Trabajo Práctico 1

Introducción

En este trabajo práctico se propone que cada grupo de alumnos se enfrente a diferentes problemas de ciencia de datos y que pueda resolverlos aplicando los contenidos que se ven en la materia.
El objetivo principal del trabajo será aplicar técnicas de análisis exploratorio, preprocesamiento de datos, agrupamiento, clasificación y regresión. En la sección enunciado se detallan los objetivos particulares de cada ejercicio.

Modalidad de entrega

Repositorio

Cada grupo deberá crear su propio repositorio en github: TA047R-2C2024-GRUPOXX
En dicho repositorio deberá estar disponible todo el contenido obligatorio de la entrega (notebooks, modelos entrenados, *datasets*, reportes, presentación) y cualquier archivo que sea necesario para la correcta ejecución del trabajo.

Notebooks

El trabajo debe ser realizado en notebooks de python, se espera que las mismas contengan **todos** los resultados de la ejecución los cuales siempre deben ser **reproducibles**. Se debe entregar una notebook para cada ejercicio, con la siguiente nomenclatura :

TA047R_TP1_GRUPOXX_ENTREGA_EJX

Las secciones del trabajo deben estar claramente diferenciadas en la notebook utilizando celdas de *markdown*. Se debe incluir una sección principal con el título del trabajo, el número de grupo y el nombre de todos los integrantes.

Todo análisis realizado debe estar acompañado de su respectiva explicación y toda decisión tomada debe estar debidamente justificada. Cualquier hipótesis que sea considerada en el desarrollo del trabajo práctico debe ser detallada y debe estar informada en la entrega. Cualquier criterio que se utilice basado en fuentes externas (*papers*, bibliografía, etc.) debe estar correctamente referenciado en el trabajo.

Todos los gráficos que se incorporen deben tener su correspondiente título, leyenda, nombres en los ejes, unidades de medidas, y cualquier referencia que se considere necesaria. Es importante que tengan presente que los gráficos son una herramienta que facilita entender el problema, por lo tanto, deben ser comprensibles por quien los vaya a leer.

Conjuntos de datos (*datasets*)

A partir de las tareas de preprocesamiento, y de las diferentes estrategias que se planteen, es posible que se generen nuevos *datasets* sobre los cuales se entrenarán los modelos. Todo conjunto de datos creado debe ser almacenado y debe estar disponible en la entrega para ser utilizado por el equipo docente.

Modelos

Todos los modelos entrenados tanto para clasificación como para regresión deben ser guardados en un archivo (joblib / pickle) y deben estar disponibles en la entrega para ser utilizados por el equipo docente.

Reportes

Se deberá confeccionar un reporte (en formato pdf) a modo de resumen de los puntos desarrollados en cada ejercicio. El documento deberá tener la siguiente nomenclatura TA047R_TP1_GRUPOXX_REPORTE y **deberá seguir el template proporcionado por la cátedra**.

Exposición del TP

Cada grupo deberá realizar una exposición de los resultados obtenidos en cada ejercicio del trabajo práctico. La misma no deberá superar los 20 minutos, podrán exponer las notebooks o confeccionar una presentación tipo PowerPoint o similar. **Deberán seguir las pautas mínimas proporcionadas por la cátedra.**

Devolución y comentarios entre grupos

Para cumplimentar la entrega del Trabajo Práctico todos los grupos deberán dar *feedback* sobre las presentaciones del resto de sus compañeros. La fecha límite se indica en la siguiente sección.

Fechas de entrega

Entrega 10/10/2024 : Deberán informar la entrega por el canal de Slack de su grupo, enviando un *link* al repositorio de **GitHub**. Allí deberá estar disponible la notebook, los modelos entrenados y el reporte con el resumen del trabajo. Para el reporte pueden tomar como referencia el modelo de reporte. **Esta fecha es obligatoria.**

Exposición 15/10/2024 - 17/10/2024: cada grupo realizará una exposición oral de los principales resultados de su trabajo práctico. El equipo docente informará la fecha y la hora en que debe exponer cada grupo. Las pautas mínimas se explican aquí. **Esta fecha es obligatoria para todos los alumnos.**

Feedback 29/10/2024 (como fecha máxima): cada alumno deberá dar su *feedback* sobre el trabajo realizado por el resto de los grupos. Para ello deberán completar aquí un formulario. **Esta actividad es obligatoria para todos los alumnos.**

Reentrega 28/11/2024: aquellos grupos que deban realizar correcciones contarán con esta fecha para volver a entregar el trabajo práctico. Esta es la última oportunidad para aprobarlo.

Enunciado

EJERCICIO 1 - *Análisis Exploratorio de Datos*

El objetivo es realizar un análisis completo del conjunto de datos, aplicar técnicas de exploración y de preprocesamiento para poder responder algunas preguntas que se planteen sobre los datos. En este ejercicio trabajaremos sobre un conjunto de datos sobre el uso de taxis Yellow Cab en USA durante el año 2023. Cada grupo deberá seleccionar un período de 3 meses indicado aquí por el equipo docente.

a) **Exploración Inicial :** analizar cada variable, considerando los siguientes aspectos
- Variables Cuantitativas: calcular medidas de resumen, media, mediana, moda, etc.
- Variables Cualitativas: mostrar cantidad de valores posibles, y frecuencias de cada uno.
- Realizar un análisis gráfico de las distribuciones de las variables más relevantes
- Analizar las correlaciones existentes entre las variables.

b) Datos Faltantes : analizar la presencia de datos faltantes en el *dataset*

- Realizar análisis de datos faltantes. Graficar para cada variable el porcentaje de datos faltantes con respecto al total del dataset. Calcular el porcentaje de datos faltantes de cada registro.
- Revisar los datos faltantes o mal ingresados y tomar una decisión sobre estos: reemplazo de valores, eliminación de registros incompletos, etc.

c) Valores atípicos : analizar la existencia de valores atípicos.

- Detectar valores atípicos en los datos tanto en forma univariada como multivariada. Realizar gráficos que permitan visualizar los valores atípicos.
- Explicar qué características poseen los datos atípicos detectados y decidir el tratamiento a aplicar sobre los mismos.

d) Nuevos Features: analizar qué nuevas variables se pueden crear ya sea que resulten derivadas de los atributos existentes o de incorporar nuevas fuentes de datos.

e) Preguntas de investigación: se deben plantear preguntas cuyas respuestas puedan ser obtenidas a partir del análisis de los datos. Ejemplo:

- ¿Existe una manera de caracterizar los lugares más recurrentes para inicio/fin de viaje?
- ¿Cómo son los viajes típicamente en distancia y tiempo?
- ¿Cómo son los viajes típicamente en distancia y tiempo según el horario y/o el día de la semana?

f) Visualización de los datos: en esta sección se espera que puedan apoyarse en visualizaciones, combinando distintas variables del dataset que los ayuden a responder preguntas sobre los datos.

Nota : Los ítems a, b, c, d, e y f son los mínimos requeridos para esta etapa, cada grupo puede sumar técnicas y/o análisis adicionales.

EJERCICIO 2 - Modelos de Clasificación Binaria

En este ejercicio vamos a usar datos de distintas estaciones meteorológicas de Australia. El objetivo es predecir si lloverá o no al día siguiente (variable **RainTomorrow**), en función de los datos meteorológicos del día actual. Cada grupo deberá seleccionar las *Locations* de acuerdo a las regiones (estados/territorios) indicados aquí por el equipo docente.

a) **Análisis Exploratorio y preprocesamiento de datos:** realizar el análisis de datos mínimo necesario para comprender el dominio del problema, limpieza de datos, generación de nuevos features, etc.

b) **Entrenamiento y Predicción:** se deberá trabajar con un 80% de datos para entrenamiento y un 20% de datos para test. Para todos los modelos se pide realizar las tareas de ingeniería de características necesarias para trabajar con cada algoritmo (*encoding*, normalización, etc). Los modelos a entrenar son los siguientes:

Modelo 1 : Árbol de decisión

- Construir un árbol de decisión y optimizar sus hiperparámetros mediante *k-fold Cross Validatio*n para obtener la mejor performance. ¿Cuántos *folds* utilizaron? ¿Qué métrica consideran adecuada para buscar los parámetros?
- Graficar el árbol de decisión con mejor performance encontrado en el punto anterior. Si es muy extenso mostrar una porción representativa.
- Analizar el árbol de decisión seleccionado describiendo los atributos elegidos, y decisiones evaluadas (explicar las primeras reglas obtenidas).
- Evaluar la performance del árbol en el conjunto de evaluación, explicar todas las métricas y mostrar la matriz de confusión. Comparar con la performance de entrenamiento.

Modelo 2: Random Forest

- Construir un clasificador RF (Random forest) y optimizar sus hiperparámetros mediante *k-fold Cross Validation* para obtener la mejor performance. ¿Cuántos *folds* utilizaron? ¿Qué métrica consideran adecuada para buscar los parámetros?
- Analizar la importancia de los atributos.
- Mostrar la conformación final de uno de los árboles generados. Si es muy extenso mostrar una porción representativa y explicar las primeras reglas.
- Evaluar la performance del modelo en el conjunto de evaluación, explicar todas las métricas y mostrar la matriz de confusión. Comparar con la performance de entrenamiento.

Modelo 3: a elección

- En este punto se debe entrenar (mediante *cross-validation*) un modelo elegido por el grupo. Se debe evaluar su performance en entrenamiento y sobre el conjunto de evaluación, explicar todas las métricas y mostrar la matriz de confusión.

c) Cuadro de resultados: armar un cuadro comparativo de los resultados obtenidos con los modelos y responder la siguiente pregunta *¿Qué modelo elegirían para predecir si lloverá o no al día siguiente?*

EJERCICIO 3 - Regresión

En este ejercicio vamos a usar datos de alojamientos de la plataforma AirBnB. El objetivo es predecir el precio de alquiler de una propiedad (**price**) en función de los datos publicados en el aviso. Cada grupo deberá seleccionar el archivo *"Detailed Listings data"* correspondiente a una ciudad específica indicada aquí por el equipo docente.

a) Análisis Exploratorio y preprocesamiento de datos: realizar el análisis de datos mínimo necesario para comprender el dominio del problema, limpieza de datos, generación de nuevos features,etc.

b) Entrenamiento y Predicción: se deberá trabajar con un 80% de datos para entrenamiento y un 20% de datos para test. Para todos los modelos se pide realizar las tareas de ingeniería de características necesarias para trabajar con cada algoritmo (encoding, normalización, etc). Los modelos a entrenar son los siguientes:

Modelo 1: Regresión lineal

- Construir un modelo Regresión Lineal Múltiple, identificando los features más importantes para predecir el precio del alquiler.
- Evaluar la performance del modelo en el conjunto de evaluación, explicar todas las métricas. Comparar con la performance de entrenamiento.

Modelo 2: XGBoost

- Construir un modelo XGBoost y optimizar sus hiperparámetros mediante *k-fold Cross Validation* para obtener la mejor performance. ¿Cuántos *folds* utilizaron? ¿Qué métrica consideran adecuada para buscar los parámetros?
- Evaluar la performance del modelo en el conjunto de evaluación, explicar todas las métricas. Comparar con la performance de entrenamiento.

Modelo 3: a elección

- En este punto se debe entrenar (mediante *cross-validation*) un modelo elegido por el grupo. Se debe evaluar su performance en entrenamiento y sobre el conjunto de evaluación explicando todas las métricas.

c) Cuadro de resultados: armar un cuadro comparativo de los resultados obtenidos con los modelos y responder la siguiente pregunta *¿Qué modelo elegirían para predecir el precio de alquiler de un Airbnb en la ciudad seleccionada?*

EJERCICIO 4 - Agrupamiento (Clustering)

En este punto se busca analizar si es posible agrupar los datos en función de algún criterio. Vamos a utilizar un conjunto de datos que contiene información sobre algunos *tracks* (canciones) de Spotify. Para conocer en detalle cada atributo del dataset pueden consultar el siguiente link:
Para esta tarea se propone utilizar el algoritmo **K-Means** y se deberán realizar los siguientes puntos:

a. Analizar la tendencia al *clustering* del dataset.
b. Estimar la cantidad apropiada de grupos que se deben formar.
c. Evaluar la calidad de los grupos formados realizando un análisis de *Silhouette*.
d. Realizar un análisis de cada grupo intentando entender en función de qué características fueron formados.