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 los mismos. Para realizar esta práctica se requiere trabajar en **grupos de dos personas**.

La entrega de esta práctica se ha de realizar según lo especificado en el apartado Formato y fecha de entrega. Se debe entregar una memoria PDF con las respuestas a los diferentes apartados, un repositorio con el código fuente y un vídeo explicativo, en el que ambos integrantes del grupo comenten los aspectos más relevantes del proyecto.

Es importante tener en cuenta las siguientes consideraciones a la hora de entregar la práctica:

- Es obligatorio y **queda como responsabilidad de cada estudiante revisar que el archivo entregado es el correcto.** Un archivo vacío o no pertinente se considerará como no entregado.
- Para que la práctica se considere como entregada, se debe completar al menos el 25% de toda la actividad.
- No podrá modificarse ningún elemento de la práctica pasada la fecha de entrega (repositorio, archivos de Google Drive, etc.).
- Asimismo, también es responsabilidad del estudiante asegurarse de que, en el momento de entregar la práctica, **se haya dado acceso al profesor a los diferentes elementos privados que se entreguen** (p. ej., repositorio GitHub privado o archivos restringidos de Google Drive). El profesor indicará en los Anuncios del aula su nombre de usuario en estas plataformas.
- No se puede hacer grupos con alumnos de aulas diferentes.

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 del dataset creado en la Práctica 1.

Importante: es importante que el dataset cuente con una amplia variedad de datos numéricos y categóricos, entre los que se encuentre al menos una variable objetivo, así como contar con datos faltantes y/o erróneos, para poder realizar un análisis más rico y poder responder a las diferentes preguntas planteadas en el enunciado de la práctica. De no ser así, se deberá buscar la forma de integrar más datos que cumplan con estos requisitos en la etapa de integración (apartado 2).

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? Resume brevemente las variables que lo forman y su tamaño.

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. Si se decide trabajar con una selección de los datos, es muy importante que esta esté debidamente justificada. Además, se recomienda mostrar un resumen de los datos que permita ver a simple vista las diferentes variables y sus rangos de valores.

3. **Limpieza de los datos.**
   3.1. ¿Los datos contienen ceros, elementos vacíos u otros valores numéricos que indiquen la pérdida de datos? Gestiona cada uno de estos casos utilizando el método de imputación que consideres más adecuado.
   3.2. Identifica y gestiona adecuadamente el tipo de dato de cada atributo (p.ej. conversión de variables categóricas en `factor`).
   3.3. Identifica y gestiona los valores extremos.
   3.4. Justifica la necesidad de otros métodos de limpieza para este dataset en particular y, de ser necesario, aplícalos.

4. **Análisis de los datos.**
   4.1. Aplica un modelo supervisado y uno no supervisado a los datos y comenta los resultados obtenidos.
   4.2. Aplica una prueba por contraste de hipótesis. Ten en cuenta que algunas de estas pruebas requieren verificar previamente la normalidad y homocedasticidad de los datos.

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 apartado.
   Se debe representar tanto el contenido del dataset para observar las proporciones y distribuciones de las diferentes variables una vez aplicada la etapa de limpieza, como los resultados obtenidos tras la etapa de análisis.

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	1	2	2	2	1	1	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. **La memoria de la práctica,** que deberá ser **un único documento PDF**, con las respuestas a las preguntas, los nombres de los componentes del grupo y el enlace al repositorio Git de la práctica. 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

Este documento se entregará, por cada uno de los integrantes del grupo, en el **Espacio de Entrega PR2 del aula Canvas.**

2. Un **repositorio Git** con las soluciones de la práctica. El repositorio Git se creará en Github (https://github.com/) y deberá ser un **repositorio privado**, por lo que se deberá dar acceso al profesor en el momento de la entrega, mediante el nombre de usuario que se indicará en el tablón de Anuncios 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:

   a) Un **README.md** con los nombres de los componentes del grupo y una descripción de los ficheros que componen el repositorio.
   b) Una carpeta con el **código generado** para analizar los datos.
   c) El **fichero CSV con los datos originales.**
   d) 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.** El enlace del mismo se debe indicar en el apartado 8 del documento PDF.

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.

Some conclusions:

Este documento de entrega final de la Práctica 2 se debe entregar en el aula antes de las **23:59h CET** del día **8 de enero del 2025**. 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.