# Report on

**Healthcare Data Exploration**

**By**

**Prabhav Kansal – 202410116100143**

**Priyanshu Kansal – 202410116100155**

**Sagar Gupta – 202410116100174**

# Session:2024-2026 (II Semester)

Under the supervision of

# **Ms. Komal Salgotra (Assistant Professor)**

****

**DEPARTMENT OF COMPUTER APPLICATIONS**

**KIET GROUP OF INSTITUTIONS,DELHI-NCR,**

**GHAZIABAD- 201206**

**INTRODUCTION**

Healthcare is a data-rich industry, from patient data and medical imaging to genomic data and real-time feedback from wearable sensors. As the adoption of electronic health records (EHRs), Internet-of-Things (IoT)-based medical devices, and artificial intelligence intensifies, healthcare data has reached unprecedented levels of complexity and volume.

Healthcare data analysis entails the exploration of medical data for insights to inform better patient outcomes, enhanced hospital operations, and medical research. Using methods like statistical modeling, machine learning, and big data analytics, clinicians and data scientists can determine trends, find anomalies, and inform evidence-based decision-making.

This task is most important in the context of disease prediction, optimizing patient care, drug discovery, and planning for healthcare policy. Nonetheless, there are challenges accompanying healthcare data exploration, including concerns related to privacy, missing and inconsistent data, and the need for domain expertise in interpreting the findings appropriately.

Throughout this investigation, we will explore fundamental datasets, methods, and software used to analyze healthcare data, guaranteeing that insights drawn help yield more effective and efficient healthcare solutions.

**OBJECTIVES OF THE PROJECT**

This project aims to achieve the following objectives:

1. **Understanding the Data**

* Identifying data structure (rows, columns, data types)
* Checking for missing or inconsistent values
* Understanding feature distributions

2. **Data Cleaning & Preprocessing**

* Handling missing values
* Removing duplicates and outliers
* Encoding categorical variables

3. **Statistical Analysis**

* Calculating summary statistics (mean, median, standard deviation)
* Identifying trends and patterns

4. **Data Visualization**

* Using libraries like Matplotlib and Seaborn to create histograms, scatter plots, heatmaps
* Identifying relationships between variables

5. **Feature Engineering & Selection**

* Creating new meaningful features
* Selecting relevant features for predictive modeling

**METHODOLOGY**

The methodology for exploring healthcare data in Python involves a structured approach to understanding, cleaning, and analyzing the dataset. Here’s a step-by-step guide:

### **1. Data Acquisition**

* **Sources**: Healthcare data is collected from hospitals, clinical trials, EHR (Electronic Health Records), wearable devices, or open datasets (Kaggle, UCI, WHO).
* **Formats**: Data is often in CSV, Excel, JSON, SQL, or API format.
* **Python Libraries**: pandas, numpy, requests for loading data.

### **2. Data Cleaning & Preprocessing**

* **Handle Missing Values**: Impute missing values with mean, median, or mode.
* **Remove Duplicates**: Drop duplicate records to avoid bias.
* **Convert Data Types**: Ensure numerical and categorical features are correctly formatted.

### **3. Exploratory Data Analysis (EDA)**

* **Descriptive Statistics**: Mean, median, mode, standard deviation.
* **Distribution Analysis**: Histograms, box plots, KDE plots.
* **Correlation Analysis**: Heatmaps to identify relationships.

### **4. Feature Engineering & Selection**

* **Create New Features**: Combine existing columns (e.g., BMI from weight & height).
* **Handle Categorical Data**: One-hot encoding or label encoding.
* **Feature Scaling**: Normalize or standardize data for better performance.

### **5. Data Visualization**

* **Trends & Patterns**: Time series analysis for disease outbreaks.
* **Comparisons**: Bar charts, pie charts for categorical variables.
* **Geospatial Data**: Map visualizations for disease spread.

### **6. Predictive Modeling Preparation**

* **Train-Test Split**: Split dataset for model evaluation.
* **Choose Model**: Classification (e.g., diabetes prediction) or regression (e.g., hospital readmission rates).

### **7. Insights & Decision Making**

* **Actionable Insights**: Identify risk factors for diseases.
* **Operational Improvements**: Optimize hospital resource allocation.
* **Policy Recommendations**: Support public health strategies.

**CONCLUSION**

Healthcare data exploration in Python is a crucial step in understanding, analyzing, and deriving meaningful insights from complex medical datasets. By following a structured methodology—ranging from data acquisition, cleaning, and preprocessing to exploratory analysis, feature engineering, and predictive modeling—healthcare professionals and data scientists can make data-driven decisions that improve patient outcomes and optimize healthcare operations.

Through exploratory data analysis (EDA), statistical insights, and visualizations, patterns and trends in diseases, treatment effectiveness, and patient demographics can be uncovered. Additionally, feature engineering and predictive modeling pave the way for advancements in early disease detection, personalized treatment plans, and efficient resource allocation.

Ultimately, healthcare data exploration serves as the foundation for data-driven innovations in medicine, aiding in better patient care, public health strategies, and clinical decision-making. By leveraging Python’s powerful libraries and tools, professionals can transform raw healthcare data into actionable insights, leading to more efficient and effective healthcare systems.

**CODE AND OUTPUT**

import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns

# Load the dataset

df = pd.read\_csv('/content/healthcare\_data (1).csv')

# Display the first few rows of the dataset

print(df.head())

# Get basic information about the dataset

print(df.info())

# Get summary statistics

print(df.describe())

# Check for missing values

print(df.isnull().sum())

# Fill missing values with the mean (for numerical columns)

df.fillna(df.mean(), inplace=True)

# Drop rows with missing values (if necessary)

# df.dropna(inplace=True)

# Set the aesthetic style of the plots

sns.set(style="whitegrid")

# Example: Distribution of Age

plt.figure(figsize=(10, 6))

sns.histplot(df['Age'], bins=30, kde=True)

plt.title('Distribution of Age')

plt.xlabel('Age')

plt.ylabel('Frequency')

plt.show()

# Example: Correlation Heatmap

plt.figure(figsize=(12, 8))

sns.heatmap(df.corr(), annot=True, cmap='coolwarm')

plt.title('Correlation Heatmap')

plt.show()

# Example: Count of Gender

plt.figure(figsize=(8, 6))

sns.countplot(x='PatientID', data=df)

plt.title('Count of Gender')

plt.xlabel('PatientID')

plt.ylabel('Count')

plt.show()

# Example: Boxplot of Cholesterol by Gender

plt.figure(figsize=(10, 6))

sns.boxplot(x='PatientID', y='SugarLevel', data=df)

plt.title('SugrLevel By Patient ID')

plt.xlabel('PatientID')

plt.ylabel('SugarLevel')

plt.show()









