Flujo completo

Objetivo de la sesión

Repaso de transformación de datos

Predicción de la calidad de vinos

Mercado Vitivinícola

El tamaño del mercado global en 2019 fue de $364.25 bdd y se proyecta que alcance los $443.93 bdd para el 2027, lo que representa una tasa anual compuesta de crecimiento de 6.06%.

Industria en México.

- La industria genera empleos para 500 mil jornaleros.
- El principal estado vitivinícola de México es Baja California, con el Valle de Guadalupe.
- El 67% del vino que se consume en México es importado.
- En 2017, se importaron 71.9 millones de litros.
- El consumo per cápita en México es de 0.78 litros vs 20.08 en España.
- El 61.4% del vino que se consume es tinto.

Factores que inciden en la calidad del vino

- Medio en el que crece la planta. Clima, tiempo, temperatura, luz y suelo.
- Especies y variedades de la uva.
- Prácticas de viticultura.
- Prácticas enológicas.

Pasos de implementación

Cargar datos y Explorarlos	Transformación de variables	Selección de variables	Modelación	Evaluación / Resultados
Se utilizó el data set de uci machine learning repository wine-quality. Se realizó un análisis exploratorio de los datos, que mostró algunas variables con sesgo en su distribución	Se binarizó la variable objetivo utilizando calidad 6 como criterio de corte, es decir, una calidad mayor o igual a 6 es buena, y una menor a 6 es mala. Se separó el data set en entrenamiento y validación. Se utilizó la transformación WOE para discretizar las variables independientes	Utilizando árboles de decisión, random forest y Information Value (IV) se obtuvieron métricas que indican el poder de predicción de las variables, al final de este proceso se seleccionaron todas las variables	Se desarrolló una clase que permitió correr diferentes modelos asociados a una variable objetivo y sus variables independientes, en este paso se probó utilizando las variables independientes del dataset original vs las transformadas usando WOE	Se generaron las métricas de evaluación de los modelos usando el conjunto de datos de prueba

Registrate en

https://rstudio.cloud/

usaremos el script flujo_completo.R