PROYECTO FINAL DATA SCIENCE

ALUMNA: PERETTI NADIA

PROFESOR: FERNANDO CARABEDO

TUTOR: GUILLERMO MONTERO

- **INDICE:**

1. CONTEXTO Y AUDIENCIA

Me motiva el siguiente trabajo la inquietud que tenemos en nuestro emprendimiento personal de elaboración y venta de barras de chocolate.

Nuestra audiencia puede ser cualquier emprendedor que quiera incursionar en este negocio.

2. METADATA

Data Set: https://www.kaggle.com/datasets/evangower/chocolate-bar-ratings?select=chocolate_bars.csv.

Tamaño: 2530 rows × 11 columns

Calificaciones de más de 2500 barras de chocolate de todo el mundo.
Cada chocolate se evalúa a partir de una combinación de cualidades objetivas e interpretación subjetiva.
Sistema de calificación de sabores de cacao: 4.0 - 5.0 = Sobresaliente 3.5 - 3.9 = Altamente recomendado 3.0 - 3.49 = Recomendado 2.0 - 2.9 = Decepcionante 1.0 - 1.9 = Desagradable.

2. METADATA

VARIABLES

Nombre	Descripción
id	Es un identificador numérico del registro
manufacturer	Nombre de la barra de chocolate
company_location	País de origen de los granos de cacao
year_reviewed	El año de revisión
bean_origin	Ubicación del fabricante
bar_name :	Nombre del fabricante de la barra de chocolate
cocoa_percent :	Contenido de cacao de la barra de chocolate (%)
num_ingredients :	Número de ingredientes en la barra de chocolate.
ingredients	Ingredientes utilizados (Frijoles), S (Azúcar), S* (Edulcorante distinto del azúcar o de remolacha), C (Manteca de cacao), (V) Vainilla, (L) Lecitina, Sa (Sal)
review	Resumen de las características más recordadas de la barra de chocolate
rating	Sistema de calificación de sabores de cacao: 4.0 - 5.0 = Sobresaliente 3.5 - 3.9 = Altamente recomendado 3.0 - 3.49 = Recomendado 2.0 - 2.9 = Decepcionante 1.0 - 1.9 = Desagradable

2. METADATA

VARIABLES NULAS:

- Num-ingredients : había 87 nulos y se reemplazaron por la media.

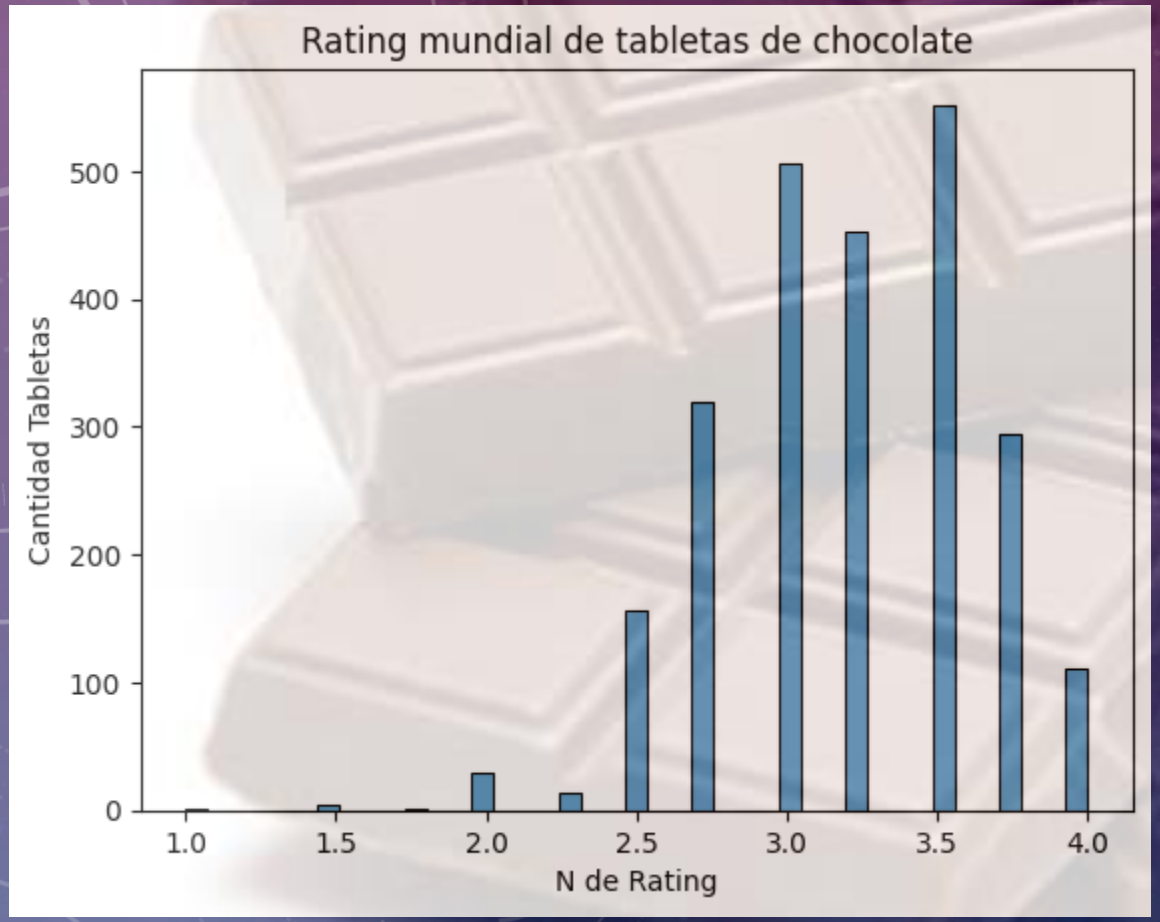
- Ingredients: también con 87 nulos se eliminaron porque consideré que su valor no era relevante para el análisis.

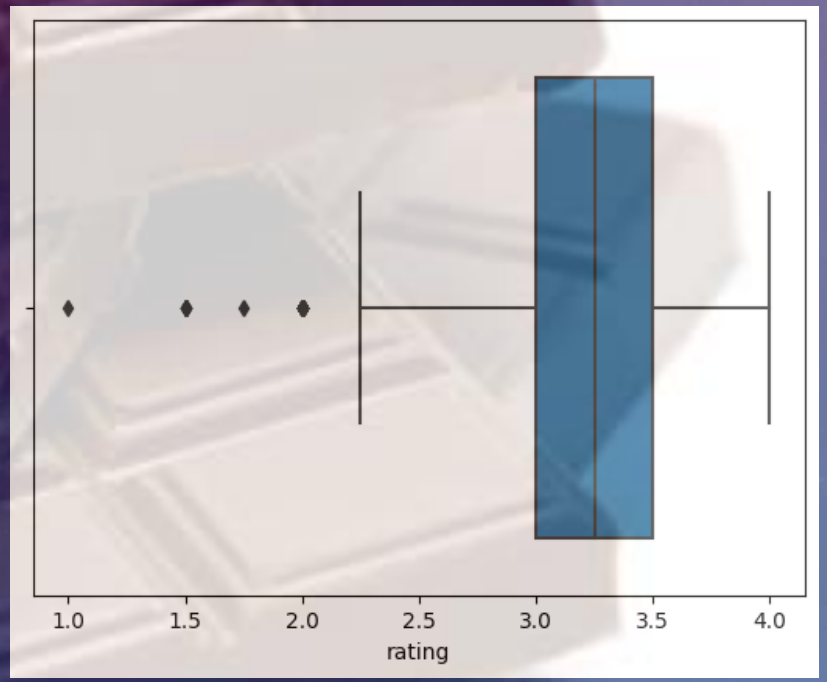
3. PREGUNTAS HIPÓTESIS

- ¿ Qué quiere la gente?
- ¿ Qué posiciona a un buen producto?
- ¿Conviene diversificar los productos?
- ¿Qué variables modifican la percepción de los clientes?

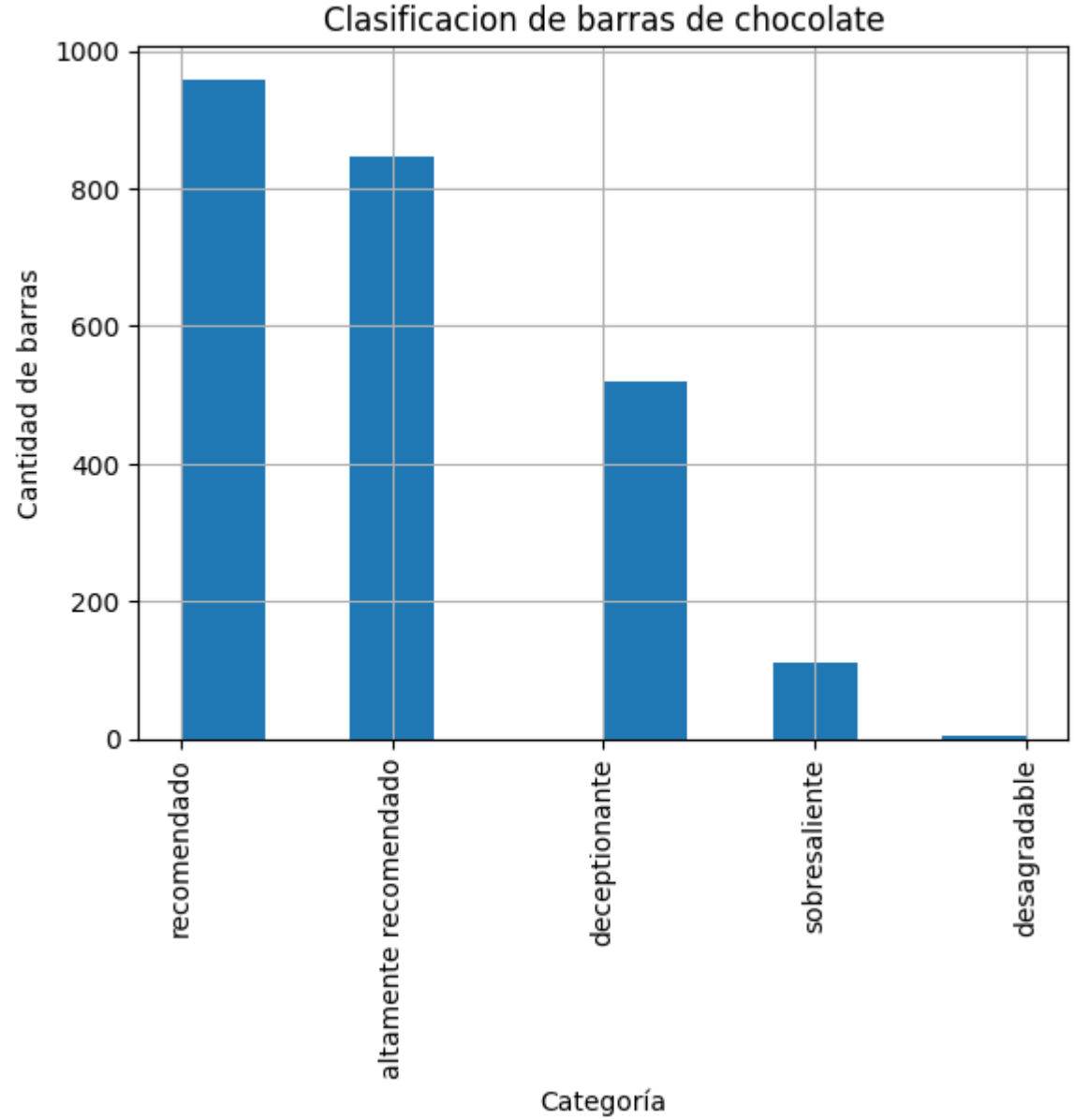
- Armar un modelo para **predecir el ranking** de una barra de chocolate en función de otras variables .
- Evaluar cuáles son las variables más importantes y analizar cuál es la más rentable para nuestro negocio.

4. VISUALIZACIONES





4. VISUALIZACIONES



Sistema de calificación de sabores de cacao:

4.0 - 5.0 = Sobresaliente
3.5 - 3.9 = Altamente recomendado
3.0 - 3.49 = Recomendado
2.0 - 2.9 = Decepcionante
1.0 - 1.9 = Desagradable

5. INSIGHTS

La variable target "rating" tiene 5 posibles valores, por lo que realizo una tabla multinivel, agregando una *tercera dimensión* llamada "Clasificación“ para categorizar dicha variable.
Sistema de calificación de sabores de cacao: 4.0 - 5.0 = Sobresaliente 3.5 - 3.9 = Altamente recomendado 3.0 - 3.49 = Recomendado 2.0 - 2.9 = Decepcionante 1.0 - 1.9 = Desagradable.

6. MODELADO

1- REGRESIÓN LINEAL:

A- REGRESIÓN LINEAL con todas las variables: Presenta `OVERFITTING` por lo que debo eliminar variables.

B- REGRESIÓN LINEAL ajustando a la variable de mayor correlación, emepeoró con respecto al modelo anterior.

C- REGRESIÓN LINEAL reduciendo la dimensionabilidad, reagrupando por año el rating de acuerdo al % de cacao veo que mejora mucho el modelo con respecto a los anteriores pero sigue siendo baja la precisión del modelo.

D- FORWARD SELECTION iteramos la selección de variables del modelo anterior para ver si mejora pero no es así.

6. MODELADO

2- ÁRBOL DE DECISIÓN:

A- REGRESIÓN LINEAL:
- SIN PODA
- Con 3 NIVELES
- Con 6 NIVELES

El que mejor predice es SIN PODA.

B- CLASIFICACIÓN:
- SIN PODA
- Con 3 NIVELES
- Con 6 NIVELES
- Con 8 NIVELES
- Con 30 NIVELES

EL MODELO CON 3 HOJAS ES EL QUE MEJOR PREDICE (mejor cross validation). Y MEJORA CONSIDERABLEMENTE CON RESPECTO A LA APLICAIÓN DE ESTE METODO TOMANDO EL MODELO COMO REGRESIÓN LINEAL. ES DECIR EN LA SEGUNDA PRUEBA DONDE SE TOMÓ EL MODELO COMO CLASIFICACIÓN Y SE PREDIJO CON ÁRBOL DE DECISIÓN DE 3 NIVELES ES CUANDO MEJOR PERFORMÓ.

6. MODELADO

PARA LOS SIGUIENTES MODELOS EVALÚO EL MODELO DE CLASIFICACIÓN, ELIGIENDO UNA VARIABLE PARA PODER REALIZAR LA MATRIZ DE CONFUSIÓN:

3- KNN:

- Con 3 NIVELES
- Con 6 NIVELES

4- GaussianNB

5- DecisionTreeClassifier

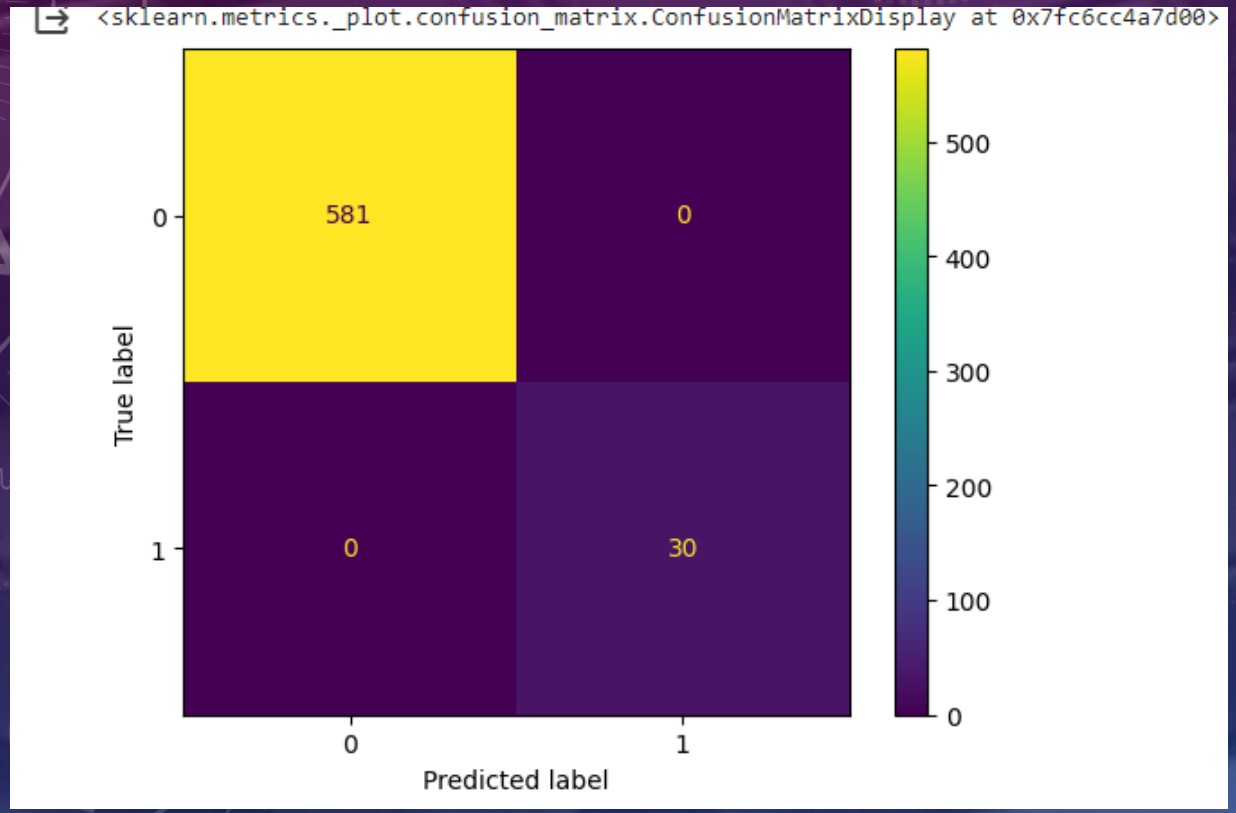
6. MODELADO

- MÉTRICAS DE DESEMPEÑO MODELADO: Comparativa de los resultados obtenidos de los modelos que mejor performaron:

MÉTRICA/ MODELO	KNN (3)	KNN (5)	GaussianNB	DecisionTreeClasifier(3)
ACCURACY	0,959	0,956	0,979	1
RECALL	0,172	0,061	0,53	1

6. MODELADO

- MATRIZ DE CONFUSIÓN:



SE LOGRÓ MINIMIZAR LOS FALSOS NEGATIVOS AL 100%.

7. CONCLUSIONES

EL MODELO GANADOR CON MEJORES MÉTRICAS ES:

DecisionTreeClassifier(max_leaf_nodes=3)

Este modelo tiene un valor de Recall de 1, lo que significa que el 100% de valores positivos fueron clasificados correctamente, tambien tiene un Accuracy de 1 lo que significa que el 100% de predicciones del modelo fueron realizada de manera correcta.