Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2022.Doi Number

Predecir el origen de un vino a partir de sus componentes químicos

Daniel Felipe Ospina Pérez1, Sebastián Castellanos Mejía1

1Departamento de Ingeniería de Sistemas, Universidad de Antioquia, Medellín, Colombia

Daniel Felipe Ospina Pérez (e-mail: daniel.ospinap@udea.edu.co).

Sebastián Castellanos Mejía (e-mail: sebastian.castellanosm@udea.edu.co).

RESUMEN En este informe, se aborda el problema de clasificación de diferentes tipos de vinos utilizando la base de datos de vinos de la Universidad de Irvine, California (UCI, por sus siglas en inglés). Se describen las variables de entrada, la variable a predecir y se exploran las características del conjunto de datos. Además, se realiza una revisión de la literatura sobre trabajos relacionados con el mismo problema de clasificación.

1. INTRODUCCIÓN

El problema de clasificación de vinos tiene aplicaciones significativas en la industria vitivinícola y en la enología. El objetivo es desarrollar un modelo de aprendizaje automático que pueda clasificar los vinos en diferentes categorías basadas en sus características químicas. En este informe, se analizarán las variables de entrada, la variable a predecir y se revisarán trabajos relacionados.

1. DESCRIPCIÓN DEL PROBLEMA

El problema de predicción abordado implica clasificar diferentes tipos de vino en función de sus propiedades químicas. Esta tarea consiste en identificar el tipo de vino al que pertenece la muestra en función de sus propiedades químicas específicas. Los tres tipos de vino en cuestión se denominan “Tipo 1”, “Tipo 2” y “Tipo 3”.

Este problema de clasificación es relevante para el campo de la enología (el estudio del vino) y la industria del vino. Clasificar los vinos en función de sus propiedades químicas podría tener aplicaciones prácticas en la producción y comercialización del vino. Por ejemplo, un enólogo o una bodega puede utilizar un modelo de clasificación para identificar y etiquetar automáticamente los vinos en función de su calidad o características distintivas.

En este caso, se está resolviendo un problema de clasificación. La razón es que se asigna cada muestra (cada botella de vino en el conjunto de datos) a una de tres categorías distintas: “Tipo 1”, “Tipo 2” o “Tipo 3”. Las tareas de clasificación implican predecir la pertenencia a una categoría o clase particular, a diferencia de las tareas de regresión, en las que se predice un valor numérico continuo.

1. BASE DE DATOS

Para este problema, la base de datos utilizada consta de 178 muestras con 13 características, junto con una variable de salida que se utiliza para clasificar en uno de los tres tipos de vino. En la tabla 1 se describen las características de las muestras.

TABLA 1

CARACTERÍSTICAS QUÍMICAS PARA IDENTIFICAR UN VINO

|  |  |  |  |
| --- | --- | --- | --- |
| Nombre | Rol | Tipo | |
| class | Objetivo | | Categórica |
| Alcohol | Característica | | Continuo |
| Malicacid | Característica | | Continuo |
| Ash | Característica | | Continuo |
| Alcalinity\_of\_ash | Característica | | Continuo |
| Magnesium | Característica | | Entero |
| Total\_phenols | Característica | | Continuo |
| Flavanoids | Característica | | Continuo |
| Nonflavanoid\_phenols | Característica | | Continuo |
| Proanthocyanins | Característica | | Continuo |
| Color\_intensity | Característica | | Continuo |
| Hue | Característica | | Continuo |
| 0D280\_0D315\_of\_diluted\_wines | Característica | | Continuo |
| Proline | Característica | | Entero |

La base de datos de vinos de UCI no contiene valores faltantes, lo que significa que no se requiere ningún proceso de imputación de datos o eliminación de instancias con datos faltantes en este caso.

1. EXPERIMENTOS

En el contexto de los experimentos, se empleó un conjunto de datos que comprende 178 muestras distribuidas en tres categorías de vino, como se detalla en la tabla 2.

TABLA 2

DISTRIBUCIÓN DE VINOS EN TIPOS

|  |  |
| --- | --- |
|  | Número de muestras |
| Tipo 1 | 59 |
| Tipo 2 | 71 |
| Tipo 3 | 48 |

Dicho conjunto de datos se estandarizó y se dividió en dos subconjuntos: el primero se destinó al proceso de entrenamiento, abarcando el 70% de las muestras, mientras que el segundo se reservó para la evaluación de los modelos, con un 30% de las muestras.

Para abordar la resolución de este problema de clasificación, se consideraron tres modelos potenciales: Redes Neuronales Artificiales, Gradiente Boosting Tree y Análisis Discriminante Cuadrático. Se usó Grid Search como método de comparación y como métrica principal de evaluación del rendimiento de cada modelo, se empleó la precisión (accuracy).

Para el modelo de Redes Neuronales Artificiales, se configuraron los siguientes hiperparámetros: el número máximo de iteraciones se estableció en 10,000, y se llevaron a cabo experimentos con diferentes configuraciones para el tamaño de las capas ocultas, incluyendo 1 y 2 capas con 10, 20, 30, 40 y 50 neuronas en cada capa. Estos resultados se pueden ver en la tabla 3.

TABLA 3

RESULTADOS MODELO DE REDES NEURONALES ARTIFICIALES

|  |  |  |
| --- | --- | --- |
| Número de capas | Número de neuronas por capa | Precisión (accuracy) |
| 1 | 10 | 0.967 |
| 1 | 20 | 0.975 |
| 1 | 30 | 0.975 |
| 1 | 40 | 0.967 |
| 1 | 50 | 0.967 |
| 2 | 10 | 0.967 |
| 2 | 20 | 0.975 |
| 2 | 30 | 0.983 |
| 2 | 40 | 0.975 |
| 2 | 50 | 0.967 |

Para el modelo de Gradiente Boosting Tree, se realizaron experimentos con tres configuraciones diferentes: 100, 200 y 300 estimadores, y se exploraron dos opciones de características, considerando 3 y 5 características en cada conjunto de datos. Los resultados de estos experimentos se pueden ver en la tabla 4.

TABLA 4

RESULTADOS MODELO DE GRADIENTE BOOSTING TREE

|  |  |  |
| --- | --- | --- |
| Número máximo de características | Número de estimadores | Precisión (accuracy) |
| 3 | 100 | 0.975 |
| 3 | 200 | 0.975 |
| 3 | 300 | 0.975 |
| 5 | 100 | 0.967 |
| 5 | 200 | 0.959 |
| 5 | 300 | 0.959 |

Para el modelo de Análisis Discriminante Cuadrático, se llevaron a cabo experimentos variando el hiperparámetro de "guardado de covarianza", probando tanto con el valor "True" como "False". Los resultados de estos experimentos se encuentran en la tabla 5.

TABLA 5

RESULTADOS MODELO DE GRADIENTE BOOSTING TREE

|  |  |
| --- | --- |
| Guardado de covarianza | Precisión (accuracy) |
| False | 0.951 |
| True | 0.951 |

En todas las configuraciones de hiperparámetros para los modelos, se observa que la precisión supera el 95%. Además, la variación entre las diferentes configuraciones es mínima. Por lo tanto, se opta por elegir la configuración que proporciona el modelo más sencillo, ya que logra un alto rendimiento sin complicaciones innecesarias.

Al evaluar estos modelos con el conjunto de datos de evaluación, se obtienen los resultados que se presentan en la tabla 6.

TABLA 6

RESULTADOS MEJOR MODELO

|  |  |
| --- | --- |
| MODELO | Precisión (accuracy) |
| Redes Neuronales Artificiales  (1 capa, 10 neuronas) | 0.926 |
| Gradiente Boosting Tree  (3 características, 100 estimadores) | 0.981 |
| Análisis Discriminante Cuadrático  (Guardado de covarianza: False) | 1.000 |

La evidencia de estos experimentos puede ser hallada en el repositorio de este, ver [1].

REFERENCIAS

1. Danielospinap, “DANIELOSPINAP/Wines,” GitHub, https://github.com/danielospinap/wines/tree/main (accessed Nov. 6, 2023).
2. SebastianCastellanosM, ClasificacionVinosProyecto. GitHub. “https://github.com/SebastianCastellanosM/ClasificacionVinosProyecto.git”