Práctica 2 M2.851 - Tipología y ciclo de vida de los datos

Autores:

Waziri Ajibola Lawal
David Fernández González

Índice

1. Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende responder?

El conjunto de datos para el análisis se ha obtenido en Kaggle. El dataset está relacionado con las variantes de vino rojo del tipo "Vinho Verde". Debido a restricciones de seguridad y logística, se han excluído datos relativos a tipo de uvo, marca, precio de venta, etc.

El conjunto de dataset se compone de los siguientes campos:

- **fixed acidity:** la mayoría de los ácidos involucrados con el vino o fijos o no volátiles (no se evaporan fácilmente).
- **volatile acidity:** la cantidad de ácido acético en el vino, que en niveles demasiado altos puede producir un desagradable sabor a vinagre.
- **citric acid:** encontrado en pequeñas cantidades, el ácido cítrico puede añadir "frescura" y sabor a los vinos.
- **residual sugar:** la cantidad de azúcar que queda después de que la fermentación se detenga, es raro encontrar vinos con menos de 1 gramo/litro y los vinos con más de 45 gramos/litro se consideran dulces.
- **chlorides:** la cantidad de sal en el vino.
- **free sulfur dioxide:** la forma libre de SO2 que existe en equilibrio entre el SO2 molecular (como gas disuelto) y el ión bisulfito; impide el crecimiento microbiano y la oxidación del vino.
- **total sulfur dioxide:** cantidad de formas libres y ligadas de S02; en bajas concentraciones, el SO2 es en su mayoría indetectable en el vino, pero en concentraciones de SO2 libre superiores a 50 ppm, el SO2 se hace evidente en la nariz y el sabor del vino.
- **density:** la densidad del vino se acerca a la del agua dependiendo del porcentaje de alcohol y del contenido de azúcar.
- **PH:** describe cuán ácido o básico es un vino en una escala de 0 (muy ácido) a 14 (muy básico); la mayoría de los vinos están entre 3 y 4 en la escala de pH.
- **sulphates:** un aditivo para el vino que puede contribuir a los niveles de dióxido de azufre (S02), que actúa como antimicrobiano y antioxidante.
- **alcohol:** el porcentaje de contenido de alcohol del vino.
- **quality:** variable de salida (basada en datos sensoriales, puntuación entre 0 y 10).

El objetuvo principal es encontrar que variables ofrecen más información sobre la calidad del vino. También intentaremos hacer predicciones de la calidad de un vino, y comprobar si se corresponde con su calidad real.

2. Integración y selección de los datos de interés a analizar

3. Limpieza de datos.

3.1. ¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías cada uno de estos casos?

3.2. Identificación y tratamiento de valores extremos

4. Análisis de los datos.

4.1. Selección de los grupos de datos que se quieren analizar/comparar (planificación de los análisis a aplicar).

4.2. Comprobación de la normalidad y homogeneidad de la varianza.

4.3. Aplicación de pruebas estadísticas para comparar los grupos de datos. En función de los datos y el objetivo del estudio, aplicar pruebas de contraste de hipótesis, correlaciones, regresiones, etc. Aplicar al menos tres métodos de análisis diferentes.

5. Representación de los resultados a partir de tablas y gráficas.

6. Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?

Contribuciones	Firmas
Investigación previa	WAjibolaL,DFdezGlez
Redacción de las respuestas	WAjibolaL,DFdezGlez
Desarrollo código	WAjibolaL,DFdezGlez