PRÁTICA 2: Limpieza y validación de datos

Integrantes:

- **Luis Leandro Jiménez**
- **Edna Espejo**

Solución

1. **Descripción del dataset**

   El dataset es obtenido de un estudio de minería de datos sobre las características de los vinos rojo y blanco de Portugal. Las variables que se utilizan aquí son sólo algunas de todo el conjunto de datos de ese proyecto que pretendía hacer una predicción de preferencia de sabor de los vinos. Se puede encontrar más información en la siguiente fuente:

   P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.

   Descripción de variables:

   Variables basadas en test psicometricos:

   1 - fixed acidity: la mayoría de los ácidos involucrados en el vino o fijos o no volátiles

   2 - volatile acidity: Describe el nivel de ácido acético en el vino, en niveles altos provoca un sabor desagradable

   3 - citric acid: Nivel de ácido cítrico en el vino

   4 - residual sugar: Cantidad de azúcar que queda después de que se detiene el proceso de fermentación

   5 – chlorides: Cantidad de sal en el vino

6 - free sulfur dioxide: Cantidad de formas libres dioxido de azufre, este componente previene el crecimiento microbiano y la oxidación del vino

7 - total sulfur dioxide: Cantidad de formas libres y unidas de dioxido de azufre

8 – density: Describe el nivel de densidad del vino

9 – pH: describe qué tan ácido o básico es un vino en una escala de 0 (muy ácido) a 14 (muy básico)

10 – sulphates: aditivo del vino que puede contribuir a los niveles de dioxido de azufre

11 – alcohol: el porcentaje de contenido de alcohol del vino

Variables basadas en datos por sensores

12 - quality (score between 0 and 10)

13 - color: describe el tipo de vino (variable incluida manualmente)

2. Importancia y objetivos del análisis

El objetivo del análisis es determinar cuales son las variables más características según el tipo de vino. Analizar la relación entre ellas, aplicar las técnicas de preparación y limpieza de datos para futuros estudios.

Este tipo de análisis es de suma importancia para las empresas que tienen una relación con las compañías de vino y requieren conocer las características, por ejemplo, ofrecer vino para beber, recetas, almacenamiento, transporte, entre otros.

3. Limpieza de datos y análisis de datos

Etapas se encuentran en el archivo PRAC2.Rmd

4. Conclusiones

Para un analisis futuro de estos datos intentaria contar con datos más equitativos porque tenemos más datos de vino blanco que tinto. Y esta diferencia afecta en proporción los resultados que se obtienen.

Cuando el porcentaje de alcohol disminuye, la densidad aumenta. El nivel de alcohol del vino tinto es más alto que el del vino blanco.

Existen muchas herramientas, técnicas o métodos que se pueden utilizar para hacer un análisis exploratorio de los datos y que son importantes para entender la distribución de los datos y que variables se pueden utilizar o son más significativas en los tipos de vinos. Por ejemplo, utilizar un random forest o una reducción de dimensionalidad. Los tratamientos son necesarios para reducir el sesgo o problemas en interpretación de los datos.