**Carrera:**

**Ingeniería en Software**

**Asignatura:**

**Minería de Datos**

**Practica 1. Conociendo tus datos**

**Alumno:**

**Profesor:**

**Dr. Sergio Valadez Godínez**

**Fecha de entrega: dd/mm/aa**

**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** |
| Describir la base de datos a utilizar, así como la fuente de la misma. |
| **b) Tipos de atributos** |
| Describir los atributos de la base de datos y el tipo de dato de cada uno de los atributos.  Graficar la distribución de los datos. |
| **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** |
|  |