Trabajo práctico integrador.
Análisis de datos

1. Introducción y motivación

Les proponemos para este trabajo final realizar el análisis completo para un set de datos, para ello les vamos a proponer varios de estos y la idea es que ustedes elijan uno.

Para dar un poco de noción sobre a qué corresponden estos datos, también vamos a presentar a que tipo de problema/etapa del análisis de datos corresponden.

1.1 Datasets disponibles

Vamos a proponer 6 datasets:

1. Datos de distintas estaciones meteorológicas de Australia.
    - Preguntas interesantes para considerar aquí: El objetivo es predecir si lloverá o no al día siguiente (variable *RainTomorrow*), en función datos meteorológicos del día actual.
2. Datos de distintas canciones en Spotify.
    - Preguntas interesantes para considerar aquí: El objetivo aquí es poder estimar si un tema nuevo será del gusto de la persona que tiene esta playlist activa. En este caso la variable *label* corresponde a nuestra variable de salida a analizar.
3. Uso de taxis *Yellow Cab* en USA en el año 2020.
    - Preguntas interesantes para considerar aquí: (elija una o dos)
        - ¿Existe una manera de caracterizar los lugares más recurrentes para inicio/fin de viaje?
        - ¿Cómo son los viajes típicamente en distancia y tiempo?
        - ¿Podremos segmentar los viajes de alguna manera? (clusterización)
4. Dataset de bandas de metal a través de los años.
    - Preguntas interesantes para considerar aquí:
        - ¿Cuándo fue el auge del género de Metal?
        - Según estos datos, ¿cómo evoluciono el género?
5. Dataset de piezas creadas por año de LEGO.
    - Preguntas interesantes para considerar aquí:
        - ¿Cómo evolucionaron los sets de lego en tamaño a través de los años?
        - ¿Existe alguna asociación entre los colores y las temáticas?
        - ¿Podría predecir a que temática pertenece un set basado en el contenido de este? (recomendado)
        - A través de los años, ¿Cuál o cuáles son los sets que tienen las piezas más raras?
        - ¿Cómo evolucionaron los colores en los sets de lego a través de los años?
6. Dataset de avistamiento de UFO's en el mundo.

- Preguntas interesantes para considerar aquí:
    - ¿Cuáles son los lugares de avistamiento más comunes?
    - ¿Cuál es la descripción más común de UFO?
    - ¿Existe una manera de predecir donde podría ser avistado un UFO basado en estos datos?

2. Análisis exploratorio inicial

- Visualizar las primeras filas.
- Realizar un resumen de 5 números.
- Identificar los tipos de datos: categórico, ordinal, etc. Responder para cada variable su tipo y si es informativa para un problema de clasificación (por ejemplo si se trata de un código, como una matrícula, o un nombre propio).
- Identificar las variables de entrada y de salida del problema.
- Variables de entrada:
    - Realizar los siguientes análisis por tipo de variable:
        - Numéricas: Obtener conclusiones acerca de la distribución de los datos.
        - Categóricas: Obtener conclusiones acerca de la cardinalidad, representación de cada categoría, etc.
        - Compuestas: ¿Pueden tratarse para utilizarse en el problema a resolver?
- Variables de salida (en caso de aplicar):
    - ¿Están balanceadas las clases?
    - (en caso de aplicar) ¿Qué técnicas consideraría para codificar la variable de salida? Justifique.

3. Limpieza y preparación de datos / ingeniería de features

Datos faltantes. Indicar cantidad de observaciones y valores faltantes para cada variable.

¿Qué supuestos puede realizar acerca de los datos faltantes? ¿Qué técnicas de imputación recomendaría? Ensayar distintas técnicas y analizar los resultados.

En función del estudio inicial de las variables que se hizo en la sección anterior, elegir una técnica de codificación para cada variable. Cuando lo considere apropiado, ensayar distintas técnicas y comparar los resultados, teniendo en cuenta el tipo de clasificador a utilizar. Nota: para tipos de datos compuestos o estructurados, considerar la obtención de variables de tipo numérico/categórico.

¿Qué puede decir acerca de las relaciones entre las variables de entrada?

Antes de entrenar un modelo de aprendizaje automático, ¿Podría identificar las variables de entrada de mayor importancia? Considerar por lo menos dos técnicas para cada variable. Explique brevemente los métodos utilizados.

4. Entrenamiento de modelos (opcional)

Recurriendo a los modelos que conozca, defina una lista de modelos candidatos a entrenar (puede ser el mismo tipo de clasificador con distintos hiperparámetros). Nota: no se contemplará el

desempeño del modelo elegido, sino las conclusiones que puedan establecerse a partir de la preparación previa de los datos.

Ensaye distintas cadenas de procesamiento con las técnicas consideradas en la sección 3 (por ejemplo, distintas técnicas de imputación, selección de variables de entrada, codificación de variables categóricas, transformación, etc.).

4.1 Separación de datos

Los pasos siguientes comprenden las etapas de preparación de datos y evaluación de resultados.

Para ello, se debe particionar el dataset en entrenamiento y validación.

4.2 Evaluación de resultados

¿Qué puede concluir acerca de los modelos y preparaciones de datos ensayadas? Tener en cuenta como cada preparación afecta a los distintos modelos.

5. Presentación de resultados

Como presentación de este trabajo, deberán realizar una presentación para lo cual tendrán un bloque de tiempo de máximo 15 min por grupo. Durante la cursada también les recomendamos fuertemente presenten un informe de varianza muy cortito para poder guiarlos en el desarrollo.