Actividad 2 – Proyecto de Ciencia de Datos

Título: *Grado en Ciencia de Datos e Inteligencia Artificial*
Asignatura: *Introducción a la Ciencia de Datos*
Código: *07GIAR*
Créditos: *6 ECTS*
Curso: *Octubre 2024-2025*

Índice

1. Introducción de la práctica 2

DESCRIPCIÓN	
Introducción	En esta actividad propone desarrollar un proyecto de Ciencia de Datos respecto a el uso de bicicletas de alquiler en la ciudad de Valencia. En el proyecto se desarrollarán diferentes etapas de un proyecto de Ciencia de Datos, en el que se detallan los pasos a seguir en el Apartado 2. La calificación de esta actividad está asociada a la rúbrica disponible en el Apartado 4 de este documento.
Objetivo	El objetivo de la práctica es mostrar el desarrollo y aplicación de técnicas de Ciencia de Datos en el contexto de un problema específico determinado por el equipo docente.
Trabajo previo	Para realizar esta actividad es recomendable visualizar los vídeos docentes y el manual de la asignatura que se encuentran en la pestaña: Recursos y materiales > 01.Materiales docentes y 02. Vídeos de la asignatura.
Metodología	El profesor expondrá en cada sesión los contenidos a desarrollar en la práctica. Se compartirá en campus los ficheros necesarios para iniciar la práctica y el código necesario para trabajar sobre ella. En la sección de "*Actividades*" del aula se dispondrá de un documento plantilla donde cada estudiante irá recogiendo los resultados esperados según las instrucciones asignadas por cada uno.
Entrega	Se entregará la memoria de prácticas en formato pdf usando como plantilla el documento adjunto en la actividad de campus. Se entregará en un fichero comprimido (zip) los notebooks ejecutados para la evaluación.

2. Enunciado de la práctica

La práctica consiste en aplicar los conceptos vistos en la asignatura sobre un proyecto de Ciencia de Datos. Se parte de un conjunto de datos, proporcionados por el profesor, sobre alquiler de bicicletas en la ciudad de Valencia. El problema a resolver consiste en conocer y modelar la demanda del alquiler de bicicletas según una zona concreta de la ciudad de Valencia. El objetivo es que la empresa de alquiler de bicicletas pueda disponer de recursos suficientes para atender la demanda en cualquier momento y conocer el comportamiento de los usuarios.

2.1. Introducción

Antes de comenzar la práctica, cada estudiante tendrá que registrar su participación en la práctica mediante el siguiente formulario que, tras cumplimentarlo, recibirá un correo con una serie de datos que deberá aplicar en cada uno de los apartados correspondientes a esta actividad:

Formulario de recepción de parámetros de la práctica 2

En el correo se extraerán los siguientes datos junto con atributos:

Barrio de estudio, ejemplo: AIORA

Barrio de comparación, ejemplo: RUSSAFA

Visualización: lineplot

Preprocesamiento, ejemplo: StandardScaler

Algoritmo de agrupación o Clustering, ejemplo: KMeans

Algoritmo de regresión, ejemplo: Árbol de decisión

Las variables proporcionadas en el presente enunciado sirven solamente de ejemplo para ilustrar el procedimiento de la práctica. Cada estudiante debe utilizar los datos recibidos por correos y adaptar la práctica a su caso, de forma que la entrega de la actividad quede constancia de la aplicación de dichas variables en su memoria de entrega.

Los datos están divididos en diferentes carpetas correspondientes a cada fase de la práctica. En las primeras fases se usará los datos de la carpeta "*raw*", luego, los datos de la carpeta "*interim*", y finalmente, la carpeta "*processed*" que serán los datos a usar en la fase de modelado.

2.2. Análisis exploratorio de datos

El primer paso consiste en conocer los datos, se cuenta con carpeta "*raw*" con una muestra de los datos disponibles en la fuente de origen (en crudo) que permite conocer el problema de tratar con datos directamente de la fuente. Por ello, se pide un análisis exploratorio y descriptivo de los datos sin aplicar transformaciones transformaciones.

Por otro lado, se pide analizar el "Barrio de estudio" (variables aplicando las operaciones vistas en clase respondiendo a la siguiente pregunta: ¿cómo es la distribución por horas del uso de bicicletas entre dos barrios concretos y si existe relación con alguna variable meteorológica?. Para ello se pide aplicar la siguiente operación en *pandas* relacionado con el análisis exploratorio:

- *pivot_table* con las columnas los barrios a comparar, en las filas las horas, “uso_bici” y otra variable meteorológica como valores a agregar.

INFORME: Se pide una captura del resultado de la consulta (*pivot_table*) realizada según el barrio de estudio y el de comparación. Analizar y comentar los resultados de las diferentes operaciones de la exploración.

VARIABLES A UTILIZAR: Barrio de estudio y Barrio de comparación

2.3. Visualización de datos

La visualización de datos es un proceso que aplica a lo largo de todas las fases de un proyecto de Ciencia de Datos. En este apartado se aborda la visualización como un análisis complementario de la fase de Análisis Exploratorio de Datos. Se pide que el alumno experimente con las diferentes propuestas de visualización para la exploración de datos en el caso de uso de bicicletas de alquiler resolviendo las siguientes cuestiones:

- Distribución de las estaciones y uso de bicicletas.

- Uso de bicicletas por horas en el barrio de estudio.

- Relación de variables meteorológicas con el uso de bicicletas.

- Otras cuestiones no definidas previamente y valoradas por el estudiante.

También, se pide aplicar la función asignada en el correo personal en la variable “Visualización”. Para ello tendrá que aplicar la función a los datos y variables adecuadas a la dicha función asignada (puede revisar los argumentos de entrada y ejemplos de dicha función).

INFORME: Se pide que un extracto de las visualizaciones y una captura de la figura relacionada con la función de visualización asignada. También, se pide comentar aspectos relevantes que puedan aportar valor al estudio inicial y pueda servir para toma de decisiones en siguientes procesos o iteraciones del proyecto.

VARIABLES A UTILIZAR: Barrio de estudio y Visualización.

2.4. Investigación y referencias

Tras la fase de exploración de datos, es necesario conocer qué otros trabajos similares se han hecho y cómo han sido llevados a cabo. Para ello, la búsqueda en internet de trabajos académicos, proyectos publicados de Ciencia de Datos, retos de programación, blogs u otras referencias, pueden servir para abordar el problema desde otra perspectiva o consolidar la hipótesis de partida.

INFORME: Se pide hacer una búsqueda de, al menos, una referencia sobre proyectos similares al problema de la presente práctica, describir y proponer posibles acciones que pudieran aplicarse en la presente práctica.

VARIABLES A UTILIZAR: Ninguna.

2.5. Preprocesamiento de datos

La mayor parte de transformaciones o preprocesamiento de datos se ha hecho previamente, de forma que pueda trabajarse aspectos más conceptuales de un proyecto de Ciencia de

Datos. En este apartado, se debe trabajar con los datos del fichero "*estaciones.csv*" *y "usobarriosmeteo.csv"* de la carpeta *"interim"*, para aplicar:

- "*estaciones.csv":* Aplicar la técnica de transformación de datos asignada por estudiante (como *StandardScaler,* entre otras) para aplicar a las variables: *lat, lon y uso_bici*. Se guardará los datos de todas las estaciones en un nuevo fichero. El objetivo de esta operación es conocer cómo es la distribución de estaciones y la demanda de bicis tras aplicar esta transformación.

- *"usobarriosmeteo.csv"*. Extracción de características como hora y fin de semana y aplicar selección de datos para el barrio a estudiar. Tras analizar en la fase de exploración qué variables pueden tener interés en la obtención de conocimiento, se quiere incluir estas nuevas variables para modelado.

Por otro lado, se busca hacer un conjunto de datos adaptado y preparado para minería o modelado. Se pide guardar en la carpeta *"processed"* los resultados de aplicar transformaciones en ambos conjuntos de datos teniendo dos resultados en formato *csv*.

INFORME: Se pide al estudiante que describa la técnica asignada y un análisis de los resultados tras aplicar la técnica a los datos de "*estaciones.csv"*. Se pide que se indiquen los metadatos (registros, variables, …) o estructura de los nuevos ficheros guardados en la carpeta *"processed"*.

VARIABLES A UTILIZAR: Barrio de estudio y Preprocesamiento.

2.6. Aprendizaje no supervisado

Como parte del entendimiento del problema, se pueden aplicar técnicas de agrupamiento o *clustering* que forman parte de aprendizaje automático no supervisado. Las estaciones donde se pueden alquilar y depositar las bicicletas tienen una distribución homogénea por toda la ciudad. Sin embargo, se quiere analizar posibles grupos en función de su uso y posición.

Para ello se proponen dos algoritmos como es el *clustering* jerárquico o *KMeans* en función a lo asignado para la práctica de forma individual. En tarea se trabaja con *dataset "estaciones.csv"* de la carpeta *"interim" y "processed"* de forma que se evaluará el resultado de aplicar el algoritmo a los datos procesados y sin procesar. El objetivo es encontrar y evaluar si una agrupación de estaciones en función del uso puede ser más conveniente que la agrupación original de estaciones en función del barrio.

INFORME: Se pide al estudiante que aplique el algoritmo asignado a los dos conjuntos de datos. Defina un valor de *k* agrupaciones que pueda ser adecuado en función del barrio asignado. También, pide una gráfica que muestre el resultado de la agrupación escogida.

VARIABLES A UTILIZAR: Barrio de estudio y Algoritmo de agrupación o *Clustering*.

2.7. Aprendizaje supervisado

En esta fase se desarrolla el modelado de datos para la predicción de la demanda de uso de bicicletas de alquiler. Para ello, se hace uso de fichero *"usobarriosmeteo.csv"* de la carpeta *"processed"* donde se han generado nuevas características a incluir en el entrenamiento del modelo. Primero, se debe dividir la muestra en entrenamiento y prueba (*train/test*). Luego se aplicará el algoritmo de regresión para el barrio (ambos asignados para cada estudiante). En esta fase, se debe experimentar con la combinación de variables hasta encontrar un resultado

que se considere el óptimo teniendo en cuenta las métricas de evaluación y el desempeño en la predicción. Se tomará una referencia o métrica base de evaluación (*baseline*) para evaluar los modelos.

INFORME: Se pide hacer modelado de datos en base al algoritmo asignado para la regresión de uso de bicicletas. Se pide indicar las variables usadas, métricas y gráficas de resultado para, al menos, dos experimentos de modelado.

VARIABLES A UTILIZAR: Barrio de estudio y Algoritmo de regresión.

2.8. Resultados y conclusiones

INFORME: Se pide un análisis de los resultados de la fase de modelado en la fase de aprendizaje no supervisado y supervisado. También, en este apartado, se debe analizar el proceso completo llevado en todos los apartados, para definir nuevas acciones que puedan mejorar los resultados del proyecto y aplicar nuevas iteraciones sobre una o varias fases llevadas en el proyecto.

2.9. Informe y *notebooks*

Se pide subir el informe con la plantilla proporcionada en la actividad de campus en formato *pdf*. Cada uno de los apartados de esta práctica no puede superar la extensión de una página. En caso de el estudiante quiera extender su práctica aportando más de lo solicitado, puede utilizar la plantilla cambiando el nombre como “anexos_p2” de forma que se entregará en otro fichero el desarrollo adicional.

Los *notebooks* tienen que subirse en una misma carpeta comprimidas (formato *zip* solo *notebooks*) con los nombres asociados a cada apartado. Todos los *notebooks* tienen que estar ejecutados en la entrega, de forma que se muestre todos los resultados.

Referencias necesarias para la práctica:

Véase guía de estudio de la asignatura

3. Entrega

Fecha de entrega	
1ª Convocatoria	29/04/2025 hasta las 23.59
2ª Convocatoria	25/06/2025 hasta las 23.59

4. Rúbrica de evaluación

	Suspenso (0)	Aprobado (5)	Sobresaliente (10)
Análisis Exploratorio de Datos (10%)*	No se responden a las preguntas del análisis.	Se muestra las operaciones asignadas. Falta responder a algunas descripciones de los datos.	Se responden a las preguntas de la fase de análisis exploratorio.
Visualización de datos (10%)*	No se aplica la técnicas asignadas.	Las gráficas son correctas. No hay descripción o comentarios relevantes.	Se aplica la técnicas asignadas. Se desarrolla y comenta el proceso completamente.
Investigación y referencias (20%)*	No hay referencias o no se justifica la relación.	Se incluye una o varias referencias, pero no llega a comentarse o justificarse debidamente.	La/las referencias están justificadas y alineadas al objetivo de la práctica.
Preprocesamiento de datos (10%)*	No se aplica la técnicas asignadas.		Se aplica la técnicas asignadas. Se desarrolla el proceso completamente.
Aprendizaje no supervisado (10%)*	No se aplica la técnicas asignadas.		Se aplica la técnicas asignadas. Se desarrolla el proceso completamente.
Aprendizaje supervisado (10%)*	No se aplica la técnicas asignadas.		Se aplica la técnicas asignadas. Se desarrolla el proceso completamente.
Resultados y conclusiones (10%)*	No se comparan resultados. No se comentan conclusiones.	Hay comparación de resultados. No se conectan los diferentes resultados con los objetivos. No se demuestra aportar nuevo conocimiento en el problema.	Se muestra la comparación de resultados. Las conclusiones hablan de los objetivos del proyecto. Se proponen mejoras.
Informe y notebooks (20%)	No se sigue las instrucciones de entrega.	Se sigue parcialmente las instrucciones de entrega.	Se entrega en formato pdf, la extensión de cada apartado es de una página. Los notebooks entregados muestran el resultado de ejecutar las celdas.

* Apartado de rúbrica que se valora con el documento *pdf* entregado usando el formato de plantilla proporcionado en el enunciado de la práctica.