Actividad: PRA2: Limpieza y análisis de datos

Reison A. Torres Urina

Diciembre 2019

Índice

1. Miembros del equipo

La actividad ha sido realizada de manera individual por:
Reison Arturo Torres Urina.

2. Detalles de la actividad

2.1. Descripción

En esta practica se elabora un caso práctico orientado a aprender a identificar los datos relevantes para un proyecto analítico y usar las herramientas de integración, limpieza, validación y análisis de las mismas.

2.2. Objetivos

Los objetivos concretos de esta práctica son:

- Aprender a aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios o multidisciplinares.
- Saber identificar los datos relevantes y los tratamientos necesarios (integración, limpieza y validación) para llevar a cabo un proyecto analítico.
- Aprender a analizar los datos adecuadamente para abordar la información contenida en los datos.

- Identificar la mejor representación de los resultados para aportar conclusiones sobre el problema planteado en el proceso analítico.
- Actuar con los principios éticos y legales relacionados con la manipulación de datos en función del ámbito de aplicación.
- Desarrollar las habilidades de aprendizaje que les permitan continuar estudiando de un modo que tendrá que ser en gran medida autodirigido o autónomo.
- Desarrollar la capacidad de búsqueda, gestión y uso de información y recursos en el ámbito de la ciencia de datos.

2.3. Competencias

En esta práctica se desarrollan las siguientes competencias del Máster de Data Science:

- Capacidad de analizar un problema en el nivel de abstracción adecuado a cada situación y aplicar las habilidades y conocimientos adquiridos para abordarlo y resolverlo.
- Capacidad para aplicar las técnicas específicas de tratamiento de datos (integración, transformación, limpieza y validación) para su posterior análisis.

3. Solución

3.1. Descripción del dataset

Los datos seleccionados, fueron obtenidos del sitio de data science, www.Kaggle.com, en el encontramos una variedad de dataset Open Data. El conjunto de datos seleccionados para desarrollar esta actividad es Red Wine Quality, en este dataset encontramos la pruebas fisicoquimica y sensoriar de la variedad de vino roja y blanco del vino Portuges "Vinho Verde".

En este dataset encontramos mas de 4898 observaciones, y 12 variables. Las característica presenten en este dataset, nos permitirá cumplir los objetivos propuestos en esta actividad.

A continuación describimos el conjunto de variables que describen este dataset:

- **fixed acidity**: Nivel de acides que se encuentra el vino. Los ácidos son los principales componentes del vino y contribuyen en gran medida a su sabor.
- **volatile acidity:** Nivel de acides volátil presente en el vino.
- **citric acid:** Nivel de ácido cítrico. Aumentar la acidez, o le da un sabor específico.
- **residual sugar:** Cantidad de azúcar en el vino luego que se detiene la fermentación.
- **chlorides:** Cantidad de sal en el vino.
- **free sulfur dioxide:** Niveles de conservantes o antioxidante en el vino.
- **total sulfur dioxide:** Cantidad de free sulfur dioxide unidads al SO2 del vino.
- **density:** Cantidad de azúcar y alcohol contenida en el vino.

- **pH**: Describe que tan ácido o básico es un vino, en una escala de 0 (muy ácido) a 14 (muy básico); la mayoría de los vinos están entre 3-4 en la escala de pH.
- **sulphates**: Cantidad de conservante para prevenir la oxidación y mantener la frescura del vino.
- **alcohol**: Cantidad de alcohol contenida en el vino.
- **quality**: Indica la calidad del vino, determinada por datos sensoriales (escala entre 0 y 10).

3.2. Importancia y objetivos de los análisis

Descripción: Este conjunto de datos contiene el listado de productos y precios que se comercializan diariamente en los puntos de venta de las diferentes Bodegas de la corporación CORABASTOS y dichos precios están dirigido al consumidor.

3.3. Contenido

La base datos de productos y precios que se comercializan a diario al consumidor, publicados por la entidad CORABASTOS, está constituido por variables numéricas y de textos. A continuación daremos una pequeña descripción, acerca de lo que representa, cada una de las variables, que se encuentra en este conjunto de datos.

1. **Grupo**: Nombre del grupo al que pertenece un producto (HORTALIZAS, FRUTAS, TUBERCULOS, PLATANOS, GRANOS Y PROCESADOS, LACTEOS, CARNICOS y HUEVOS) . (Texto).
2. **Nombre**: Nombre o descripción del producto. (Texto).
3. **Presentación**: Presentaciones de venta en la que viene el producto. (Texto).
4. **Cantidad**: Número de unidades que vienen por presentación. (Numérico).
5. **Unidad**: Unidad de medida en la que se vende el producto. (Texto).
6. **Cal_Extra**: Precio máximo de venta. (Numérico).
7. **Cal_Primera**: Precio mínimo de venta. (Numérico).
8. **Valor**: Precio de venta. (Numérico).
9. **Fecha_publicacion**: Fecha de publicacion del precio del producto. (Fecha).

Estos datos son publicados diariamente solo en los días hábiles de L-V.

3.4. Agradecimientos

Los datos han sido recolectados desde la base de datos online AppCoraPrecios. Para ello, se ha hecho uso del lenguaje de programación Python y de técnicas de *Web Scraping* para extraer la información alojada en las páginas HTML.

3.5. Inspiración

El presente conjunto de datos podría utilizarse en el ámbito comercial, donde se podría elaborar modelos predictivos, que nos ayuden a predecir el precio del producto en el futuro, y con esto poder preparar estrategias

de markiting.

También podría ser de gran utilidad en el campo de la *Agricultura*, para informar al pequeño productor de las temporadas de mayor de manda de sus productos en el mercado nacional, para que puedan preparar sus cosechas para suplir esta demanda.

3.6. Licencia

Para la publicación de este conjunto de datos se seleccionó la licencia **CC BY-SA 4.0 License**. Los motivos por los cual se selecciono esta licencia son:

- *Los trabajos derivadas del conjunto de datos generado, su distribución se debe hacer con una licencia igual a la que regula el trabajo original.* Con esto garantizamos que los trabajos derivados del trabajo original, seguirán distribuyéndose bajo los mismo términos que el autor original planteo.
- *Se debe dar crédito de manera adecuada al creador del conjunto de datos generado, e indicar si se han realizado cambios.* De esta manera damos crédito del trabajo del autor original, y mantenemos una transparencia en la medida que se indican las aportaciones/cambios realizados al trabajo original.
- *Explotación de los datos generados, incluyendo una finalidad comercial.* Con esto garantizamos que la utilización de los datos generados, no serán de uso privativo, dando la posibilidad que otras empresas utilicen los datos generados y realicen trabajos de calidad que mantenga su competitividad en el mercado.

3.7. Código fuente y dataset

- En la siguiente ruta (**src/foodPriceCorabasto.py**) encontraras el código Python desarrollado.
- En la siguiente ruta(**src/foodPriceCorabasto.ipynb**) encontraras el código Python desarrollado en formato Jupyter Notebook.
- En la siguiente ruta(**data/**) encontraras los dataset generados en formato CSV.

4. Recursos

1. Subirats, L., Calvo, M. (2018). Web Scraping. Editorial UOC.

2. Creative Commons. (2016). "Attribution-ShareAlike 4.0 International (CC BY-SA 4.0)". Creative Commons public licenses [artículo en línea]. [Fecha de consulta: 22 de octubre del 2019]. https://creativecommons.org/licenses/by-sa/4.0/legalcode.

3. Lawson, R. (2015). Web Scraping with Python. Packt Publishing Ltd.

4. Masip, D. El lenguaje Python. Editorial UOC.

5. Willems, Karlijn. (19 Marzo 2019). "Python Numpy Array Tutorial". DataCamp [artículo en línea]. [Fecha de consulta: 26 de octubre del 2019]. https://www.datacamp.com/community/tutorials/python-numpy-tutorial#make.

6. Willems, Karlijn. (17 Enero 2019). "Pandas Tutorial: DataFrames in Python". DataCamp [artículo en línea]. [Fecha de consulta: 27 de octubre del 2019]. https://www.datacamp.com/community/tutorials/pandas-tutorial-dataframe-python#question3.