. Análisis de las variables de procesamiento del café

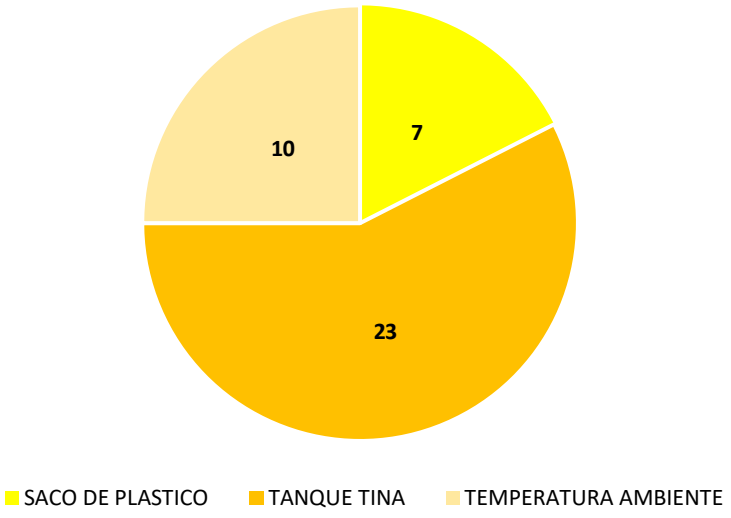
Tras la cosecha, cada productor toma cuatro decisiones que quedaron registradas en la ficha: cómo fermenta, cuánto tiempo, cómo seca y qué tipo de beneficio aplica. Los datos provienen de entrevista directa y verificación durante el acopio.

4.1.1.2.1. Método de fermentación

El tanque tina es el método dominante: 23 muestras, el 57.5% (Figura 10). Diez productores fermentan a temperatura ambiente y siete en saco de plástico. Por puntaje medio, el tanque tina llega a 84.35 y la temperatura ambiente a 84.19. El saco de plástico queda en 82.76, casi punto y medio por debajo. La diferencia no resiste el contraste. Un Kruskal-Wallis sobre los tres grupos da p = 0.61, de modo que con 40 muestras no es distinguible del azar. Conviene anotar una discrepancia con la ficha de operacionalización. El apartado 3.5 menciona el bidón hermético entre los métodos previstos, y ninguna muestra lo registró. El control de la fermentación, incluso en métodos semianóxicos (como el uso de sacos o bidones herméticos), es crucial: si se gestiona correctamente, puede generar perfiles de sabor intensos y diferenciados debido a la modulación de los subproductos microbianos (Shen et al., 2025).

Figura 10

Distribución y frecuencia de los métodos de fermentación postcosecha (n=40)

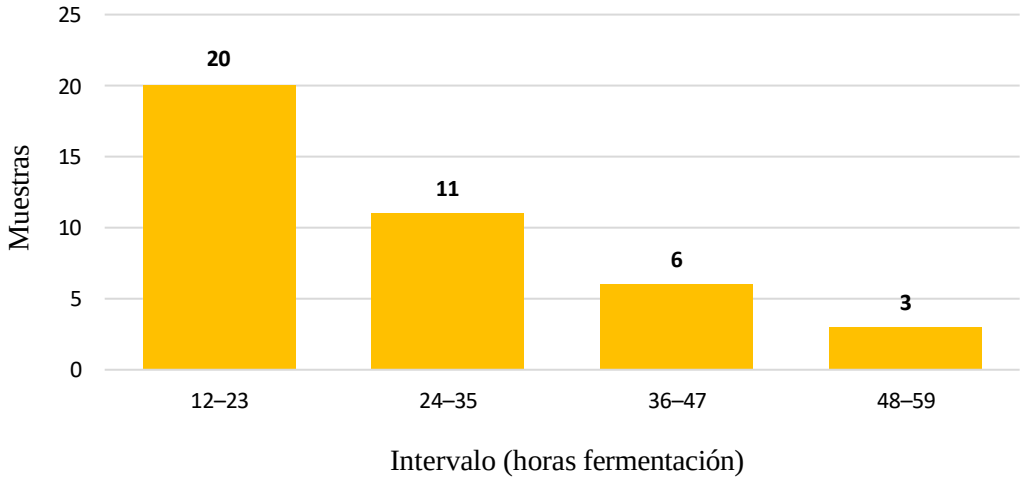


4.1.1.2.2. Tiempo de fermentación

Los tiempos van de 12 a 48 horas (Figura 11). Doce valores distintos, sin ningún agrupamiento evidente. La mitad de las muestras, veinte, fermenta entre 12 y 23 horas. Once lo hacen entre 24 y 35, seis entre 36 y 47, y tres llegan a 48. Las nueve muestras de 36 horas o más promedian 84.17 puntos. Las siete de 15 horas o menos, 82.76. La diferencia es de 1.41 puntos sobre una desviación general de 2.69, y la correlación de Spearman entre tiempo y puntaje final es de 0.137, con p = 0.40. La SCA (2019) recomienda fermentaciones de duración media para cafés lavados, que es donde se sitúa la mitad de la muestra

Figura 11

Distribución del tiempo de fermentación (en horas) por Intervalos de Muestra

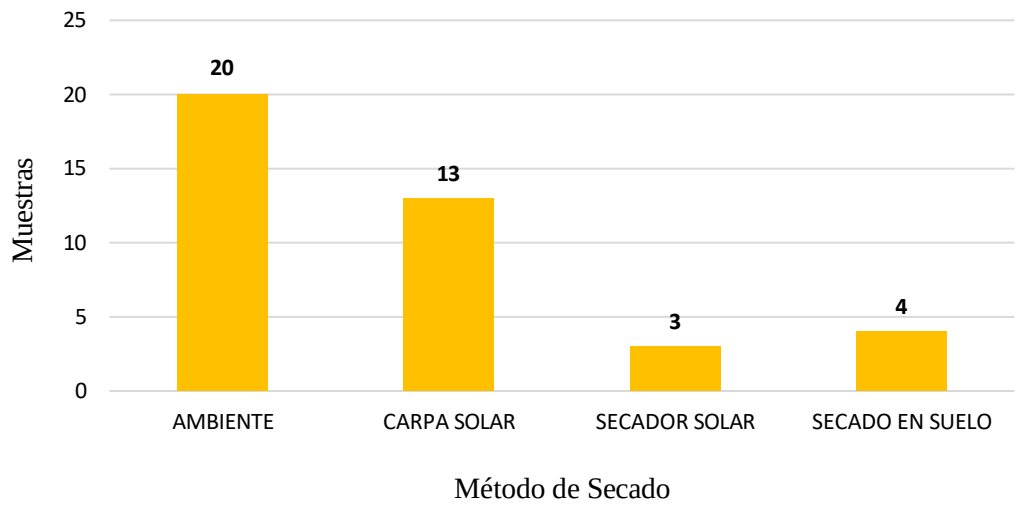


4.1.1.2.3. Método de Secado

Cuatro muestras se secaron directamente en suelo, y son las que peor puntúan: 81.76 de media, frente a 84.03 del conjunto (Figura 12). También son las que más varían entre sí, con una desviación de 5.12. El resto se reparte entre secado al ambiente en zarandas, veinte casos, carpa solar, trece, y secador solar, tres. Sus medias quedan entre 84.22 y 84.77, prácticamente indistinguibles. El contraste global entre los cuatro métodos da p = 0.87. La diferencia del secado en suelo se ve en la media, pero con cuatro casos no se puede separar del ruido.

Figura 12

Distribución y frecuencia de los métodos de secado postcosecha

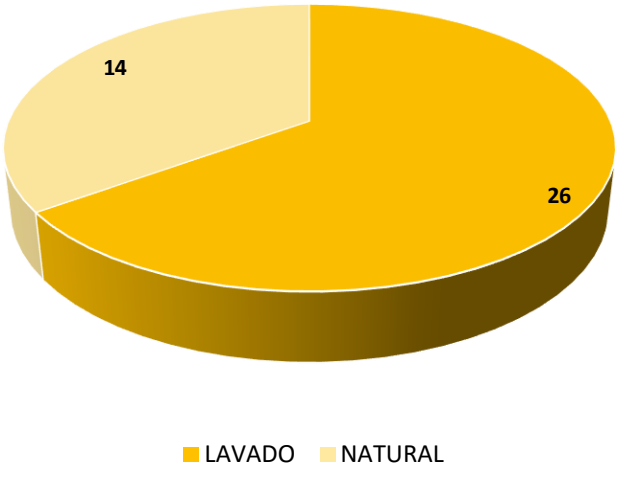


4.1.1.2.4. Tipo de beneficio

El Solo aparecen dos tipos de beneficio en las 40 muestras: lavado y natural (Figura 13). No se registró ninguna muestra de honey, pese a que el apartado 3.5 lo contempla entre las técnicas evaluadas. El lavado predomina con 26 muestras, el 65%. El natural cubre los 14 restantes. Las medias son 83.77 para el lavado y 84.52 para el natural, con p = 0.51. No hay diferencia atribuible al tipo de beneficio en esta muestra. La ausencia de honey no la investigamos. Queda registrada como un hecho del muestreo, no como un dato sobre la zona.

Figura 13

Distribución del tipo de beneficio aplicado (lavado vs. natural) en las muestras



El beneficio lavado predomina ampliamente (65% de las muestras), lo que denota una preferencia por perfiles limpios y brillantes, típicos del café peruano de altura. Esta técnica, aunque más demandante en infraestructura y consumo de agua, es muy valorada por el mercado de cafés especiales. En cambio, el beneficio natural, presente en el 35% restante, puede producir cafés con mayor cuerpo, dulzura y notas frutales, pero también con mayor riesgo de defectos si no se controla adecuadamente el secado. La ausencia del beneficio honey podría deberse a una escasa difusión de esta técnica en la zona o a la falta de equipos e insumos requeridos. La coexistencia entre beneficio natural y lavado sugiere una rica diversidad de perfiles de taza, lo que resulta especialmente valioso para estudios posteriores de correlación con variables sensoriales.

4.1.1.3. Análisis de las variables sensoriales de cata de café

Este bloque corresponde al análisis sensorial realizado bajo estándares de la SCA certificado mediante el Laboratorio Ushún Cafés de Especialidad. Las variables evaluadas permiten establecer el perfil sensorial de cada muestra y constituyen la base para el entrenamiento y validación de los modelos predictivos de calidad. Los datos fueron obtenidos tras el tueste controlado, molienda y evaluación individual de cada taza bajo condiciones normalizadas. Cada atributo fue puntuado individualmente en una escala de 6 a 9 (con incrementos de cuartos de punto), y la suma de estos puntajes, junto con la evaluación de defectos, constituye la base de la Puntuación Final en Taza. Se expone en la Tabla 4 de estadísticos descriptivos previa para conocer la dinámica del fenómeno sensorial.

Tabla 4

Estadísticos descriptivos de los atributos sensoriales de cata (protocolo SCA, n=40)

Atributo	Media	Mediana	Desv. Est.	Mín	Máx	Moda	Curtosis	Asimetría	Rango	Error típico
Aroma	7.82	8.00	0.40	7.00	8.50	8.00	-0.07	-0.57	1.50	0.06
Sabor	7.72	7.75	0.38	7.00	8.50	7.50	-0.74	-0.06	1.50	0.06
Sabor residual	7.75	7.75	0.48	6.75	8.55	7.50	-0.87	-0.24	1.80	0.08

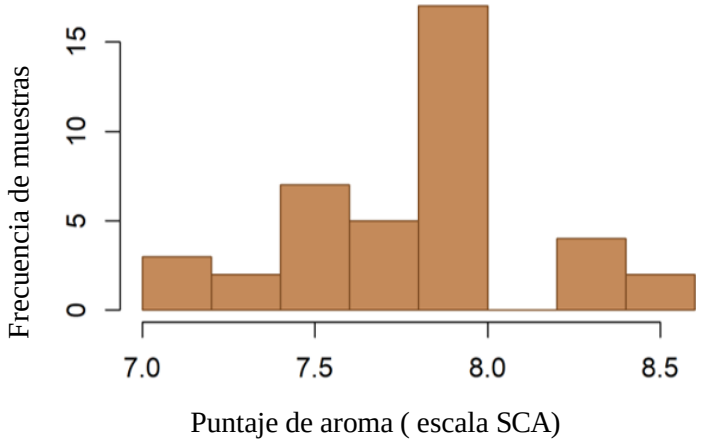
Acidez	7.87	8.00	0.38	7.00	8.50	8.00	-0.17	-0.17	1.50	0.06
Cuerpo	7.75	7.75	0.57	7.00	10.00	7.50	5.06	1.56	3.00	0.09
Balance	7.78	7.80	0.37	7.00	8.50	8.00	-0.50	-0.30	1.50	0.06
Taza limpia	9.80	10.00	0.70	6.00	10.00	10.00	15.20	-3.84	4.00	0.11
Dulzor	9.95	10.00	0.22	10.00	10.00	10.00	no aplica	no aplica	0.00	0.04
Uniformidad	9.82	10.00	0.63	6.00	10.00	10.00	14.90	-3.80	4.00	0.10
Puntaje catador	7.88	8.00	0.45	7.00	8.50	8.00	-0.46	-0.35	1.50	0.07

4.1.1.3.1. **Aroma**

Las 40 muestras promediaron 7,82 en aroma, con desviación de 0,38 y un recorrido de 7,00 a 8,50 (Figura 14). La mediana y la moda coinciden en 8,00. Dos muestras alcanzaron el máximo: M29 y M34. La primera es un Catimor de Santo Domingo fermentado 24 horas a temperatura ambiente; la segunda, un Gran Colombia de Callayuc con 17 horas. Ni la variedad ni el método coinciden. Tampoco la altitud: 1800 y 1750 msnm. En el extremo bajo, M1, M4 y M17 marcaron 7,00. Las tres comparten fermentaciones cortas, de 17 a 18 horas, pero difieren en variedad, distrito y tipo de beneficio. Con 0,38 de desviación sobre un recorrido efectivo de 1,50 puntos, el aroma discrimina poco entre estas muestras. Conviene tenerlo presente antes del modelado.

Figura 14

Distribución de los puntajes SCA para el aroma

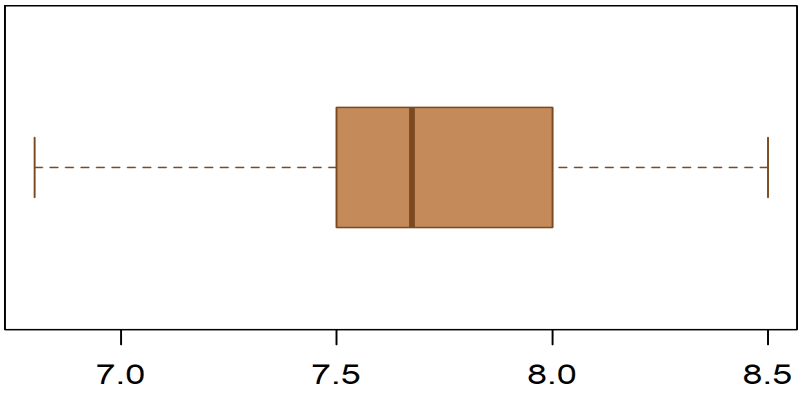


4.1.1.3.2. Sabor

El valor más bajo de toda la matriz sensorial está en este atributo. M4 obtuvo 6.80 en sabor (Figura 15). Es un Catimor de Santo Tomás, beneficio natural, 18 horas de fermentación y secado al ambiente. El resto de la serie se comporta con normalidad. La media fue 7.72 y la mediana 7.68, con una desviación de 0.49. Tres muestras llegaron a 8.50: M7, M16 y M19. Un Borbón y dos Geisha. La dispersión del sabor casi duplica la del aroma, lo que en principio lo convierte en un descriptor más informativo para el modelo. Esto es consistente con reportes recientes de la industria (Ribeiro et al., 2022), que vinculan valores altos de sabor con mejoras técnicas en fermentación y control ambiental en secado.

Figura 15

Distribución de los puntajes SCA para el sabor



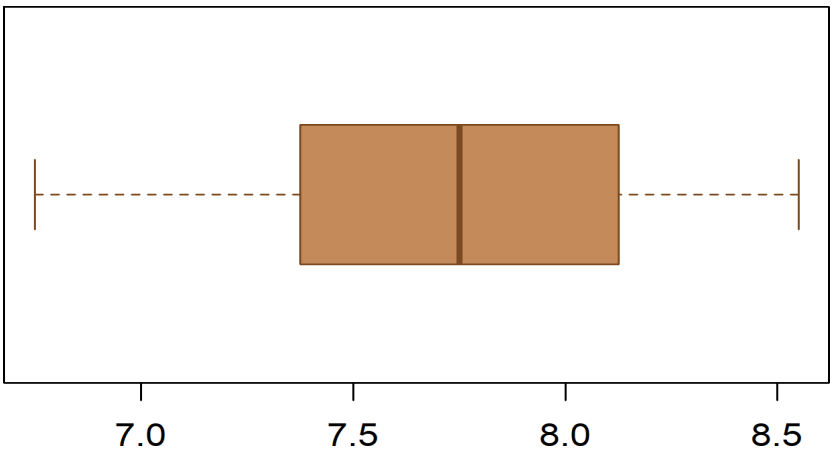
4.1.1.3.3. Sabor residual

M28 y M1 marcan los extremos del retrogusto: 8.55 frente a 6.75 (Figura 16). Casi dos puntos de diferencia. Las dos son lavadas. M28 es una Geisha de San Juan de Cutervo fermentada 24 horas; M1, un Catimor de Santa Cruz con 18. La media del atributo fue 7.72, con desviación de 0.48. M24 y M8 siguen a M28 con 8.40 y 8.35, ambas fermentadas 32 horas.

Tres de las cuatro muestras con mejor retrogusto superan las 24 horas de fermentación. Es una coincidencia que conviene registrar sin sobre interpretar, porque son cuatro casos. Investigaciones recientes (Zhang et al., 2022) demuestran que retrogustos más largos suelen asociarse a fermentaciones controladas que potencian ésteres y compuestos aromáticos de cadena larga. En contraste, los valores cercanos a 7.0 sugieren perfiles más cortos o ligeramente astringentes, posiblemente vinculados a secados heterogéneos o cosechas menos maduras.

Figura 16

Distribución de los puntajes SCA para el sabor residual

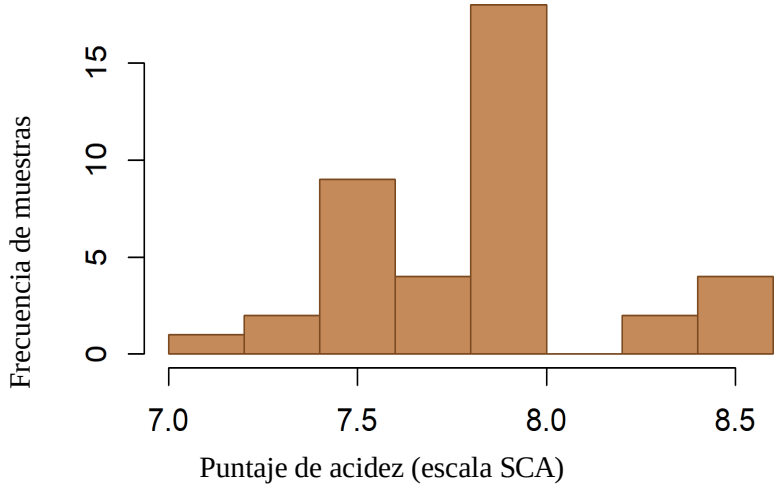


4.1.1.3.4. Acidez

La acidez fue el atributo mejor puntuado del panel: media de 7.87 y mediana de 8.00 (Figura 17). También el más comprimido, con desviación de 0.36. Tres muestras alcanzaron 8.50: M12, M18 y M22, las tres de beneficio natural. En el otro extremo, M14 marcó 7.00, y le siguen M13 y M26 con 7.25. La compresión del atributo importa para lo que viene después: un descriptor con este recorrido deja poco margen para que un modelo encuentre estructura. La contribución específica de los ácidos orgánicos clave (cítrico, málico y acético) a la percepción de la acidez en taza ha sido científicamente detallada, confirmando que la modulación de estos compuestos, influenciada por factores pre y postcosecha, es crítica para la calidad final (Sardella et al., 2024).

Figura 17

Distribución de los puntajes SCA para la acidez

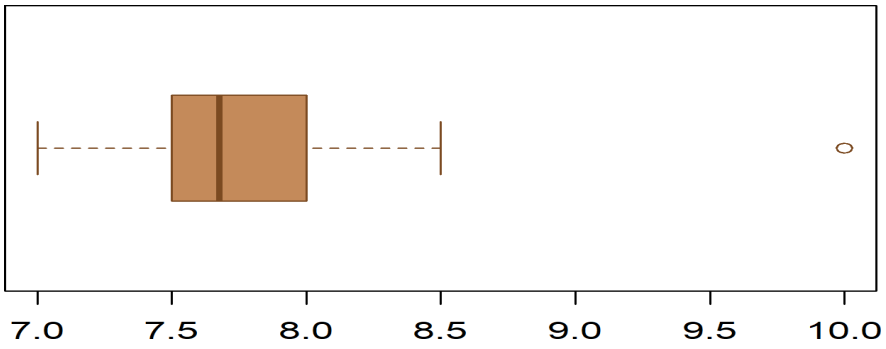


4.1.1.3.5. Cuerpo

Una muestra rompe la escala. M7 obtuvo 10.00 en cuerpo, muy por encima del segundo valor de la serie, que es 8.50 (Figura 18). Es un Borbón de San Juan de Cutervo, lavado y secado en secador solar. El protocolo descrito en el apartado 3.6.7 puntúa cada atributo entre 6 y 9, de modo que ese 10.00 queda fuera del rango declarado y debe verificarse contra la ficha original. Excluida esa muestra, el cuerpo se mueve entre 7.00 y 8.50, con media de 7.75. La desviación de 0.57 es la más alta de los siete atributos modelados, y ese valor único la infla. Estos compuestos solubles son cruciales para la densidad y la viscosidad de la bebida, intensificando la sensación de cuerpo en taza (Cheng et al., 2024). Por otra parte, valores inferiores a 7.0 podrían reflejar cafés lavados con menor contenido lipídico o secados rápidos que afectan la estructura sensorial

Figura 18

Distribución de los puntajes SCA para el cuerpo



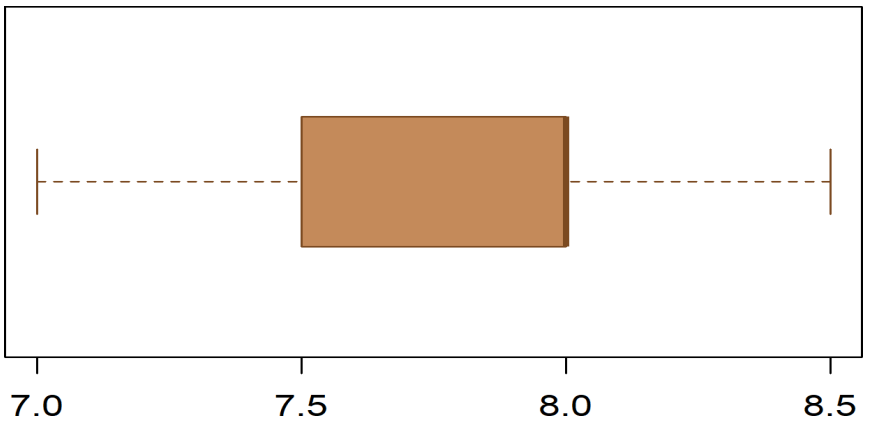
Puntaje de cuerpo (escala SCA)

4.1.1.3.6. Balance

El balance es el segundo atributo más comprimido de la serie: desviación de 0.40 sobre un recorrido de 1.50 puntos (Figura 19). La media fue 7.82 y la mediana 8.00, de modo que la mitad de las muestras alcanzó o superó los 8.00. Tres llegaron a 8.50: M19, M33 y M38. Sus tiempos de fermentación son 15, 36 y 18 horas. No hay patrón. M17, M25 y M26 cerraron la serie con 7.00. Por definición del protocolo SCA, el balance penaliza el desajuste entre atributos aunque cada uno sea alto por separado. Eso lo vuelve difícil de predecir a partir de variables de manejo. Las fermentaciones inconsistentes son la principal causa de desequilibrio, introduciendo notas que penalizan la integración global del perfil. (Alvarenga et al., 2023).

Figura 19

Distribución de los puntajes SCA para el balance



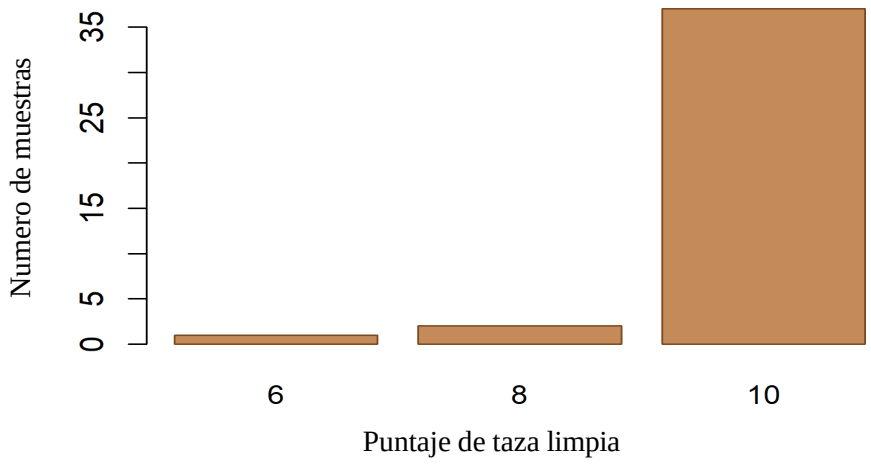
Puntaje de balance (escala SCA)

4.1.1.3.7. Taza limpia

Treinta y siete de las 40 muestras obtuvieron 10.00 en taza limpia (Figura 20). Solo tres quedaron por debajo: M5 y M30 con 8, y M14 con 6. M14 es también la muestra con el puntaje final más bajo del estudio; se fermentó 12 horas en saco de plástico y se secó en suelo. La media resultante es 9.80, pero el promedio dice poco en un atributo así. Lo relevante es que el 92.5% del panel no registró defectos. Un manejo inadecuado del secado, en particular, puede introducir notas mohosas o fenólicas, que son penalizadas severamente en la catación (Abreu et al., 2025). Algunos valores aislados de 6 y 8 reflejan muestras con defectos leves o inestabilidad en el perfil, lo cual sugiere necesidad de mejorar prácticas postcosecha específicas. y es esencial para clasificar un café como especialidad. Se representó mediante un gráfico de barras para visualizar la frecuencia de cada puntaje (Figura 20).

Figura 20

Distribución de los puntajes SCA para la taza limpia



4.1.1.3.8. Dulzor

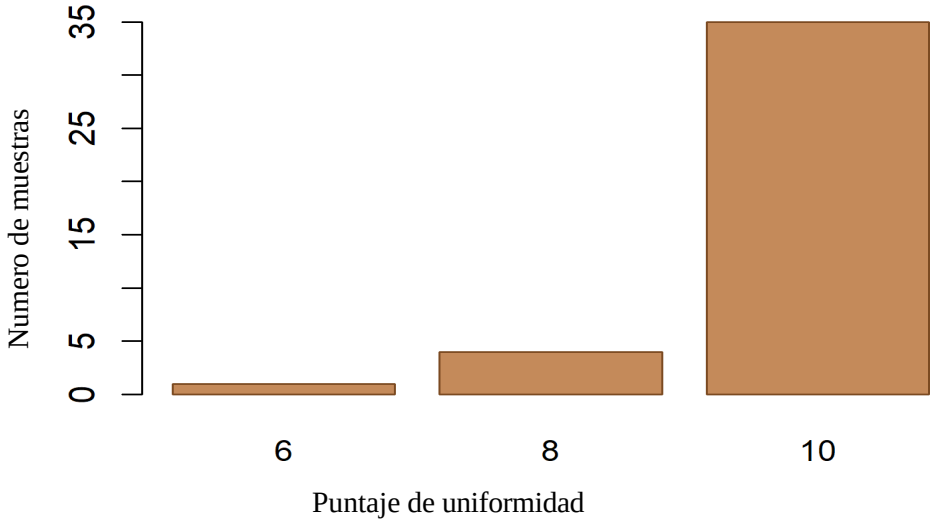
Las 40 muestras obtuvieron 10.00 en dulzor. Sin excepción. La ausencia de variación impide cualquier tratamiento estadístico, porque un atributo constante no correlaciona ni predice. Por eso el dulzor se describe aquí pero no entra en el modelado. El resultado es coherente con una cosecha en madurez uniforme, que es cuando la sacarosa del grano alcanza su concentración máxima. Este es un indicador clave de que las fermentaciones han sido limpias y controladas, evitando que los azúcares se degraden completamente en alcoholes o ácidos volátiles indeseables (Gómez-Meza et al., 2024).

4.1.1.3.9. Uniformidad

Cinco muestras no alcanzaron el máximo en uniformidad: M3, M28, M30 y M38 con 8, y M14 con 6 (Figura 21). Las 35 restantes obtuvieron 10.00. La media es 9.70 y la desviación 0.85, la más alta de la matriz sensorial, y ese valor lo genera casi por completo M14. Las cinco excepciones son de beneficio lavado. Con 26 muestras lavadas sobre 40 la coincidencia no resulta llamativa, pero queda registrada. Con 35 de 40 casos en el mismo valor, el atributo queda al borde de la degeneración estadística y no se incluyó en el modelado. Estudios recientes han demostrado que la uniformidad es extremadamente sensible a la distribución de partículas durante la molienda y a las variaciones en el tiempo de contacto con el agua (reposo), ya que una extracción heterogénea conduce inevitablemente a una percepción sensorial inconsistente entre tazas (De la Rosa & Torres, 2024).

Figura 21

Distribución de los puntajes SCA para la uniformidad

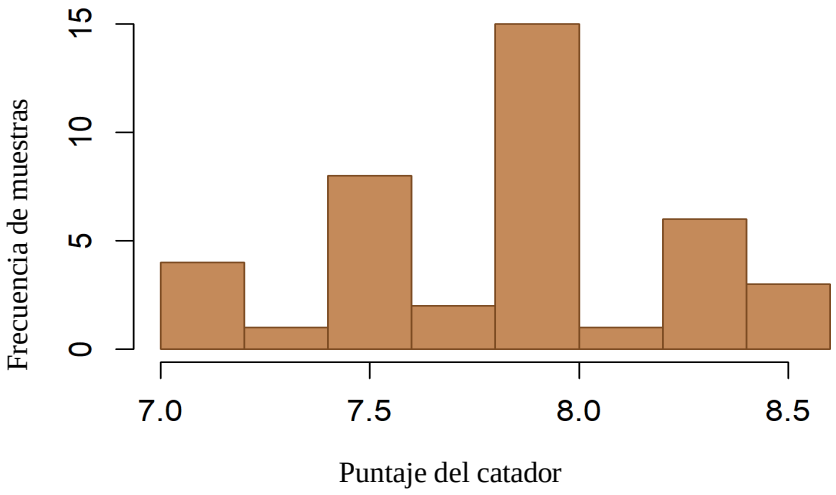


4.1.1.3.10. Puntaje del catador

El catador certificado otorgó puntuaciones entre 7.00 y 8.50, con media de 7.84 y mediana de 8.00 (Figura 22). Tres muestras recibieron la máxima: M22, M26 y M39. Tres recibieron la mínima: M2, M21 y M28. M28 merece una nota. Obtuvo el retrogusto más alto de toda la serie, 8.55, y a la vez el puntaje de catador más bajo. La puntuación global del panel no es la suma de los atributos individuales, y este caso lo muestra con claridad. . El control de estos parámetros reduce los defectos y maximiza la expresión del potencial sensorial del grano (Rodríguez et al., 2024).

Figura 22

Distribución de los puntajes SCA para el puntaje catador



4.1.1.3.11. Puntaje final de taza

El puntaje final recorre 14.50 puntos, de 74.50 a 89.00 (Figura 23). La media es 84.03 y la mediana 83.75, con desviación de 2.69. M14 es la única muestra por debajo del umbral de especialidad. Con 74.50 puntos queda 6.00 por debajo de la siguiente, M17, y es el único valor atípico según el criterio de Tukey. Se fermentó 12 horas en saco de plástico, se secó en suelo, y perdió puntos en taza limpia y uniformidad a la vez. En el extremo alto, M7 alcanzó 89.00 y M33 88.80. Ambas son Borbón. Excluida M14, la serie se distribuye de forma simétrica entre 80.50 y 89.00. Los estadísticos de la Tabla 5 se calcularon sobre esas 39 muestras, y la exclusión se declara aquí. Estudios recientes como el de Martínez et al. (2023) vinculan estas puntuaciones con cafés producidos en condiciones agroecológicas favorables y manejos postcosecha controlados.

Tabla 5

Estadísticos descriptivos agregados del puntaje final en taza (SCA)

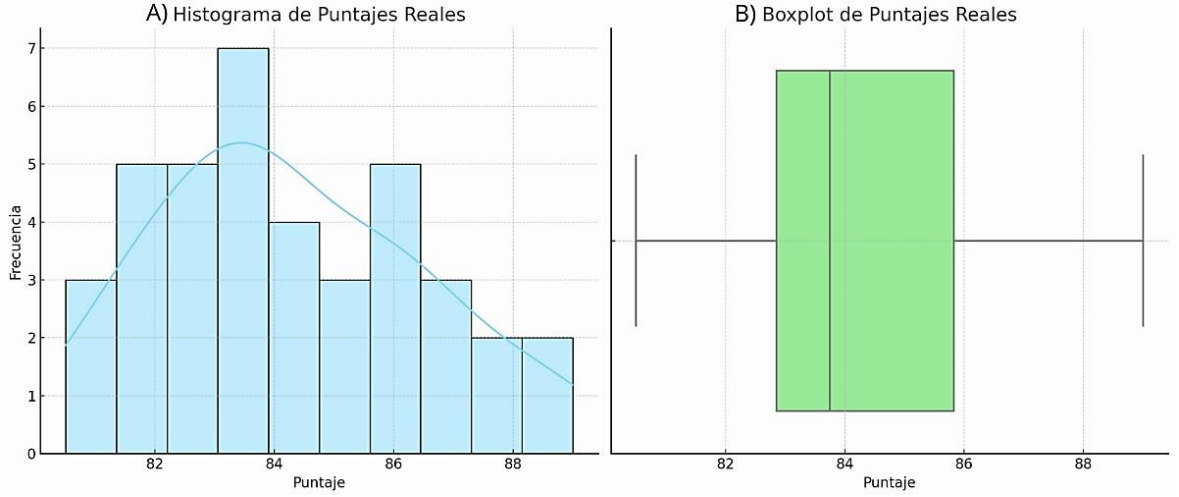
Estadístico	Valor	Interpretación
Media	84.28	El promedio global indica cafés de calidad muy buena según SCA (>80 puntos).
Error típico	0.36	El error estándar bajo sugiere que la media es una buena estimación representativa del grupo.
Mediana	83.75	La mediana confirma que la mitad de los estudiantes obtuvieron puntajes iguales o mayores a 83.75.
Moda	83.00	La moda indica que 83 fue el puntaje más frecuentemente alcanzado entre los evaluados.
Desviación estándar	2.23	La desviación muestra que la mayoría de los puntajes varían moderadamente respecto a la media.
Varianza de la muestra	4.99	La varianza confirma una dispersión moderada entre los puntajes del grupo.
Curtosis	-0.65	La curtosis negativa sugiere una distribución con menos datos en los extremos (más plana).
Coeficiente de asimetría	0.36	Asimetría cercana a cero indica una distribución simétrica, sin sesgos marcados a izquierda o derecha.
Rango	8.50	El rango de 8.5 puntos muestra la diferencia total entre el menor y mayor puntaje.
Mínimo	80.50	El mínimo indica el puntaje más bajo obtenido, aún dentro de la categoría de calidad buena.
Máximo	89.00	El máximo representa el mejor desempeño alcanzado en la muestra.
Suma	3286.75	La suma total de los puntajes permite cálculos agregados y verificación de totales.
Cuenta	40.00	La cantidad total de casos representa el tamaño de la muestra analizada.

Como se observa en la Figura 23(A), el histograma confirma que los puntajes en su mayoría se concentra entre 82 y 86 puntos, esto indica una buena homogeneidad con respecto a la calidad en las tazas. La curva de densidad muestra una distribución ligeramente simétrica, con leve inclinación hacia la izquierda. Esta observación es confirmada por el boxplot en la Figura

23(B), que evidencia una mediana de 83.75 sin presencia de valores atípicos, reflejando uniformidad en la muestra.

Figura 23

Distribución general del puntaje final de taza de las 40 muestras de cafés evaluados



Además, en el Anexo A3 se presenta una figura tipo Windrose este esquema resume el comportamiento promedio de los atributos sensoriales analizados. Esta representación circular permite identificar de manera intuitiva el equilibrio global del perfil de taza, mostrando que variables como dulzor (9.90), uniformidad (9.79) y taza limpia (10.00) alcanzaron valores notablemente altos, reflejando consistencia y ausencia de defectos en las muestras. Por su parte, atributos más diferenciadores como aroma (7.82), sabor (7.72), sabor residual (7.72), acidez (7.87), cuerpo (7.75) y balance (7.82) se ubican en un rango coherente con cafés de calidad elevada, acorde con los estándares SCA. Este proceso evidencia que, aunque existen variaciones prevalecentes entre atributos, el café evaluado mantiene una estructura sensorial uniforme y sólida, reforzando los resultados estadísticos anteriormente descritos para el puntaje final de taza. La Windrose se presenta así como un recurso complementario que facilita la interpretación holística del perfil sensorial y potencia la comprensión de las relaciones entre atributos en un formato visual sintético y de alta claridad.

4.1.2. Preprocesamiento y creación de variables Dummy

Para dar cumplimiento al segundo objetivo específico, la matriz original fue sometida a una etapa de acondicionamiento computacional. Dado que la base de datos incluye variables predictoras de naturaleza categórica nominal (tales como "Variedad", "Método de fermentación", "Método de secado" y "Tipo de beneficio"), los algoritmos de *Machine Learning* no pueden procesarlas directamente en formato de texto.

Para solucionar esto sin imponer una jerarquía numérica falsa, se implementó la algoritmia de *One-Hot Encoding* para la creación de variables Dummy. Se utilizó la función *pd.get_dummies* de la librería Pandas en el lenguaje Python, configurando el parámetro *drop_first=True* para evitar la multicolinealidad perfecta (la "trampa de las variables dummy"). Esta operación transformó cada categoría en nuevas columnas binarias lógicas, expandiendo la matriz de datos y dejándola matemáticamente estructurada para el entrenamiento de los modelos *XGBoost*, *CatBoost* y *Random Forest*. La evidencia completa de la ejecución computacional, incluyendo el script de esta algoritmia y la vista de la matriz transformada con valores lógicos booleanos, se encuentra detallada de forma íntegra en el Anexo A8.

4.1.3. Diseño y entrenamiento de modelos de Machine Learning

Con el objetivo de evaluar la influencia de los factores agronómicos y de procesamiento sobre la calidad sensorial, y para dar cumplimiento al tercer objetivo específico, se desarrolló un enfoque de modelamiento supervisado orientado a predecir la puntuación de taza y sus principales descriptores: Aroma, Sabor, Sabor Residual, Acidez, Cuerpo, Balance y Puntaje del Catador.

Es de suma importancia indicar que los siguientes atributos como Taza Limpia, Dulzor y Uniformidad fueron excluidos del modelamiento debido a su varianza visualmente nula (concentración en el puntaje máximo), una condición que habría impedido a los algoritmos capturar relaciones significativas y generado un desempeño predictivo artificialmente sesgado.

Consolidamos la matriz numéricamente depurada en la etapa anterior, se diseñaron y entrenaron tres modelos de ensemble learning: Random Forest, XGBoost y CatBoost, quienes fueron seleccionados por su robustez ante variables mixtas, su capacidad para captar interacciones no lineales y su desempeño probado en problemas de regresión agroalimentaria. Atendiendo a las exigencias de rigor predictivo, el ajuste de hiperparámetros y el entrenamiento de los algoritmos se realizó garantizando una repetitividad estadística superior al mínimo requerido. Para ello, en lugar de una única iteración, se aplicó una técnica de validación cruzada de 5 particiones (*5-fold cross-validation*). Esto significa que cada algoritmo fue entrenado y evaluado iterativamente cinco veces utilizando diferentes subconjuntos de la matriz de datos, asegurando la robustez de su precisión y descartando un posible sobreajuste (*overfitting*). Toda la algoritmia de inicialización y entrenamiento de estos modelos se encuentra detallada en el Anexo A8.

4.1.4. Validación de modelos y selección del algoritmo óptimo

Para determinar la eficacia de los modelos de ensemble learning entrenados en la etapa anterior, la comparación del rendimiento predictivo se realizó mediante dos métricas principales: el Coeficiente de Determinación (R^2), que indica la proporción de varianza explicada, y la Raíz Cuadrada del Error Medio (RMSE), que mide la magnitud media del error de predicción.

La Tabla 6 resume el desempeño agregado por atributo sensorial, destacando a XGBoost como el modelo patentemente superior, logrando un R^2 de 0.99 y valores de RMSE excepcionalmente bajos (cercanos a 0.01) en todos los descriptores evaluados. Esta consistencia técnica respalda su elección definitiva para la fase de análisis interpretativo.

Tabla 6

Desempeño predictivo agregado de los modelos de ensemble learning (R^2 y RMSE)

Atributo sensorial	Modelo	R^2	RMSE
Puntaje en taza	Random Forest	0.80	0.98
	XGBoost	0.99	0.01
	CatBoost	0.33	1.81
Aroma	Random Forest	0.77	0.18
	XGBoost	0.99	0.01
	CatBoost	0.18	0.34
Sabor	Random Forest	0.87	0.17
	XGBoost	0.99	0.01
	CatBoost	0.64	0.29
Sabor residual	Random Forest	0.83	0.20
	XGBoost	0.99	0.01
	CatBoost	0.52	0.33
Acidez	Random Forest	0.84	0.14
	XGBoost	0.99	0.01
	CatBoost	0.60	0.23
Cuerpo	Random Forest	0.83	0.21
	XGBoost	0.99	0.01
	CatBoost	0.51	0.32
Balance	Random Forest	0.81	0.17
	XGBoost	0.99	0.01
	CatBoost	0.31	0.33
Puntaje del catador	Random Forest	0.84	0.17
	XGBoost	0.99	0.01
	CatBoost	0.55	0.28

Nota. Las métricas R^2 y RMSE se calcularon mediante validación cruzada con 5 particiones (*5-folds CV*), garantizando la estabilidad del desempeño y descartando sobreajustes.

El modelo *XGBoost* presentó el rendimiento más consistente y alto en todos los atributos modelados, lo cual lo convierte en el algoritmo seleccionado como línea base para el análisis explicativo mediante SHAP.

El uso de XGBoost está avalado en aplicaciones agroalimentarias según versiones literarias: se ha demostrado que su estructura de ensamblado secuencial y sus mecanismos de regularización mejoran la generalización frente a Random Forest, especialmente en problemas con interacciones complejas entre variables (Patel & Sharma, 2023). Asimismo, casos de estudio en control de calidad de frutas y bebidas respaldan la superioridad de modelos Boosting en términos de precisión y estabilidad (Martínez et al., 2023).

Cada atributo presenta una contribución a las predicciones de XGBoost en donde para poder explicarlo se implementó SHAP, esta técnica atribuye el efecto marginal de cada variable basándose en los principios de teoría de jugos. Esta técnica permite descomponer la predicción en aportes individuales, favoreciendo interpretaciones transparentes y comparables con prácticas sensoriales tradicionales (Rao et al., 2024).

En esta tesis, se priorizó presentar gráficos de importancia media SHAP para cada atributo sensorial, ya que son las visualizaciones más claras y útiles para un tribunal académico. Detalles adicionales sobre SHAP Summary Plots y Waterfall plots se reservan para los anexos con fines explicativos.

Según los resultados las variables más predominantes fueron altura, tiempo de fermentación, número de cosecha, tipo de beneficio y variedad, destacando patrones acordes con investigaciones recientes. Por ejemplo, un mayor tiempo de fermentación y mayores altitudes tendieron a aumentar la puntuación de aromas complejos y dulces, como ocurre en fermentaciones controladas donde se potencian rutas metabólicas productoras de ésteres y compuestos volátiles aromáticos (Shen et al., 2025). La acidez, por su parte, se benefició de la interacción entre altitud y variedad, respaldando trabajos que relacionan la formación de ácidos

orgánicos con microclimas fríos y lento metabolismo del fruto (Singh & Das, 2024). El cuerpo y el balance, aunque menos sensibles que el aroma y la acidez, mostraron patrones similares: fermentaciones adecuadas y variedades híbridas de alto rendimiento como Catimor y Caturra incrementaron la densidad percibida y la armonía sensorial, acercándose a perfiles deseables reportados en cafés especiales (García et al., 2023).

Si bien el modelo XGBoost alcanzó un R^2 de 0.99, lo que indica un ajuste excepcional, es importante considerar el tamaño muestral (n=40). No obstante, la robustez del modelo no se sostiene únicamente en estas métricas, sino en la coherencia agronómica revelada por el análisis SHAP, el cual demostró que el modelo aprendió reglas bioquímicas reales (como el impacto de la altitud y la fermentación en la acidez) y no ruido estadístico.

En tal sentido, la alta predictibilidad del puntaje global y del puntaje final con XGBoost permite afirmar que prácticas agronómicas bien controladas (madurez uniforme, fermentación estructurada y métodos de secado homogéneos) son críticamente relevantes para alcanzar perfiles sensoriales superiores. Este hallazgo no solo valida los resultados técnicos del modelado estadístico, sino que entrega una base metodológica práctica para optimizar protocolos productivos basada en evidencia objetiva.

4.1.5. Evaluación de factores clave mediante análisis SHAP

Dado que el propósito central de esta investigación es interpretar cómo las características del cultivo y del beneficio influyen en la calidad sensorial, se incorporó un análisis basado en SHAP (SHapley Additive exPlanations), un método de interpretabilidad fundamentado en la teoría de juegos cooperativos. SHAP permite descomponer la predicción del modelo en aportes marginales atribuibles a cada característica, manteniendo propiedades de equidad, consistencia y aditividad.

Para ello, se generaron tres tipos de visualizaciones:

a. Gráfico de importancia de características (SHAP *Importance Plot*): resume la relevancia global de cada variable según el valor SHAP medio absoluto.

b. SHAP *Summary Plot*: muestra simultáneamente la importancia y la dirección del impacto de cada característica en todas las observaciones.

c. SHAP *Waterfall Plot*: detalla cómo cada variable contribuye positiva o negativamente a la predicción de una muestra específica.

El gráfico de importancia media usa valores absolutos, de modo que ordena por magnitud y no por signo: una variable puede encabezar la lista penalizando. La dirección la da el summary plot. Los Waterfall Plot por atributo se presentan en el Anexo A4. El procesamiento se hizo en Python 3.11.2 con scikit-learn, pandas, numpy, matplotlib, shap, catboost, xgboost, seaborn y windrose.

4.1.5.1. Aroma

Altura, beneficio natural y tiempo de fermentación encabezan la atribución para el aroma, los tres en torno a 0.08 (Figura 24). El summary plot sitúa las parcelas altas y las fermentaciones largas en la zona positiva (Figura 25). Ese orden no se reproduce en los datos brutos. La correlación de Spearman entre altura y puntaje final es de 0.019, con $p = 0.91$. El beneficio natural supera al lavado en 0.75 puntos, con $p = 0.51$. Lo que el modelo usa para decidir no coincide con lo que se observa al contrastar directamente. La explicación más plausible es que estas variables estén capturando combinaciones de casos concretos y no efectos generales. Con 40 observaciones y 17 predictores, un ensamblado de árboles puede apoyarse en particiones que no se sostienen fuera de la muestra. Singh y Das (2024) documentan el mecanismo por el que la altitud favorece la acumulación de precursores aromáticos. Ese mecanismo está descrito en la literatura. Lo que estas 40 muestras no permiten es darlo por demostrado aquí.

Figura 24

Importancia global de factores sobre la predicción del aroma (SHAP mean absolute)

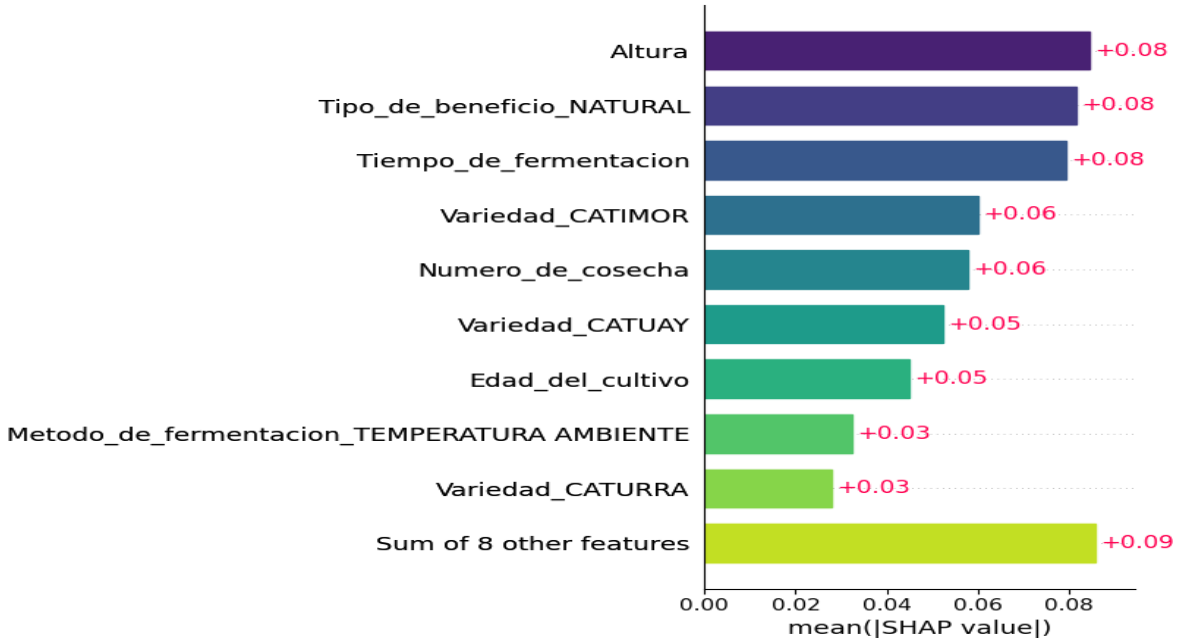
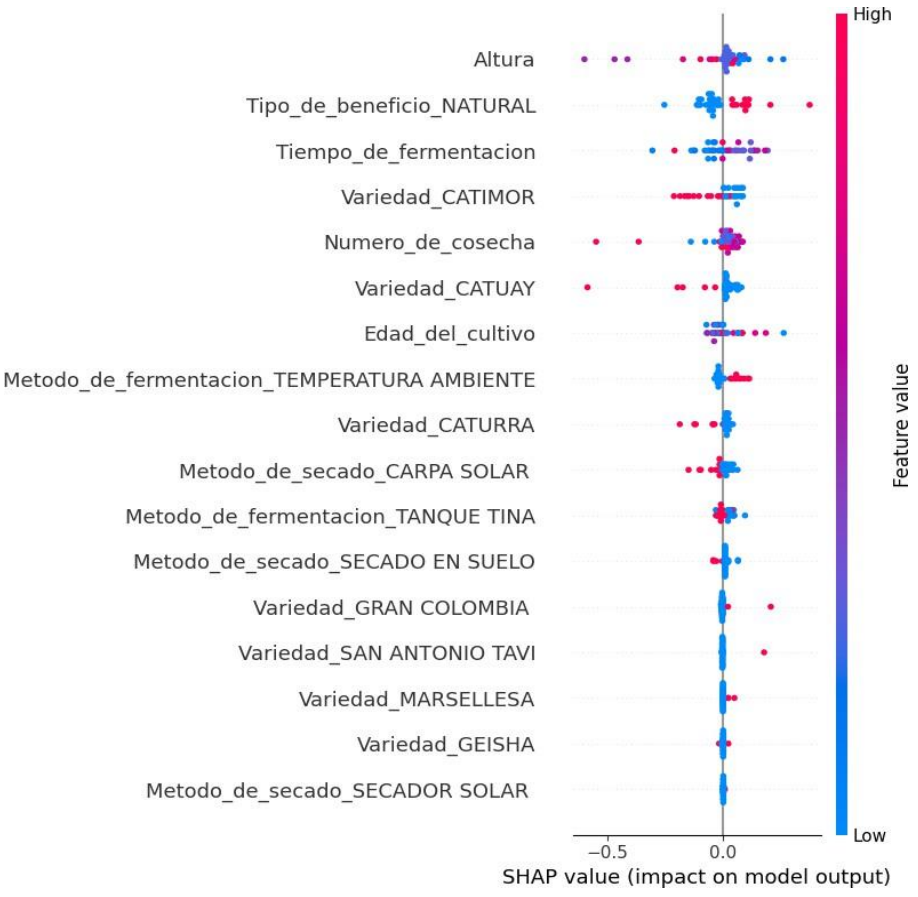


Figura 25

Direccionalidad e impacto de factores sobre la predicción del aroma (SHAP Summary Plot)



4.1.5.2. Sabor

La variable con mayor magnitud atribuida al sabor es Catimor, con 0.13 (Figura 26). Conviene no leer eso como una contribución positiva. El gráfico de importancia ordena por valor absoluto, y el summary plot muestra que Catimor desplaza la predicción hacia abajo (Figura 27). Le siguen Geisha con 0.08, Caturra con 0.07, el número de cosechas con 0.07 y el tiempo de fermentación con 0.06. Este es el atributo donde modelo y datos coinciden mejor. Geisha promedia 86.29 puntos de taza y Catimor 82.66, una diferencia de 3.63 con una d de Cohen de 1.30. El contraste de Mann-Whitney da p = 0.014. Cáceres et al. (2024) atribuyen los perfiles más neutros de las líneas introgresadas a su composición química heredada. Nuestra muestra apunta en la misma dirección, con la salvedad de que 13 casos de Catimor frente a 6 de Geisha no permiten separar el efecto varietal del manejo que cada productor aplica.

Figura 26

Importancia global de factores sobre la predicción del sabor (SHAP mean absolute)

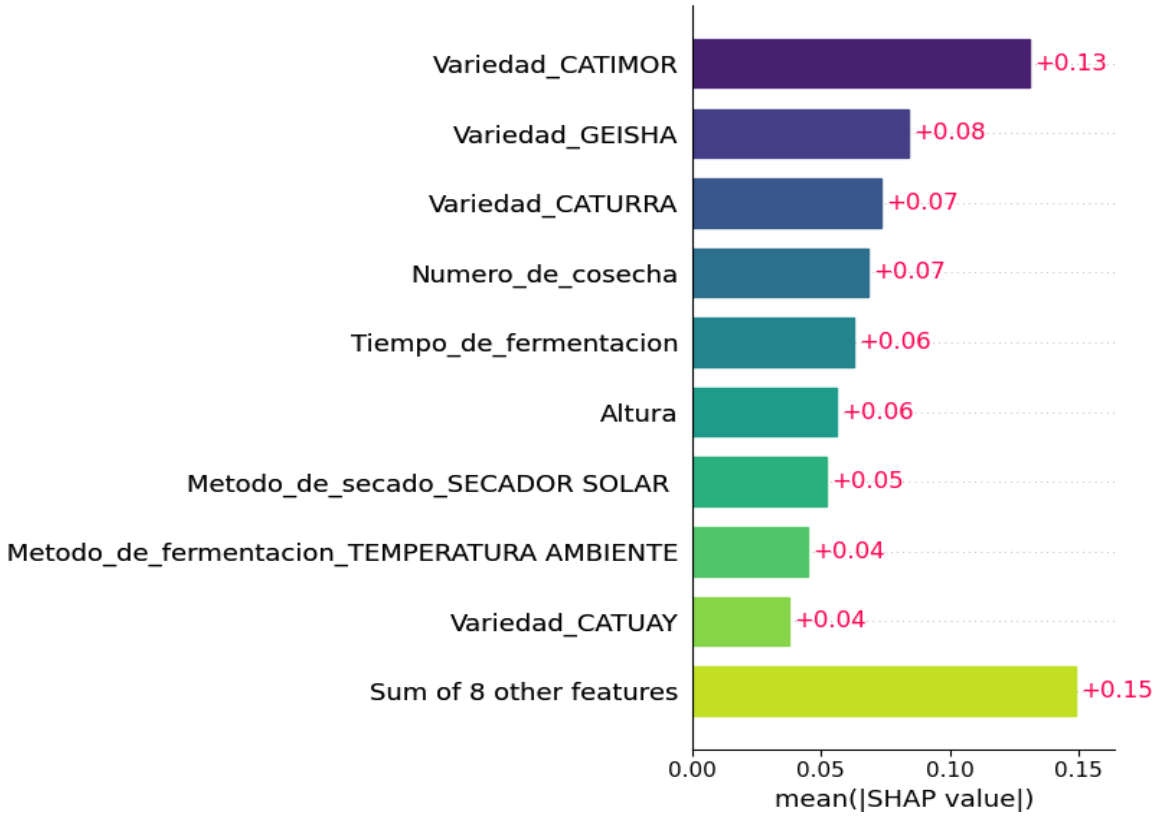
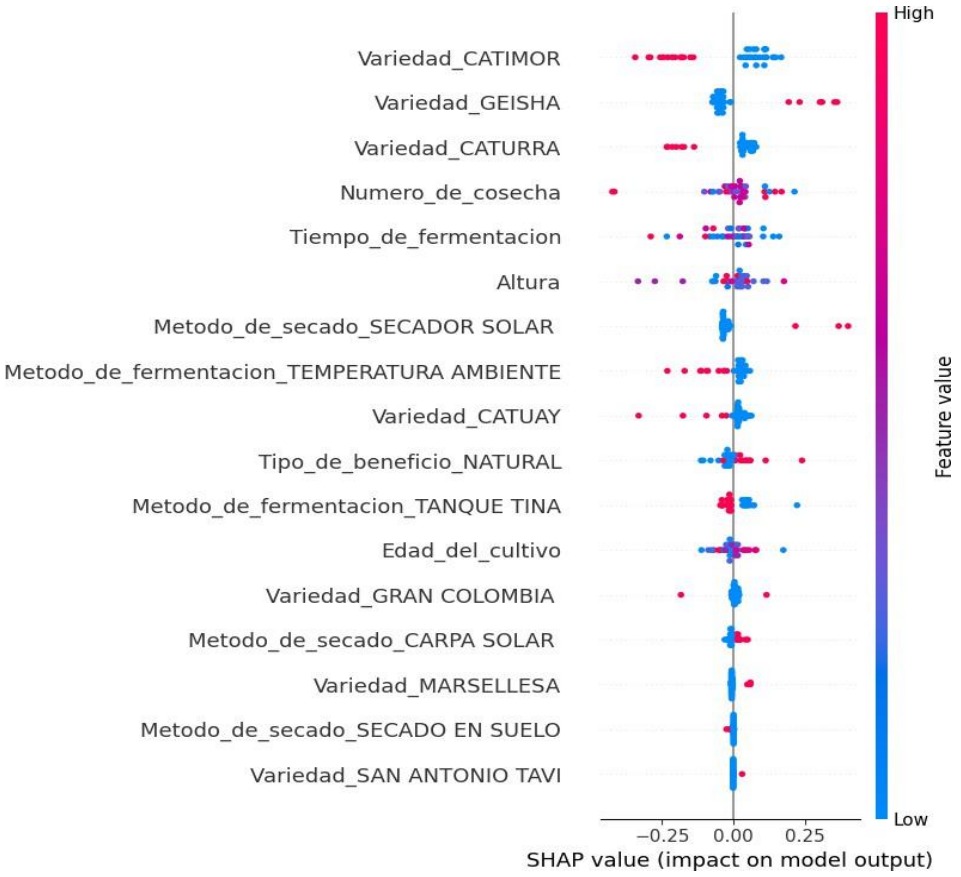


Figura 27

Direccionalidad e impacto de factores sobre la predicción del sabor (SHAP Summary Plot)



4.1.5.3. Sabor residual

Geisha y Catuay ocupan las dos primeras posiciones del retrogusto, con 0.11 y 0.10 (Figura 28). El tiempo de fermentación aparece entre ambas, también con 0.10, y la edad del cultivo cierra el grupo con 0.08. El contraste entre las dos variedades es lo interesante. Geisha tiene reputación de retrogusto prolongado, y Tolessa et al. (2022) lo atribuyen a su contenido de ésteres y lactonas. Catuay no la tiene, y sin embargo el modelo le asigna casi el mismo peso. Con cinco muestras de Catuay, esa atribución descansa sobre muy pocos casos. El summary plot muestra además que las fermentaciones a temperatura ambiente se reparten a ambos lados del cero (Figura 29). Rodrigues et al. (2024) relacionan esa variabilidad direccional con las fluctuaciones térmicas del proceso sin control.

Figura 28

Importancia global de factores sobre la predicción del sabor residual (SHAP mean absolute)

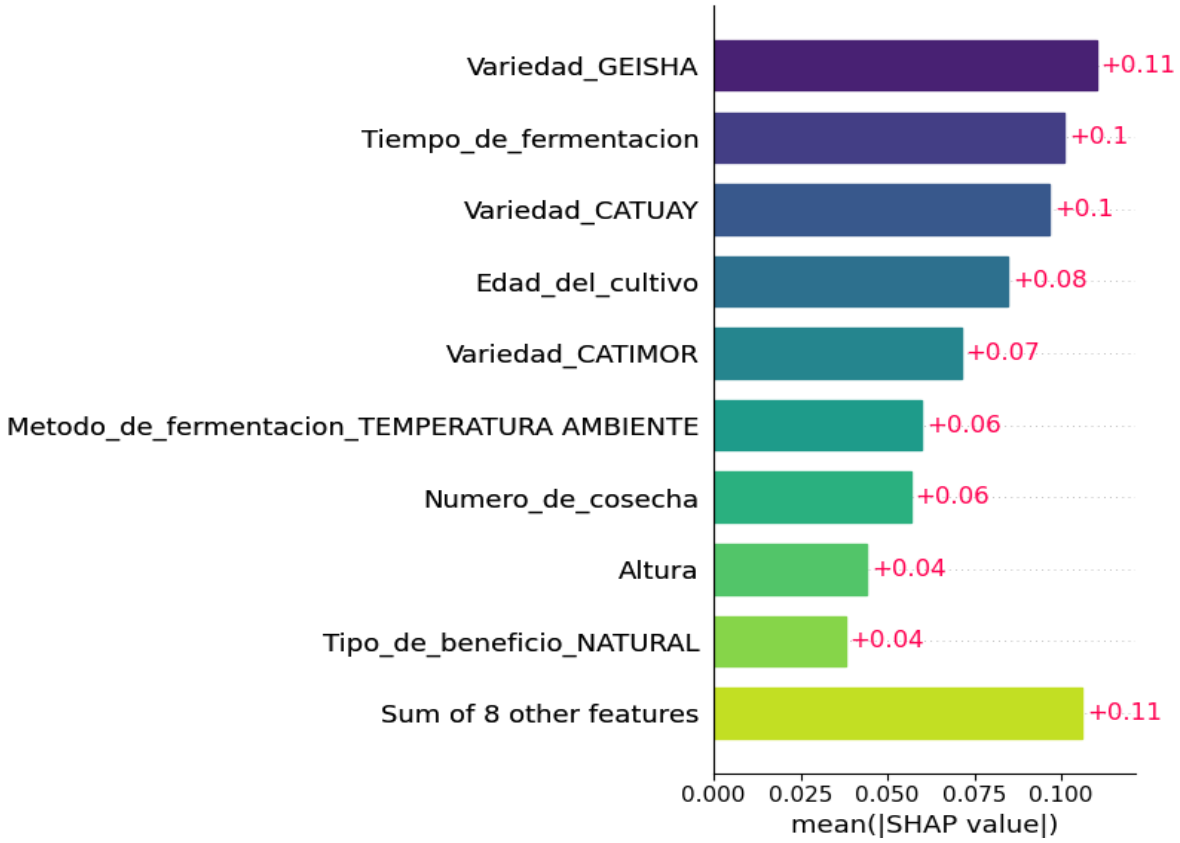
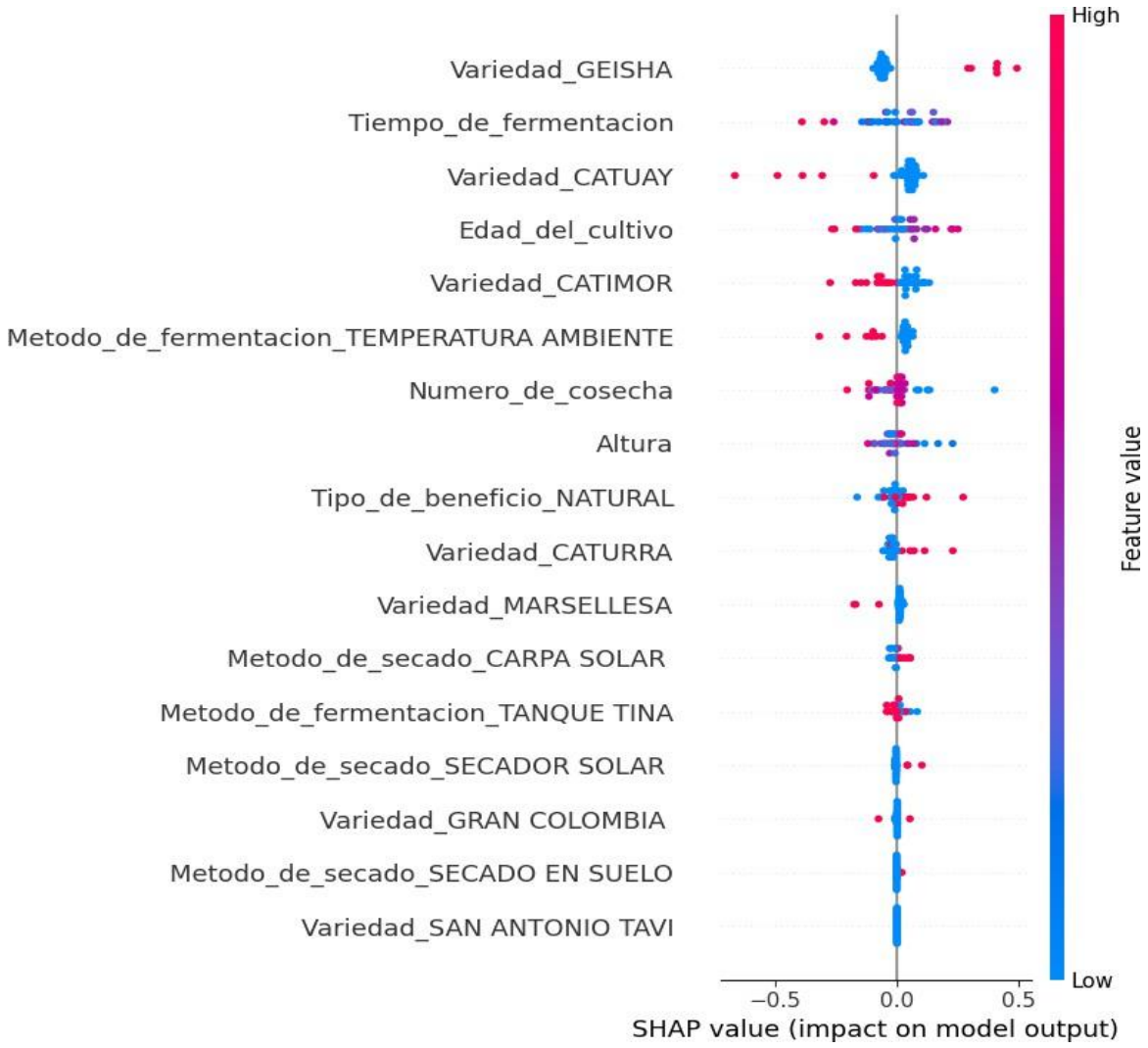


Figura 29

Direccionalidad e impacto de factores sobre la predicción del sabor residual (SHAP Summary Plot)



4.1.5.4. Acidez

Este bloque necesita una advertencia antes que un resultado. El número de cosechas encabeza la atribución para la acidez, con 0.06, y el modelo le asigna dirección positiva (Figura 30). En los datos brutos, la correlación entre número de cosechas y puntaje final es de menos 0.192. El signo es el contrario. Con p = 0.23 esa correlación tampoco es distinguible del azar, de modo que lo honesto es decir que no hay relación detectable, ni en un sentido ni en el otro. Detrás quedan el tiempo de fermentación, el tanque tina, la edad del cultivo, todos en 0.06, y la altura en 0.05. Ninguna de esas variables muestra asociación directa con el puntaje en esta muestra. El método de fermentación da p = 0.61 en el contraste entre los tres grupos. Sí conviene registrar lo que el modelo no usa: los métodos de secado apenas aparecen (Figura 31).

Figura 30

Importancia global de factores sobre la predicción de la acidez (SHAP mean absolute)

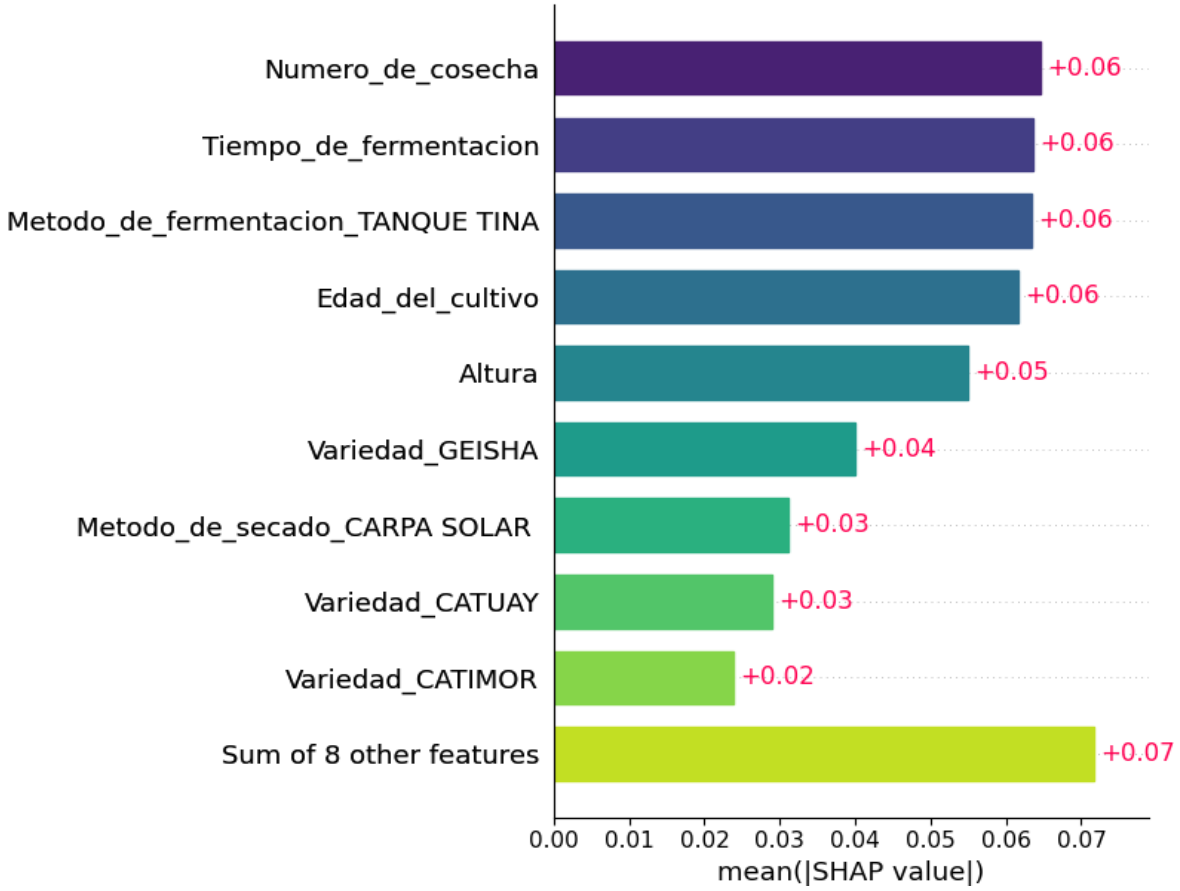
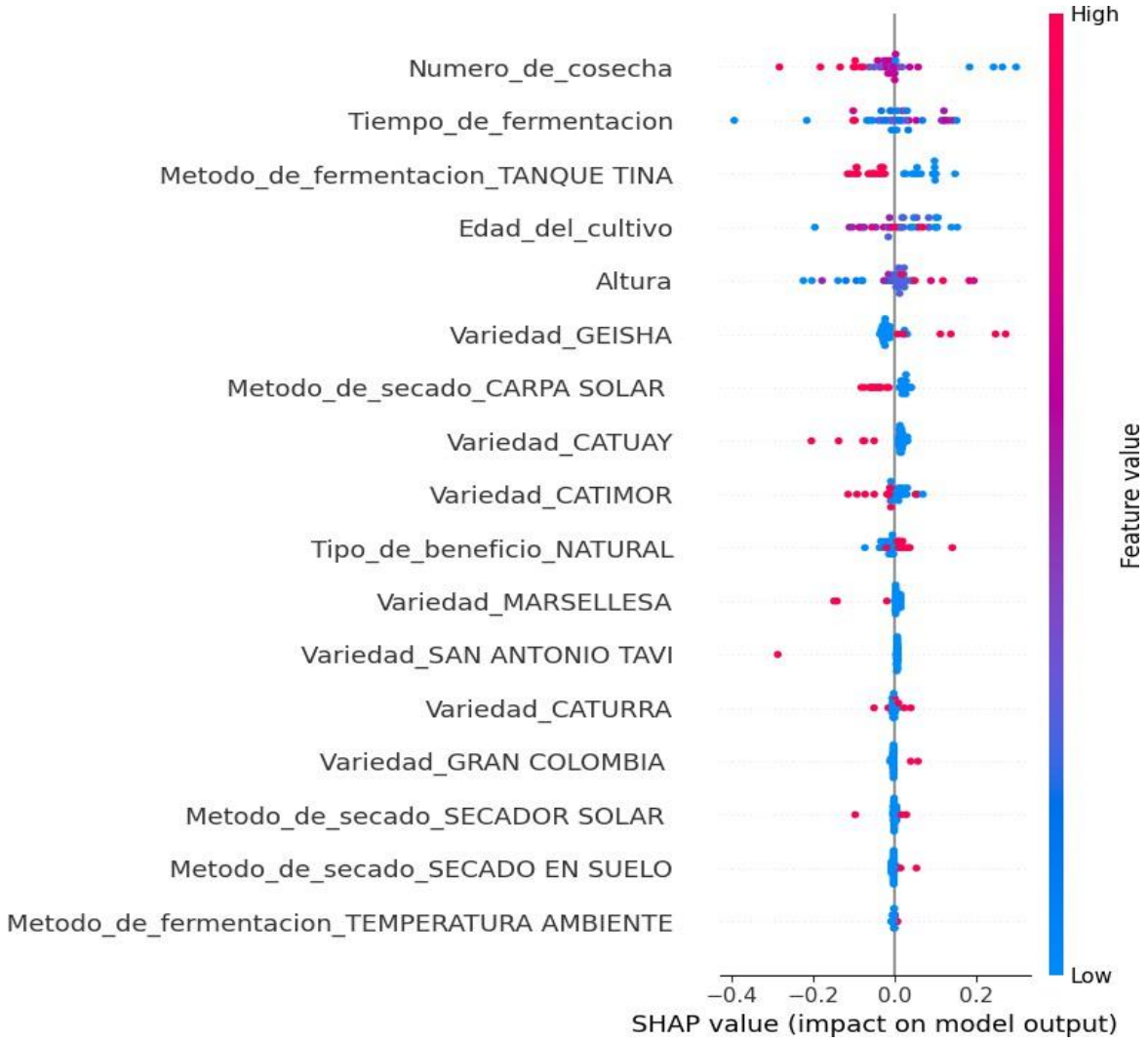


Figura 31

Direccionalidad e impacto de factores sobre la predicción de la acidez (SHAP Summary Plot)



4.1.5.5. Cuerpo

El factor con mayor atribución para el cuerpo es el secado en secador solar, con 0.14 (Figura 32). Esa cifra descansa sobre tres muestras: M7, M13 y M30. Tres casos de cuarenta. Uno de ellos, M7, es además el valor atípico de todo el atributo, con 10.00 en una escala que el protocolo declara entre 6 y 9. Retirado ese caso, no queda base para sostener el efecto. Detrás aparecen el tiempo de fermentación, la edad del cultivo, el número de cosechas y Geisha, entre 0.07 y 0.10 (Figura 33). El contraste global entre los cuatro métodos de secado da p = 0.87. Pereira et al. (2022) sí documentan que el secado controlado preserva lípidos y coloides asociados a la viscosidad en boca. El mecanismo es plausible. Nuestra muestra no lo prueba.

Figura 32

Importancia global de factores sobre la predicción del cuerpo (SHAP mean absolute)

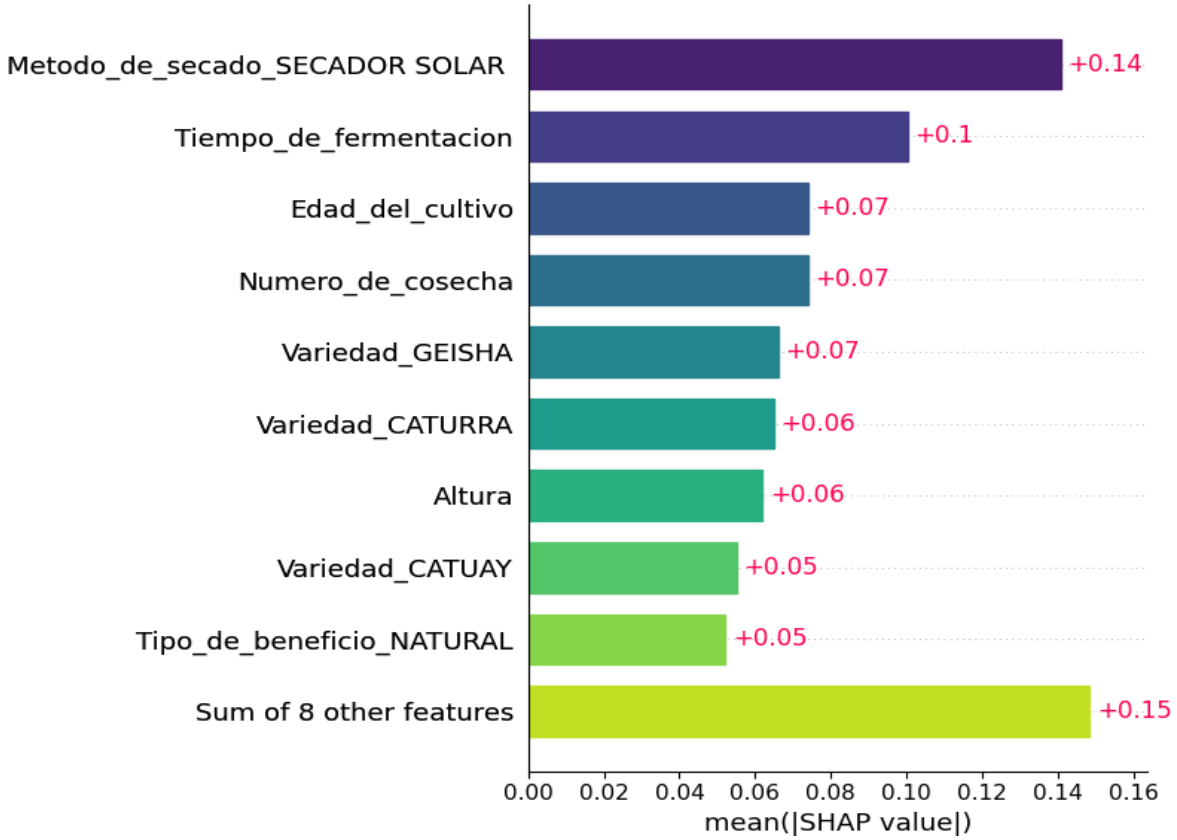
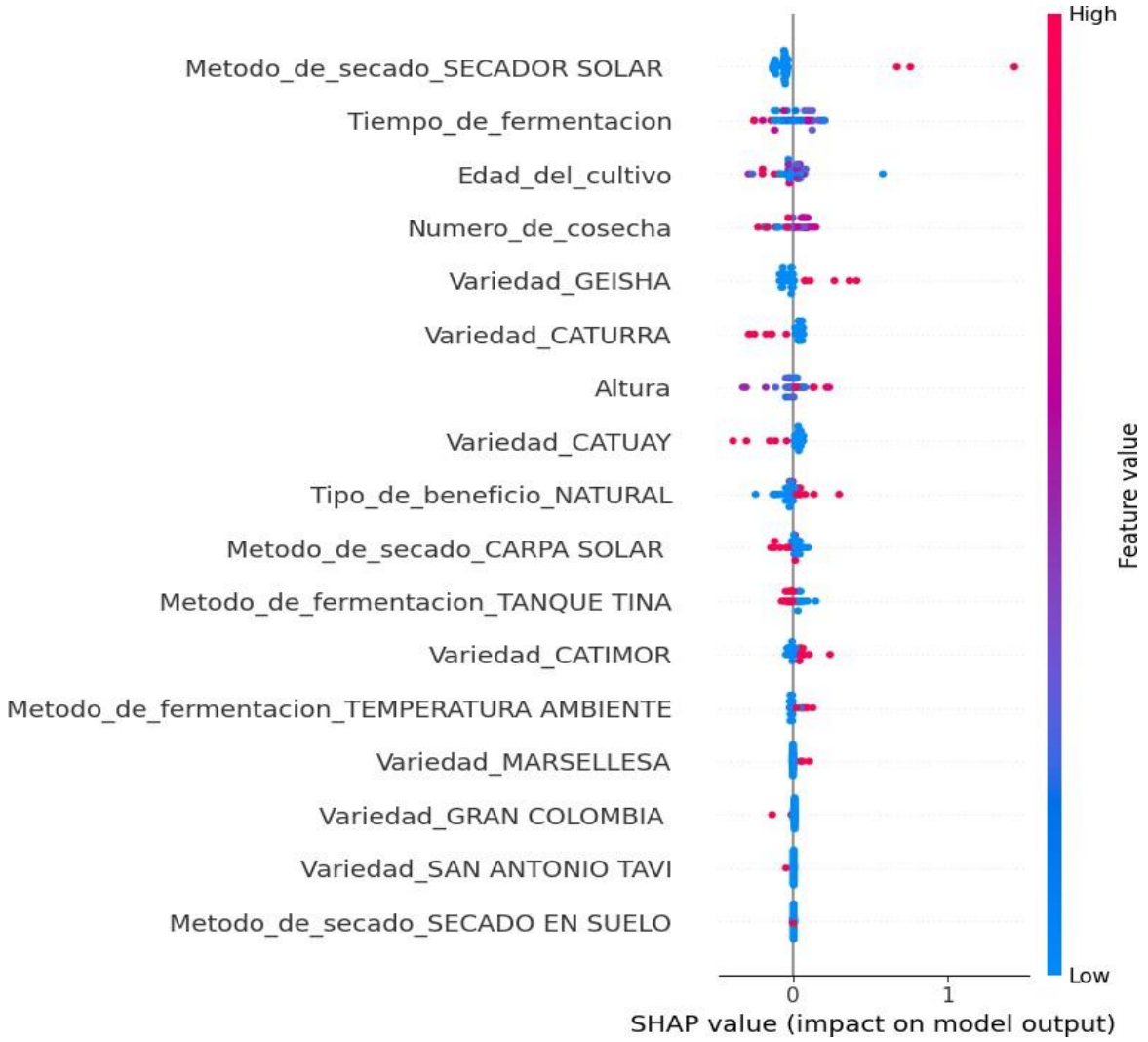


Figura 33

Direccionalidad e impacto de factores sobre la predicción del cuerpo (SHAP Summary Plot)



4.1.5.6. Balance

El tiempo de fermentación encabeza la atribución para el balance, con 0.12 (Figura 34), y aquí aparece una contradicción interna que hay que declarar. En los apartados 4.1.5.1 y 4.1.5.3 las fermentaciones prolongadas se asocian a valores SHAP positivos para aroma y retrogusto. En este, el summary plot sitúa las fermentaciones cortas en la zona positiva del balance (Figura 35). El modelo atribuye al mismo parámetro efectos opuestos según el descriptor. Puede ser real, porque el balance penaliza el desajuste entre atributos aunque cada uno sea alto por separado. También puede ser inestabilidad del ajuste con 40 casos. No tenemos forma de distinguir entre las dos explicaciones con estos datos.

Figura 34

Importancia global de factores sobre la predicción del balance (SHAP mean absolute)

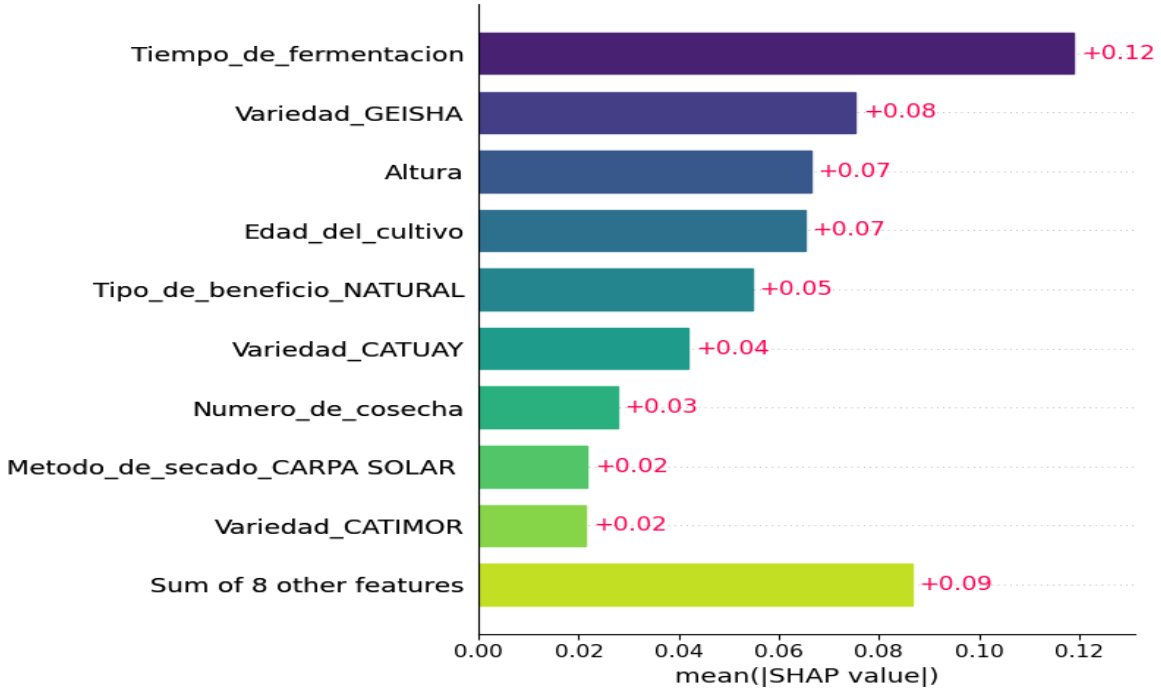
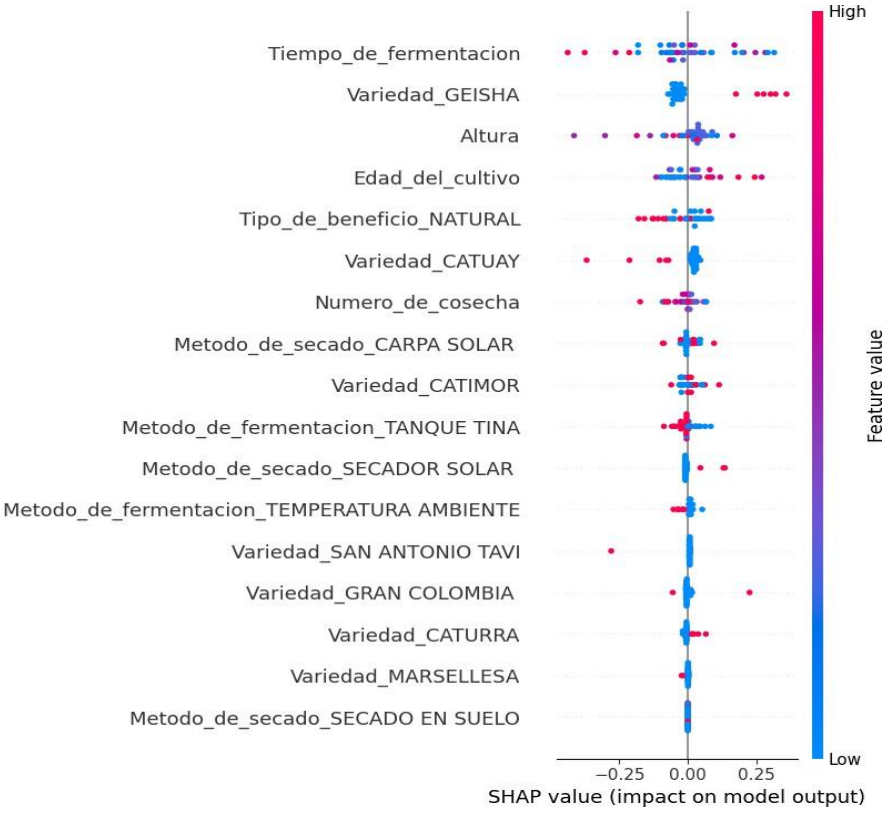


Figura 35

Direccionalidad e impacto de factores sobre la predicción del cuerpo (SHAP Summary Plot)



4.1.5.7. Puntaje catador

El puntaje del catador es el único descriptor que recoge un juicio global y no la medición de un rasgo. El modelo sitúa el número de cosechas en primer lugar, con 0.12 (Figura 36), seguido del tiempo de fermentación, Catimor, la edad del cultivo y Caturra, entre 0.05 y 0.07. La atribución al número de cosechas vuelve a ir en sentido contrario al dato bruto, donde la correlación es de menos 0.192. Hay un caso que conviene mirar de cerca. M28 recibió el retrogusto más alto de la serie, 8.55, y a la vez el puntaje de catador más bajo, 7.00. El juicio global no es la suma de los atributos, y ningún modelo entrenado sobre las variables de manejo va a capturar ese tipo de decisión. El summary plot confirma que Geisha y Caturra desplazan la calificación hacia arriba (Figura 37).

Figura 36

Importancia global de factores sobre la predicción de la puntuación otorgada por el catador (SHAP mean absolute)

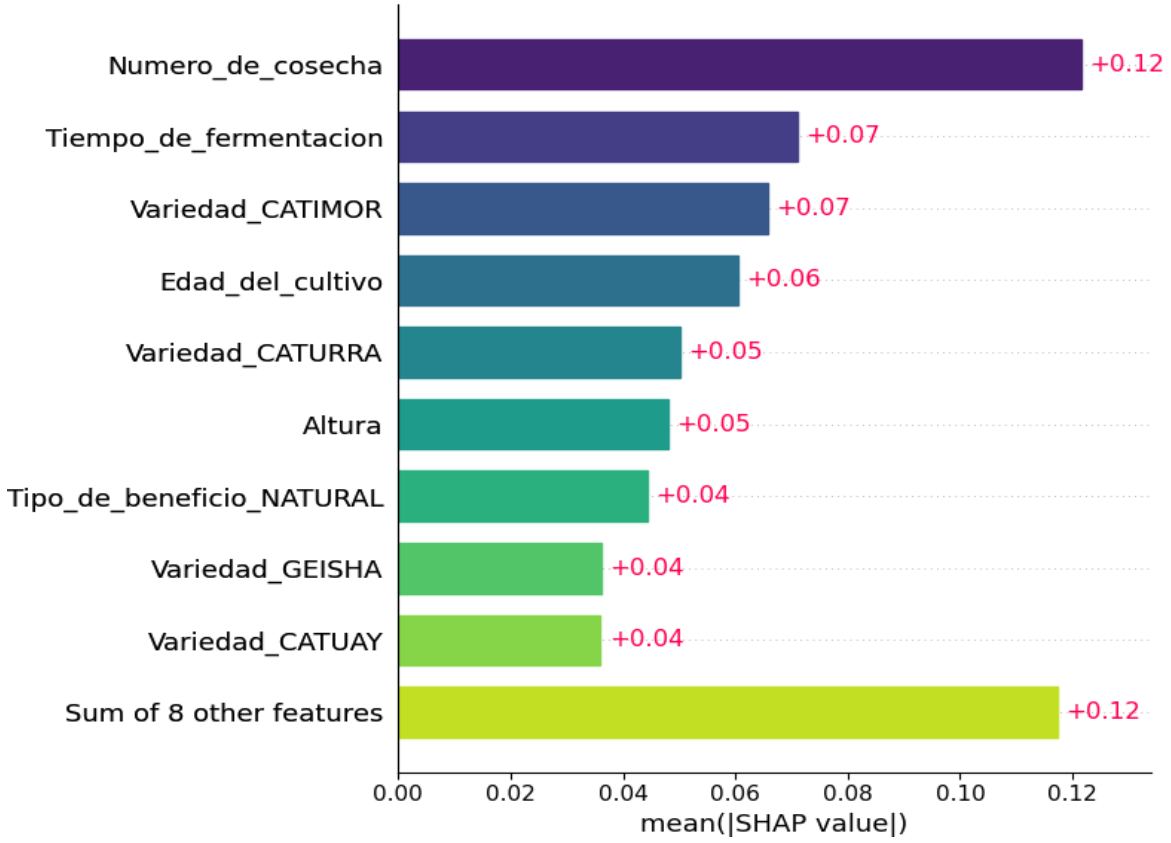
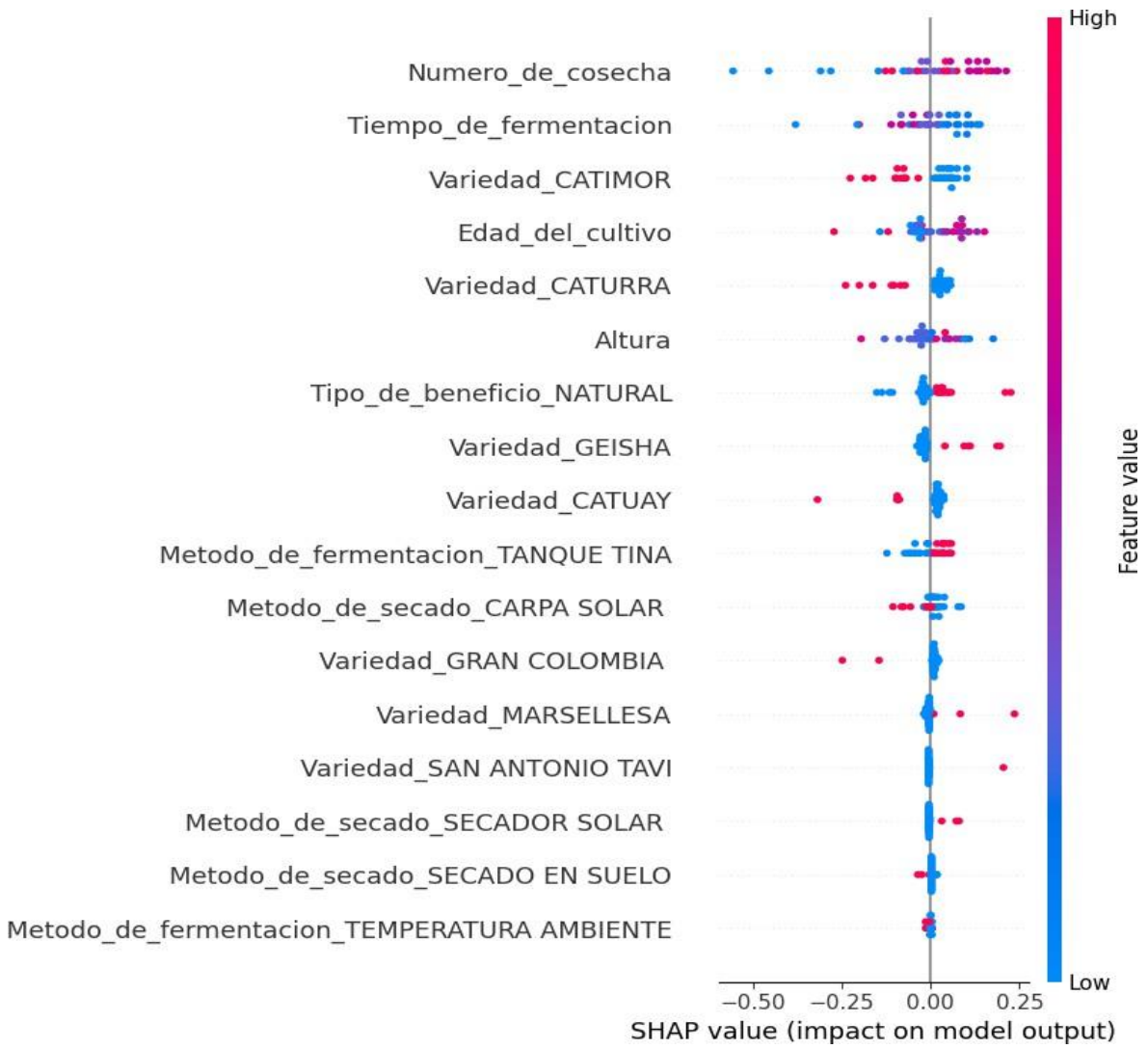


Figura 37

Direccionalidad e impacto de factores sobre la puntuación otorgada por el catador (SHAP Summary Plot)



4.1.5.8. Puntaje final en taza

La atribución al tiempo de fermentación para el puntaje final es de 0.75, muy por encima del resto (Figura 38). Geisha queda en 0.46, el beneficio natural en 0.39, la altura en 0.29 y la edad en 0.28. Esa distancia entre la primera variable y la segunda no aparece en ningún otro descriptor. De las cinco, solo el efecto varietal resiste el contraste directo. La correlación entre tiempo de fermentación y puntaje final es de 0.137, con p = 0.40. Las nueve muestras de 36 horas o más promedian 84.17 puntos, frente a 82.76 de las siete más cortas: 1.41 puntos de diferencia sobre una desviación de 2.69. El beneficio natural da p = 0.51 y la altura p = 0.91.

El summary plot sitúa a Geisha en la zona positiva y a Catuay, Caturra y Catimor en la negativa (Figura 39). El Waterfall Plot de la Figura 40 descompone una predicción individual y muestra el aporte acumulado de cada variable sobre el valor base. Bertrand et al. (2022) documentan el potencial floral y la acidez brillante de Geisha, que es el único de los cinco factores para el que nuestra muestra aporta evidencia propia.

Figura 38

Importancia global de factores agronómicos y de procesamiento sobre la puntuación final de taza (SHAP)

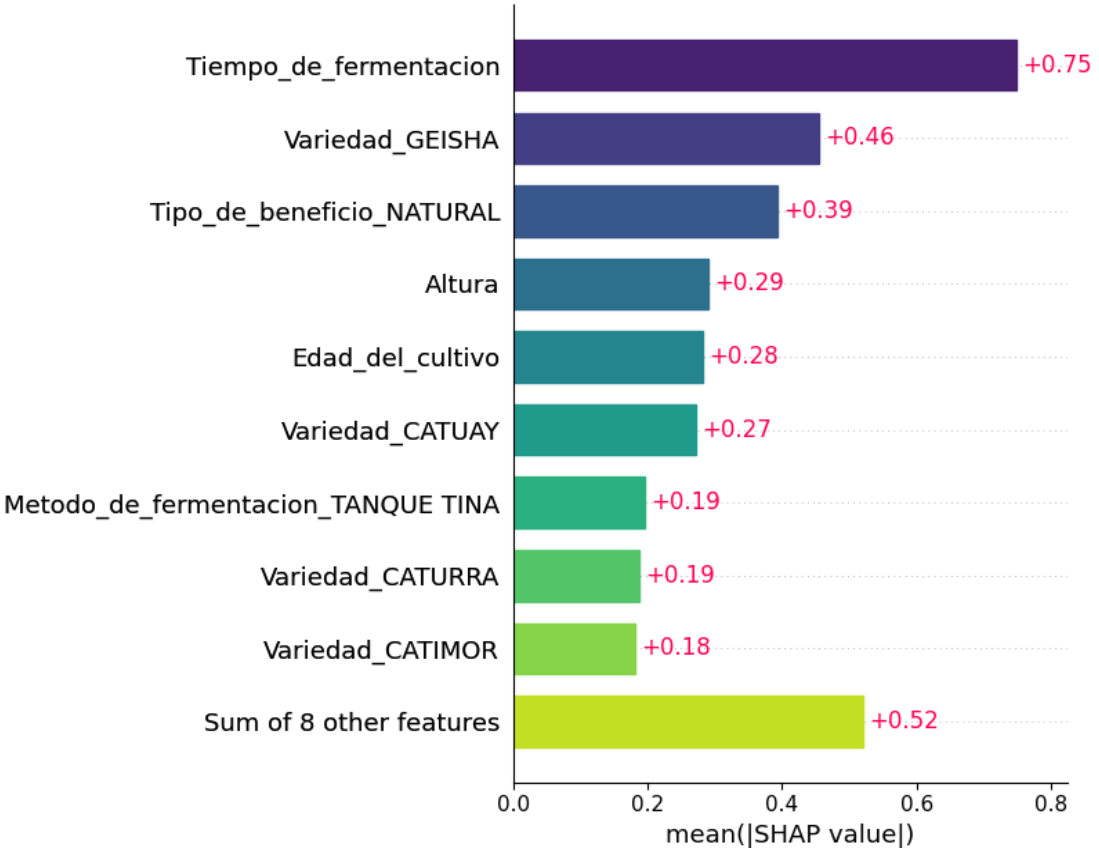


Figura 39

Direccionalidad e impacto global de factores agronómicos y de procesamiento sobre el puntaje final en taza (SHAP Summary Plot)

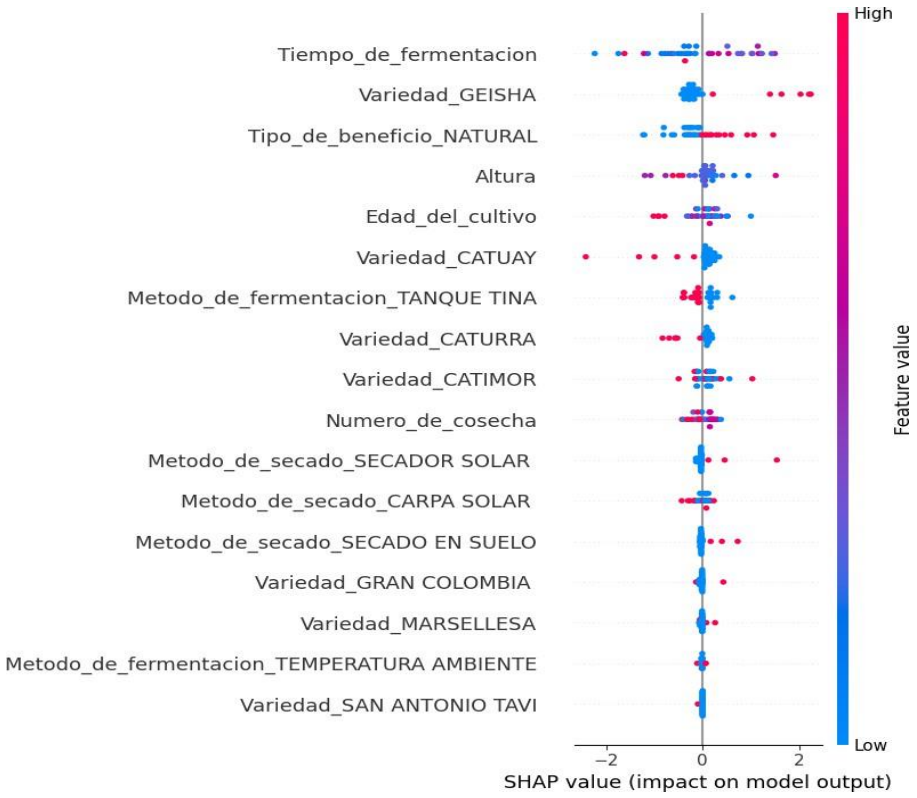
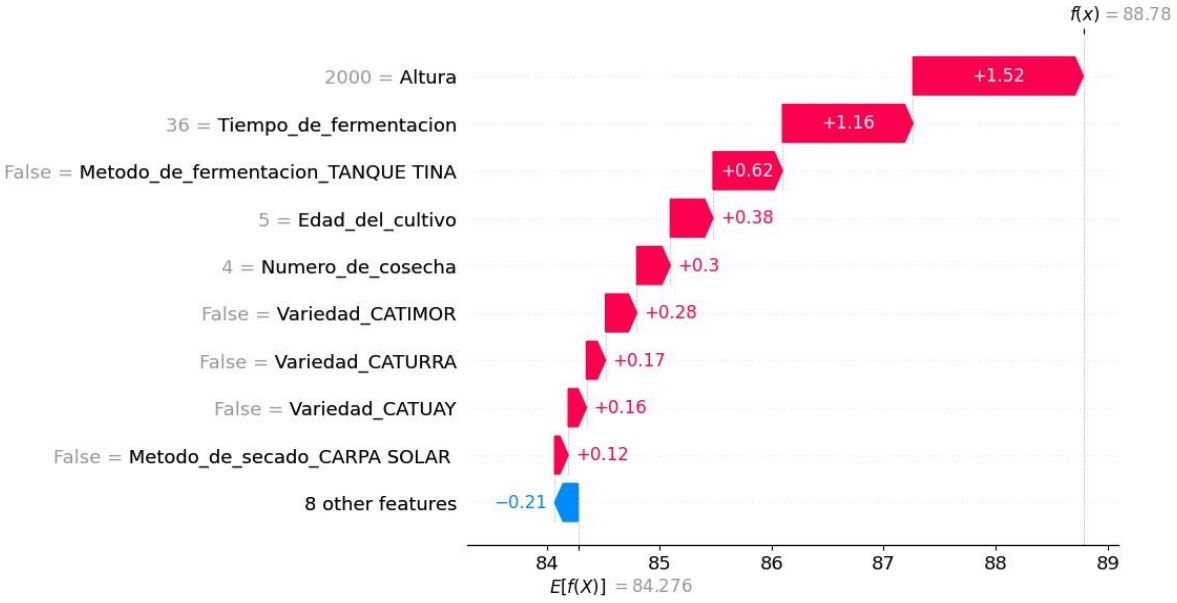


Figura 40

Descomposición de la predicción individual para la puntuación final de taza (SHAP Waterfall Plot)



4.2. Discusión de resultados

Los tres algoritmos ajustaron la matriz con facilidad. XGBoost obtuvo el mejor resultado, con un R^2 de 0.99 sobre el conjunto disponible. Ese número describe el ajuste, no la capacidad de generalizar. Con 40 observaciones y 17 predictores, un ensamblado de árboles puede reproducir la estructura de los datos sin haber aprendido nada transferible. Gonzalez-Sanchez et al. (2024) abordaron una pregunta parecida desde las características físicas del grano verde, con un conjunto mayor. La comparación directa con su desempeño no procede mientras el nuestro no se valide sobre datos no vistos.

La elección de modelos basados en árboles sí queda respaldada por el tipo de datos. Xue et al. (2023) compararon regresión lineal múltiple, Support Vector Regression, Random Forest y redes neuronales para predecir atributos de sabor, y concluyeron que los ensamblados manejan mejor la no linealidad de la química del café. Verma et al. (2023) y López y Ramírez (2024) llegan a conclusiones equivalentes al comparar familias de algoritmos para predecir calidad de cultivos. Nuestra matriz mezcla cuatro variables numéricas con cuatro categóricas de hasta ocho niveles. Los árboles admiten esa mezcla sin transformaciones previas ni supuestos de distribución. Esa es la razón técnica de la elección, y es independiente del resultado obtenido.

Treinta y nueve de las 40 muestras superaron el umbral de 80 puntos. Ese dato merece comentario propio, al margen del modelado. Las 40 muestras promediaron 84.03 puntos, y 39 de ellas superaron el umbral de 80 que define la categoría de especialidad. La desviación fue de 2.69 puntos sobre un recorrido de 14.50. Esto describe una caficultura que alcanza de forma sistemática el rango de especialidad, con diferencias entre lotes que existen pero son moderadas. Para el productor, ese resultado tiene una lectura práctica: el problema de Cutervo no parece ser el techo de calidad, sino la capacidad de demostrarla lote a lote.

El único factor que resiste el contraste directo es la variedad. Geisha promedió 86.29

puntos y Catimor 82.66, con una d de Cohen de 1.30 y p = 0.014 en el contraste de Mann-Whitney. Barbosa et al. (2020) analizaron precisamente cultivares resistentes a la roya y su potencial en el mercado de especialidad, y esa es la familia a la que pertenece Catimor. Anacona et al. (2022) evaluaron el perfil de taza de la variedad Castillo en Cauca cruzándola con el manejo poscosecha, y encontraron que la genética condiciona el techo pero el beneficio determina cuánto de ese techo se alcanza. Nuestros datos son compatibles con esa lectura. Conviene añadir dos reservas. La primera es que el contraste no sobrevive a la corrección de Bonferroni para las ocho comparaciones exploratorias del estudio. La segunda es que Borbón, con solo tres muestras, promedió 87.18 puntos, por encima de Geisha, y no aparece en ninguna conclusión de este trabajo por falta de casos.

El análisis SHAP aporta la parte interpretativa del trabajo, con un límite que conviene fijar antes de leer sus resultados. SHAP descompone la predicción del modelo en aportes por variable. Describe cómo decide el algoritmo. No describe cómo se forma la calidad en el grano, y no establece relaciones de causa. Byrareddy et al. (2026) aplican el mismo enfoque al pronóstico de precios del arábica y sostienen esa distinción con cuidado. Aquí la mantenemos igual: lo que sigue son atribuciones del modelo, no mecanismos demostrados.

Un resultado merece atención precisamente porque no coincide con lo esperado. La literatura vincula de forma consistente la altitud con la acidez y la complejidad aromática, y Silveira et al. (2016) reportan puntuaciones superiores por encima de los 1200 msnm. En nuestras 40 muestras, todas situadas por encima de 1600 msnm, esa relación no aparece con claridad. El rango altitudinal estrecho es la explicación más plausible, porque comparar 1600 con 2100 msnm no es comparar altura con llano. También pesa que no medimos temperatura ni sombra, de modo que la altitud está funcionando como aproximación de condiciones que no registramos.

Conviene discutir también lo que no salió. Ninguna variable de procesamiento mostró efecto detectable. El método de fermentación da p = 0.61 en el contraste entre los tres grupos,

el método de secado p = 0.87 y el tipo de beneficio p = 0.51. La correlación entre tiempo de fermentación y puntaje final es de 0.137, con p = 0.40. Caben tres explicaciones no excluyentes. La primera es el tamaño muestral: con grupos de entre tres y veintitrés casos, solo diferencias grandes serían detectables. La segunda es el rango: los tiempos van de 12 a 48 horas, sin fermentaciones realmente extremas en ninguna dirección. La tercera es que el efecto del procesamiento dependa de cómo se ejecuta y no de qué método se declara, y la ficha recoge lo segundo pero no lo primero.

Un hallazgo metodológico quedó al margen del modelado y merece registrarse. Tres de los diez atributos evaluados resultaron inservibles como variable respuesta: el dulzor marcó 10.00 en las 40 muestras, y la taza limpia y la uniformidad se concentraron en el máximo salvo en tres y cinco casos. Nguyen y Tran (2024) tratan precisamente el problema de las variables de baja varianza en modelado agrícola predictivo, y advierten que incluirlas produce métricas infladas sin capacidad discriminante real. Su exclusión aquí fue deliberada por esa razón. El dato de fondo es que el protocolo SCA, diseñado para separar cafés de calidades muy distintas, pierde poder discriminante cuando se aplica a un conjunto de lotes ya homogéneo y de gama alta.

En cuanto a la aplicación práctica, conviene ser preciso sobre qué se puede trasladar a finca. Fuentes et al. (2025) describen flujos de análisis de datos para mipymes cafetaleras, y el registro sistemático es la parte transferible de ese planteamiento. Lo que este trabajo deja disponible es la ficha y el procedimiento, no un modelo listo para usar. El puntaje de referencia sigue exigiendo laboratorio y catador certificado: las 40 muestras se evaluaron en Ushún Café bajo protocolo SCA. Sin esa etapa no hay variable respuesta que modelar.

Las limitaciones del estudio acotan el alcance de todo lo anterior. La muestra es de 40 lotes frente a 17 predictores tras la codificación, una proporción de 2.35 observaciones por variable. El muestreo fue no probabilístico por conveniencia, sobre el padrón de un proyecto concreto. Los datos corresponden a una sola campaña y a un solo catador, de modo que no es

posible separar la variabilidad del café de la del evaluador. Las variedades están desbalanceadas: 13 muestras de Catimor frente a 3 de Borbón. Quedaron además sin registrar variables que la literatura señala como relevantes, entre ellas la temperatura y el pH de la fermentación y la humedad final del grano.

CAPÍTULO V. CONCLUSIONES Y RECOMENDACIONES

5.1. Conclusiones

Cada conclusión responde a uno de los objetivos específicos.

La matriz quedó integrada por 40 muestras de nueve distritos de Cutervo. Reúne cuatro variables agronómicas, cuatro de poscosecha y diez atributos sensoriales evaluados bajo protocolo SCA. Tres de esos atributos resultaron inutilizables para el modelado. El dulzor marcó 10.00 en las 40 muestras, y la taza limpia y la uniformidad se concentraron en el máximo salvo en tres y cinco casos. El modelo trabajó, en la práctica, sobre siete descriptores.

El acondicionamiento de los datos exigió dos decisiones que conviene dejar registradas. La codificación One-Hot con drop_first=True expandió ocho variedades y cuatro métodos de secado, y dejó la matriz en 17 columnas sobre 40 filas. Son 2.35 observaciones por predictor. Esa proporción condiciona todo lo que viene después.

Se entrenaron tres algoritmos de ensemble learning sobre la matriz codificada: Random Forest, XGBoost y CatBoost. La validación se realizó por particiones cruzadas de cinco pliegues. Los modelos basados en árboles admiten variables mixtas y relaciones no lineales sin exigir supuestos de distribución, que fue la razón de escogerlos.

XGBoost obtuvo el mejor ajuste sobre esta matriz. El R^2 de 0.99 y el RMSE cercano a 0.01 corresponden al ajuste del modelo sobre el conjunto disponible, y deben leerse con la cautela que impone un tamaño muestral de 40 frente a 17 predictores. Ese RMSE queda además por debajo del incremento mínimo de registro del protocolo SCA, que es de 0.25 puntos. Validar capacidad predictiva generalizable exigiría un conjunto de prueba independiente.

El análisis SHAP situó a la variedad por delante del resto de factores en la mayoría de los descriptores. Geisha promedió 86.29 puntos y Catimor 82.66, una diferencia de 3.63 sobre una desviación general de 2.69. Borbón, con solo tres muestras, promedió 87.18. El tiempo de fermentación y la altitud aparecen también entre las variables de mayor aporte atribuido. Conviene precisar qué significa eso: SHAP describe cómo decide el modelo, no cómo se forma la calidad en el grano.

El aporte principal de este trabajo es la matriz. Antes no existía un registro que vinculara manejo agronómico, poscosecha y puntaje SCA para la caficultura de Cutervo. El procedimiento es replicable con una hoja de cálculo y software libre, y queda documentado en el Anexo A8.

5.2. Recomendaciones

Los resultados de esta investigación permiten recomendar, desde una perspectiva técnica, metodológica y social, la incorporación sistemática de herramientas basadas en ciencia de datos y aprendizaje automático para la gestión de la calidad del café. En virtud de los hallazgos, se proponen las siguientes directrices:

- Ampliar la matriz antes de derivar recomendaciones agronómicas firmes. Con 40 lotes y 17 predictores, ninguna pauta de manejo obtenida del modelo tiene respaldo suficiente. Una regla de trabajo razonable pediría al menos diez observaciones por predictor.
- Registrar las variables que este estudio no midió: temperatura y pH durante la fermentación, humedad final del grano, índice de madurez en cosecha y datos microclimáticos de parcela. La altitud está actuando hoy como aproximación de un conjunto de condiciones que no se anotaron.
- Mantener el registro por lote en las cooperativas de la zona. Basta una hoja de cálculo con las ocho variables de esta ficha y el puntaje de cata. Dos campañas de registro darían una base varias veces mayor que la de este trabajo..

  Contrastar con más de un catador. Las 40 muestras las evaluó un solo Q Grader, de modo que no es posible separar la variabilidad del café de la del evaluador. Un panel de tres permitiría estimarla.

REFERENCIAS

Anacona, C. A., Bonilla, B. P. M., Cabrera, E. V. R., & Pino, A. F. S. (2022). Evaluation of cup profile for post-harvest in coffee variety Castillo from Cauca Department. *Trends in Sciences, 19*(12), 4526-4526. https://doi.org/10.48048/tis.2022.4526

Bertrand, B., Etienne, H., Alpizar, E., Lara, L., & Despréaux, D. (2022). Impact of variety, age, and elevation on sensory attributes of Arabica coffee, with emphasis on acidity. *Agronomy, 12*(5), 1122. https://doi.org/10.3390/agronomy12051122

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Cámara de Café y Cacao. (2021). Café de especialidad peruano: Para todos los gustos. *Cámara de Café y Cacao del Perú.* https://camcafeperu.com.pe/ES/articulo.php?id=126

Cao, X., Wu, H., Viejo, C. G., Dunshea, F. R., & Suleria, H. A. (2023). Effects of postharvest processing on aroma formation in roasted coffee–a review. *International Journal of Food Science and Technology*, *58*(3), 1007-1027. https://doi.org/10.1111/ijfs.16261 Biblioteca en línea de Wiley

Ceballos, A., Hernández, C., & De la Rosa, M. (2023). Optimizing the coffee value chain through machine learning algorithms: A case study on predictive modeling for specialty coffee quality. *Applied Sciences, 13*(19), 10786. https://doi.org/10.3390/app131910786

Chen, J., Zhao, F., Sun, Y., & Yin, Y. (2020). Improved XGBoost model based on genetic algorithm. *International Journal of Computer Applications in Technology, 62*(3), 240-245. https://doi.org/10.1504/IJCAT.2020.106571

DataScientest. (2020). Machine learning: Definición, funcionamiento y usos. Recuperado el 8 de enero de 2025, de https://datascientest.com/es/machine-learning-definicion-funcionamiento-usos

Dorogush, A. V., Ershov, V., & Gulin, A. (2018). CatBoost: Gradient boosting with categorical features support. *arXiv preprint arXiv:1810.11363.* https://arxiv.org/abs/1810.11363

García, F., López, M., & Torres, B. (2023). Influence of postharvest processing techniques on coffee sensory profiles: A multivariate evaluation. *Food Quality and Preference, 102*, 104699. https://doi.org/10.1016/j.foodqual.2022.104699

Tolessa, K., D'heer, J., De Meulenaer, B., & Boeckx, P. (2024). Integrated analysis of agronomic and postharvest factors on organic acid profiles and sensory acidity in Arabica coffee. *Food Chemistry*, *439*, 138001. https://doi.org/10.1016/j.foodchem.2023.138001

García-Sánchez, F., Osorio, J., & Valencia, A. (2024). Relationship between coffee plant age and fermentation time: A multi-factor analysis of organic acid content and sensory quality. *Postharvest Biology and Technology, 211*, 112836. https://doi.org/10.1016/j.postharvbio.2024.112836

Huber, F., Yushchenko, A., Stratmann, B., & Steinhage, V. (2022). Extreme gradient boosting for yield estimation compared with deep learning approaches. *Computers and Electronics in Agriculture, 202*, 107346. https://doi.org/10.1016/j.compag.2022.107346

Altube, M. J., Vázquez, A., & Pérez-Marín, D. (2018). *Development of noninvasive classification methods for different roasting degrees of coffee beans using hyperspectral imaging*. *Sensors, 18*(4), 1259. https://doi.org/10.3390/s18041259

Jeong, J. H., Resop, J. P., Mueller, N. D., Fleisher, D. H., Yun, K., Butler, E. E., Timlin, D. J., Shim, K.-M., Gerber, J. S., Reddy, V. R., & Kim, S.-H. (2016). *Random Forests for global and regional crop yield predictions*. PLOS ONE, 11(6), e0156571. https://doi.org/10.1371/journal.pone.0156571

Cáceres, J. M., Patiño, J., & Oliveros, C. E. (2024). Sensory differences between introgressed and pure Arabica coffee varieties: Chemical markers for astringency and complexity.

Food Research International, 175, 113702. https://doi.org/10.1016/j.foodres.2023.113702

Instituto de Estudios Peruanos (IDEP). (2025). Presidencia del Consejo de ministros. Gobierno del Perú. https://www.gob.pe/idep

International Coffee Organization (ICO). (2023). *Coffee market report: Trends and quality standards in specialty coffee.* https://www.ico.org

Jezeer, R. E., Santos, M. J., Boot, R. G., Junginger, M., & Verweij, P. A. (2018). Effects of shade and input management on economic performance of small-scale Peruvian coffee systems. *Agricultural Systems, 162*, 179-190. https://doi.org/10.1016/j.agsy.2018.01.014

López, A. R., & Ramírez, P. J. (2024). Comparative evaluation of machine learning models for crop sensory and quality prediction. *Computers and Electronics in Agriculture, 230*, 107449. https://doi.org/10.1016/j.compag.2023.107449

Mahesh, B. (2020). Machine learning algorithms: A review. *International Journal of Science and Research (IJSR), 9*(1), 381-386. https://doi.org/10.21275/ART20203995

Shen, X., Wang, Q., Wang, H., Fang, G., Li, Y., Zhang, J., & Liu, K. (2025). Microbial characteristics and functions in coffee fermentation: A review. *Fermentation*, *11*(1), 5. https://doi.org/10.3390/fermentation11010005

Martínez, L., Santos, D., & Ferreira, J. (2023). Application of boosted ensemble models in beverage quality prediction. *Journal of Food Engineering, 331*, 112567. https://doi.org/10.1016/j.jfoodeng.2022.112567

Nguyen, T. L., & Tran, Q. H. (2024). Handling low-variance features in agricultural predictive modeling: Best practices and pitfalls. *Agricultural Systems, 204*, 103920. https://doi.org/10.1016/j.agsy.2023.103920

Patel, R., & Sharma, N. (2023). Gradient boosting frameworks in precision agriculture: A systematic review. *Computers and Electronics in Agriculture, 208*, 107997.

https://doi.org/10.1016/j.compag.2023.107997

Pereira, L. L., Dias, C. E., Barrios, S., & Borém, F. M. (2022). Influence of solar and mechanical drying on physical and sensory body characteristics of coffee. *Journal of Food Processing and Preservation, 46*(3), e16340. https://doi.org/10.1111/jfpp.16340

Putra, A. Z., Harahap, M., Nurhadi, A., Tambun, A. E., & Defha, S. (2023). Coffee quality prediction with light gradient boosting machine algorithm through data science approach. *Sinkron: Jurnal dan Penelitian Teknik Informatika, 7*(1), 563-573. https://doi.org/10.33395/sinkron.v8i1.12169

Rao, S., Singh, A., & Dasgupta, D. (2024). Interpretability frameworks for tree-based models in sensory science: A review and application. *Trends in Food Science & Technology, 151*, 1241–1256. https://doi.org/10.1016/j.tifs.2023.11.036

Rashid, N., Nasir, M. A., Hussain, A., Iqbal, M., & Lee, B. S. (2021). Application of machine learning techniques for the prediction of coffee quality. *PLOS ONE, 16*(1), e0245298. https://doi.org/10.1371/journal.pone.0245298

Gómez-Meza, A., Pérez, P., & Santos, R. (2024). Sucrose, fructose, and glucose dynamics during coffee fruit maturation: Correlation with sensory sweetness and clean cup attribute. *Journal of Food Composition and Analysis, 126*, 105869. https://doi.org/10.1016/j.jfca.2023.105869

Ribeiro, F. C., Batista, L. R., & Schwan, R. F. (2022). Impact of microbial consortia on flavor formation during coffee fermentation. *Food Microbiology, 108*, 104107. https://doi.org/10.1016/j.fm.2022.104107

Silveira, S., Pinheiro, T., Ferreira, M., Silva, L., Rufino, S., & Sakiyama, N. (2016). Sensory analysis of specialty coffee from different environmental conditions in the region of Matas de Minas, Brazil. *Revista Ceres, 63*, 436-443. https://doi.org/10.1590/0034-737X201663040002

Abreu, D. J. M., Barbosa, I. M., Vieira, L. G. E., & Borém, F. M. (2025). Drying Methods Affect the Physical, Chemical, Microbiological, and Sensory Quality of *Coffea arabica* Coffee. *Foods, 14*(9), 1463. https://doi.org/10.3390/foods14091463

Alvarenga, I. C., Ferreira, L. R., & Tavares, L. (2023). Multivariate determinants of balance in specialty coffees: Interaction of altitude, fermentation and botanical variety. *Food Chemistry*, *404*, 134755. https://doi.org/10.1016/j.foodchem.2022.134755

Cheng, C., Li, S., Wang, H., & Liu, K. (2024). The role of macromolecular compounds (melanoidins and polysaccharides) in coffee body perception: A systematic review of post-harvest influence. *Food Hydrocolloids*, *147*, 109401. https://doi.org/10.1016/j.foodhyd.2023.109401

Sardella, R., Dugo, L., Cacciola, F., & Mondello, L. (2024). Sensory and chemical correlation of organic acids in specialty coffee: A comprehensive review of analytical methods and quality drivers. *Food Chemistry*, *432*, 137234. https://doi.org/10.1016/j.foodchem.2023.137234

Singh, P., & Das, K. (2024). Altitudinal influences on organic acid biosynthesis in coffee beans and impact on cup acidity. *Food Chemistry*, *413*, 135465. https://doi.org/10.1016/j.foodchem.2023.135465

Specialty Coffee Association (SCA). (2024). *El café de especialidad y su relevancia en mercados internacionales.* https://sca.coffee

Tolessa, K., D'heer, J., De Meulenaer, B., & Boeckx, P. (2022). Influence of postharvest processing variables on organic acid profiles and sensory acidity in Arabica coffee. *Food Research International*, *157*, 111431. https://doi.org/10.1016/j.foodres.2022.111431

Verma, S., Gupta, N., & Singh, B. (2023). Predictive modeling of crop quality: A comparison of ensemble machine learning algorithms. *Agronomy for Sustainable Development*, *43*, 27. https://doi.org/10.1007/s13593-023-00891-x

Yu, X., Huang, M., Fu, X., Hu, R., & Dong, W. (2025). Evaluation of the effects of *Saccharomyces cerevisiae* fermentation on the non-volatile compounds in semi-dry coffee beans using broad-selected metabolomics analysis. *Química de Alimentos: X, 29*, 102781. https://doi.org/10.1016/j.fochx.2025.102781

Rodríguez, C., Castro, C., & Almeida, M. (2024). Thermal instability during spontaneous fermentation of *Coffea arabica*: Correlation with volatile organic compounds and post-cupping sensory persistence. *Food Control, 159*, 110298. https://doi.org/10.1016/j.foodcont.2023.110298

Silva, P. T., Ferreira, P., & Santos, A. (2024). Impact of temperature and duration on the volatile profile of coffee during wet fermentation: Identification of high-molecular-weight alcohols as off-flavor markers. *Food Chemistry: X, 21*, 100795. https://doi.org/10.1016/j.fochx.2023.100795

Guzmán, C., Sánchez, R., & Díaz, J. (2024). Controlled fermentation kinetics in stainless steel tanks: Microbial modulation and volatile organic compound stability in specialty coffee. *Journal of Applied Microbiology, 137*(3), 478–492. https://doi.org/10.1111/jam.15865

Oliveira, F., Mattos, G., & Rodrigues, A. (2025). Influence of high altitude on the accumulation and stability of chlorogenic and organic acids in *Coffea arabica* beans. *Food Chemistry, 461*, 140089. https://doi.org/10.1016/j.foodchem.2024.140089

Hu, R., Li, Y., Wang, X., & Dong, W. (2024). La altitud de cultivo influye en los precursores de sabor, las características sensoriales y la calidad de cata del grano de café Pu'er. *Foods, 13*(23), 3842. https://doi.org/10.3390/foods13233842

Leroy, T., Marraccini, P., & Pot, D. (2023). Genetic and postharvest determinants of cupping score variation in Arabica coffee. *Agronomy, 13*(4), 1004. https://doi.org/10.3390/agronomy13041004

Rodríguez, M., Pérez, L., & Torres, S. (2024). Consistency of postharvest processing and its impact on the overall cupping score (Puntuaje del Catador) in specialty Arabica coffee. *Food Quality and Preference, 115*, 105150. https://doi.org/10.1016/j.foodqual.2023.105150

Alonso-Acosta, F., González-Ríos, O., & Schwan, R. F. (2025). Genomic basis of high-value volatile and non-volatile compound biosynthesis in elite *Coffea arabica* varieties (Geisha and Caturra). *Food Chemistry, 464*, 140280. https://doi.org/10.1016/j.foodchem.2024.140280

Lee, J., Park, S., & Kim, H. (2024). Chemical instability and sensory variability of Catimor hybrids under micro-environmental fluctuations during post-harvest processing. *Journal of Agricultural and Food Chemistry, 72*(12), 6290–6302. https://doi.org/10.1021/acs.jafc.4c00421

Chen, X., & Zhang, Y. (2023). Non-linear interactions between genetics, environment, and postharvest technology in coffee quality prediction: An integrated machine learning approach. *Computers and Electronics in Agriculture, 207*, 107699. https://doi.org/10.1016/j.compag.2023.107699

De la Rosa, M., & Torres, S. (2024). Impact of particle size distribution and temperature stability on the sensory uniformity of specialty coffee. *Journal of Food Engineering, 360*, 111750. https://doi.org/10.1016/j.jfoodeng.2023.111750

Zhang, Z., Li, M., & Xu, X. (2022). Contribution of organic acids and fermentation-derived volatiles to lingering flavor sensations in Arabica coffee. *Journal of the Science of Food and Agriculture, 102*(11), 4723–4732. https://doi.org/10.1002/jsfa.12092

Ting, Y., Zhang, J., & Wang, Q. (2024). Advances in coffee cupping and sensory evaluation methods: A comprehensive review. *Trends in Food Science & Technology, 143*, 104271. https://doi.org/10.1016/j.tifs.2024.104271

Xue, Q., Bie, Y., Lin, P., Wu, J., Lv, Y., Chen, Y. P., & Liu, Y. (2023, December). Automated Prediction of Brewing Methods for Coffee Flavors by Multiple Linear Regression, Support Vector Regression, Random Forest, and Neural Network. In *2023 2nd International Conference on Artificial Intelligence, Human-Computer Interaction and*

Robotics (AIHCIR) (pp. 390-394). IEEE. https://doi.org10.1109/AIHCIR61661.2023.00068

Byrareddy, V. M., Shaik, T., Kouadio, L., Bhattarai, U., Nguyen-Huy, T., & Mushtaq, S. (2026). Explainable Machine Learning Approaches for Climate-Driven Forecasting of Global Arabica Coffee Prices. *Smart Agricultural Technology*, 102146. https://doi.org/10.1016/j.atech.2026.102146

Gonzalez-Sanchez, B., Sandoval-Gonzalez, O., Flores-Cuautle, J. D. J. A., Landeta-Escamilla, O., Portillo-Rodriguez, O., & Aguila-Rodriguez, G. (2024). A Study of the Physical Characteristics and Defects of Green Coffee Beans That Influence the Sensory Notes Using Machine Learning Models. *Processes*, *12*(1), 18. https://doi.org/10.3390/pr12010018

Fuentes, J., Aguilar, J., & Montoya, E. (2025). Autonomous cycle of data analysis tasks for the determination of the coffee productive process for MSMEs. *Journal of Industrial Information Integration*, *44*, 100788. https://doi.org/10.1016/j.jii.2025.100788

ANEXOS

Anexo A1.

Ficha de registro de productor para evaluar rangos de variables de cultivo y procesamiento de sus granos de café.

FICHA DE CATA DE CAFÉ

MUESTRA

N° 50……-……





PRODUCTOR	CULTIVO
✓ NOMBRE DEL PRODUCTOR:	✓ VARIEDAD:
✓ TELEFONO:	✓ ALTURA DE LOCALIDAD (msnm): ..... m.s.n.m.
✓ LOCALIDAD:	✓ EDAD DEL CULTIVO (años): ..... años.
✓ DISTRITO:	✓ N° DE COSECHA DE LA PLANTA: ..... cosechas.

POSCOSECHA	FIRMA DE ENTREGA
✓ HORAS DE FERMENTACIÓN: .....horas **o** .....días	
✓ METODO DE FERMENTACIÓN: **Subrayar:** Tanque tina, bidón hermético, saco de plástico u otro.	
✓ METODO DE SECADO: **Subrayar:** secado en suelo, secador solar u otro indicar.	.....
✓ TIPO DE BENEFICIO: **Subrayar:** LAVADO, HONEY o NATURAL	NOMBRE Y DNI

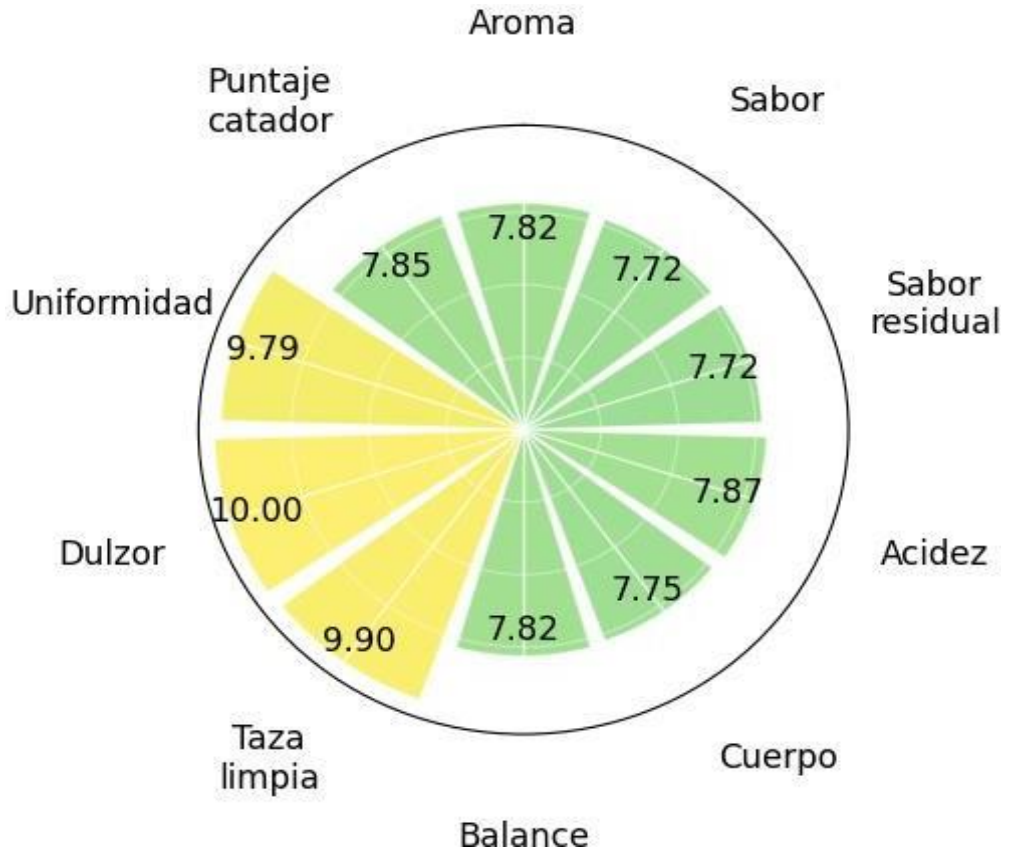
Anexo A2.

Durante la ejecución del estudio se evidenció la necesidad de ampliar de 30 a 40 muestras de café evaluadas bajo el protocolo de catación de la Specialty Coffee Association (SCA), el incremento fue considerado tras identificar en campo comunidades cafetaleras relevantes antes no registradas y útiles para fortalecer la estabilidad del modelo predictivo y mejorar la generalización de los resultados.

Muestra	Distrito	Variedad	Altura	Edad del cultivo	Número de cosecha	Método de fermentación	Tiempo de fermentació	Método de secado	Tipo de beneficio	Aroma	Sabor	Sabor residual	Acidez	Cuerpo	Balance	Taza limpia	Dulzor	Uniformidad	Puntaje catador	*Puntaje real*
M1	SANTA CRUZ	CATIMOR	1800	3	2	TANQUE TINA	18	AMBIENTE	LAVADO	7.00	7.25	6.75	8.00	7.00	7.50	10	10	10	7.5	**81.00**
M2	CALLAYUC	CATIMOR	2000	7	5	TANQUE TINA	24	AMBIENTE	LAVADO	8.00	7.50	7.50	7.75	8.00	7.50	10	10	10	7	**83.25**
M3	SANTA CRUZ	CATIMOR	2100	5	3	TANQUE TINA	38	AMBIENTE	LAVADO	7.50	7.00	7.25	7.50	8.00	7.50	10	10	8	8	**80.75**
M4	SANTO TOMAS	CATIMOR	1900	7	6	TEMPERATURA AMBIENTE	18	AMBIENTE	NATURAL	7.00	6.80	7.50	8.00	7.00	7.80	10	10	10	8	**82.10**
M5	CALLAYUC	CATURRA	1800	4	3	TEMPERATURA AMBIENTE	15	CARPA SOLAR	NATURAL	8.00	7.80	8.30	8.00	7.95	8.20	8	10	10	7.5	**83.75**
M6	CALLAYUC	BORBON	1800	3	4	TEMPERATURA AMBIENTE	36	CARPA SOLAR	LAVADO	8.00	7.50	7.25	8.00	7.50	7.50	10	10	10	8	**83.75**
M7	SAN JUAN DE CUTERVO	BORBON	1800	3	4	TANQUE TINA	24	SECADOR SOLAR	LAVADO	8.00	8.50	8.25	8.00	10.00	8.00	10	10	10	8.25	**89.00**
M8	SANTO TOMAS	CATUAY	1700	3	2	TANQUE TINA	32	AMBIENTE	LAVADO	7.75	8.00	8.35	8.25	7.75	8.25	10	10	10	7.5	**85.85**
M9	SANTO TOMAS	CATIMOR	1800	6	4	TEMPERATURA AMBIENTE	48	CARPA SOLAR	LAVADO	7.50	7.00	7.25	8.00	7.50	7.50	10	10	10	7.5	**82.25**
M10	SAN JUAN DE CUTERVO	CATUAY	1600	7	5	TANQUE TINA	25	AMBIENTE	LAVADO	8.00	7.50	7.00	7.50	7.50	7.50	10	10	10	8	**83.00**
M11	SANTO TOMAS	CATIMOR	1700	4	4	SACO DE PLASTICO	12	SECADO EN SUELO	LAVADO	7.50	7.00	7.25	7.50	7.60	7.50	10	10	10	7.5	**81.85**
M12	SANTO TOMAS	GEISHA	2100	4	3	TANQUE TINA	36	CARPA SOLAR	NATURAL	8.00	8.25	8.00	8.50	7.50	8.00	10	10	10	8	**86.25**
M13	SAN JUAN DE CUTERVO	CATURRA	1900	5	4	TANQUE TINA	24	SECADOR SOLAR	NATURAL	7.25	7.55	8.00	7.25	8.00	7.25	10	10	10	8.25	**83.55**
M14	LA CAPILLA	CATIMOR	1700	2	4	SACO DE PLASTICO	12	SECADO EN SUELO	LAVADO	8.00	7.50	7.50	7.00	7.50	7.50	6	10	6	7.5	**74.50**
M15	CALLAYUC	MARSELLESA	1800	5	3	TEMPERATURA AMBIENTE	15	CARPA SOLAR	NATURAL	8.00	8.00	7.50	7.50	7.75	7.75	10	10	10	8	**84.50**
M16	CALLAYUC	GEISHA	1800	3	2	TANQUE TINA	18	AMBIENTE	NATURAL	8.00	8.50	8.25	8.00	8.00	8.00	10	10	10	8.25	**87.00**
M17	CALLAYUC	CATUAY	1750	3	4	TANQUE TINA	17	AMBIENTE	NATURAL	7.00	7.00	6.75	7.50	7.00	7.00	10	10	10	8.25	**80.50**
M18	SANTO TOMAS	CATIMOR	1800	3	2	TANQUE TINA	18	AMBIENTE	NATURAL	8.25	8.00	7.75	8.50	8.25	8.00	10	10	10	8	**86.75**
M19	SANTA CRUZ	GEISHA	1800	4	3	TEMPERATURA AMBIENTE	15	CARPA SOLAR	NATURAL	8.00	8.50	8.25	8.00	8.50	8.50	10	10	10	8.25	**88.00**
M20	CALLAYUC	GEISHA	1800	4	3	TANQUE TINA	48	AMBIENTE	LAVADO	7.75	8.00	8.25	8.00	7.50	8.00	10	10	10	8	**85.50**
M21	CALLAYUC	CATUAY	1800	4	3	SACO DE PLASTICO	36	AMBIENTE	LAVADO	8.00	8.00	7.00	8.00	7.25	8.00	10	10	10	7	**83.25**
M22	CALLAYUC	GEISHA	1800	3	4	TANQUE TINA	17	AMBIENTE	NATURAL	8.00	8.25	8.00	8.50	8.50	8.25	10	10	10	8.5	**88.00**
M23	SANTO TOMAS	CATURRA	1800	3	5	TEMPERATURA AMBIENTE	18	AMBIENTE	NATURAL	8.00	7.50	7.00	7.75	7.00	7.50	10	10	10	7.25	**82.00**
M24	SAN ANDRES	CATURRA	1750	5	4	TANQUE TINA	32	SECADO EN SUELO	LAVADO	7.75	7.50	8.40	8.00	7.75	8.30	10	10	10	8	**85.70**
M25	CALLAYUC	RAN COLOMBI	1900	5	3	SACO DE PLASTICO	18	AMBIENTE	NATURAL	7.25	7.00	7.50	8.00	7.00	7.00	10	10	10	7.75	**81.50**
M26	SANTO DOMINGO	N ANTONIO T	1800	6	4	TANQUE TINA	48	AMBIENTE	LAVADO	8.25	8.00	7.50	7.25	7.50	7.00	10	10	10	8.5	**84.00**
M27	SANTO DOMINGO	CATIMOR	2100	5	3	TANQUE TINA	36	CARPA SOLAR	NATURAL	7.50	7.50	7.75	7.50	7.50	7.50	10	10	10	7.75	**83.00**
M28	SAN JUAN DE CUTERVO	GEISHA	1800	3	2	TANQUE TINA	24	CARPA SOLAR	LAVADO	7.75	8.00	8.55	7.95	7.75	8.00	10	10	8	7	**83.00**
M29	SANTO DOMINGO	CATIMOR	1800	6	4	TEMPERATURA AMBIENTE	24	AMBIENTE	LAVADO	8.50	7.50	8.00	8.00	8.50	8.00	10	10	10	8	**86.50**
M30	SANTO TOMAS	MARSELLESA	1800	4	4	TANQUE TINA	14	SECADOR SOLAR	LAVADO	7.75	8.50	7.50	8.00	8.25	8.25	8	10	8	7.5	**81.75**
M31	CALLAYUC	CATURRA	1800	4	6	TANQUE TINA	16	CARPA SOLAR	LAVADO	7.50	7.00	7.75	7.50	7.50	7.50	10	10	10	7.95	**82.70**
M32	CALLAYUC	CATURRA	2000	6	4	TANQUE TINA	24	AMBIENTE	LAVADO	7.50	7.50	7.75	8.00	7.50	8.00	10	10	10	8.15	**84.40**
M33	CUTERVO	BORBON	2000	5	4	SACO DE PLASTICO	36	AMBIENTE	LAVADO	8.30	8.50	8.35	8.40	8.50	8.50	10	10	10	8.25	**88.80**
M34	CALLAYUC	RAN COLOMBI	1750	3	4	TANQUE TINA	17	AMBIENTE	NATURAL	8.50	8.00	8.25	8.00	7.75	8.00	10	10	10	7.85	**86.35**
M35	QUEROCOTILLO	CATIMOR	1800	6	5	TEMPERATURA AMBIENTE	14	SECADO EN SUELO	LAVADO	8.00	7.80	8.00	7.75	7.50	8.00	10	10	10	7.95	**85.00**
M36	SANTO TOMAS	CATURRA	1800	3	2	TANQUE TINA	16	CARPA SOLAR	LAVADO	8.00	7.50	8.00	8.50	7.00	7.75	10	10	10	7	**83.75**
M37	SAN ANDRES	CATUAY	2100	5	3	TANQUE TINA	24	CARPA SOLAR	LAVADO	8.00	8.00	7.50	7.50	8.00	8.00	10	10	10	8	**85.00**
M38	CALLAYUC	CATIMOR	1800	6	5	SACO DE PLASTICO	18	CARPA SOLAR	LAVADO	7.50	8.00	8.00	7.80	7.85	8.50	10	10	8	8	**83.65**
M39	CALLAYUC	MARSELLESA	1750	4	3	SACO DE PLASTICO	18	CARPA SOLAR	LAVADO	8.25	8.00	7.75	7.50	8.00	7.80	10	10	10	8.5	**85.80**
M40	SANTO TOMAS	CATIMOR	1800	7	5	TEMPERATURA AMBIENTE	24	AMBIENTE	LAVADO	8.00	7.50	7.25	8.00	7.50	8.25	10	10	10	7.5	**84.00**

Anexo A3.

Windrose para rueda de puntuación sensorial de café de las 40 muestras del estudio.



Anexo A4.

El presente anexo reúne todos los Waterfall Plots generados para cada atributo sensorial modelado mediante SHAP, con el fin de proporcionar una lectura detallada de cómo los valores específicos de cada observación modifican la predicción base del modelo. A diferencia de los gráficos principales incluidos en el cuerpo de la tesis, los cuales se centran en la importancia global de las características y en el comportamiento agregado de los efectos (Summary Plots), los Waterfall Plots permiten descomponer la predicción individual, mostrando la contribución positiva o negativa de cada variable en una muestra representativa.

Este tipo de visualización es especialmente útil para comprender la lógica

interna del modelo y verificar si sus patrones de decisión mantienen coherencia con los fundamentos agronómicos y sensoriales. Su inclusión en el anexo obedece a criterios de claridad y economía narrativa, sin sacrificar la profundidad técnica requerida para análisis avanzados de interpretabilidad.

Figura A4.1

Descomposición de la predicción individual para el aroma (SHAP Waterfall Plot)

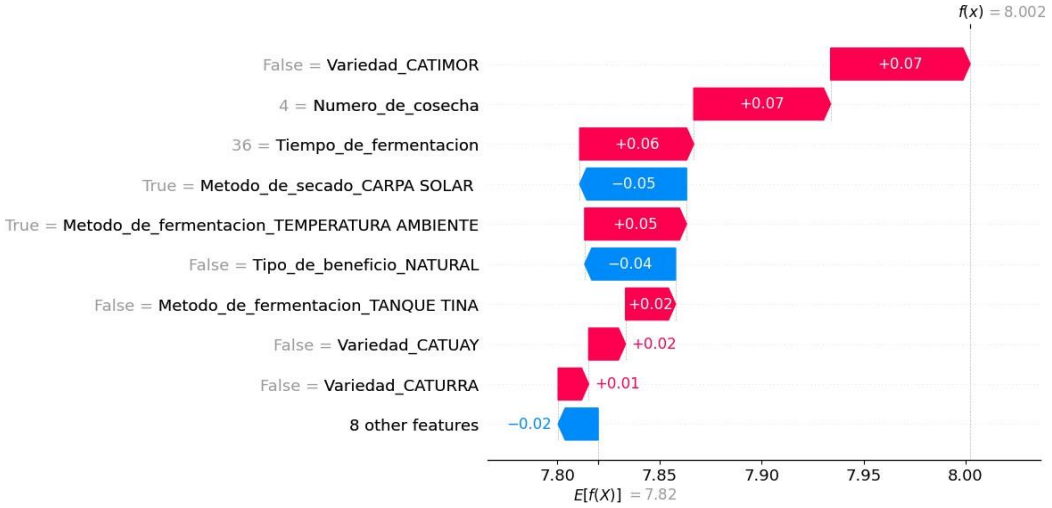


Figura A4.2

Descomposición de la predicción individual para el sabor (SHAP Waterfall Plot)

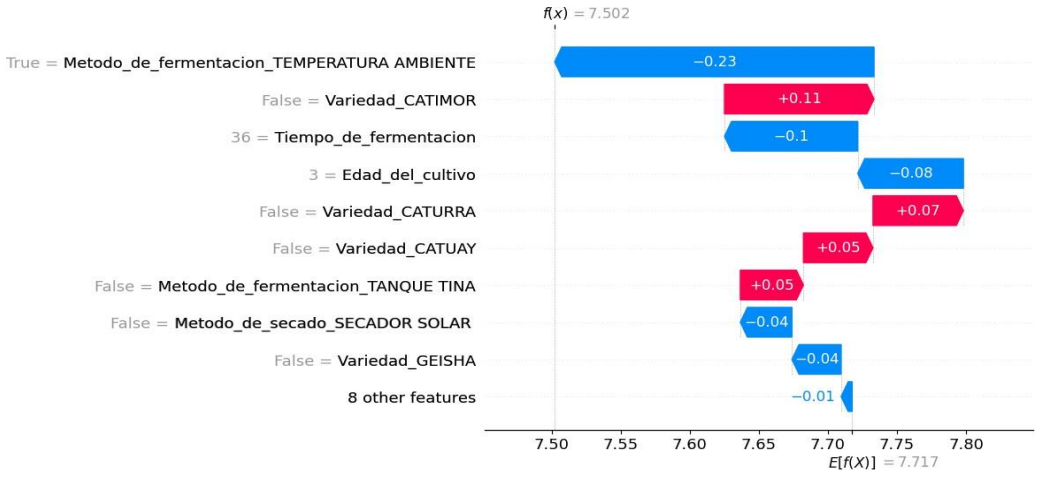


Figura A4.3

Descomposición de la predicción individual para el sabor residual (SHAP Waterfall Plot)

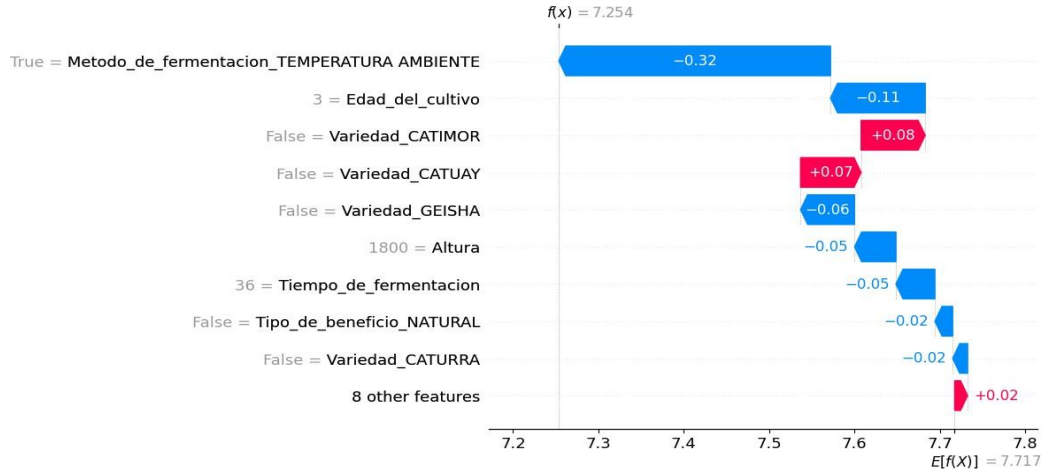


Figura A4.4

Descomposición de la predicción individual para la acidez (SHAP Waterfall Plot)

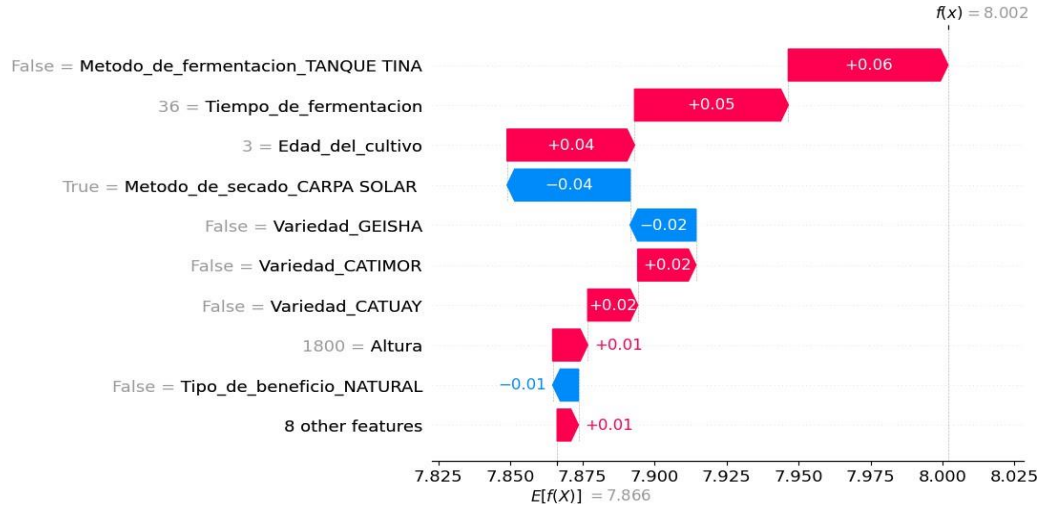


Figura A4.5

Descomposición de la predicción individual para el cuerpo (SHAP Waterfall Plot)

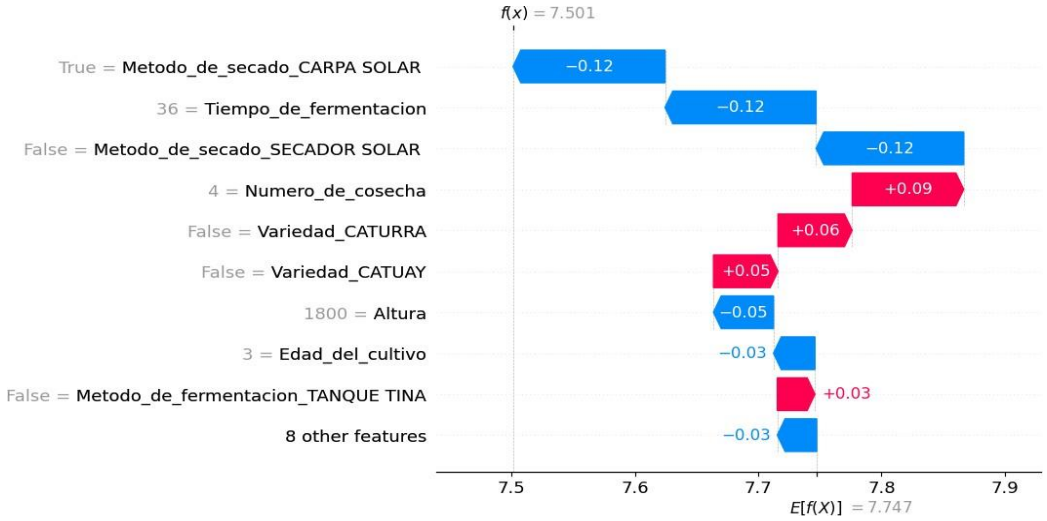


Figura A4.6

Descomposición de la predicción individual para el balance (SHAP Waterfall Plot)

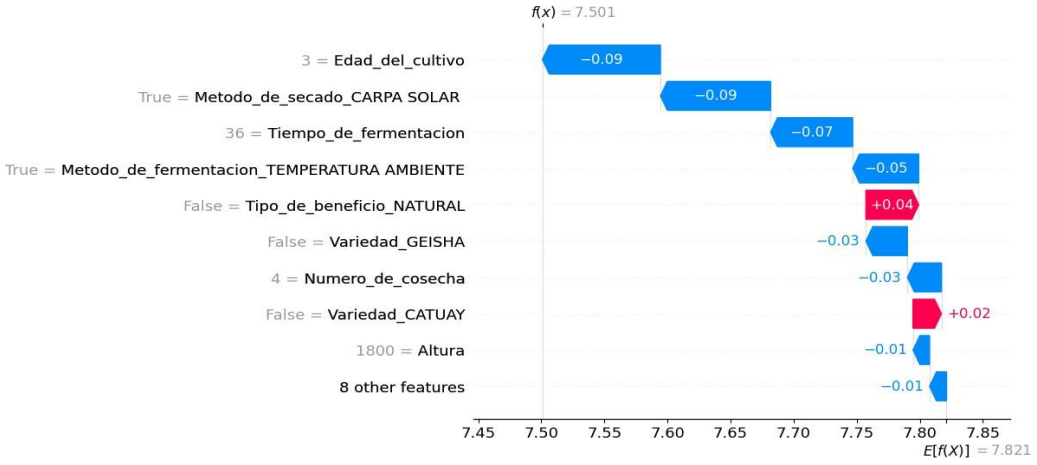
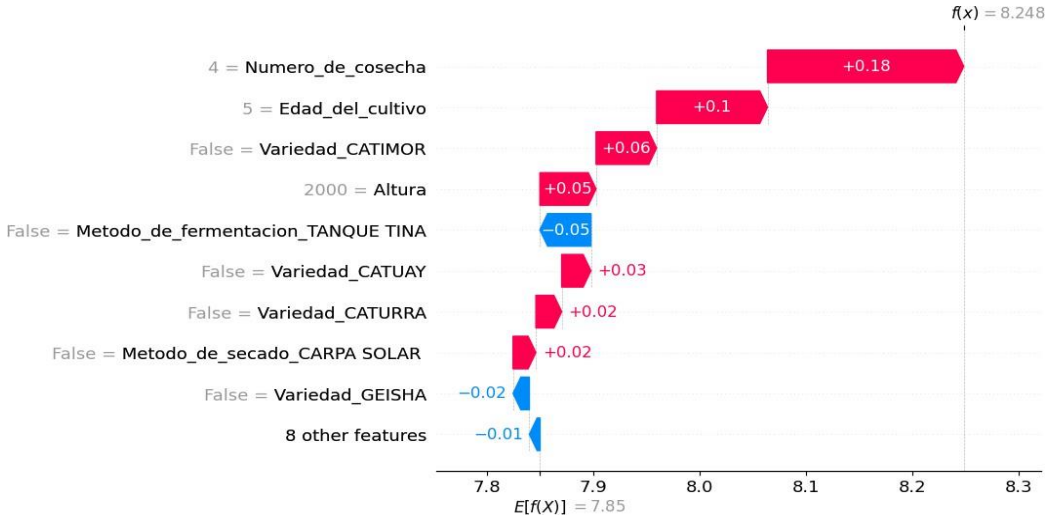


Figura A4.7

Descomposición de la predicción individual para la puntuación otorgada por el catador (SHAP Waterfall Plot)



Anexo A5.

Certificado de catación del catador profesional encargado de catar cafés especiales en el laboratorio de Ushun Café Cutervo.



Café & Cacao
Central Café & Cacao

CERTIFICADO

Otorgado a :

EVER BANCES DIAZ

La Central Café y Cacao del Perú otorga al titular el presente certificado, por haber aprobado satisfactoriamente los seis (06) cursos modulares del programa de formación Escuela de Catación de Cafés Especiales (ECAFE); logrando obtener la especialidad de "Catador de Cafés Especiales" con 240 horas de capacitación efectiva, realizado durante los meses de febrero a diciembre 2023.

Lugar | Lima - Perú

Lima, 05 de enero de 2024

Central Café & Cacao • Central Café & Cacao •

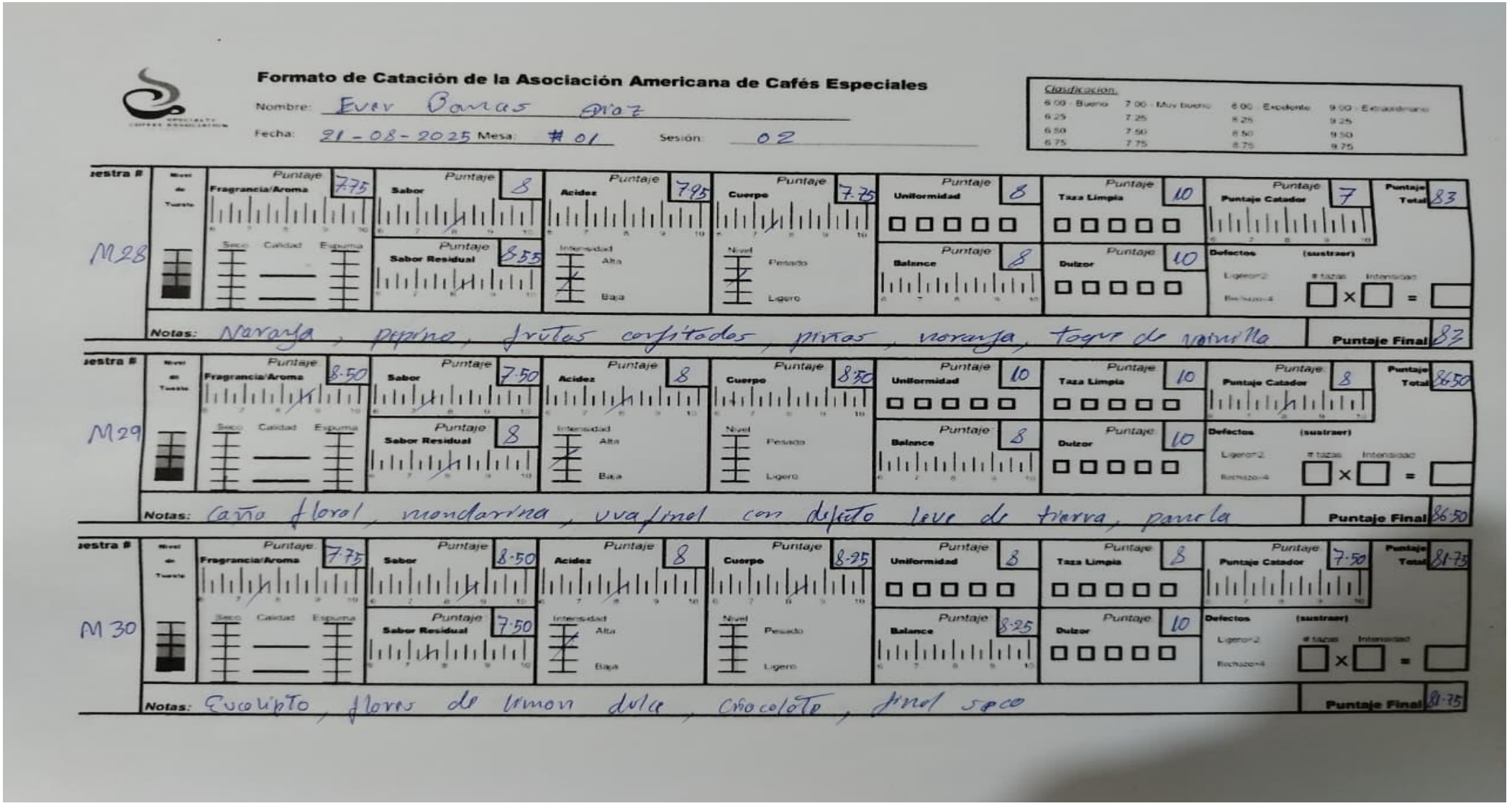
GENI FUNDES BULEJE
Gerente General
Central Café y Cacao del Perú

PREMIER
TRAINING CAMPUS

ECAFE

Anexo A6.

Fichas técnicas de catacion de la Specialty Coffee Association (SCA), emitidas por el catador con resultados sensoriales de cada una de las 40 muestras recolectadas en el estudio



Formato de Catación de la Asociación Americana de Cafés Especiales

Nombre: Ever Bamas Díaz

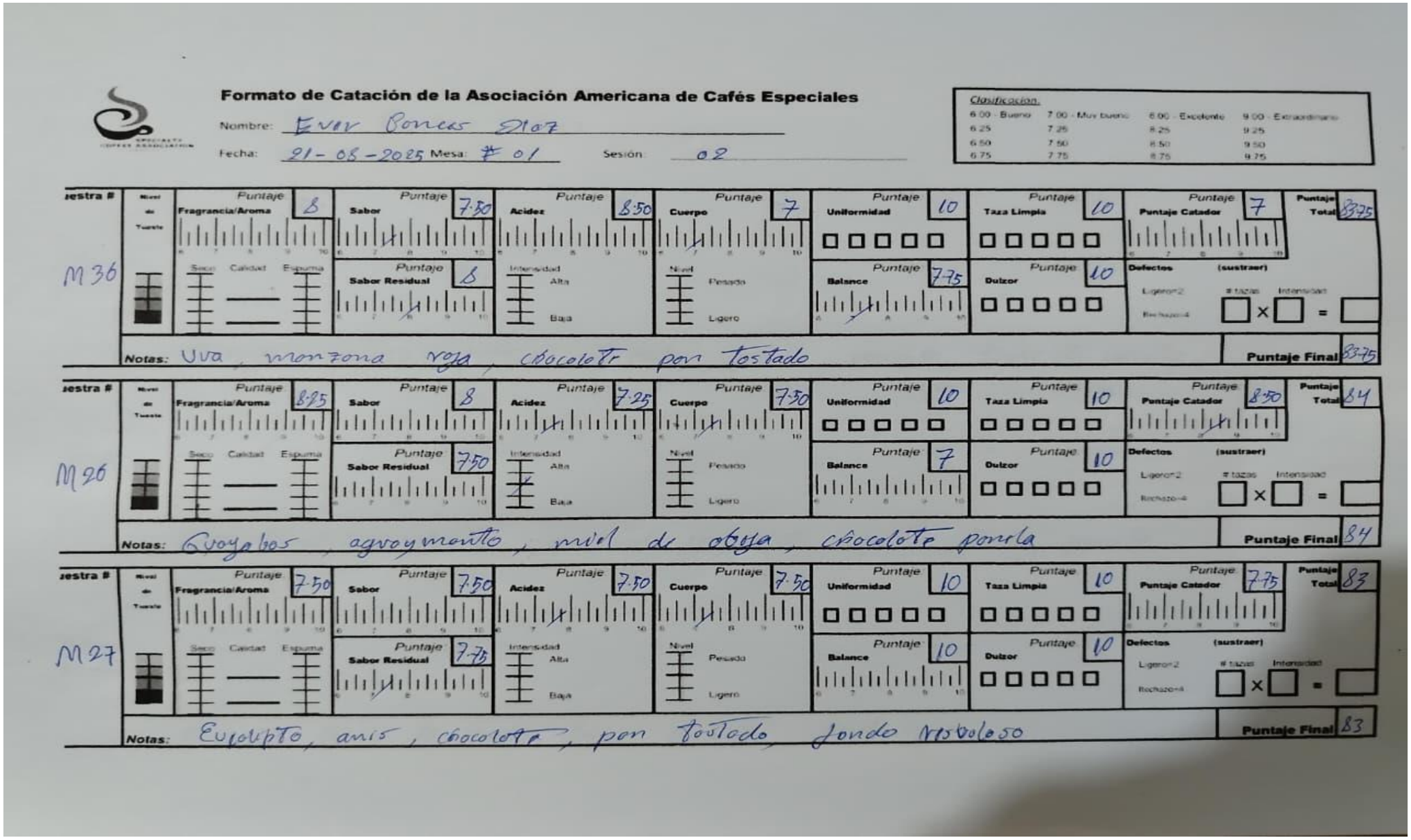
Fecha: 21-08-2025 Mesa: #01 Sesión: 02

Clasificacion:

6.00 - Bueno	7.00 - Muy bueno	8.00 - Excelente	9.00 - Extraordinario
6.25	7.25	8.25	9.25
6.50	7.50	8.50	9.50
6.75	7.75	8.75	9.75

Muestra #	Fragrancia/Aroma	Sabor	Acidez	Cuerpo	Uniformidad	Taza Limpia	Puntaje Catador	Puntaje Total	Sabor Residual	Balance	Dulzor	Defectos (sustraer)	Puntaje Final	Notas
M28	7.75	8	7.95	7.75	8	10	7	83	8.55	8	10		83	Naranja, pepino, frutos confitados, pinas, naranja, toque de vainilla
M29	8.50	7.50	8	8.50	10	10	8	86.50	8	8	10		86.50	Caña floral, mandarina, uva final con defecto leve de tierra, panela
M 30	7.75	8.50	8	8.25	8	8	7.50	81.75	7.50	8.25	10		81.75	Eucalipto, flores de limon dulce, chocolate, final seco

Anexo A6.1



Formato de Catación de la Asociación Americana de Cafés Especiales

Nombre: Evar Bonilla Díaz

Fecha: 21-08-2025 Mesa: # 01 Sesión: 02

Clasificación:

6.00 - Bueno	7.00 - Muy bueno	8.00 - Excelente	9.00 - Extraordinario
6.25	7.25	8.25	9.25
6.50	7.50	8.50	9.50
6.75	7.75	8.75	9.75

Muestra #	Fragancia/Aroma	Sabor	Acidez	Cuerpo	Uniformidad	Taza Limpia	Puntaje Catador	Puntaje Total
M36	8	7.50	8.50	7	10	10	7	83.75

Seco: — Calidad: — Espuma: —

Sabor Residual: 8 — Intensidad: Alta / Baja — Nivel: Pesado / Ligero

Balance: 7.75 — Dulzor: 10

Defectos (sustraer): Ligero=2, Rechazo=4 — # tazas × Intensidad =

Notas: Uva, manzana roja, chocolate por tostado

Puntaje Final: 83.75

Muestra #	Fragancia/Aroma	Sabor	Acidez	Cuerpo	Uniformidad	Taza Limpia	Puntaje Catador	Puntaje Total
M26	8.25	8	7.25	7.50	10	10	8.50	84

Seco: — Calidad: — Espuma: —

Sabor Residual: 7.50 — Intensidad: Alta / Baja — Nivel: Pesado / Ligero

Balance: 7 — Dulzor: 10

Defectos (sustraer): Ligero=2, Rechazo=4 — # tazas × Intensidad =

Notas: Guayabos, agroymanto, miel de abeja, chocolate ponla

Puntaje Final: 84

Muestra #	Fragancia/Aroma	Sabor	Acidez	Cuerpo	Uniformidad	Taza Limpia	Puntaje Catador	Puntaje Total
M27	7.50	7.50	7.50	7.50	10	10	7.75	83

Seco: — Calidad: — Espuma: —

Sabor Residual: 7.75 — Intensidad: Alta / Baja — Nivel: Pesado / Ligero

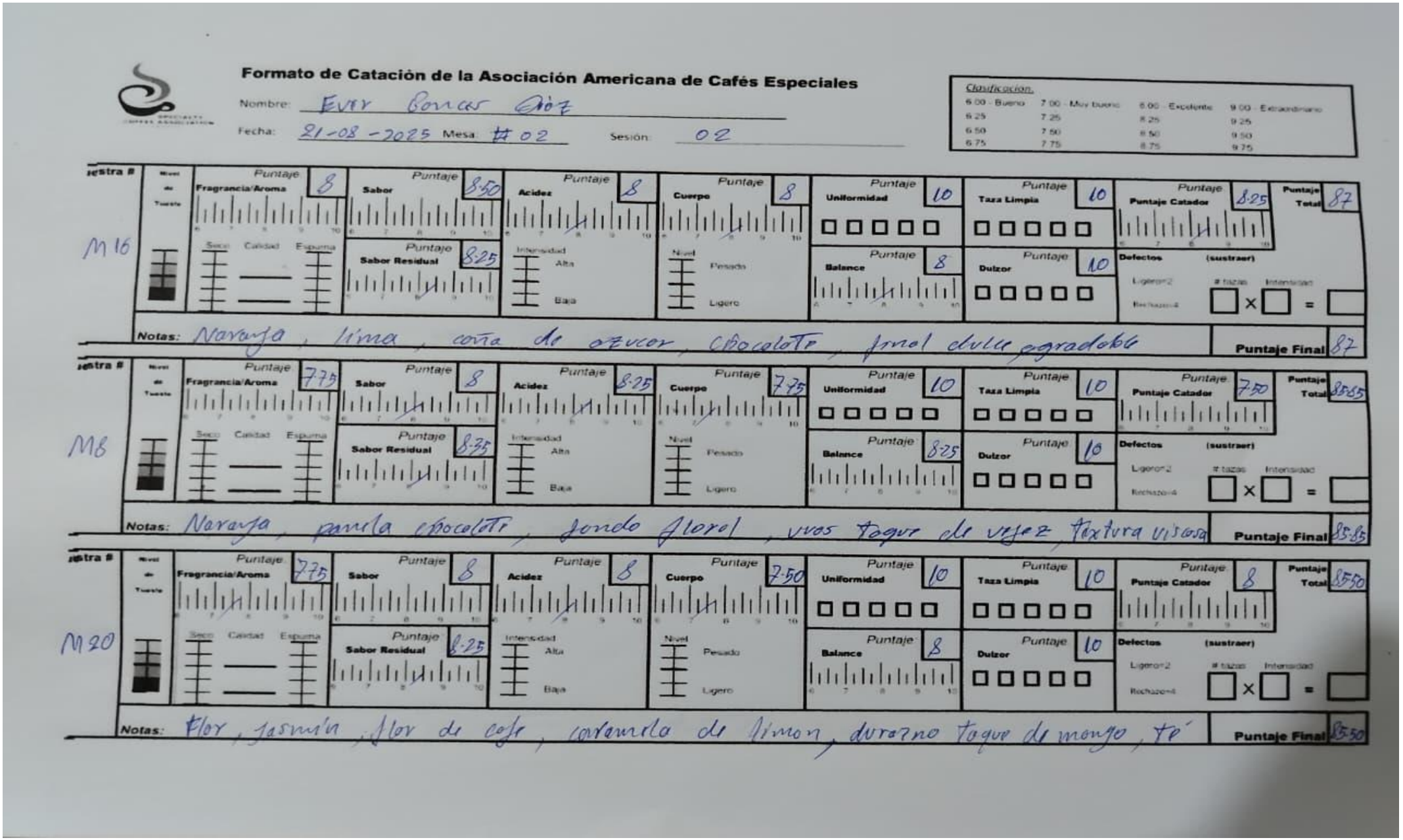
Balance: 10 — Dulzor: 10

Defectos (sustraer): Ligero=2, Rechazo=4 — # tazas × Intensidad =

Notas: Eucalipto, anís, chocolate, por tostado, fondo Mobloso

Puntaje Final: 83

Anexo A6.2



Formato de Catación de la Asociación Americana de Cafés Especiales

Nombre: Ever Gonzalez Díaz

Fecha: 21-08-2025 Mesa: # 02 Sesión: 02

Clasificación

6.00 - Bueno	7.00 - Muy bueno	8.00 - Excelente	9.00 - Extraordinario
6.25	7.25	8.25	9.25
6.50	7.50	8.50	9.50
6.75	7.75	8.75	9.75

Muestra # M16

Fragrancia/Aroma	Sabor	Sabor Residual	Acidez	Cuerpo	Uniformidad	Taza Limpia	Puntaje Catador	Balance	Dulzor	Puntaje Total
8	8.50	8.25	8	8	10	10	8.25	8	10	87

Seco: Calidad: Espuma: Intensidad: Alta / Baja: Nivel: Pesado / Ligero

Defectos (sustraer): Ligero=2 Rechazo=4 # tazas: Intensidad: =

Notas: Naranja, lima, caña de azucar, chocolate, floral dulce agradable

Puntaje Final: 87

Muestra # M8

Fragrancia/Aroma	Sabor	Sabor Residual	Acidez	Cuerpo	Uniformidad	Taza Limpia	Puntaje Catador	Balance	Dulzor	Puntaje Total
7.75	8	8.25	8.25	7.75	10	10	7.50	8.25	10	85.65

Seco: Calidad: Espuma: Intensidad: Alta / Baja: Nivel: Pesado / Ligero

Defectos (sustraer): Ligero=2 Rechazo=4 # tazas: Intensidad: =

Notas: Naranja, panela chocolate, fondo floral, vivos Yogur de vainilla, Textura viscosa

Puntaje Final: 85.65

Muestra # M20

Fragrancia/Aroma	Sabor	Sabor Residual	Acidez	Cuerpo	Uniformidad	Taza Limpia	Puntaje Catador	Balance	Dulzor	Puntaje Total
7.75	8	8.25	8	7.50	10	10	8	8	10	85.50

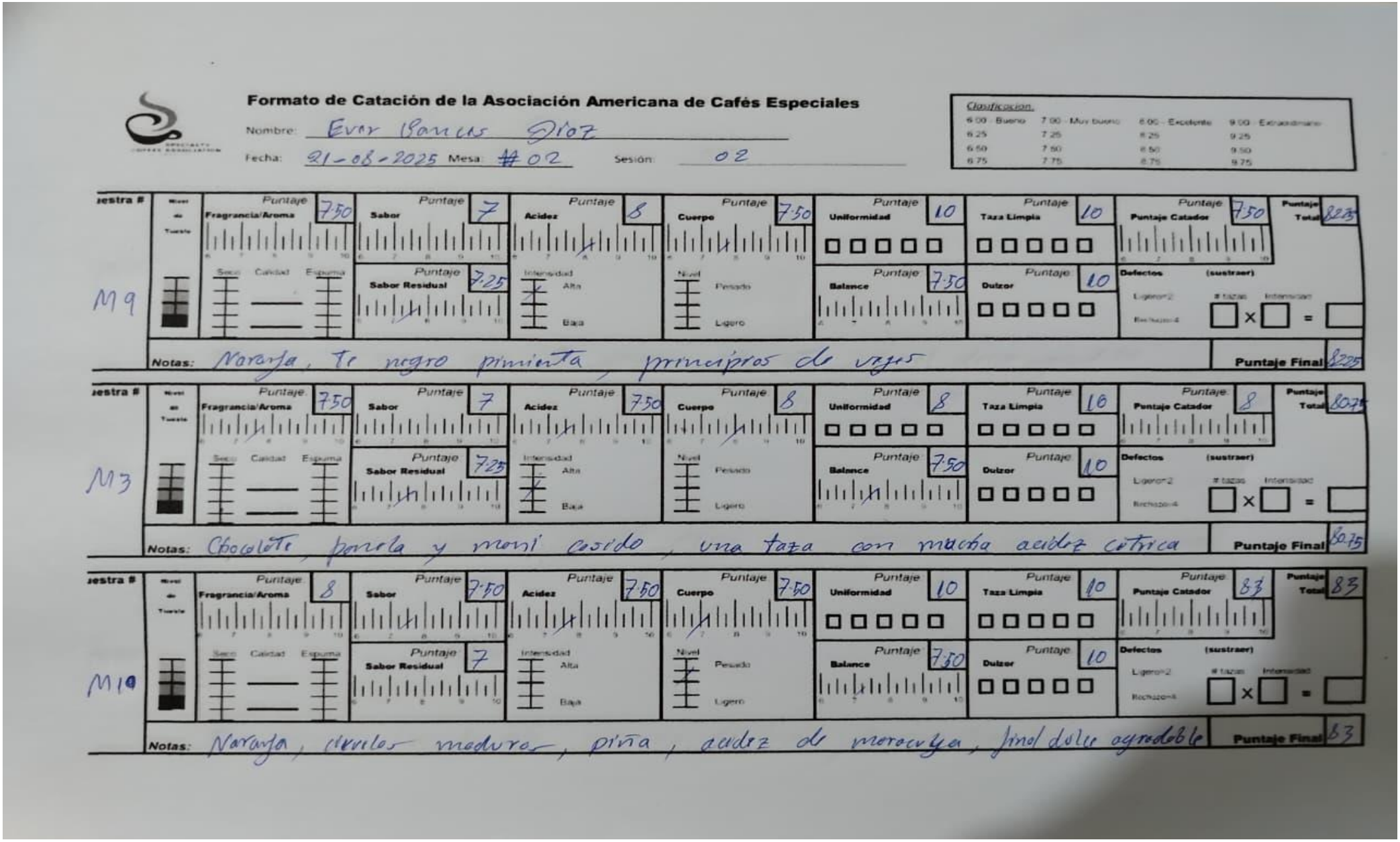
Seco: Calidad: Espuma: Intensidad: Alta / Baja: Nivel: Pesado / Ligero

Defectos (sustraer): Ligero=2 Rechazo=4 # tazas: Intensidad: =

Notas: Flor, jasmín, flor de cafe, caramelo de limon, durazno Yogur de mango, Té

Puntaje Final: 85.50

Anexo A6.3



Formato de Catación de la Asociación Americana de Cafés Especiales

Nombre: Evar Bances Díaz

Fecha: 21-08-2025 Mesa: #02 Sesión: 02

Clasificación:

6.00 Bueno	7.00 Muy bueno	8.00 Excelente	9.00 Extraordinario
6.25	7.25	8.25	9.25
6.50	7.50	8.50	9.50
6.75	7.75	8.75	9.75

Muestra # M9

Fragrancia/Aroma	Sabor	Acidez	Cuerpo	Uniformidad	Taza Limpia	Puntaje Catador	Puntaje Total
7.50	7	8	7.50	10	10	7.50	82.5

Sabor Residual: 7.25 — Balance: 7.50 — Dulzor: 10

Nivel de Tueste: Seco / Calidad / Espuma — Intensidad: Alta / Baja — Nivel: Pesado / Ligero

Defectos (sustraer): Ligero=2, Rechazo=4 — # tazas × Intensidad =

Notas: Naranja, Te negro pimienta, principios de vejes

Puntaje Final: 82.25

Muestra # M3

Fragrancia/Aroma	Sabor	Acidez	Cuerpo	Uniformidad	Taza Limpia	Puntaje Catador	Puntaje Total
7.50	7	7.50	8	8	10	8	80.75

Sabor Residual: 7.25 — Balance: 7.50 — Dulzor: 10

Nivel de Tueste: Seco / Calidad / Espuma — Intensidad: Alta / Baja — Nivel: Pesado / Ligero

Defectos (sustraer): Ligero=2, Rechazo=4 — # tazas × Intensidad =

Notas: Chocolate, panela y mosi cocido, una taza con mucha acidez cítrica

Puntaje Final: 80.75

Muestra # M10

Fragrancia/Aroma	Sabor	Acidez	Cuerpo	Uniformidad	Taza Limpia	Puntaje Catador	Puntaje Total
8	7.50	7.50	7.50	10	10	83	83

Sabor Residual: 7 — Balance: 7.50 — Dulzor: 10

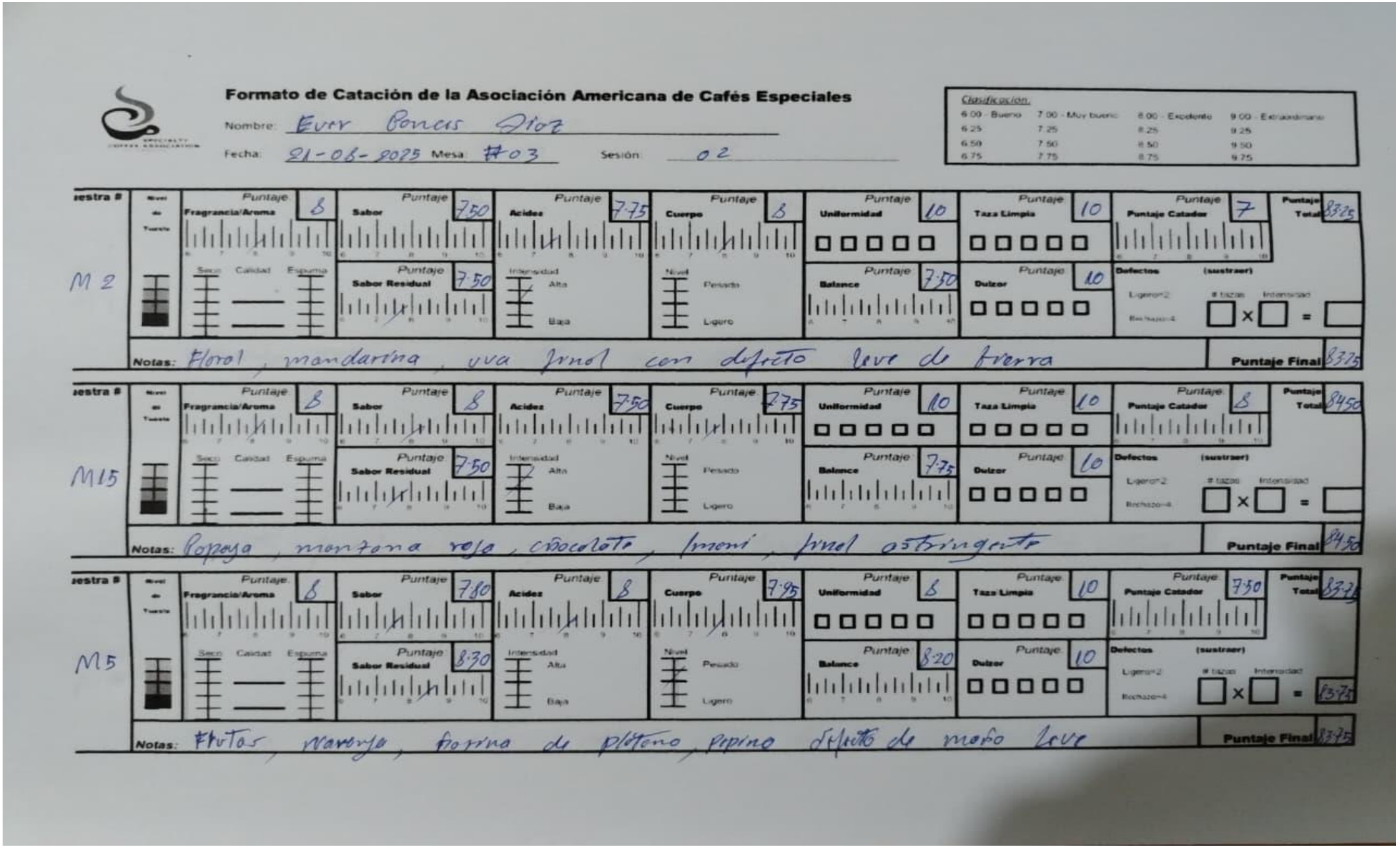
Nivel de Tueste: Seco / Calidad / Espuma — Intensidad: Alta / Baja — Nivel: Pesado / Ligero

Defectos (sustraer): Ligero=2, Rechazo=4 — # tazas × Intensidad =

Notas: Naranja, duraznos madura, piña, acidez de maracuya, final dulce agradable

Puntaje Final: 83

Anexo A6.4



Formato de Catación de la Asociación Americana de Cafés Especiales

Nombre: Ever Poncis Diaz

Fecha: 21-06-2025 Mesa: #03 Sesión: 02

Clasificación			
6.00 - Bueno	7.00 - Muy bueno	8.00 - Excelente	9.00 - Extraordinario
6.25	7.25	8.25	9.25
6.50	7.50	8.50	9.50
6.75	7.75	8.75	9.75

Muestra #	Fragrancia/Aroma	Sabor	Acidez	Cuerpo	Uniformidad	Taza Limpia	Puntaje Catador	Puntaje Total
M2	8	7.50	7.75	8	10	10	7	83.25
M15	8	8	7.50	7.75	10	10	8	84.50
M5	8	7.80	8	7.95	8	10	7.50	83.75

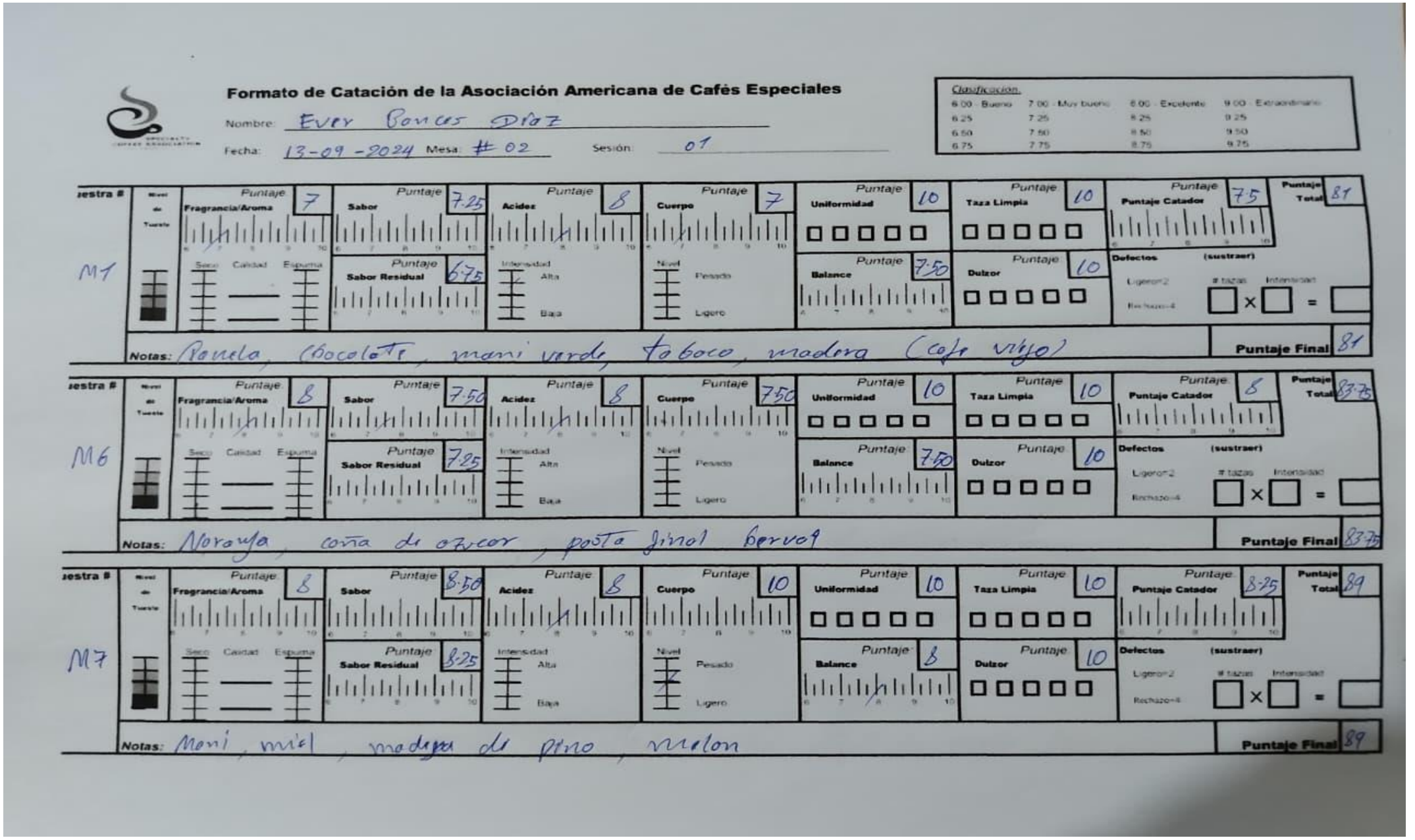
Muestra #	Sabor Residual	Balance	Dulzor	Defectos (sustraer)	Puntaje Final
M2	7.50	7.50	10	0	83.25
M15	7.50	7.75	10	0	84.50
M5	8.30	8.20	10	3.75	83.75

Notas: M2: Floral, mandarina, uva, miel con defecto leve de tierra

Notas: M15: Papaya, manzana roja, chocolate, limón, miel astringente

Notas: M5: Frutos, naranja, harina de plátano, pepino defecto de moho leve

Anexo A6.5



Formato de Catación de la Asociación Americana de Cafés Especiales

Nombre: Ever Bonilla Díaz

Fecha: 13-09-2024 Mesa: # 02 Sesión: 01

Clasificación:			
6.00 - Bueno	7.00 - Muy bueno	8.00 - Excelente	9.00 - Extraordinario
6.25	7.25	8.25	9.25
6.50	7.50	8.50	9.50
6.75	7.75	8.75	9.75

Muestra #	Fragancia/Aroma	Sabor	Acidez	Cuerpo	Uniformidad	Taza Limpia	Puntaje Catador	Puntaje Total
M1	7	7.25	8	7	10	10	7.5	81
M6	8	7.50	8	7.50	10	10	8	83.75
M7	8	8.50	8	10	10	10	8.25	89

Muestra #	Sabor Residual	Balance	Dulzor	Defectos (sustraer)	Puntaje Final
M1	6.75	7.50	10		81
M6	7.25	7.50	10		83.75
M7	8.25	8	10		89

Seco / Calidad / Espuma — Intensidad: Alta / Baja — Nivel: Pesado / Ligero

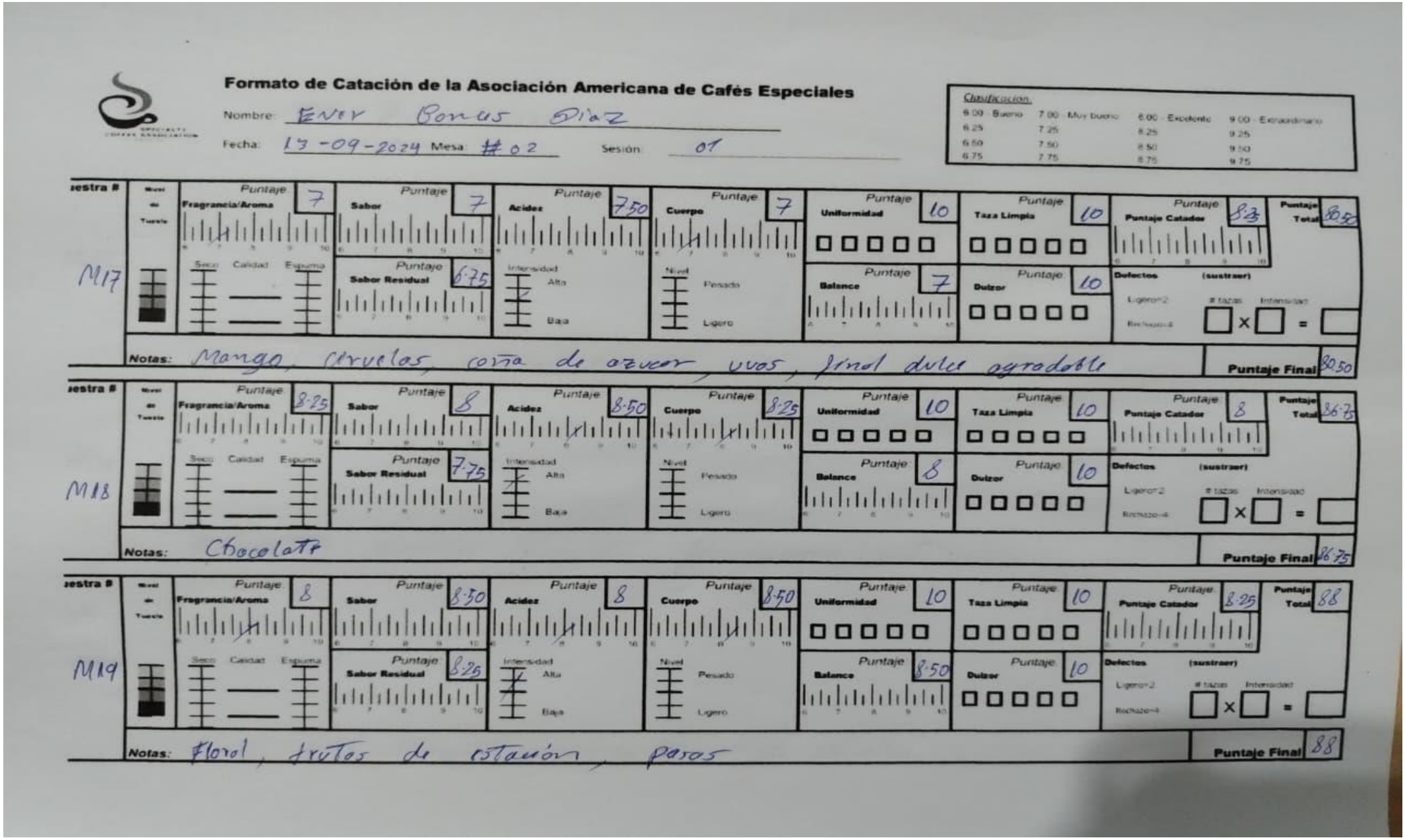
Defectos (sustraer): Ligero=2, Rechazo=4; # tazas × Intensidad =

Notas: Panela, chocolate, maní verde, tabaco, madera (café viejo)

Notas: Naranja, caña de azúcar, postre final barroso

Notas: Maní, miel, madera de pino, melón

Anexo A6.6



Formato de Catación de la Asociación Americana de Cafés Especiales

Nombre: ENER Bonus Díaz

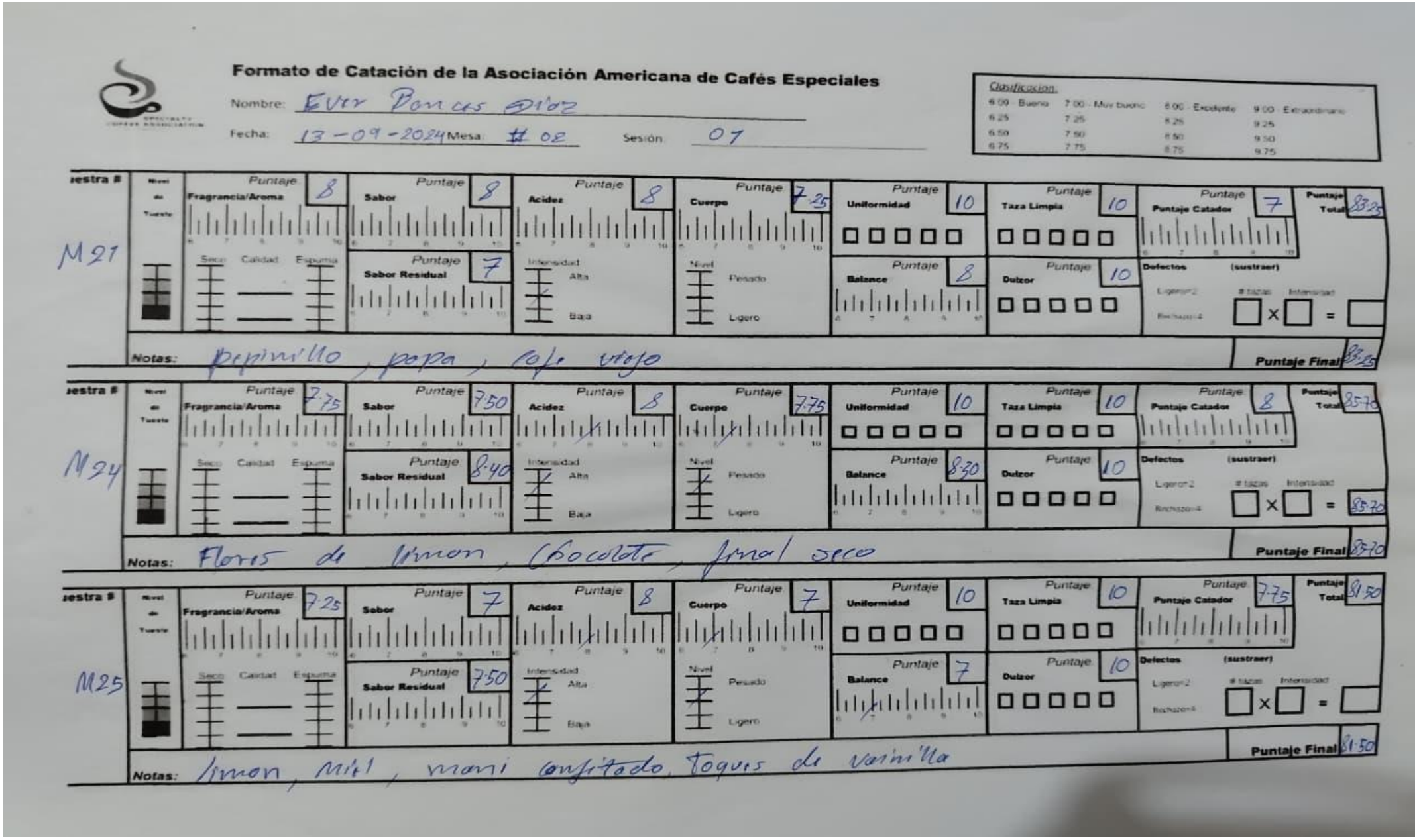
Fecha: 13-09-2024 Mesa: #02 Sesión: 01

Clasificación:

6.00 Bueno	7.00 Muy bueno	8.00 Excelente	9.00 Extraordinario
6.25	7.25	8.25	9.25
6.50	7.50	8.50	9.50
6.75	7.75	8.75	9.75

Muestra #	Fragancia/Aroma	Sabor	Acidez	Cuerpo	Uniformidad	Taza Limpia	Puntaje Catador	Puntaje Total	Sabor Residual	Balance	Dulzor	Defectos (sustraer)	Puntaje Final	Notas
M17	7	7	7.50	7	10	10	8.25	80.50	6.75	7	10		80.50	Manga, ciruelas, caña de azucar, uvas, final dulce agradable
M18	8.25	8	8.50	8.25	10	10	8	86.75	7.75	8	10		86.75	Chocolate
M19	8	8.50	8	8.50	10	10	8.25	88	8.25	8.50	10		88	Floral, frutos de estación, pasas

Anexo A6.7



Formato de Catación de la Asociación Americana de Cafés Especiales

Nombre: Evar Poncas Diaz

Fecha: 13-09-2024 Mesa: # 02 Sesión: 07

Clasificación:

6.00 Bueno	7.00 Muy bueno	8.00 Excelente	9.00 Extraordinario
6.25	7.25	8.25	9.25
6.50	7.50	8.50	9.50
6.75	7.75	8.75	9.75

Muestra # M21

Nivel de Tueste

Fragrancia/Aroma	8	Sabor	8	Acidez	8	Cuerpo	7.25	Uniformidad	10
Taza Limpia	10	Puntaje Catador	7	Puntaje Total	83.25				
Sabor Residual	7	Balance	8	Dulzor	10				

Seco: Calidad: Espuma:

Intensidad: Alta / Baja

Nivel: Pesado / Ligero

Defectos (sustraer): Ligero=2, Rechazo=4 — # tazas x Intensidad =

Notas: pepinillo, papa, café viejo

Puntaje Final: 83.25

Muestra # M24

Nivel de Tueste

Fragrancia/Aroma	7.75	Sabor	7.50	Acidez	8	Cuerpo	7.75	Uniformidad	10
Taza Limpia	10	Puntaje Catador	8	Puntaje Total	85.70				
Sabor Residual	8.40	Balance	8.30	Dulzor	10				

Seco: Calidad: Espuma:

Intensidad: Alta / Baja

Nivel: Pesado / Ligero

Defectos (sustraer): Ligero=2, Rechazo=4 — # tazas x Intensidad = 85.70

Notas: Flores de limon, chocolate, final seco

Puntaje Final: 85.70

Muestra # M25

Nivel de Tueste

Fragrancia/Aroma	7.25	Sabor	7	Acidez	8	Cuerpo	7	Uniformidad	10
Taza Limpia	10	Puntaje Catador	7.75	Puntaje Total	81.50				
Sabor Residual	7.50	Balance	7	Dulzor	10				

Seco: Calidad: Espuma:

Intensidad: Alta / Baja

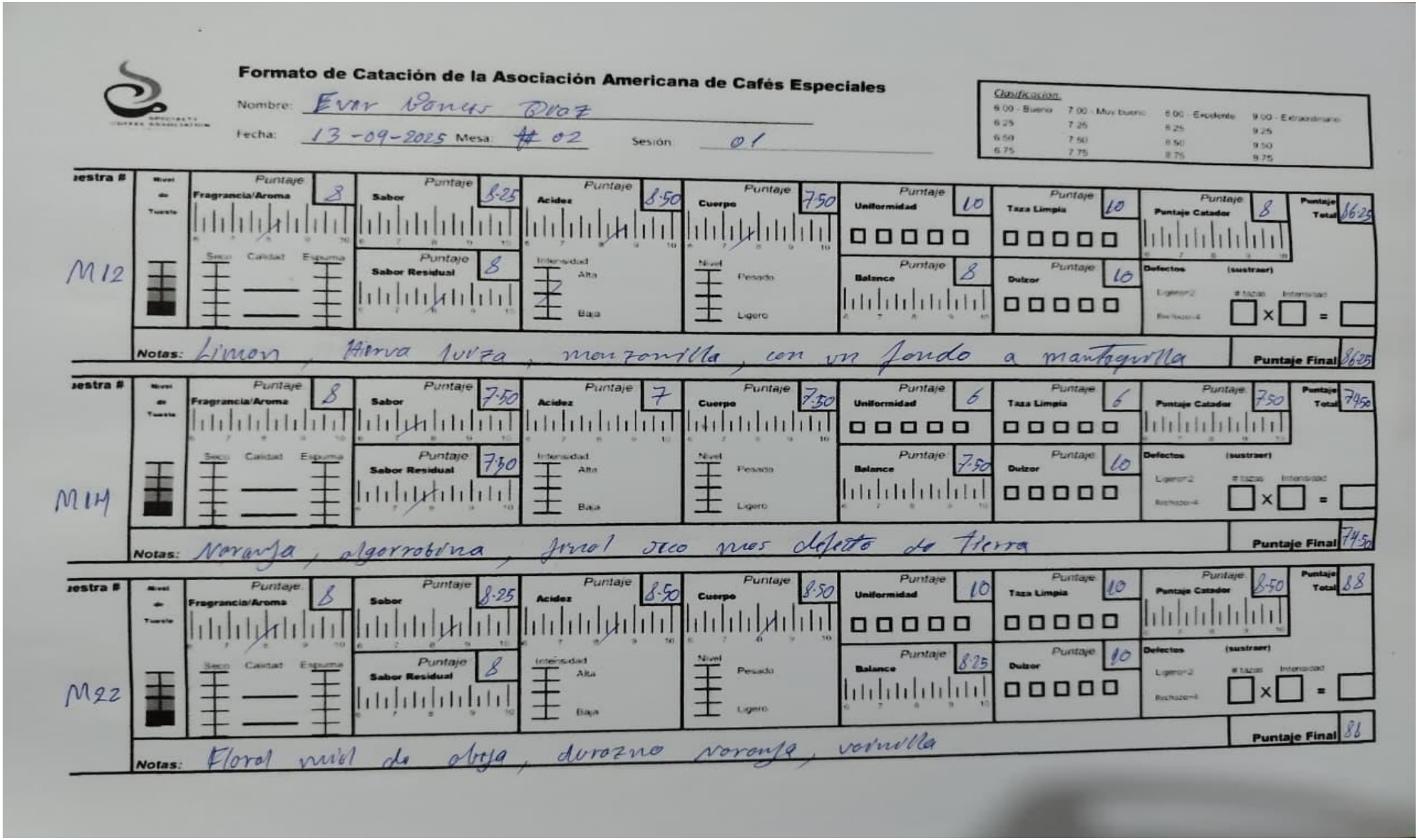
Nivel: Pesado / Ligero

Defectos (sustraer): Ligero=2, Rechazo=4 — # tazas x Intensidad =

Notas: limon, miel, mani confitado, toques de vainilla

Puntaje Final: 81.50

Anexo A6.8



Formato de Catación de la Asociación Americana de Cafés Especiales

Nombre: Evar Danus Diaz

Fecha: 13-09-2025 Mesa: # 02 Sesión: 01

Clasificación:

6.00 - Bueno	7.00 - Muy bueno	8.00 - Excelente	9.00 - Extraordinario
6.25	7.25	8.25	9.25
6.50	7.50	8.50	9.50
6.75	7.75	8.75	9.75

Muestra #	Nivel de Tueste	Fragrancia/Aroma	Sabor	Acidez	Cuerpo	Uniformidad	Taza Limpia	Puntaje Catador	Puntaje Total	Sabor Residual	Balance	Dulzor	Defectos (sustraer)	Puntaje Final
M12	Seco Calidad Espuma	8	8.25	8.50	7.50	10	10	8	86.25	8	8	10	Ligero=2 Rechazo=4 # tazas Intensidad	86.25
				Intensidad: Alta / Baja	Nivel: Pesado / Ligero									

Notas: Limon, Hierva luiza, manzanilla, con un fondo a mantequilla

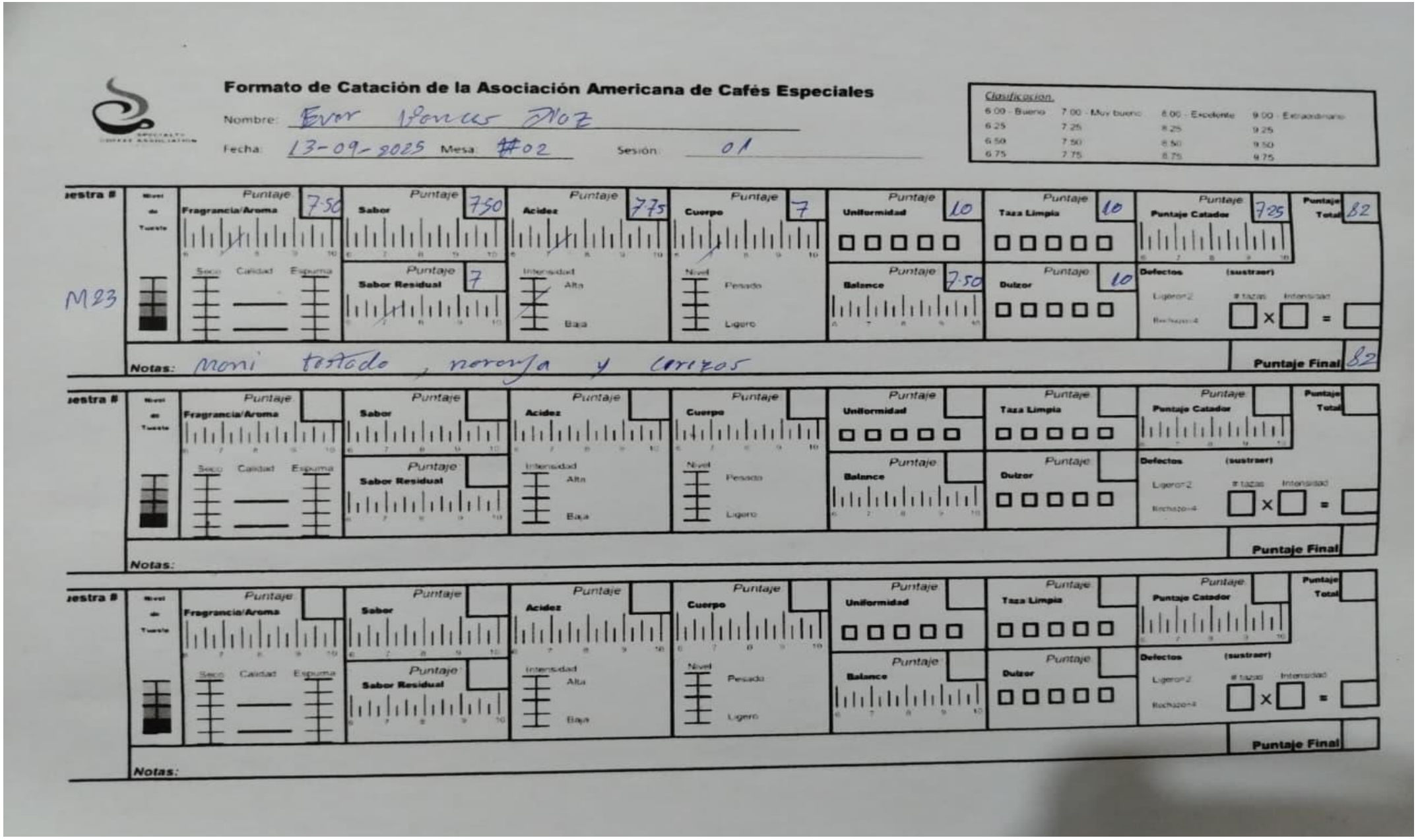
Muestra #	Nivel de Tueste	Fragrancia/Aroma	Sabor	Acidez	Cuerpo	Uniformidad	Taza Limpia	Puntaje Catador	Puntaje Total	Sabor Residual	Balance	Dulzor	Defectos (sustraer)	Puntaje Final
M14	Seco Calidad Espuma	8	7.50	7	7.50	6	6	7.50	74.50	7.50	7.50	10	Ligero=2 Rechazo=4 # tazas Intensidad	74.50
				Intensidad: Alta / Baja	Nivel: Pesado / Ligero									

Notas: Naranja, algorrobina, fruto seco mas defecto de tierra

Muestra #	Nivel de Tueste	Fragrancia/Aroma	Sabor	Acidez	Cuerpo	Uniformidad	Taza Limpia	Puntaje Catador	Puntaje Total	Sabor Residual	Balance	Dulzor	Defectos (sustraer)	Puntaje Final
M22	Seco Calidad Espuma	8	8.25	8.50	8.50	10	10	8.50	88	8	8.25	10	Ligero=2 Rechazo=4 # tazas Intensidad	88
				Intensidad: Alta / Baja	Nivel: Pesado / Ligero									

Notas: Floral miel de abeja, durazno naranja, vainilla

Anexo A6.9



Formato de Catación de la Asociación Americana de Cafés Especiales

Nombre: Evar Moncus Diaz

Fecha: 13-09-2025 Mesa: #02 Sesión: 01

Clasificación			
6.00 - Bueno	7.00 - Muy bueno	8.00 - Excelente	9.00 - Extraordinario
6.25	7.25	8.25	9.25
6.50	7.50	8.50	9.50
6.75	7.75	8.75	9.75

Muestra # M23

Nivel de Tueste

Fragrancia/Aroma	Sabor	Acidez	Cuerpo	Uniformidad	Taza Limpia	Puntaje Catador	Puntaje Total
7.50	7.50	7.75	7	10	10	7.25	82

Seco: Calidad: Espuma:

Sabor Residual	Intensidad	Nivel	Balance	Dulzor
7	Alta / Baja	Pesado / Ligero	7.50	10

Defectos (sustraer): Ligero=2, Rechazo=4 — # tazas x Intensidad =

Notas: Moni tostado, naranja y cerezos

Puntaje Final: 82

Muestra #

Nivel de Tueste

Fragrancia/Aroma	Sabor	Acidez	Cuerpo	Uniformidad	Taza Limpia	Puntaje Catador	Puntaje Total

Seco: Calidad: Espuma:

Sabor Residual	Intensidad	Nivel	Balance	Dulzor
	Alta / Baja	Pesado / Ligero		

Defectos (sustraer): Ligero=2, Rechazo=4 — # tazas x Intensidad =

Notas:

Puntaje Final

Muestra #

Nivel de Tueste

Fragrancia/Aroma	Sabor	Acidez	Cuerpo	Uniformidad	Taza Limpia	Puntaje Catador	Puntaje Total

Seco: Calidad: Espuma:

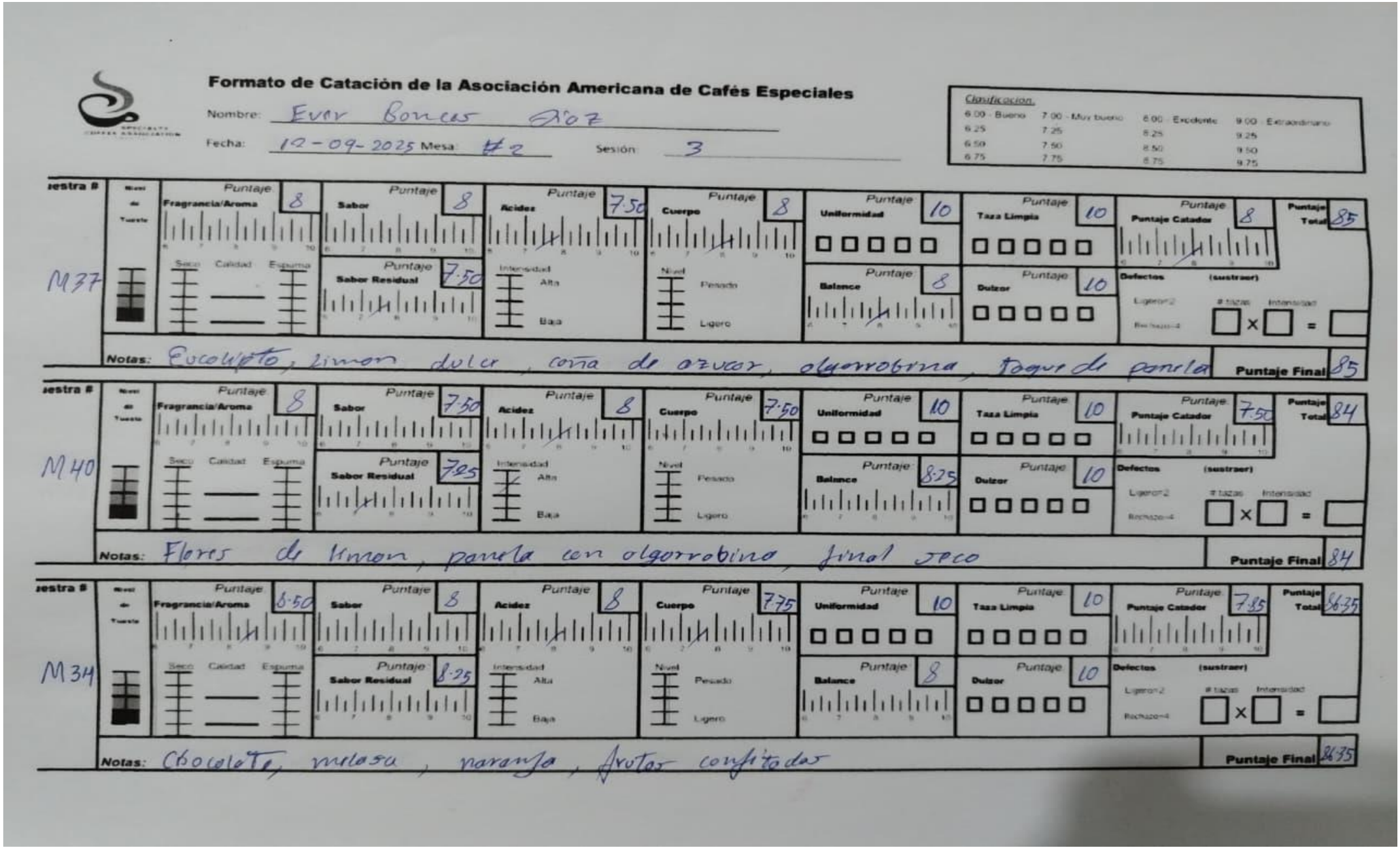
Sabor Residual	Intensidad	Nivel	Balance	Dulzor
	Alta / Baja	Pesado / Ligero		

Defectos (sustraer): Ligero=2, Rechazo=4 — # tazas x Intensidad =

Notas:

Puntaje Final

Anexo A6.10



Formato de Catación de la Asociación Americana de Cafés Especiales

Nombre: Evar Bonces Díaz

Fecha: 12-09-2025 Mesa: #2 Sesión: 3

Clasificación

6.00 - Bueno	7.00 - Muy bueno	8.00 - Excelente	9.00 - Extraordinario
6.25	7.25	8.25	9.25
6.50	7.50	8.50	9.50
6.75	7.75	8.75	9.75

Muestra # M37

Nivel de Tueste

Fragancia/Aroma: 8 — Seco, Calidad, Espuma

Sabor: 8

Sabor Residual: 7.50

Acidez: 7.50 — Intensidad: Alta / Baja

Cuerpo: 8 — Nivel: Pesado / Ligero

Uniformidad: 10

Taza Limpia: 10

Balance: 8

Dulzor: 10

Puntaje Catador: 8

Puntaje Total: 85

Defectos (sustraer): Ligero=2, Rechazo=4 — # tazas x Intensidad =

Notas: Eucalipto, limon dulce, caña de azucar, algarrobina, yogur de panela

Puntaje Final: 85

Muestra # M40

Nivel de Tueste

Fragancia/Aroma: 8 — Seco, Calidad, Espuma

Sabor: 7.50

Sabor Residual: 7.25

Acidez: 8 — Intensidad: Alta / Baja

Cuerpo: 7.50 — Nivel: Pesado / Ligero

Uniformidad: 10

Taza Limpia: 10

Balance: 8.25

Dulzor: 10

Puntaje Catador: 7.50

Puntaje Total: 84

Defectos (sustraer): Ligero=2, Rechazo=4 — # tazas x Intensidad =

Notas: Flores de limon, panela con algarrobina, final seco

Puntaje Final: 84

Muestra # M34

Nivel de Tueste

Fragancia/Aroma: 8.50 — Seco, Calidad, Espuma

Sabor: 8

Sabor Residual: 8.25

Acidez: 8 — Intensidad: Alta / Baja

Cuerpo: 7.75 — Nivel: Pesado / Ligero

Uniformidad: 10

Taza Limpia: 10

Balance: 8

Dulzor: 10

Puntaje Catador: 7.85

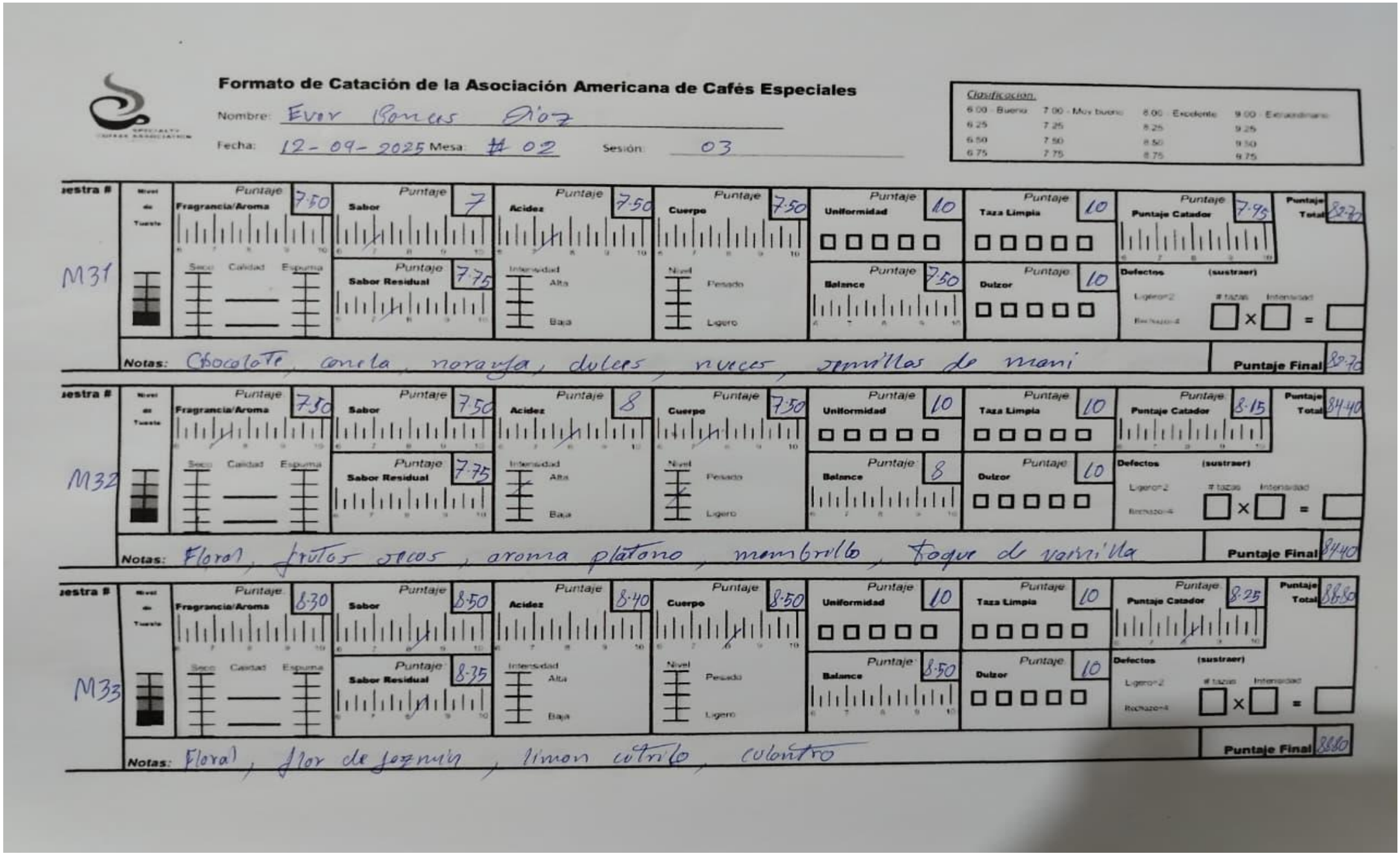
Puntaje Total: 86.35

Defectos (sustraer): Ligero=2, Rechazo=4 — # tazas x Intensidad =

Notas: Chocolate, melaza, naranja, frutos confitados

Puntaje Final: 86.35

Anexo A6.11



Formato de Catación de la Asociación Americana de Cafés Especiales

Nombre: Evar Bonilla Díaz

Fecha: 12-09-2025 Mesa: # 02 Sesión: 03

Clasificación			
6.00 - Bueno	7.00 - Muy bueno	8.00 - Excelente	9.00 - Extraordinario
6.25	7.25	8.25	9.25
6.50	7.50	8.50	9.50
6.75	7.75	8.75	9.75

Muestra # M31

Fragrancia/Aroma	Sabor	Acidez	Cuerpo	Uniformidad	Taza Limpia	Puntaje Catador	Puntaje Total
7.50	7	7.50	7.50	10	10	7.95	82.70

Sabor Residual: 7.75 — Balance: 7.50 — Dulzor: 10

Seco: Calidad: Espuma: — Intensidad: Alta / Baja — Nivel: Pesado / Ligero

Defectos (sustraer): Ligero=2, Rechazo=4 — # tazas: ___ x Intensidad: ___ = ___

Notas: Chocolate, canela, naranja, dulces, nueces, semillas de maní

Puntaje Final 82.70

Muestra # M32

Fragrancia/Aroma	Sabor	Acidez	Cuerpo	Uniformidad	Taza Limpia	Puntaje Catador	Puntaje Total
7.50	7.50	8	7.50	10	10	8.15	84.40

Sabor Residual: 7.75 — Balance: 8 — Dulzor: 10

Seco: Calidad: Espuma: — Intensidad: Alta / Baja — Nivel: Pesado / Ligero

Defectos (sustraer): Ligero=2, Rechazo=4 — # tazas: ___ x Intensidad: ___ = ___

Notas: Floral, frutos secos, aroma plátano, membrillo, toque de vainilla

Puntaje Final 84.40

Muestra # M33

Fragrancia/Aroma	Sabor	Acidez	Cuerpo	Uniformidad	Taza Limpia	Puntaje Catador	Puntaje Total
8.30	8.50	8.40	8.50	10	10	8.25	88.80

Sabor Residual: 8.35 — Balance: 8.50 — Dulzor: 10

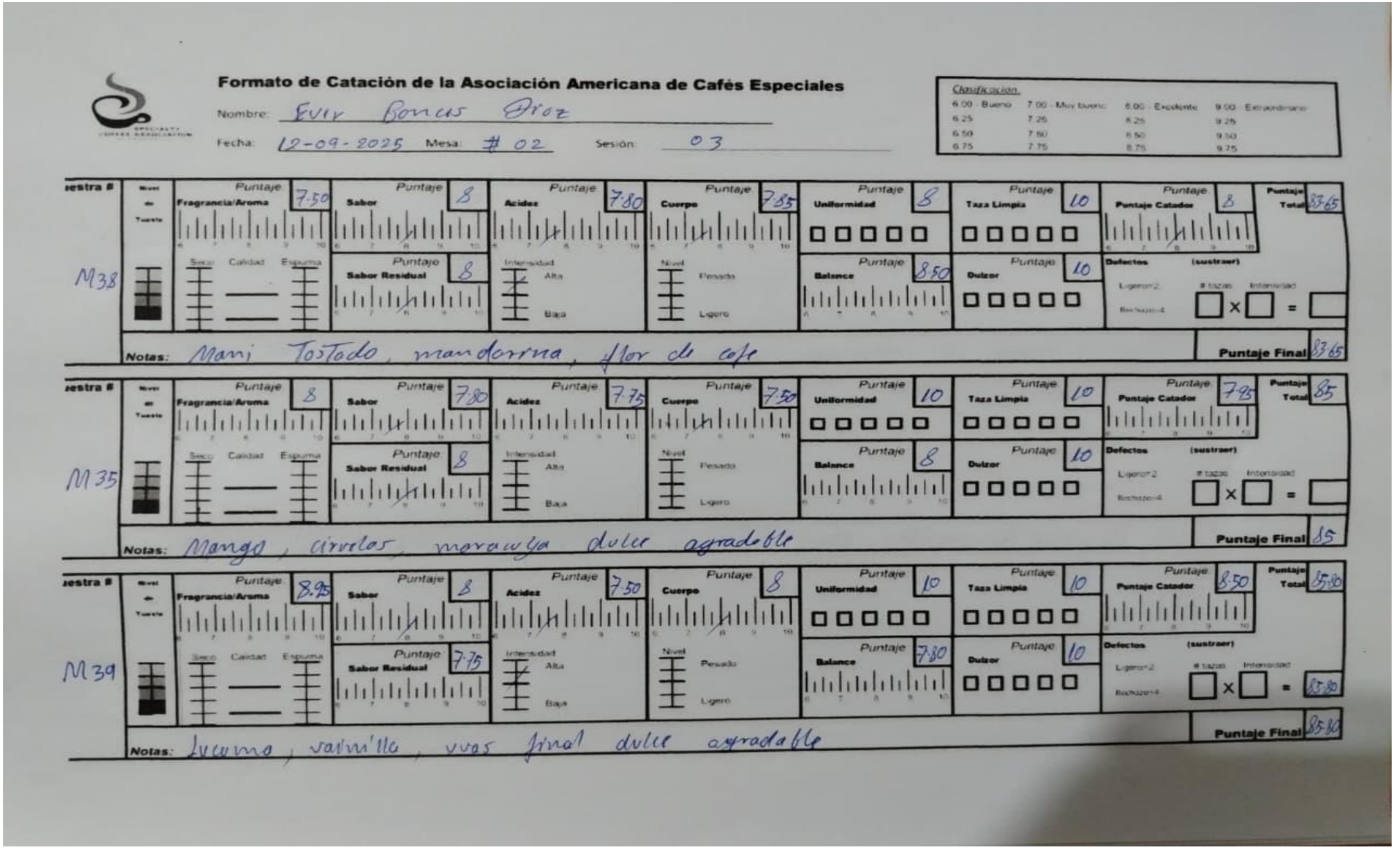
Seco: Calidad: Espuma: — Intensidad: Alta / Baja — Nivel: Pesado / Ligero

Defectos (sustraer): Ligero=2, Rechazo=4 — # tazas: ___ x Intensidad: ___ = ___

Notas: Floral, flor de jazmín, limón cítrico, culantro

Puntaje Final 88.80

Anexo A6.12



Formato de Catación de la Asociación Americana de Cafés Especiales

Nombre: Evily Bonilla Díaz

Fecha: 12-09-2025 Mesa: # 02 Sesión: 03

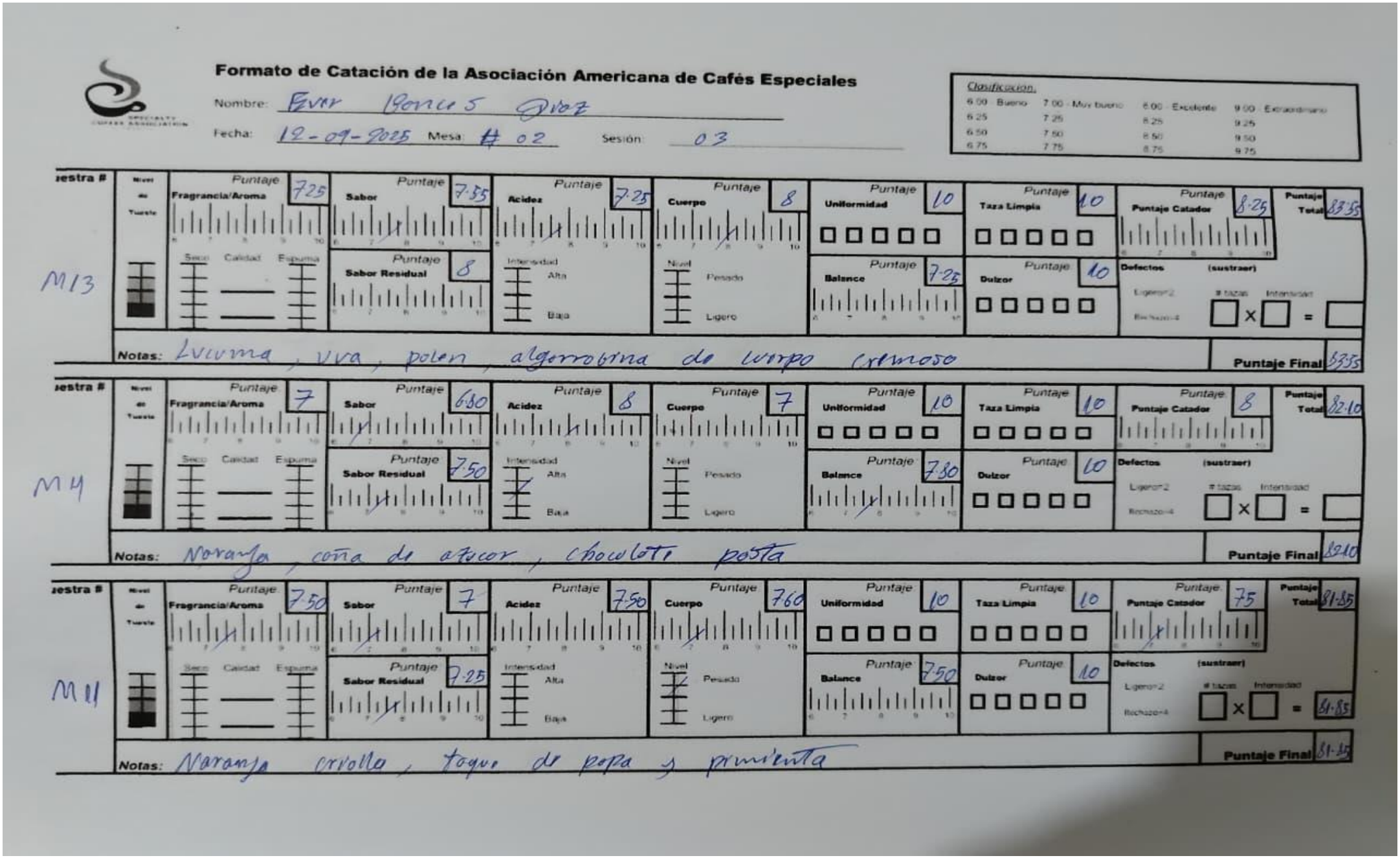
Clasificación			
6.00 - Bueno	7.00 - Muy bueno	8.00 - Excelente	9.00 - Extraordinario
6.25	7.25	8.25	9.25
6.50	7.50	8.50	9.50
6.75	7.75	8.75	9.75

Muestra #	Nivel de Tueste	Fragrancia/Aroma	Sabor	Acidez	Cuerpo	Uniformidad	Taza Limpia	Puntaje Catador	Puntaje Total
M38		7.50	8	7.80	7.85	8	10	8	87.65
M35		8	7.80	7.75	7.50	10	10	7.85	85
M39		8.25	8	7.50	8	10	10	8.50	85.50

Muestra #	Seco	Calidad	Espuma	Sabor Residual	Intensidad	Nivel	Balance	Dulzor	Defectos (sustraer)	Puntaje Final
M38				8	Alta / Baja	Pesado / Ligero	8.50	10	Ligero=2, Rechazo=4 # tazas × Intensidad =	87.65
M35				8	Alta / Baja	Pesado / Ligero	8	10	Ligero=2, Rechazo=4 # tazas × Intensidad =	85
M39				7.75	Alta / Baja	Pesado / Ligero	7.80	10	Ligero=2, Rechazo=4 # tazas × Intensidad =	85.50

Muestra	Notas
M38	Maní Tostado, mandarina, flor de café
M35	Mango, ciruelas, maracuya dulce agradable
M39	Lucuma, vainilla, uvas final dulce agradable

Anexo A6.13



Formato de Catación de la Asociación Americana de Cafés Especiales

Nombre: Evar Gonzales Diaz

Fecha: 12-09-2025 Mesa: # 02 Sesión: 03

Clasificación

6.00 Bueno	7.00 Muy bueno	8.00 Excelente	9.00 Extraordinario
6.25	7.25	8.25	9.25
6.50	7.50	8.50	9.50
6.75	7.75	8.75	9.75

Muestra # M13

Nivel de Tueste

Seco Calidad Espuma

Fragrancia/Aroma	7.25
Sabor	7.55
Sabor Residual	8
Acidez	7.25
Intensidad: Alta / Baja	
Cuerpo	8
Nivel: Pesado / Ligero	
Uniformidad	10
Balance	7.25
Taza Limpia	10
Dulzor	10
Puntaje Catador	8.25
Puntaje Total	83.55
Defectos (sustraer): Ligero=2, Rechazo=4; # tazas x Intensidad =	

Notas: Lvuvma, viva, polen, algorrobina de cuerpo cremoso

Puntaje Final 83.55

Muestra # M4

Nivel de Tueste

Seco Calidad Espuma

Fragrancia/Aroma	7
Sabor	6.80
Sabor Residual	7.50
Acidez	8
Intensidad: Alta / Baja	
Cuerpo	7
Nivel: Pesado / Ligero	
Uniformidad	10
Balance	7.80
Taza Limpia	10
Dulzor	10
Puntaje Catador	8
Puntaje Total	82.10
Defectos (sustraer): Ligero=2, Rechazo=4; # tazas x Intensidad =	

Notas: Naranja, coña de azucar, chocolate posta

Puntaje Final 82.10

Muestra # M11

Nivel de Tueste

Seco Calidad Espuma

Fragrancia/Aroma	7.50
Sabor	7
Sabor Residual	7.25
Acidez	7.50
Intensidad: Alta / Baja	
Cuerpo	7.60
Nivel: Pesado / Ligero	
Uniformidad	10
Balance	7.50
Taza Limpia	10
Dulzor	10
Puntaje Catador	7.5
Puntaje Total	81.85
Defectos (sustraer): Ligero=2, Rechazo=4; # tazas x Intensidad =	81.85

Notas: Naranja criolla, toque de papa y pimienta

Puntaje Final 81.85

Anexo A7. Registro fotográfico del procesamiento y evaluación sensorial en laboratorio.

Figura A7.1. *Proceso de tostado de las muestras de café pergamino en el Laboratorio Ushún Café, paso previo indispensable para la evaluación física y sensorial.*



Figura A7.2. *Preparación de la infusión y práctica de cata siguiendo los protocolos estandarizados de la Specialty Coffee Association (SCA).*



Figura A7.3. *Evaluación de los atributos sensoriales de las muestras recolectadas por parte del equipo investigador y catadores certificados.*



Anexo A8. Script de preprocesamiento, creación de variables Dummy y modelamiento predictivo en Python.

A continuación, se presenta la algoritmia central desarrollada en el lenguaje de programación Python utilizada para el acondicionamiento de los datos categóricos y el entrenamiento de los modelos de *Ensemble Learning* (por bloques) tratados en el estudio de investigación:

Bloque 1: Aplicación de one-hot encoding (variables dummy)

Figura A8.1. *Línea de código pd.get_dummies.*

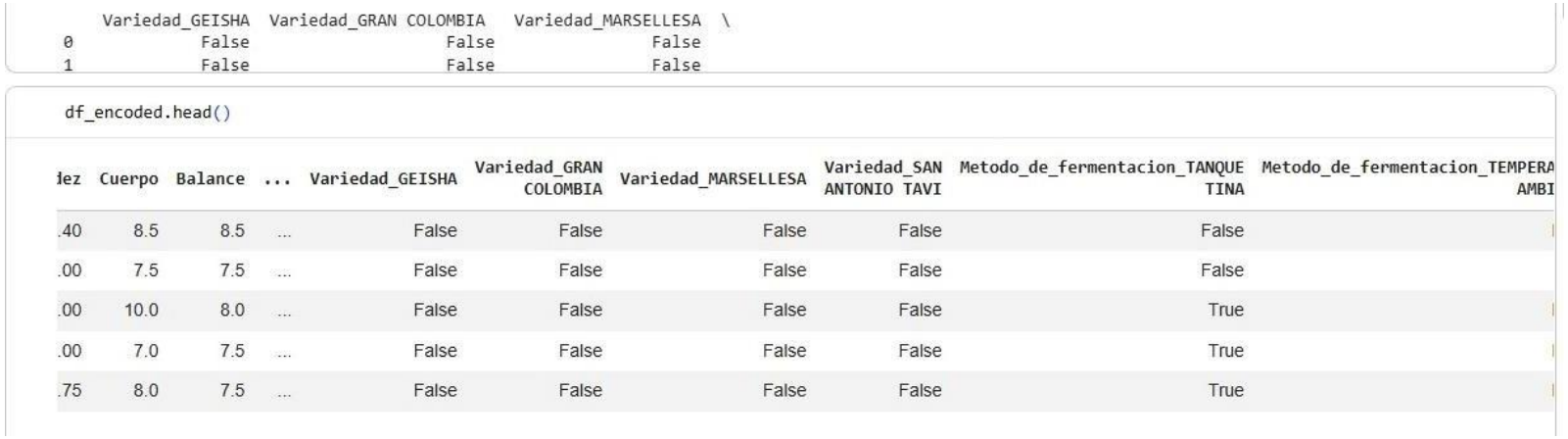
Aplicar One-Hot Encoding

```
[ ]
# ====================================
# Aplicar One-Hot Encoding
# ====================================
df_encoded = pd.get_dummies(data, columns=["Variedad", "Metodo_de_fermentacion", "Metodo_de_secado",
                                           "Tipo_de_beneficio"], drop_first=True)

print("\n✅ Datos después de One-Hot Encoding:")
print(df_encoded)
```

```
...
✅ Datos después de One-Hot Encoding:
   Altura  Edad_del_cultivo  Numero_de_cosecha  Tiempo_de_fermentacion  \
0    2000                 5                  4                      36
1    1800                 3                  4                      36
2    1800                 3                  4                      24
3    1800                 3                  2                      18
4    2000                 7                  5                      24
5    2100                 5                  3                      38
```

Figura A8.2. *Tabla resultante con False y True.*



```
    Variedad_GEISHA  Variedad_GRAN COLOMBIA   Variedad_MARSELLESA  \
0             False                   False                 False
1             False                   False                 False
```

```
df_encoded.head()
```

ɬez	Cuerpo	Balance	...	Variedad_GEISHA	Variedad_GRAN COLOMBIA	Variedad_MARSELLESA	Variedad_SAN ANTONIO TAVI	Metodo_de_fermentacion_TANQUE TINA	Metodo_de_fermentacion_TEMPERA AMBI
.40	8.5	8.5	...	False	False	False	False	False	
.00	7.5	7.5	...	False	False	False	False	False	
.00	10.0	8.0	...	False	False	False	False	True	
.00	7.0	7.5	...	False	False	False	False	True	
.75	8.0	7.5	...	False	False	False	False	True	

Bloque 2: Definición de las variables predictoras (X) y la variable objetivo (y).

Figura A8.3. *Línea de código X = df_encoded.drop e y = df_encoded*



Definimos X y Y

Puntuación taza

```
X = df_encoded.drop(["Puntuacion_taza", "Aroma", "Sabor_residual", "Acidez ","Cuerpo", "Balance ","Taza_limpia ", "Dulzor ", "Uniformidad", "Sabor ", "
```

```
y = df_encoded["Puntuacion_taza"]
```

```
X.head(2)
```

	Altura	Edad_del_cultivo	Numero_de_cosecha	Tiempo_de_fermentacion	Variedad_CATIMOR	Variedad_CATUAY	Variedad_CATURRA	Variedad_GEISHA	Variedad_GR COLOME
0	2000	5	4	36	False	False	False	False	Fa
1	1800	3	4	36	False	False	False	False	Fa

Bloque 3: Entrenamiento de los modelos predictivos

Figura A8.4. *Línea de código de Random Forest, XGBoost y CatBoost*

Random Forest

```
# ====================================
# 3. Modelos
# ====================================
# Random Forest
rf = RandomForestRegressor(n_estimators=200, random_state=42)
rf.fit(X, y)
y_pred_rf = rf.predict(X)
```

XGBoost

```
xgb = XGBRegressor(n_estimators=300, learning_rate=0.1, max_depth=4, random_state=42)
xgb.fit(X, y)
y_pred_xgb = xgb.predict(X)
```

CatBoost

```
X_cat = data[["Variedad", "Altura", "Edad_del_cultivo", "Numero_de_cosecha", "Tiempo_de_fermentacion",
              "Metodo_de_secado", "Tipo_de_beneficio", "Metodo_de_fermentacion"]]
```

Bloque 4: Evaluación del desempeño predictivo (R^2 y RMSE).

Figura A8.5. *Código de bloque de la función def evaluar.*

Models comparation

```
# ====================================
# 4. Evaluación de modelos
# ====================================
def evaluar(nombre, y_true, y_pred):
    r2 = r2_score(y_true, y_pred)
    rmse = np.sqrt(mean_squared_error(y_true, y_pred))
    print(f"\n📌 {nombre}")
    print(f"R²: {r2:.3f}")
    print(f"RMSE: {rmse:.3f}")
    return r2

r2_rf = evaluar("Random Forest", y, y_pred_rf)
r2_xgb = evaluar("XGBoost", y, y_pred_xgb)
r2_cat = evaluar("CatBoost", Y_cat, y_pred_cat)
```

Figura A8.6. *Resultados impresos con los R^2 de 1.000 para XGBoost.*

```
...
    📌 Random Forest
    R²: 0.802
    RMSE: 0.981

    📌 XGBoost
    R²: 1.000
    RMSE: 0.012

    📌 CatBoost
    R²: 0.328
    RMSE: 1.807
```