

Reporte de Predicción de Calidad del Café 'Campesino'

Miguel Angel Fonseca Aldana

15 de julio de 2025

1. Introducción

El objetivo de este proyecto es predecir la calidad del café, medida por el 'Puntaje de Taza', utilizando datos históricos de la tostadora 'Campesino'. Se analizarán los datos proporcionados en tres archivos Excel, se realizará un preprocesamiento adecuado, se entrenarán modelos de regresión y se analizará la explicabilidad de los resultados para identificar las variables más influyentes.

2. Visión General de los Datos y Preprocesamiento

Se proporcionaron tres archivos Excel:

- `CCFT17FormatodeControldeCalidadCafédeTrillado(1).xlsx`: Información de calidad y puntaje de taza
- `CCFT18FormatodeTostión(1).xlsx`: Datos del proceso de tostión
- `CCFT21FormatodeControldeDespachos(1).xlsx`: Información de despachos (no utilizada)

Limpieza y Combinación de Datos

Los archivos Excel presentaban filas de encabezado no estándar y metadatos en las primeras filas. Se cargaron los datos especificando la fila correcta del encabezado y se renombraron las columnas para mayor claridad. Se incluyeron los datos de la segunda hoja de cada documento Excel, concatenándolos con los datos de la primera hoja para ampliar el conjunto de datos disponible. Los DataFrames de 'CalidadTrillado' y 'Tostión' se unieron utilizando la columna 'Lote' como clave común. Las filas con valores nulos en la variable objetivo 'Puntaje_Taza' fueron eliminadas.

Selección y Transformación de Variables

Variables de entrada (X) seleccionadas:

- **Numéricas:** Humedad, Mallas, Peso_Verde, Merma, Peso_Tostado, Temp_Inicio, Temp_Final, Tiempo_Tueste_Minutos
- **Catóricas:** Origen, Variedad, Proceso, Beneficio, Perfil

Transformaciones realizadas:

- Conversi3n de columnas numéricas con manejo de errores
- Imputaci3n de valores nulos con la media
- Extracci3n de Temp_Inicio y Temp_Final
- Conversi3n de Tiempo_Tueste a Tiempo_Tueste_Minutos

Preprocesamiento para Modelado

- ColumnTransformer con:
 - StandardScaler para características numéricas
 - OneHotEncoder para características catatóricas
- Divisi3n de datos: 80 % entrenamiento, 20 % prueba

3. Entrenamiento y Evaluaci3n de Modelos de Regresi3n

Se entrenaron dos modelos:

Resultados del Modelo de Regresi3n Lineal

Métrica	Valor
R2 Score	0.9795
Mean Absolute Error (MAE)	0.1292
Mean Squared Error (MSE)	0.0426

Resultados del Modelo Random Forest Regressor

Métrica	Valor
R2 Score	0.9863
Mean Absolute Error (MAE)	0.0636
Mean Squared Error (MSE)	0.0284

Conclusi3n: El modelo Random Forest Regressor muestra mejor rendimiento y ser3 utilizado para el an3lisis de explicabilidad.

4. Análisis de Explicabilidad (Feature Importance en RandomForest)

Variables más influyentes en la predicción:

Feature	Importance
Variedad_Wush Wush	0.448046
Proceso_Honey	0.103838
Beneficio_Honey	0.102180
Temp_Final	0.071245
Variedad_Bourbon Sidra	0.032757
Beneficio_Lavado	0.031865
Origen_Acevedo	0.031306
Tiempo_Tueste_Minutos	0.029402
Origen_Jerico	0.022422

Observaciones Clave

- La Variedad_Wush Wush es, con diferencia, la característica más influyente, lo que sugiere que esta variedad específica tiene un impacto dominante en el Puntaje de Taza. Esto refuerza la importancia de la genética del grano en la calidad final.
- El Proceso_Honey y el Beneficio_Honey también muestran una alta importancia, indicando que el método de procesamiento post-cosecha es un factor crítico en la calidad del café.
- La Temp_Final del tueste es otra variable significativa, lo que resalta la importancia del perfil de tueste en el resultado final.
- Otras variedades como Variedad_Bourbon Sidra y Variedad_Gesha también son relevantes, aunque con menor impacto que la Wush Wush.
- El Origen del café (Origen_Acevedo, Origen_Jerico, Origen_Planadas) sigue siendo un factor importante, destacando la influencia del terruño.
- El Tiempo_Tueste_Minutos, Proceso_Natural, Beneficio_Natural, Temp_Inicio, Humedad, Peso_Verde, Variedad_Red Bourbon, Peso_Tostado y Merma también contribuyen a la predicción, aunque con menor peso.

5. Conclusiones

Los modelos **Random Forest Regressor** y **Linear Regression** han demostrado ser muy efectivos en la predicción del Puntaje de Taza del café '*Campesino*' a partir de los datos proporcionados. El análisis de importancia de características revela que en el modelo **Random Forest Regressor** la **variedad del café**, el **proceso** y **beneficio**

y la **temperatura final del tueste** son los factores más determinantes en la calidad, seguidos por el **origen** y el **tiempo de tueste**.