Trabajo Practico N° 2
Minería de datos

Objetivo

El objetivo de este trabajo practico es integrar los conocimientos adquiridos en las unidades 4 y 5 en dos problemas reales asociados uno al comportamiento financiero de 1000 empresas y otro a al tipo de droga farmacéutica.

Actividades

1. Descargar el conjunto de datos, 1000_Companies.csv[1], para realizar el trabajo práctico. Realizar un análisis exploratorio de datos: visualizar distribuciones, valores faltantes, correlaciones, etc. Limpiar el conjunto de datos (manejar valores faltantes, eliminar outliers) si es necesario. Codificar variables categóricas (si es necesario). Normalizar o estandarizar las características.

2. Analizar los atributos del conjunto de datos (distribuciones, valores, outliers, tipos de datos, etc.) y elegir un método de estandarización.

3. Realizar la estimación del atributo Profit utilizando árboles de decisión (Regresión) analizando los parámetros máximo profundidad, número mínimo de observaciones, número mínimo de observaciones por separación y criterio de separación. Graficar el árbol obtenido en el proceso de entrenamiento y mostrar los resultados sobre dos conjuntos de test (Error Absoluto Medio, Error Cuadrático Medio y Raíz del Error Cuadrático Medio).

4. Descargar el conjunto de datos, drugType.csv[2], para realizar el trabajo práctico. Realizar un análisis exploratorio de datos: visualizar distribuciones, valores faltantes, correlaciones, etc. Limpiar el conjunto de datos (manejar valores faltantes, eliminar outliers) si es necesario. Codificar variables categóricas (si es necesario). Normalizar o estandarizar las características. Generar dos conjuntos de datos considerando 80-20 y 70-30 para entrenar y evaluar los modelos.

5. Realizar la estimación del atributo Droga utilizando árboles de decisión (Clasificación) analizando los parámetros máximo profundidad, número mínimo de observaciones, número mínimo de observaciones por separación y criterio de separación. Graficar un árbol obtenido en el proceso de entrenamiento y luego a aplicar una poda. Mostrar los resultados sobre ambos conjuntos de test (Precisión, Exhaustividad y Exactitud).

6. Realizar la estimación del atributo Droga utilizando Bayes Ingenuo. Aquí deberá considerar un criterio de división de los atributos continuos para discretizarlos. Mostrar los resultados sobre ambos conjuntos de test (Precisión, Exhaustividad y Exactitud).

7. Realizar la estimación del atributo Droga utilizando k-NN analizando los parámetros cantidad de vecinos, métrica y valor de p. Mostrar los resultados sobre ambos conjuntos de test (Precisión, Exhaustividad y Exactitud).

[1] https://www.kaggle.com/datasets/subhamp7/company-profit-and-expenditures
[2] https://www.kaggle.com/datasets/lykin22/drug-data

Trabajo Practico N° 2
Minería de datos

Presentación

La entrega es por grupos de dos estudiantes y se entrega un archivo por grupo.

Cualquier integrante del grupo puede hacer la entrega mediante el campus de la materia.

El informe deberá tener una cabecera en la que se indique: año, materia, integrantes. Además, deberá contar con una sección de conclusiones al final del mismo.

El formato del informe deberá ser en formato ipynb y no debe contener las definiciones teóricas ni el significado de los parámetros de los métodos dados en clase.

Las gráficas mostradas en el informe deben contener una explicación de lo observado y si es coherente que su hipótesis previa, por tanto, la cantidad de graficas debe estar acotadas y ser representativas.

Las entregas fuera del plazo establecido no serán consideradas salvo excepciones previamente justificadas por el grupo.