

Reto para Categoría Avanzada

Data Challenge Industrial DCI 2.0

Introducción

Veracruz es el segundo estado con mayor producción de café en México, y se reconoce como uno de los mejores cafés del mundo, habiendo ganado cuatro premios en el concurso Taza de Excelencia organizado por el Alliance for Coffee Excellence. El estimulante sabor, aroma y consistencia de esta bebida dependen de las características agro-climáticas, precipitación y temperatura de cada región.


La descripción de las características de cuerpo acidez, dulzura y notas aromáticas del café de especialidad se conoce como perfil de taza. Estos atributos se dan gracias a las condiciones agroclimáticas de las regiones, la variedad de la planta de café y los procesos en su producción. El café veracruzano, consistentemente presenta un perfil de sabor destacadamente dulce, con una base consistente a chocolate y una variación entre frutos rojos.

El reto

En este Data Challenge Industrial tendremos oportunidad de trabajar con datos asociados a los perfiles de taza de café de excelencia del estado de Veracruz correspondientes a muestras de diversos actores de la industria en el estado.

Cada registro contiene datos asociados a un perfil de taza realizado por catadores certificados en protocolos y estándares internacionales, además de información sobre la localidad de procedencia del producto y condiciones agroclimáticas de dicha región. Adicionalmente, se incluyen muestras sin un perfil de taza asociado, únicamente con los datos de procedencia y condiciones agro-climáticas.

El reto consiste en

1. determinar regiones de producción de café de especialidad con base en las características agro-climáticas (temperaturas, precipitación, etc.) y discutir las diferencias con respecto a las regiones basadas en condiciones geopolíticas.
 2. generar un modelo que describa cómo influyen las condiciones agro-climáticas en las características que describen el perfil de taza de las muestras.
 3. generar un modelo para predecir, a partir del perfil de taza de una muestra, la región de procedencia (agro-climática o geopolítica).
- 

Los datos

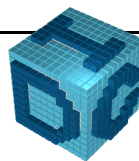
La tabla consisten de 5157 muestras de café de las distintas regiones productoras del estado. Además de las coordenadas geográficas, denotadas por **POINT_X,POINT_Y**, se incluyen 19 parámetros que representan promedios anuales de diversos factores agroclimáticos y cuyos nombres empiezan por el prefijo **bio_n** (ver Tabla 1). Adicionalmente los valores de temperatura minima y maxima mensuales, denotadas por las variables con el prefijo **tmin_n, tmax_n**, y precipitación promedio mensual que inician con el prefijo **prec_n, donde** n representa el número de mes (1 a 12)

El **perfil de taza** está representado por 10 atributos de características organolépticas nombrados con el prefijo OL_n, OL_D y OL_TOT este último corresponde a la suma de los valores de atributos OL_n excepto OL_D, valor que se sustrae del total. Un café de excelencia tiene un valor de OL_TOT por arriba de 81 puntos. Los primeros cuatro atributos de características organolépticas OL_1 a OL_4 tienen asociadas k=12 notas, nombradas N_1_1_k, N_1_2_k, N_2_k y N_3_k, el primero corresponde al atributo OL_1, etc. Los valores de notas N_n_k de cada uno de los atributos NOT_n indican el número de veces que esa nota particular se detecta en la muestra.

BIO1 = Annual Mean Temperature
BIO2 = Mean Diurnal Range (Mean of monthly (max temp - min temp))
BIO3 = Isothermality (BIO2/BIO7) (* 100)
BIO4 = Temperature Seasonality (standard deviation *100)
BIO5 = Max Temperature of Warmest Month
BIO6 = Min Temperature of Coldest Month
BIO7 = Temperature Annual Range (BIO5-BIO6)
BIO8 = Mean Temperature of Wettest Quarter
BIO9 = Mean Temperature of Driest Quarter
BIO10 = Mean Temperature of Warmest Quarter
BIO11 = Mean Temperature of Coldest Quarter
BIO12 = Annual Precipitation
BIO13 = Precipitation of Wettest Month
BIO14 = Precipitation of Driest Month
BIO15 = Precipitation Seasonality (Coefficient of Variation)
BIO16 = Precipitation of Wettest Quarter



Laboratorio Nacional de Informática Avanzada A.C.



DATA
CHALLENGE
INDUSTRIAL

BIO17 = Precipitation of Driest Quarter

BIO18 = Precipitation of Warmest Quarter

BIO19 = Precipitation of Coldest Quarter

DEM= Altitud

¡Mucha suerte y esperamos que se diviertan!