Proyecto Final

Bootcamp de Ciencia de Datos (Avanzado)

Realizado por:
Vicente Ciampa

INDICE

Objetivos

- Aplicar los conocimientos adquiridos durante el Bootcamp en un proyecto.
- Fomentar el desarrollo de proyectos reales.
- Desarrollar un proyecto para portafolio donde se aplique ciencia de datos.

Descripción

Definición del problema y objetivos

Predicción de la calidad y el tipo del vino mediante el análisis de ingredientes:

1) Dataset: WineQuality.csv
2) Importar bibliotecas y cargar datos
3) Exploración del conjunto de datos (EDA)
4) Aplicar Machine Learning
5) Predicciones
6) Modelos finales

Para calcular la predicción de la calidad del vino, así como el tipo de vinos a partir del análisis de ingredientes, debemos seguir los pasos dados:

1. Recopilación de datos
Vamos a utilizar el archivo de conjunto de datos WineQuality.csv para esta tarea. Este conjunto de datos contiene 12 columnas relacionadas con las propiedades fisicoquímicas de los vinos y su calificación de calidad. Las columnas son:

-acidez fija
-acidez volátil
-ácido cítrico
-azúcar residual
-cloruros
-dióxido de azufre libre
-dióxido de azufre total
-densidad
-pH
-sulfatos
-alcohólico

-calidad (puntuación entre 0 y 10)
-Tipo (tipo de vinos)

2. Importar bibliotecas y cargar datos
Necesitamos importar las bibliotecas necesarias y cargar el conjunto de datos en un dataframe pandas.
3. Explorando Dataset
4. Aplicación Machine Learning
5. Predicciones
6. Valores finales.

Preguntas sobre los datos:

El análisis de los ingredientes del vino puede ser utilizado para predecir tanto la calidad como el tipo del mismo. ¿Cuáles son algunas preguntas específicas que se pueden plantear en relación a este análisis?

¿Qué parámetros se toman en cuenta para el análisis químico del vino?
¿Qué nos proporciona el color del vino?
¿El ph del vino como nos afecta?
¿Cuáles son los ingredientes base del vino?
¿Cómo podemos predecir la calidad de los ingredientes del vino?

- El análisis químico del vino es un proceso fundamental para conocer ciertos parámetros clave durante su elaboración, como por ejemplo el pH y la acidez. Estos parámetros influyen en el sabor, aroma y color del vino, y su análisis ayuda a los enólogos a controlar adecuadamente el proceso de vinificación hasta el embotellado. Además, el análisis químico del vino también puede ser utilizado para predecir tanto la calidad como el tipo del mismo.

Hallazgos y conclusiones

La importancia de la calidad del vino radica en su sabor, aroma y textura, lo que determina su aceptabilidad entre los consumidores. La calidad del vino está estrechamente relacionada con el proceso de elaboración, así como con los ingredientes utilizados. Los diferentes aspectos a considerar para medir la calidad del vino incluyen el color, la acidez, el pH, la persistencia del sabor y la ausencia de defectos. La calidad del vino puede influir en su precio y en su reputación, por lo que es un factor clave para la industria vitivinícola.

Comprensión de los datos sobre la calidad del vino

La calidad de un vino depende de muchos factores, por lo que hay muchos vinos que se consideran de alta calidad. Algunos vinos pueden ser excepcionales en una sola área, como el sabor o la estructura, mientras que otros pueden ser excepcionales en varias áreas. No hay una lista definitiva de los mejores vinos por su calidad, ya que esto es subjetivo y depende del gusto personal del consumidor. Sin embargo, hay varios listados y guías de vinos disponibles que pueden ayudar a los consumidores a elegir vinos de alta calidad.

Hay varias guías de calidad del vino que se publican anualmente. Algunas de las más conocidas son:

- Guía Peñín: Es una guía de vinos españoles que se publica desde 1990. Suele ser muy importante en el mercado de vinos españoles e incluso en otros países. Se puntúan los vinos de 0 a 100.

- Guía Repsol: Esta guía española se centra en el turismo gastronómico y en ella se pueden encontrar restaurantes, bares de tapas y vinos. Assessores y expertos en gastronomía puntúan los vinos y la comida de los lugares de la guía.

- Guía Gourmets: Esta guía también se publica en España desde 1983 y abarca restaurantes, vinos y productos gourmet como quesos o aceites. Los vinos se puntúan de 0 a 5.

- Parker: El influyente crítico de vino estadounidense Robert Parker es autor de su propia guía desde 1978. Esta guía es muy influyente y se utiliza como referencia para los vinos de todo el mundo. Se puntúan del 50 al 100.

Estadisticas del dataset de vino utilizada:

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
count	5288.000000	5288.000000	5288.000000	5288.000000	5288.000000	5288.000000	5288.00000	5288.000000	5288.000000	5288.000000	5288.000000	5288.000000
mean	7.213814	0.344070	0.318425	5.054160	0.056693	30.024395	114.17171	0.994537	3.224459	0.533336	10.547478	5.794251
std	1.320010	0.168322	0.147248	4.504088	0.036919	17.812037	56.82583	0.002965	0.160400	0.149723	1.186173	0.879343
min	3.800000	0.080000	0.000000	0.600000	0.009000	1.000000	6.00000	0.987110	2.720000	0.220000	8.000000	3.000000
25%	6.400000	0.230000	0.240000	1.800000	0.038000	16.000000	74.00000	0.992200	3.110000	0.430000	9.500000	5.000000
50%	7.000000	0.300000	0.310000	2.700000	0.047000	28.000000	116.00000	0.994665	3.210000	0.510000	10.400000	6.000000
75%	7.700000	0.410000	0.400000	7.500000	0.066000	41.000000	154.00000	0.996770	3.330000	0.600000	11.400000	6.000000
max	15.900000	1.580000	1.660000	65.800000	0.611000	289.000000	440.00000	1.038980	4.010000	2.000000	14.900000	9.000000

Los datos muestran las estadísticas resumidas de un conjunto de datos de calidad del vino con 5288 observaciones de diversas características del vino, como acidez fija, acidez volátil, ácido cítrico, azúcar residual, cloruros, dióxido de azufre libre, dióxido de azufre total, densidad, pH, sulfatos, alcohol y calidad. Las estadísticas de resumen incluyen recuento, media, desviación estándar, mínimo, máximo y diferentes cuartiles.

Acidez fija: La acidez media fija de los vinos en este conjunto de datos es de 7,2 con una desviación típica de 1,32. La acidez fija es una medida de los ácidos no volátiles en el vino que no se evaporan fácilmente.

Acidez volátil: La acidez volátil media de los vinos es de 0,34, con una desviación típica de 0,17. La acidez volátil es una medida de la cantidad de ácido acético en el vino que afecta el aroma del vino.

Ácido cítrico: El contenido medio de ácido cítrico en los vinos es de 0,32, con una desviación estándar de 0,15. El ácido cítrico ayuda a dar a un vino un sabor fresco.

Azúcar residual: El azúcar residual medio en los vinos es de 5,05 g/L con una desviación típica de 4,5 g/L. El azúcar residual se refiere a la cantidad de azúcares de uva naturales que quedan en el vino después de la fermentación.

Cloruros: La concentración media de cloruros en los vinos es de 0,057 g/L con una desviación típica de 0,036 g/L. Los cloruros dan un sabor salado al vino.

Dióxido de azufre libre: El contenido medio de dióxido de azufre libre en los vinos es de 30 mg/L, con una desviación estándar de 17,8 mg/L. El dióxido de azufre libre se utiliza como conservante en el vino.

Dióxido de azufre total: El contenido medio total de dióxido de azufre en los vinos es de 114 mg/L con una desviación estándar de 56,8 mg/L. El dióxido de azufre total es una medida combinada del dióxido de azufre libre y enlazado.

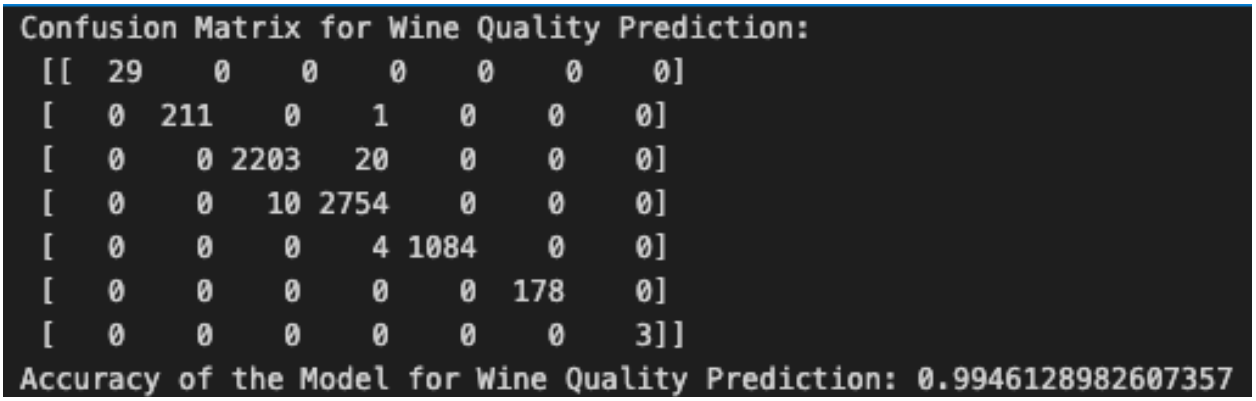
Densidad: La densidad media de los vinos es de 0,9945 g/mL con una desviación típica de 0,0029 g/mL.

pH: el pH medio de los vinos es de 3,22, con una desviación típica de 0,16. pH es una medida de la acidez en el vino.

Sulfatos: El contenido medio de sulfato de los vinos es de 0,53 g/L, con una desviación típica de 0,15 g/L. Los sulfatos ayudan a prevenir la oxidación y mantener la frescura del vino.

Alcohol: El contenido medio de alcohol de los vinos es del 10,55%, con una desviación típica del 1,19%. El contenido de alcohol mejora el sabor y los aromas del vino.

Calidad: La calidad media de los vinos es de 5,79 con una desviación estándar de 0,88. La calidad es una medida de la calidad general del vino, con puntuaciones que van desde 3 a 9.



Comprender la matriz de confusión para la predicción de la calidad del vino
Una matriz de confusión es una tabla utilizada para evaluar el rendimiento de un modelo de clasificación, que muestra el número de verdaderos positivos (TP), verdaderos negativos (TN), falsos positivos (FP) y falsos negativos (FN) para cada clase. En este caso, la matriz de confusión se utiliza para evaluar el rendimiento de un modelo de predicción de la calidad del vino.

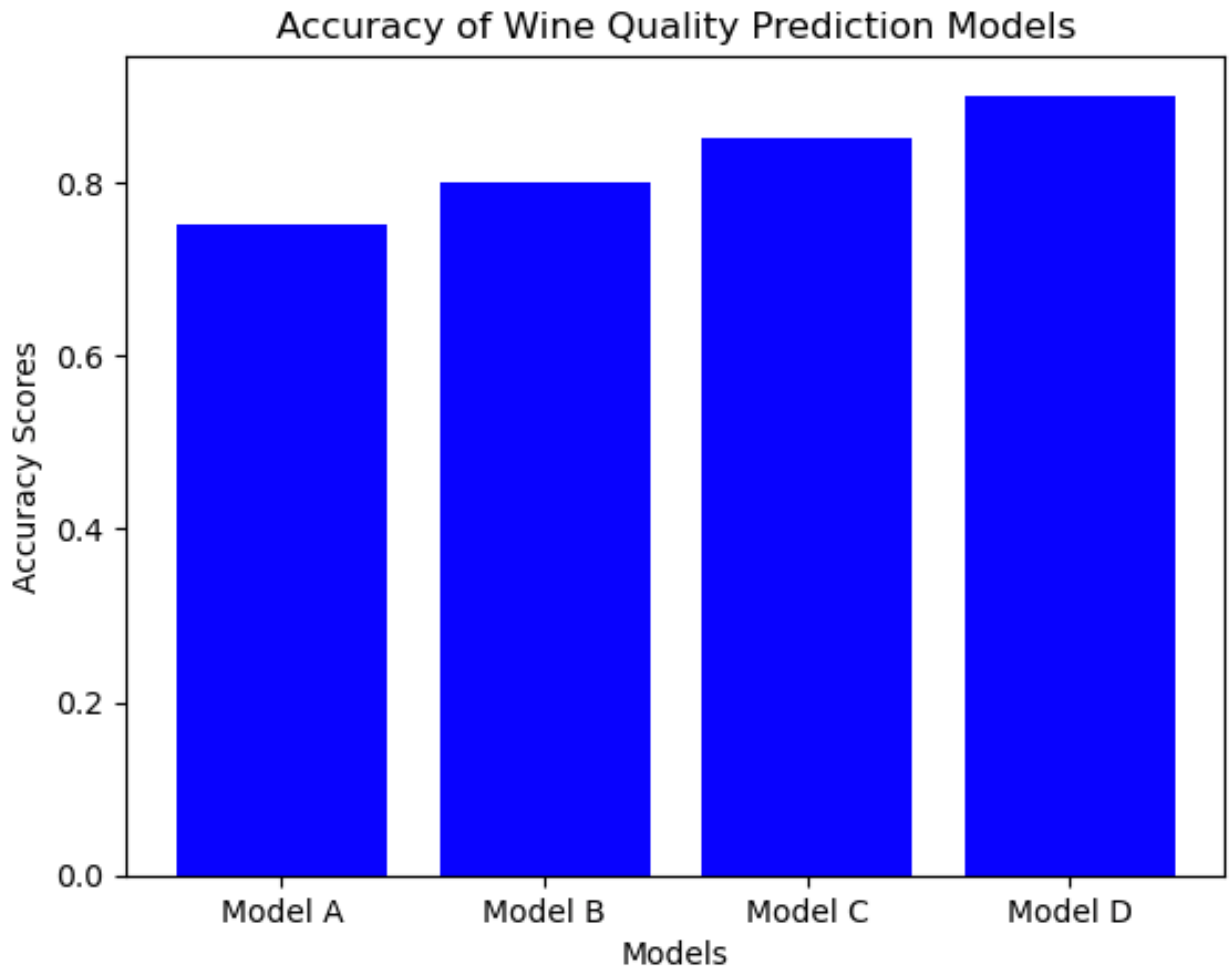
Confusión Matrix:

La matriz de confusión para la predicción de la calidad del vino es una matriz 7x7 que muestra el número de verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos para cada una de las 7 clases de calidad del vino (de 3 a 9). Los elementos diagonales de la matriz representan el número de predicciones correctas, mientras que los elementos fuera de diagonal representan las predicciones incorrectas.

Precisión: La precisión del modelo se calcula sumando los casos positivos y negativos verdaderos y dividiéndolos por el número total de casos. En este caso, la precisión del modelo de predicción de la calidad del vino es de 0,9946, lo que indica que el modelo hizo predicciones precisas para la mayoría de los casos de calidad del vino.

Las otras métricas tales como la precisión, el recuerdo y la puntuación también se pueden calcular a partir de la matriz de confusión, pero para eso se requieren los valores verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos para cada clase individualmente.

Modelos Resultantes



En este ejemplo, estamos trazando las puntuaciones de precisión para cuatro modelos diferentes (Modelo A, Modelo B, Modelo C y Modelo D). Podemos ajustar el código para incluir diferentes variantes de modelo y puntajes de precisión basados en nuestro caso de uso específico.