“PROGRAMA ESPECIALIZADO CIENCIA DE DATOS CON PYTHON”

PROYECTO GRUPAL

RETO: Desarrollar un Modelo de Machine Learning Pronóstico de Calificación de Chocolate

1. Introducción

Cuando hablamos de Chocolate debemos tener en cuenta que, según la legislación alimentaria, debe ser un producto obtenido a partir de un producto de cacao y azúcares que contenga al menos un 35% de materia seca total de cacao; un 18%, como mínimo, debe ser manteca de cacao y un 14% materia seca y desgrasada de cacao.

Para poder calificar a un chocolate como negro y que cumpla los estándares de calidad se exige que posea un 43% de materia seca total de cacao (versus el 35% del chocolate normal, según la legislación vigente), de la cual el 26% debe ser manteca de cacao (versus el 18% en otros chocolates).

Entre las variedades de chocolate negro existen desde algunas más suaves hasta otras con elevado porcentaje de cacao y con un sabor intenso del 70% o muy intenso con más del 80%.

A mayor porcentaje de cacao más puro y mejor será el chocolate y uno bueno tendrá esta información siempre visible: si el porcentaje es inferior al 10% ni siquiera podría ser llamado chocolate, según la legislación vigente.

2. Sistema de calificación de sabores de Cacao

Puntuación	Calificación	Característica
5	Elite	Trascendiendo más allá de los límites ordinarios
4	Premium	Desarrollo de sabor, carácter y estilo superior
3	Satisfactorio	3.0 satisfactorio 3.75 Digno de elogio Ambos son bien elaborados con cualidades especiales
2	Decepcionante	Aceptable, pero contiene al menos un defecto significativo
1	Desagradable	Principalmente desagradable.

3. Información de la data

El conjunto de datos con el que vamos a trabajar está compuesto por las siguientes características:

- **REF:** Valor relacionado con el momento en que se ingresó la reseña en la base de datos
- **Company:** Referencia al productor que fabrica la barra. ●

Company location: País base del fabricante.

- **Review Date:** Año de publicación de la reseña.
- **Country of Bean Origin:** País de origen del grano de cacao. ●

Specific Bean Origin or Bar Name: Región geográfica específica de origen de la barra.

- **Cocoa Percent:** Porcentaje de cacao (oscuridad) de la barra de chocolate que se está revisando.
- **Ingredients:** Ingrediente que lleva la barra de chocolate. ● **Most Memorable Characteristics:** Característica más memorable de la que está compuesta la barra de chocolate. ● **Rating:** Calificación de experto para la barra de chocolate.

4. Objetivo del trabajo

Construir un modelo de Machine Learning a partir de los datos de calificación (*rating*) que se encuentran en la data proporcionada. Este modelo debe tener la

capacidad de predecir la calificación que recibirá una empresa.

5. Actividades a desarrollar

a. Realizar las manipulaciones y tratamiento de los datos de manera que queden listos para la modelación.

Nota: Tome en cuenta los valores perdidos y los valores outliers, o alguna transformación de las variables en caso de ser necesario.

b. Hacer un análisis exploratorio de datos (EDA).
c. Defina si usará modelos de **Regresión** o **Clasificación** en base a la información suministrada u otra información de la red que le ayude a determinar qué tipo de aprendizaje supervisado vaya a realizar. Dé una breve explicación.
d. Realizar escenarios de modelos.

Nota: tome en cuenta los siguientes modelos, el orden no es determinístico. Use como mínimo 5 modelos:

- XGBoost
- Random Forest
- Gradient Boosting
- Support Vector Machine
- Naive Bayes
- KNN
- Regresión logística
- Árbol de decisión
- Extra tree decisión
- Bagging
- Regresión lineal.

e. Con los modelos que ha seleccionado construya modelos de **Regresión** o **Clasificación** en base a la información suministrada.
f. Evalué sus modelos según el tipo de modelo que usted haya elegido. **Nota:** Calcule las métricas necesarias dependiendo del tipo de modelo que haya elegido.
g. Con el mejor modelo que haya salido de la evaluación aplique una modelación de hiper parámetro, luego vuelva a calcular nuevas métricas y compare los cambios con respecto a las métricas de evaluación del paso **f**.
h. Realice un análisis PCA para reducir el número de dimensiones del conjunto de datos. Luego vuelva a modelar un sistema (el que eligió en el punto **f**) y calcule nuevas métricas. Compare con las anteriores métricas.

ENTREGABLE:

- Enviar la solución en Google Collaboratory Notebook (usar comentarios explicando brevemente lo que se hace en cada celda).

- ❖ Enviar un video en youtube sustentando el reto y solución (por el grupo).
- ❖ Máximo grupos de 4

integrantes. Data:

chocolate_ratings.csv

Subir al Proyecto - Google Classroom.

Fecha máxima de entrega: 17 de noviembre del 2024