UNIVERSIDAD DE MONTEVIDEO

CURSO DE APRENDIZAJE AUTOMÁTICO

TAREA 1

Para el presente trabajo, se utilizará el dataset Red Wine Quality [1]. Este dataset contiene datos asociadas a variantes de vino portugués. De acuerdo con el artículo [2], actualmente la calidad del vino se evalúa mediante pruebas fisicoquímicas – tal como los niveles de alcohol – y evaluaciones de expertos humanos. En el artículo, los autores proponen un enfoque basado en ciencia de datos para predecir las preferencias del vino, apoyándose en pruebas analíticas que son accesibles en las fases de certificación del vino. Las pruebas analíticas incluyen variables tales como el nivel de PH, cantidad de azúcar luego del proceso de fermentación, cantidad de sal, entre otras. De esta forma, se busca obtener herramientas de carácter objetivo para determinar la calidad de un vino.

Parte 1

Efectúe un análisis del dataset, incluyendo un estudio de correlación entre las variables de este. Explique detalladamente los métodos utilizados. Para obtener una representación gráfica de la correlación entre variables, se recomienda utilizar el método heatmap del paquete Seaborn [3]. Discuta los resultados obtenidos.

Parte 2

Partiendo de los atributos de características del vino utilice el método de regresión lineal para predecir la calidad de este. Analice los resultados obtenidos. Revise y discuta si existen mecanismos para mejorar las métricas que arroja este modelo.

Parte 3

Utilizando los métodos de clasificación vistos en el curso, se pretende encontrar el mejor modelo que permita predecir, a partir de las características disponibles en el dataset [1], la calidad del vino. La variable de calidad debe de convertirse a una clase de 2 valores: bueno y malo. Será parte del problema definir el umbral de corte para determinar esta clase, partiendo de las categorías definidas en [1].
Utilizando los métodos y métricas vistas en el curso, compare los resultados obtenidos. Justifique cuales métricas resultan más relevantes para el problema que está resolviendo.

Para todos los métodos utilizados, estudie la relevancia de cada uno de los atributos o *features*.

Parte 4

Explique las bases teóricas del algoritmo k-NN (*k-nearest neighbors*) y aplíquelo al dataset [1]. Compare con los resultados obtenidos en la parte 3.

Parte 5

Ensaye una discusión general del trabajo realizado, haciendo los comentarios y recomendaciones que considere necesarias.

Consideraciones generales

- Como ambiente de trabajo puede utilizar Jupyter Notebooks o Google Colab
- Si usa Jupyter Notebooks, el lenguaje de programación seleccionado debe ser Python
- Se recomienda fuertemente el uso del paquete scikit-learn
- Cada grupo debe entregar
    - Notebook
    - En caso de que documente fuera del notebook, entregar el documento anexo en formato PDF
    - Siempre incluya número de grupo, junto con el nombre de cada uno de los estudiantes que lo integran
- **El plazo de entrega vence el sábado 24 de setiembre a las 23:59**. La entrega se efectuará vía Moodle.

Bibliografía

[1] Kaggle, «https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009,» UCI Machine Learning. [En línea]. [Último acceso: 30/03/2021].

[2] P. &. T. J. &. C. A. &. A. F. &. M. T. &. R. J. Cortez, «Using Data Mining for Wine Quality Assessment,» 2009.

[3] Seaborn, «seaborn: statistical data visualization,» [En línea]. Available: https://seaborn.pydata.org/. [Último acceso: 30/03/2021].