Práctica 2 (25% nota final)

Presentación

En esta práctica 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.

Deberéis trabajar en grupos de 2 personas y entregar **un solo archivo** con el enlace al repositorio Git donde se encuentren las soluciones, incluyendo los nombres de los componentes del equipo. Podéis utilizar la Wiki de Github para describir vuestro equipo y los diferentes archivos que corresponden a vuestra entrega. Cada miembro del equipo tendrá que contribuir con su usuario Github.

Además, se debe entregar un **vídeo explicativo** de la práctica, donde ambos integrantes del equipo expliquen con sus propias palabras el desarrollo de la práctica, basándose en las preguntas del enunciado para justificar y explicar el código desarrollado. Este vídeo se deberá entregar a través de un enlace a Google Drive que se deberá proporcionar junto con enlace al repositorio Git.

Es importante que la entrega de esta práctica se realice en el formato especificado en el apartado Formato y fecha de entrega.

Aunque no se trata exactamente del mismo enunciado ni de una solución que obtuviera la máxima nota, el siguiente ejemplo de una edición anterior os puede servir de guía para la realización de la práctica: https://github.com/Bengis/nba-gap-cleaning

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.

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.

Descripción de la Práctica a realizar

El objetivo de esta actividad será el tratamiento de un dataset, que puede ser el creado en la Práctica 1 o bien cualquier dataset libre disponible en Kaggle https://www.kaggle.com.

Un ejemplo de dataset con el que podéis trabajar es el "Heart Attack Analysis & Prediction dataset":
https://www.kaggle.com/datasets/rashikrahmanpritom/heart-attack-analysis-prediction-dataset

Importante: si se elige un dataset diferente al propuesto es importante que este contenga una amplia variedad de datos numéricos y categóricos para poder realizar un análisis más rico y poder responder a las diferentes preguntas planteadas en el enunciado de la práctica.

Siguiendo las principales etapas de un proyecto analítico, las diferentes tareas a realizar (y **justificar**) son las siguientes:

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

2. **Integración y selección** de los datos de interés a analizar**.** Puede ser el resultado de adicionar diferentes datasets o una subselección útil de los datos originales, en base al objetivo que se quiera conseguir.

3. **Limpieza de los datos.**
   3.1. ¿Los datos contienen ceros o elementos vacíos? Gestiona cada uno de estos casos.
   3.2. Identifica y gestiona los valores extremos.

4. **Análisis de los datos.**
   4.1. Selección de los grupos de datos que se quieren analizar/comparar (p. ej., si se van a comparar grupos de datos, ¿cuáles son estos grupos y qué tipo de análisis se van a aplicar?)
   4.2. Comprobación de la normalidad y homogeneidad de la varianza.
   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.

5. **Representación de los resultados** a partir de tablas y gráficas. Este apartado se puede responder a lo largo de la práctica, sin necesidad de concentrar todas las representaciones en este punto de la práctica.

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

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.

8. **Vídeo.** Realizar un breve vídeo explicativo de la práctica **(máximo 10 minutos)**, donde ambos integrantes del equipo expliquen con sus propias palabras el desarrollo de la práctica, basándose en las preguntas del enunciado para justificar y explicar el código desarrollado. Este vídeo se deberá entregar a través de un enlace al Google Drive de la UOC (https://drive.google.com/…), junto con enlace al repositorio Git entregado.

Recursos

Los siguientes recursos son de utilidad para la realización de la práctica:

- Calvo M., Subirats L., Pérez D. (2019). Introducción a la limpieza y análisis de los datos. Editorial UOC.
- Megan Squire (2015). *Clean Data*. Packt Publishing Ltd.
- Jiawei Han, Micheine Kamber, Jian Pei (2012). *Data mining: concepts and techniques*. Morgan Kaufmann.
- Jason W. Osborne (2010). *Data Cleaning Basics: Best Practices in Dealing with Extreme Scores.* Newborn and Infant Nursing Reviews; 10 (1): pp. 1527-3369.
- Peter Dalgaard (2008). *Introductory statistics with R*. Springer Science & Business Media.
- Wes McKinney (2012). *Python for Data Analysis*. O'Reilley Media, Inc.
- Tutorial de Github https://guides.github.com/activities/hello-world.
- Herramienta para realización de gráficas: https://www.data-to-viz.com/

Criterios de valoración

Todos los apartados son obligatorios. La ponderación de los ejercicios es la siguiente:

Apartado	1	2	3	4	5	6	7	8
Puntos	0,5	0,5	2	2,5	1,5	0,5	2	0,5

Se valorará la idoneidad de las respuestas, que deberán ser claras y completas. Las diferentes etapas deberán justificarse y acompañarse del código correspondiente. También se valorará la síntesis y claridad, a través del uso de comentarios, del código resultante, así como la calidad de los datos finales analizados.

Formato y fecha de entrega

En referencia a la entrega final, se pide:

1. **Un único documento** (.txt, .pdf, .docx) que contenga **el enlace al repositorio Git** del proyecto (apartado b) y **el enlace al vídeo del proyecto** (apartado c). Este documento se entregará en el espacio de Entrega y Registro de EC del aula.

2. Un **repositorio Git** con las soluciones de la práctica. El repositorio Git se creará en Github (https://github.com/), y podrá ser un repositorio público o privado, a elección del grupo. Si se utiliza un repositorio privado, se deberá facilitar acceso al profesor, mediante el nombre de usuario que indicará en el Tablón del aula o por email. Es responsabilidad del estudiante asegurarse de que, en el momento de la entrega de la práctica, **se ha dado acceso al profesor a los diferentes elementos privados que se entreguen** (p. ej., repositorio GitHub privado o archivos restringidos de Google Drive).
   El repositorio no se podrá modificar pasada la fecha de entrega, y deberá contener:

   1. Un **README.md** con los nombres de los componentes del grupo y una descripción de los ficheros.
   2. Un **documento PDF** con las respuestas a las preguntas y los nombres de los componentes del grupo. **La extensión de este documento no debe superar las 20 páginas.** Además, al final del documento, deberá aparecer la siguiente tabla de contribuciones al trabajo, la cual debe firmar cada integrante del grupo con sus iniciales. Las iniciales representan la confirmación de que el integrante ha participado en dicho apartado. Todos los integrantes deben participar en cada apartado, por lo que, idealmente, los apartados deberían estar firmados por todos los integrantes.

Contribuciones	Firma
Investigación previa	Integrante 1, Integrante 2
Redacción de las respuestas	Integrante 1, Integrante 2
Desarrollo del código	Integrante 1, Integrante 2
Participación en el vídeo	Integrante 1, Integrante 2

3. Una carpeta con el **código generado** para analizar los datos.
4. El **fichero CSV con los datos originales.**
5. El **fichero CSV con los datos finales analizados.**

3. Un **breve vídeo** con la participación de los dos componentes del grupo, donde se realizará una presentación del proyecto, destacando los puntos más relevantes. El vídeo se deberá compartir mediante un enlace del Google Drive de la UOC. **La duración de este vídeo no debe superar los 10 minutos.**

Es responsabilidad del estudiante revisar que el fichero entregado es el correcto. Un fichero vacío o no pertinente se considerará como no entregado. Asimismo, para que una entrega se considere como realizada, se debe completar al menos el 25% de la actividad.

Este documento de entrega final de la Práctica 2 se debe entregar en el espacio de Entrega y Registro de AC del aula antes de las **23:59h CET** del día **13 de enero del 2023**. No se aceptarán entregas fuera de plazo.

Esta actividad es obligatoria. No entregarla en fecha y forma implica automáticamente el suspenso de la asignatura.

Si se estima oportuno, el profesor solicitará a los integrantes del grupo una entrevista remota (de forma conjunta o individual) mediante Google Meet, en referencia a la práctica realizada, en un día y hora acordados.