Lunes 28 de Julio, Barcelona
Ironhack

Analisis del Datasest ‘Flavors of Cacao’

Asier Novio Cara, Martí Vallhonrat Rafart

El chocolate es uno de los dulces más populares del mundo. Cada año, los residentes de Estados Unidos consumen en conjunto más de 2800 millones de libras. Sin embargo, no todas las barras de chocolate son iguales.
En este análisis vamos a intentar descifrar las razones de por las cuales cada chocolate puede tener una calificación más alta o más baja y creando un modelo que pueda predecir, a partir de ciertas características, la calidad del chocolate introducido.

1. Introducción

Nuestro objetivo principal de este análisis es, a partir de entender los datos, desglosar las razones de peso para la clasificación de calidad de las tablas de chocolate.
A partir de esa exploración el siguiente paso sería crear un modelo, de esta forma, al crear un chocolate, la empresa productora podría predecir de forma aproximada su calificación y hacer las modificaciones pertinentes para mejorarlo.

Las hipótesis formuladas para resolver la causalidad de la variable objetivo son:

1. ¿Qué país tiene la calificación más alta y cuál la más baja?
2. Si el porcentaje de chocolate es más puro, ¿se valora más?
3. ¿Qué empresa produce las barras de chocolate con mayor calidad o calificación?
4. ¿Qué tipo de semilla predomina en cada país?
5. ¿Qué país de origen de la semilla tiene la mejor calificación?
6. ¿Cuáles son las relaciones entre el país importador de cacao y el país productor de chocolate?
7. ¿Cuál es la pureza media del cacao en cada país?

El dataset proviene de Kaggle, creado por Rachael Tatman, pero los datos originales fueron compilados por Brady Brelinski, miembro fundador de la Manhattan Chocolate Society, con el objetivo de registrar valoraciones expertas de barras de chocolate oscuro.

2. Descripción y Preparación de los Datos

2.1. Fuente de los datos:

El dataset contiene calificaciones de experto de 1700 barras de chocolate individuales, junto con información sobre su origen regional, porcentaje de cacao, la variedad de grano de chocolate utilizada y dónde se cultivaron los granos, entre otras. A continuación adjuntamos una tabla con todas las variables, sus descripciones y sus tipos.

2.2 Inspección de variables:

Variable	Descripción	Tipo
company	Nombre de la empresa que produce la barra	Cualitativa, Nominal
specific_bean_origin_or_bar_name	Región o variedad específica del grano	Cualitativa, Nominal
ref	Número de la entrada a la base de datos	Cuantitativa, Discreta
review_date	Fecha en la cual se probó la barra	Cuantitativa, Discreta
cocoa_percent	Porcentaje de cacao en la barra	Cuantitativa, Continua
company_location	País donde la empresa está ubicada	Cualitativa, Nominal
rating	Valoración de la tableta de chocolate	Cuantitativa, Continua
bean_type	Tipo de especie de grano	Cualitativa, Nominal
broad_bean_origin	País de origen del cacao	Cualitativa, Nominal

La variable objetivo, 'Rating', contiene el siguiente sistema de clasificación:

1. **Desagradable** (mayormente desagradable)
2. **Decepcionante** (Pasable pero contiene al menos un defecto significativo)
3. **Satisfactorio** (3.0) o **Loable** (3.75) (bien hecho con cualidades especiales)
4. **Premium** (Desarrollo superior del sabor, carácter y estilo)

5. **Élite** (Trasciende los límites ordinarios)

Cada chocolate se califica combinando aspectos objetivos y subjetivos. La puntuación representa una experiencia específica con una tableta de un lote determinado.
El sabor es el factor principal en la evaluación, considerando su diversidad, equilibrio, intensidad y pureza, influenciado por la genética del grano, el terruño, el procesado y el almacenamiento.

2.3. Limpieza y transformación de datos:

El conjunto de datos presentaba valores e índices con errores como saltos de línea incrustados dentro de palabras, así como errores ortográficos e inconsistencias tipográficas. Por ello, se aplicó un proceso de limpieza exhaustiva, que incluyó la corrección de dichos errores, la estandarización de textos y la normalización de los valores.

Específicamente en el caso del **porcentaje del cacao** se transformaron valores porcentuales que originalmente estaban representados como cadenas de texto (por ejemplo, 76%) a su forma **numérica decimal** (0.76), con el fin de facilitar su análisis y procesamiento.

2.4. Clustering:

En nuestros datos algunas de las columnas categóricas tenían demasiadas casuísticas diferentes para poder sacar insights y hacer predicciones así que las pasamos por un proceso de agrupamiento o clusterización.

La columna de **origen de la semilla** y **ubicación de la empresa**, que originariamente eran países (por ejemplo, Francia) la agrupamos en **regiones continentales** (Europa Oeste).

En la variable de **origen específico de la semilla**, donde teníamos la variabilidad más alta pasamos de trabajar con valores varios de regiones del país a generalizar según **países**.

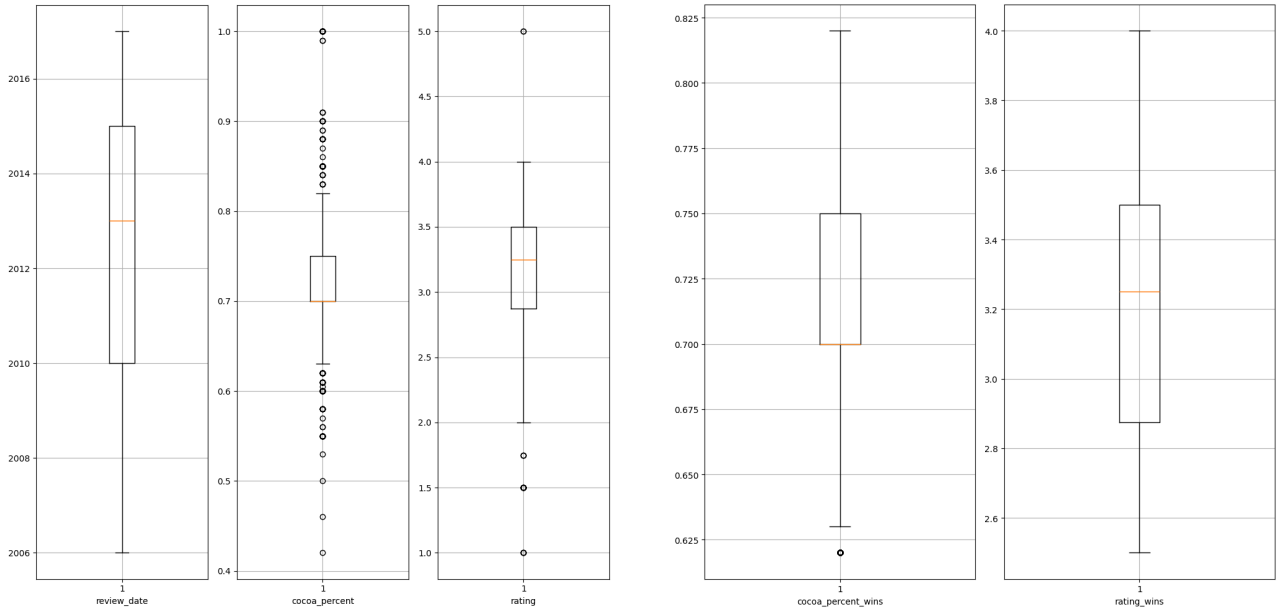
Por último los valores de **tipo de semilla** que fueran una mezcla de diferentes especies de granos han sido todos agrupados en una nueva categoría llamada **mezcla** para facilitar el análisis y el procesamiento de la columna.

2.5. Tratamiento de outliers:

La limpieza de outliers en nuestro caso ha sido un proceso delicado ya que el objetivo era suprimir los datos extremos que distorsionan el análisis pero no queríamos eliminar variación natural útil y aplanar la distribución innecesariamente.

Por estas mismas razones consideramos usar el proceso de **winsorización** solamente a los valores menores al **5%** y superiores al **95%** tanto en la columna de **calificación** como en la de **porcentaje de cacao**. En el caso de la variable **fecha de revisión** no es necesario debido a la **nula existencia de outliers**.

Boxplot de los datos antes y después del tratamiento:



2.6 Tratamiento de nulos:

En nuestros datos nos hemos encontrado con valores nulos en dos columnas. En la variable **origen de la semilla** con un **5,29%** local y un **0,66%** cómputo global, y en la columna **tipo de semilla** hay un **50,58%** en la misma columna y un **6,32%** global.

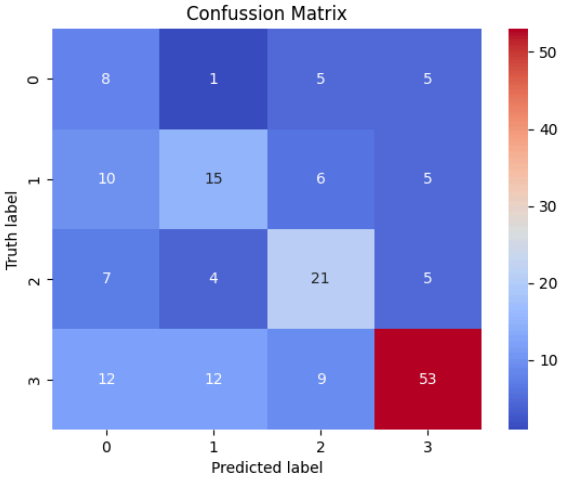
Con esta cantidad de casos con valores nulos nos planteamos reconstruir la variable ya que la posibilidad de eliminar los casos fue descartada.

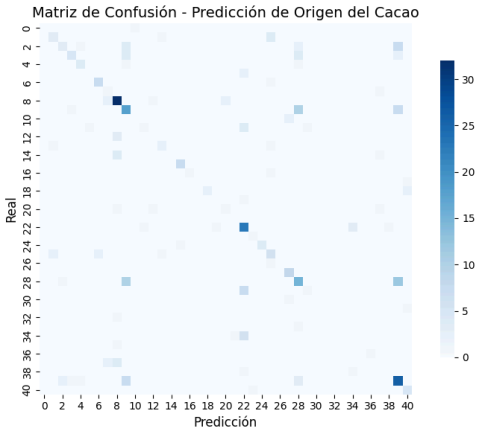
Creamos un modelo de predicción con las otras variables tanto para el caso de la columna **origen de la semilla** como para la columna **tipo de semilla**. Desgraciadamente los modelos que creamos sacaron una métrica de precisión de

47% y **53%** respectivamente, lo cual consideramos demasiado bajo para que sea una métrica fiable.

Finalmente consideramos que la opción de **imputación por más frecuente** era la más adecuada para ambas variables.

Visualizaciones de la matriz de confusión de **tipo de semilla** y **origen de semilla**:



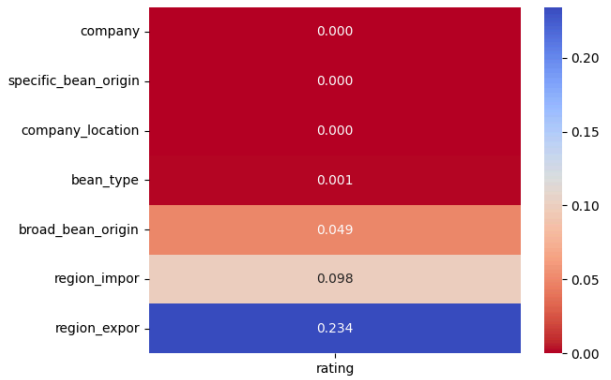


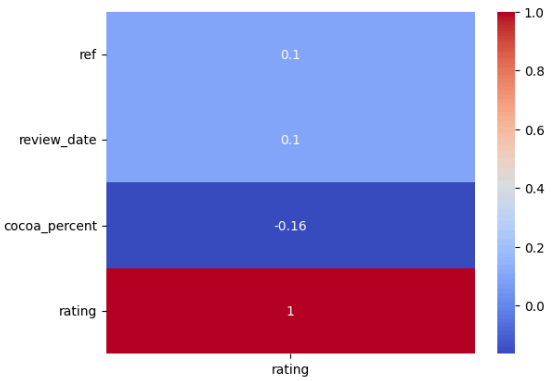
3. Análisis Exploratorio de Datos (EDA)

3.1. Análisis Bivariado:

Para iniciar el análisis, se exploraron las relaciones entre las variables independientes y la variable objetivo (**calificación** del chocolate). Se realizaron gráficos comparativos y se aplicaron pruebas estadísticas para respaldar visualmente y numéricamente los hallazgos. En particular, se utilizó un test **ANOVA** para evaluar si existían diferencias significativas en la calificación según variables **categóricas** como el tipo de semilla o el país de origen.

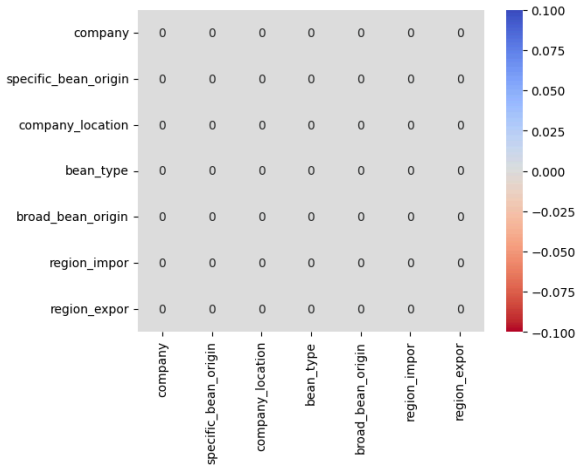
Para las variables **numéricas**, se llevó a cabo un análisis de **correlación de Pearson**, con el fin de identificar asociaciones lineales relevantes, como la posible relación entre el porcentaje de cacao y la calificación obtenida.

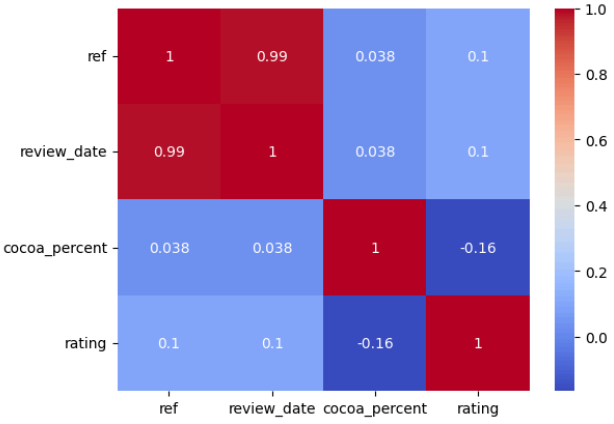




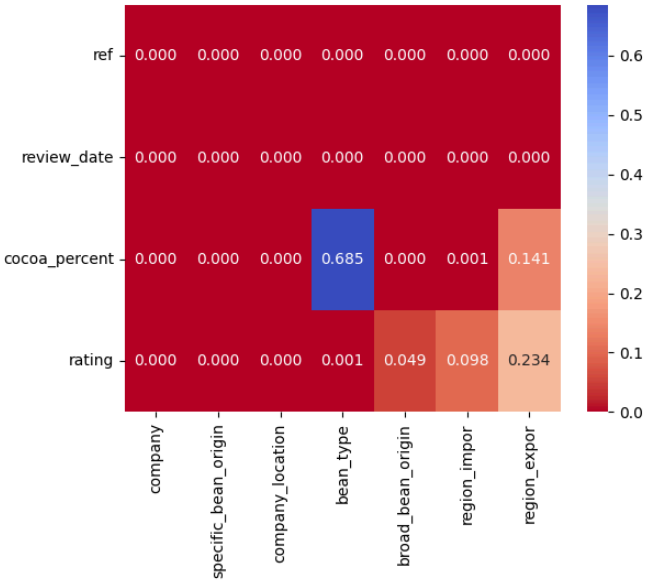
Luego, analizamos las relaciones entre variables. Para observar las relaciones entre las variables **categóricas** entre si aplicamos un **test Chi 2** con el fin de detectar asociaciones significativas.

Para las **numéricas**, se utilizó la **correlación de Pearson** con el mismo objetivo pero para identificar relaciones lineales relevantes entre variables.





Por último, se examinó la relación entre variables **categóricas** y **numéricas independientes**, utilizando pruebas de **ANOVA** y **visualizaciones comparativas**. Esto permitió identificar si ciertas categorías, como el tipo de semilla o el país de origen, se asocian con diferencias significativas en variables numéricas como el porcentaje de cacao.



4 Selección y Construcción del Modelo

4.1 Elección de modelo y justificación:

Para la etapa de modelado, decidimos utilizar un enfoque **ensamblado**, con el objetivo de combinar las fortalezas de distintos algoritmos y mejorar la capacidad predictiva del sistema.
Nuestro modelo final integra una **regresión lineal**, un **Random Forest Regressor** y un **XGBoost Regressor**, combinados mediante un esquema de promediado.
Inicialmente, probamos incluir un **MLP Regressor o red neuronal**, pero su rendimiento fue considerablemente inferior al de los demás modelos, lo que afectaba negativamente el desempeño general del ensamblado. Por esta razón, optamos por excluirlo y así mantener un mejor equilibrio entre precisión y estabilidad.

4.2 Selección de variables:

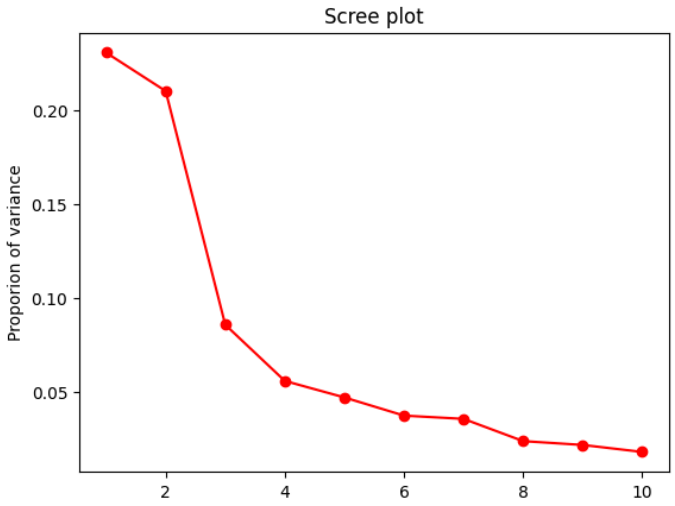
Para la selección de variables, aplicamos dos enfoques complementarios, un modelo **OLS** y un método **RFE**. Ambos nos permitieron obtener diferentes perspectivas sobre la importancia relativa de las variables y generar insights adicionales que contrastamos con nuestro análisis exploratorio previo.
Finalmente, seleccionamos las variables:

[**tipo de semilla**, **región importación**, **región exportación**, **nombre de la empresa**, **fecha de revisión**, **porcentaje de pureza**]

4.3 Estudio de Reducción de Dimensionalidad:

También exploramos la posibilidad de aplicar un **PCA** con el objetivo de reducir la dimensionalidad y capturar la variabilidad general del conjunto de datos, especialmente en variables con gran cantidad de clases como **nombre de la empresa**.
Sin embargo, el uso de PCA resultó en una pérdida de interpretabilidad y un **descenso en el rendimiento predictivo** del modelo ensamblado. Por esta razón, optamos por no incluirlo en la versión final del pipeline.

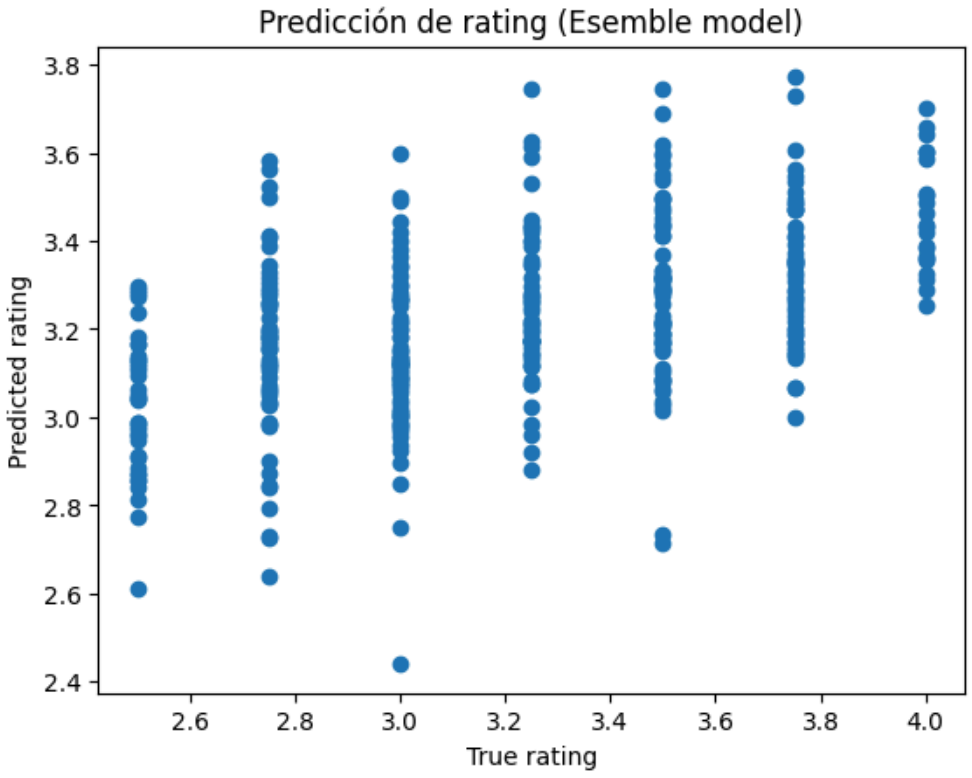
A continuación adjuntamos la visualización sobre varianza según número de agrupaciones del PCA:



5. Evaluación del Modelo

A continuación, se presenta una **tabla comparativa** con los resultados obtenidos por los distintos modelos evaluados, incluyendo los modelos individuales y el modelo ensamblado.
Para evaluar el rendimiento de cada uno, utilizamos como métricas principales el **R^2**, el **R^2 ajustado**, y el **mean squared error (MSE)**. Estas métricas nos permiten valorar tanto la capacidad explicativa como la precisión del modelo en los datos de prueba.
Al final adjuntamos también un gráfico para poder observar las diferencias entre las predicciones de nuestro modelo y el valor real.

Modelo	MSE	R^2	R^2 Ajustado
Regresión Lineal	0.1571	0.1699	0.1557
Random Forest	0.1443	0.2378	0.2248
XGBoost	0.1512	0.2013	0.1877
Ensamblado (RL + RF + GBR)	0.1408	0.2559	0.2433



6. Conclusiones

Realizamos un análisis completo del conjunto de datos, con especial esfuerzo en la limpieza y el EDA. Tras evaluar varios modelos, el ensamblado final (RL + RF + XGBoost) logró el mejor desempeño, con un R^2 cercano a 0.25.

Aunque probamos PCA y redes neuronales por separado, las descartamos por bajo rendimiento. El modelo final ofrece resultados aceptables y una buena comprensión de los factores que influyen en la calidad del chocolate.

6.1 Mejoras posibles:

Una de las principales limitaciones del modelo actual es su dificultad para predecir correctamente los valores extremos de la calificación, tanto los más altos como los más bajos. Esto se debe, en parte, a la distribución normal y concentrada del rating, algo común en datasets de evaluación, donde la mayoría de las observaciones se agrupan en torno a la media y hay pocos ejemplos representativos en los extremos.

Para mejorar este aspecto, podrían explorarse técnicas como modelos especializados en valores atípicos, o incluso enfoques de aprendizaje semi-supervisado o por cuantiles que prioricen la predicción en zonas menos representadas.

A continuación adjuntamos una visualización de la distribución de rating: