**FR. CONCEICAO RODRIGUES COLLEGE OF ENGINEERIG**

**Department of Computer Engineering**

**Experiment 1 - Collecting, Cleaning and Transforming Healthcare Data for a Speciﬁc Disease**

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** | HDSSBL701.1 Identify sources of data and methods for collecting, sharing and analyzing Healthcare data. | | |

**Aim:** Collecting, Cleaning, Integrating, and Transforming Healthcare Data for a Speciﬁc Disease:

**Objective:** The objective of this lab experiment is to familiarize students with the process of collecting, cleaning, integrating, and transforming healthcare data related to a speciﬁc disease. Students will gain hands-on experience in working with real-world healthcare datasets and preparing the data for analysis and AI applications.

# Materials:

* Data analysis software (e.g., Python, R, or any preferred tool)
* Healthcare dataset(s) related to the chosen disease (e.g., public datasets, research datasets, or simulated data)

# Procedure:

* 1. Choose a Speciﬁc Disease: Select a speciﬁc disease as the focus of the lab experiment. Consider diseases that have publicly available datasets or research data that can be accessed for analysis.

Examples of diseases could include diabetes, cardiovascular disease, cancer, respiratory disorders, etc.

* 1. Data Collection: Identify and collect relevant healthcare data related to the chosen disease. Explore public data repositories, research databases, or other reliable sources to gather datasets that contain patient information, medical records, lab results, diagnostic codes, treatment data, and any other relevant variables. Ensure compliance with ethical guidelines and data protection regulations.
  2. Data Cleaning: Clean the collected data to ensure its quality and reliability. This process may involve handling missing values, removing duplicates, standardizing formats, correcting errors, and addressing other data quality issues. Document the steps taken during the cleaning process.
  3. Data Integration: Integrate multiple datasets if available or necessary. This step involves combining data from different sources that share common variables or patient

identiﬁers. Apply appropriate techniques to merge the datasets while maintaining data integrity and ensuring consistent representations.

* 1. Data Transformation: Transform the integrated data into a suitable format for analysis and AI applications. This may involve feature engineering, scaling, normalization, encoding categorical variables, and creating derived variables. Document the transformations applied and their rationale.
  2. Exploratory Data Analysis (EDA): Perform exploratory data analysis to gain insights into the dataset and the relationships between variables. Use visualizations, statistical summaries, and other techniques to understand the distribution of data, identify patterns, and uncover any interesting ﬁndings.
  3. Summary and Documentation: Summarize the entire data preparation process, including data collection, cleaning, integration, and transformation. Document the steps taken, the challenges encountered, and the decisions made during each stage. Include any observations or insights gained from the exploratory data analysis.

Optional: Predictive Modeling: As an extension to the lab experiment, students can apply predictive modeling techniques using the prepared dataset. This can involve training machine learning models to predict disease outcomes, identify risk factors, or estimate treatment effectiveness. Students can evaluate the performance of the models and interpret their results.

Note: It is important to adhere to ethical guidelines and ensure the privacy and conﬁdentiality of patient data throughout the lab experiment. Use de-identiﬁed or simulated datasets whenever possible to avoid any privacy concerns.

**Data Repositories and Platforms:** There are various data repositories and platforms where researchers and organizations share datasets. Some popular ones include:

* Kaggle
* UCI Machine Learning Repository
* Data.gov
* NIH National Library of Medicine

**Result:**







