**Carrera:**

**Ingeniería en Software**

**Asignatura:**

**Minería de Datos**

**Practica 1. Conociendo tus datos**

**Alumno:**

**Juan Manuel Aguilar Garrido**

**Profesor:**

**Dr. Sergio Valadez Godínez**

**Fecha de entrega: 19/06/2022**

**Instrucciones:**

A partir de un caso práctico sobre minería de datos, elaborar un reporte que incluya:

a) Descripción de la base de datos

b) Tipos de atributos

c) Medidas de tendencia central

d) Medidas de dispersión

e) Resumen de cinco números, diagramas de caja y valores atípicos

f) Regresión lineal

g) Análisis de Componentes Principales (PCA)

h) Reglas de separación de patrones

i) Conclusiones

j) Bibliografía

|  |
| --- |
| **a) Descripción de la base de datos** |
| **Lenses Data Set** es una base de datos de clasificación la cual contiene 24 instancias de carácter multivariable (más de un argumento y/o atributo).  Cuenta con 3 clases, las cuales son:  1: The patient should be fitted with hard contact lenses.  2: The patient should be fitted with soft contact lenses.  3: The patient should not be fitted with contact lenses.  **Fuentes**  Cendrowska, J. "PRISM: An algorithm for inducing modular rules", International Journal of Man-Machine Studies, 1987, 27, 349-370.  Donor: Benoit Julien (Julien '@' ce.cmu.edu).  **Datos extra**  Los ejemplos son completos y libres de ruido. Los ejemplos simplificaron mucho el problema. Los atributos no describen completamente todos los factores que afectan la decisión de qué tipo encajar, si es que hay alguno.  Notas:  --Esta base de datos está completa (todas las combinaciones posibles de pares atributo-valor están representadas).  --Cada instancia es completa y correcta. |
| **b) Tipos de atributos** |
| Un atributo es un campo de datos que representa una característica o rasgo de un objeto de datos.  **Se cuenta con 4 atributos presentes dentro de la base de datos, los cuales son:**  1. age of the patient: (1) young, (2) pre-presbyopic, (3) presbyopic  2. spectacle prescription: (1) myope, (2) hypermetrope  3. astigmatic: (1) no, (2) yes  4. tear production rate: (1) reduced, (2) normal  Los datos presentes son pertenecientes a la clasificación de **atributos numéricos**, debido a que son una cantidad medible representada con valores enteros o reales.  Mas, sin embargo, estos atributos numéricos tienen una descripción correspondiente al valor que representan.  También cabe destacar que los atributos presentes son **discretos** debido a que tienen un conjunto finito de valores además de que estos son enteros.  **Distribución de los datos**    Podemos observar una serie de valores con distancias parecidas los unos de los otros, además de que muchos de los valores presentes dentro de la base de datos están traslapados de manera total los unos de los otros al crear las funciones de distribución comparando los atributos en graficas pares.  Para la graficación de los datos se ha usado la función **pairplot** proveniente de la librería de **seaborn,** esta función nos ayuda a crear un diagrama de pares trazando una relación por pares en un conjunto de datos. La función pairplot crea una cuadrícula de ejes de modo que cada variable en los datos se compartirá en el eje y en una sola fila y en el eje x en una sola columna.  Codigo fuente:  sns.pairplot(df, hue='Class', height=4)  donde df representa un data frame proveniente de la librearia Pandas, este data frame contiene la informacion de la base de datos original. |
| **c) Medidas de tendencia central** |
| Calcular la media, mediana y moda para cada uno de los atributos de la base de datos.  Poner código fuente del cálculo y el resultado. |
| **d) Medidas de dispersión** |
| Calcular el rango, los cuartiles, la varianza, la desviación estándar y el rango intercuartílico para cada uno de los atributos de la base de datos.  Calcular la matriz de correlación y covarianza de la base de datos.  Poner código fuente del cálculo y el resultado. |
| **e) Diagramas de caja y valores atípicos** |
| Calcular los diagramas de caja y los valores atípicos para cada uno de los atributos de la base de datos.  Poner código fuente del cálculo y el resultado.  Graficar los diagramas de caja para todos los atributos.  Especificar las tuplas o registros que son considerados valores atípicos. |
| **f) Regresión lineal** |
| Calcular los valores de la regresión lineal (coeficientes y cruces de las líneas, coeficiente de determinación y correlación) para los atributos de la base de datos.  Graficar la distribución de los datos con su respectiva línea de regresión.  Poner código fuente del cálculo y el resultado. |
| **g) Análisis de Componentes Principales (PCA)** |
| Calcular los componentes principales de la base de datos y su nivel de varianza.  Seleccionar el número de componentes que conservan la mayor parte de información original.  Graficar la distribución de las nuevas dimensiones de la base de datos.  Poner código fuente del cálculo y el resultado. |
| **h) Reglas de separación de patrones** |
| Crear un algoritmo con reglas IF-THEN-ELSE que permita separar lo mejor posible las clases de los patrones según el análisis estadístico realizado. |
| **i) Conclusiones** |
|  |
| **j) Bibliografía** |
|  |