**FR. CONCEICAO RODRIGUES COLLEGE OF ENGINEERIG**

**Department of Computer Engineering**

**Experiment 2 - Perform Exploratory Data Analysis of Healthcare Data**

1. **Course Details:**

|  |  |  |  |
| --- | --- | --- | --- |
| **Academic Year** | **2023 - 24** | **Estimated Time** | **Experiment No. 1 – 02 Hours** |
| **Course & Semester** | **B.E. – Sem. VII** | **Subject Name** | **Data Science for Health and Social Care Lab** |
| **Experiment Type** | **Software Performance** | **Subject Code** | **HDSSBL701** |

|  |  |  |  |
| --- | --- | --- | --- |
| **Name of Student** | Atharva Prashant Pawar | **Roll No.** | 9427 |
| **Date of Performance.:** |  | **Date of Submission.:** |  |
| **CO Mapping** | **HDSSBL701.2 Clean, integrate and transform healthcare data** | | |

**Aim:** Perform Exploratory Data Analysis of Healthcare Data

**Objective:** The objective of this experiment is to familiarize BE Computer students with the process of performing Exploratory Data Analysis (EDA) on healthcare data. Students will learn how to import, explore, visualize, and analyze a healthcare dataset to gain insights and understand the characteristics of the data

**Tools and libraries**:

* Python programming language
* Jupyter Notebook or any Python IDE
* Required Python libraries: pandas, matplotlib, seaborn

**Procedure:**

**Step 1: Data Loading and Understanding**

Start by importing the necessary Python libraries: pandas, matplotlib, and seaborn. Load the healthcare dataset into a pandas DataFrame.

Display the basic information about the dataset, such as the number of rows and columns, data types, and summary statistics

**Step 2: Data Cleaning and Preprocessing**

Check for any missing values in the dataset and decide how to handle them (e.g., imputation or removal).

Look for any duplicate entries in the dataset and handle them if found.

Convert data types if necessary (e.g., dates, categorical variables).

**Step 3: Exploratory Data Analysis (EDA)**

Generate summary statistics for relevant numerical variables (e.g., mean, median, standard deviation, etc.).

Visualize the distribution of numerical variables using histograms, box plots, or kernel density plots.

Analyze the distribution of categorical variables using bar plots or count plots

Explore the correlation between different variables using a correlation heatmap.

**Step 4: Data Visualization**

Create meaningful visualizations to understand relationships and trends in the data. Use scatter plots, bar plots, line plots, or any other appropriate visualization techniques.

Focus on speciﬁc aspects like the relationship between age and health conditions, gender-wise distribution of diseases etc.

**Step 5: Insights and Interpretation**

Based on the analysis and visualizations, derive insights and observations from the healthcare data.

Identify patterns, trends, and any interesting ﬁndings that can help in understanding the dataset better.

**Result:**









