PRÁCTICA 2: Limpieza y Análisis de Datos

Tipología y ciclo de vida de los datos

Máster en Ciencia de Datos

Daniel Laureano Cerviño Cortínez
DNI:45783258X

Universitat Oberta de Catalunya

Tabla de contenido

Asignatura	Código	Fecha inicio	Fecha fin
Tipología y ciclo de vida de los datos	M2.851	09/10/2020	09/11/2020

Preguntas de la Práctica 2

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

Me he decantado por utilizar el conjunto de datos proveniente del repositorio de datos sobre Machine Learning de UCI basado en la calidad de los vinos.

El juego de datos está compuesto por 2 conjuntos, uno relacionado con vinos rojos y otro con vinos blancos. En total, existen 1599 observaciones asociadas a vinos rojos y 4898 asociadas a vinos blancos.

Los atributos que tenemos disponibles para cada uno de ellos son los siguientes:

- ***fixed.acidity.*** Acidez fija.
- ***volatile.acidity.*** Acidez volátil.
- ***citric.acid.*** Ácido crítico.
- ***residual.sugar.*** Azúcar residual
- ***chlorides.*** Cloruros.
- ***free.sulfur.dioxide.*** Dióxido de azufre libre.
- ***total.sulfur.dioxide.*** Dióxido de azufre total.
- ***density.*** Densidad.
- ***pH.*** Potencial de Hidrógeno.
- ***sulphates.*** Sulfatos.
- ***alcohol.*** Alcohol.
- ***quality.*** Grado de calidad.

Además, para la realización de este proyecto vamos a incluir el atributo ***type*** con el fin de conocer si existen diferencias sustanciales entre vinos blancos y rojos.

- ***type***. Donde el valor 2 se trata de vinos blancos y el valor 1 se trata de un vino rojo.

Cabe destacar que las clases de calidad de los vinos se encuentran ordenadas, aunque no se encuentran equilibradas.

Se trata de un dataset que puede ser emplear en términos reales para conocer cuáles son las propiedades características de cada uno de los tipos de vinos disponibles. Esta acción es muy importante para las empresas que elaboran vinos y buscan mejorar sus productos.

A partir del conjunto de datos comentado anteriormente, se pretende conocer las características de cada uno de los tipos de vinos. Así como, cuáles son aquellas características más relevantes para determinar cuándo un vino es considerado rojo o blanco.

Para nuestro proyecto, pretendemos elaborar un clasificador a través de distintos algoritmos de clasificación para distinguir entre vinos rojos y blancos. Cabe destacar que para ello, utilizaremos los componentes ya reducidos del dataset original.

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

Para la selección de los datos, procedimos a emplear el método de reducción de dimensionalidad denominado **PCA** (Principal Component Analysis) con el fin de quedarnos con una serie de atributos representativos y 2 atributos que se podrían utilizar para tareas de clasificación, como son el grado de calidad ***quality*** y el tipo del vino que estamos tratando a través del atributo ***type***. Obtuvimos aquellos atributos que explican un 73.19% de la totalidad de la muestra.

3. Limpieza de los datos.

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

El juego de datos no presenta ningún tipo de valor nulo o vacío. Aunque depende de cuántos atributos se encuentren vacíos en la observación, a priori, completaría dichos datos a través de una media dependiendo del tipo de calidad del vino y de si es blanco o rojo.

3.2. Identificación y tratamiento de valores extremos.

En el juego de datos se encontraron valores extremos u outliers, aunque cabe destacar que estos valores son considerados legítimos, y por lo cual no se tendría que realizar ninguna modificación en ellos. La detección de estos valores se realizó a través de una serie de visualizaciones como por ejemplo, diagramas de caja o bigote (**boxplot**), gráficos de dispersión de puntos (**scatter plot**), etc. La razón de estos valores atípicos puede ser consecuencia de que los datos se encuentren desequilibrados con relación a los distintos tipos de calidad de vinos. Por lo cual, no sería necesario eliminar estas observaciones atípicas debido a que forman parte del dominio de datos de cada una de las variables y las consideramos como legítimas para su estudio. Esto se puede observar en la propia descripción del dataset. Si se eliminaran estas observaciones, se produciría un sesgo debido a que sólo obtendríamos vinos de una calidad, o baja o alta.

4. Análisis de los datos.

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

Como comenté anteriormente, se seleccionaron aquellos atributos generados a través del método **PCA**, ***quality*** y ***type***.

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

Para comprobar el test de normalidad de los atributos, se realizaron los test de **Lilliefors** y **Anderson-Darling.** Cabe destacar que todos los atributos considerados no cumplen o siguen una distribución normal.

Para comprobar la homocedasticidad de la variación en relación a los distintos atributos, se realizó el test de **Fligner-Killen**. Se realizó este test o prueba debido a que los atributos no siguen una distribución normal. Tras realizar este test, se concluyó que las variables ***PC4*** y ***quality***, y ***quality*** y ***type*** presentan homocedasticidad o igualdad de varianzas.

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.

A lo largo del proyecto, se ha utilizado la matriz de correlación de Pearson para conocer las distintas relaciones entre los atributos. Se ha empleado con los atributos iniciales del dataset, como con los del dataset con las variables ya reducidas.

Además, se han aplicado algoritmos empleados para la clasificación entre vinos rojos y blancos. Para ello, hemos utilizado la función C5.0 como algoritmo basado en árboles de decisión, y regresiones lineales y logísticas.

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

Para representar los resultados del algoritmo de clasificación, se utilizó la curva **ROC** y la **matriz de confusión** para valorar el rendimiento del mismo.

Para comprobar la correlación entre variables, se visualizó una serie de matrices de confusión enfatizando las relaciones entre variables.

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

El objetivo principal del proyecto se ha cumplido, que se trató de crear un clasificador entre vinos rojos y blancos ateniendo a sus características. Se ha podido responder al problema debido a que gracias al algoritmo basado en árboles de decisión se ha obtenido un clasificador con un 92.06% de exactitud

en sus predicciones. Cabe destacar que el rendimiento del algoritmo se ha basado en el dataset con los atributos reducidos, que a priori, ofrecen un peor desempeño que el conjunto de datos total.

Como futuras mejoras del proyecto, se podrían realizar otras acciones con el dataset que pueden mejorar el análisis del proyecto como es el uso de algoritmos de aprendizaje no supervisado para conocer las características de los vinos con mayor profundidad.

7. Código: Hay que adjuntar el código, preferiblemente en R, con el que se ha realizado la limpieza, análisis y representación de los datos. Si lo preferís, también podéis trabajar en Python.

El código se encuentra disponible en el siguiente repositorio de GitHub.

Contribuciones al trabajo

Contribuciones	Firma
Investigación previa	DLCC
Redacción de las respuestas	DLCC
Desarrollo código	DLCC

Bibliografía

Amat Rodrigo, J. (2017). *Análisis de Componentes Principales (Principal Component Analysis, PCA) y t-SNE*. https://rpubs.com/Joaquin_AR/287787

Analytics Vidhya. (2016). *PCA: A Practical Guide to Principal Component Analysis in R & Python*. https://www.analyticsvidhya.com/blog/2016/03/pca-practical-guide-principal-component-analysis-python/

Basics, T., Analysis, P. C., Procedure, V. R., Illustration, A., Redundancy, V., Component, P., Solutions, O., Analysis, N. F., Inventory, P. O., Instrument, M., Adequate, M., Size, S., Program, S. A. S., Data, T., Factor, T. P., Factor, P., Var, T., Principal, C., Analysis, C., … Description, F. (1994). *Introduction : The Basics of Principal Component Analysis*. 1–56. https://www.researchgate.net/profile/Ehsan_Khedive/post/How_many_components_can_I_retrieve_in_principal_component_analysis/attachment/59d626f2c49f478072e9b1be/AS%3A272185124425729%401441905398541/download/Principal+Component+Analysis+SAS.pdf

Buhler, P., Khattak, W., & Erl, Th. (2018). Chapter 1. Understanding Big Data. In *Big Data Fundamentals: Concepts, Drives & Techniques*. https://learning.oreilly.com/library/view/knowledge-management-toolkit/0130128538/ch01.html

Cortez, P., A., C., F., A., T., M., & J., R. (2009). *Wine Quality*. https://archive.ics.uci.edu/ml/datasets/Wine+Quality

Ngo, L. (2018). *How to read PCA biplots and scree plots*. https://blog.bioturing.com/2018/06/18/how-to-read-pca-biplots-and-scree-plots/

Rickert, J. (2019). *Some R Packages for ROC Curves*. https://rviews.rstudio.com/2019/03/01/some-r-packages-for-roc-curves/

Subirats Maté, L., Pérez Trenard, D. O., & Calvo González, M. (2019). *Introducción a la limpieza y análisis de los datos*. http://cv.uoc.edu/annotation/7206a0ab8340f28c6ecc25db3e5a4715/755822/PID_00265704/PID_00265704.html

VanderPlas, J. (2016). *In Depth: Principal Component Analysis*. https://jakevdp.github.io/PythonDataScienceHandbook/05.09-principal-component-analysis.html