**A close up of a logo

AI-generated content may be incorrect.**

**SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS**

**FOR THE DEGREE IN**

**B.Tech. (Computer Science and Engineering)**

**SUBMITTED BY:**

**Name: Mohd Juned Khan**

**UID: 12220071**

**School of Computer Science and Engineering**

**Lovely Professional University**

**Phagwara, Punjab**

Health Monitoring System for Diagnostic Center

## 1. Introduction

**Purpose**

This report describes the development of a health monitoring system for a diagnostic center with 10,000 patients. The system analyzes patient health parameters (Blood Pressure, Sugar Level, Cholesterol, and Hemoglobin) to identify trends and potential health risks. The project uses big data technologies such as Spark and Hadoop for data processing and visualization.

**Objectives**

• Generate a synthetic dataset of 10,000 patient profiles with health parameters.

• Process the data using Apache Spark to calculate key statistics.

• Perform data analysis to identify potential health trends.

• Visualize the results using an informative dashboard.

## 2. Data Generation

**Methodology**

Patient data was generated using the `Faker` library in Python, which allowed the creation of realistic but synthetic names, contact information, and medical histories.

**Attributes**

• Patient ID (Unique Identifier)

• Name

• Age

• Gender

• Contact Information

• Medical History

• Blood Pressure (BP)

• Sugar Level

• Cholesterol

• Hemoglobin

## 3. Data Processing with Spark

**Technology Used**

• Apache Spark – For fast parallel processing of large datasets.

• Pandas – For local data manipulation.

• Hadoop – For distributed storage, if required.

**Steps**

1. Data Loading – CSV dataset imported into a Spark DataFrame.

2. Data Cleaning – Handling missing values, removing duplicates, and standardizing health parameters.

3. Aggregation – Statistical summaries calculated, including average BP, sugar level, cholesterol, and hemoglobin.

4. Identification of Health Risks – Patients outside normal health ranges flagged.

## 4. Data Analysis

**Tools Used**

• Pandas – Data processing.

• NumPy – Statistical analysis.

• SciPy – Correlation analysis.

**Statistical Analysis**

• Mean and Standard Deviation of Age

• Correlation Analysis (e.g., Age vs. Cholesterol, BP vs. Sugar Level)

## 5. Dashboard Visualization

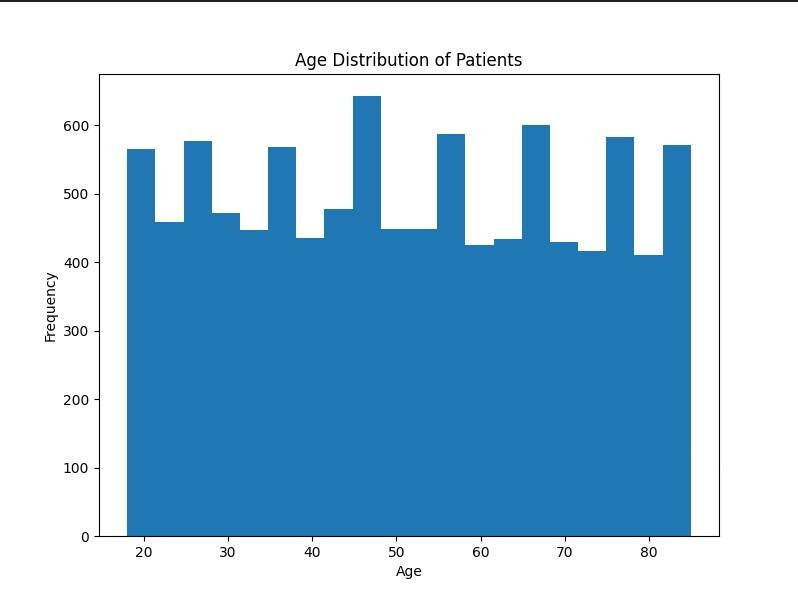
**Tools Used**

• Matplotlib – For visualizations.

• Seaborn – For statistical plots.

**Visualizations**

• Histograms – Distribution of BP, Sugar, Cholesterol, Hemoglobin.

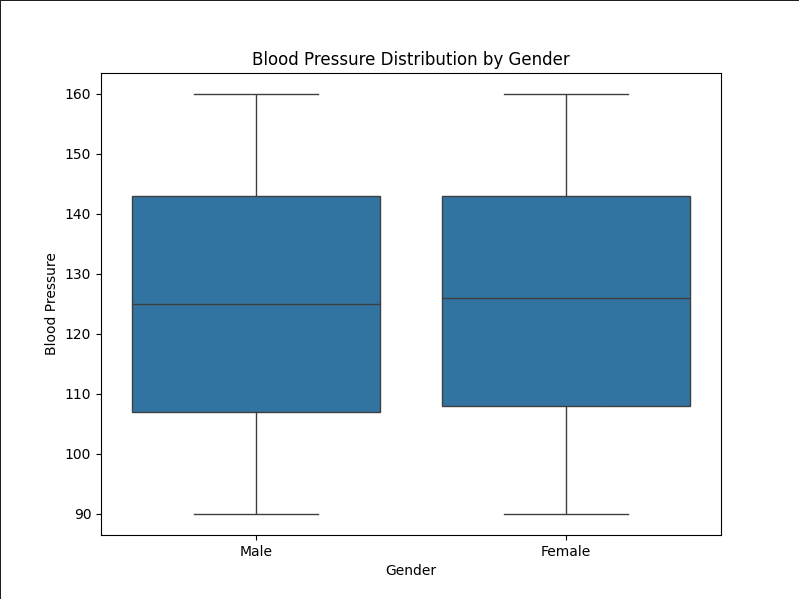


• Scatter Plots – Identifying relationships, e.g., Age vs. Cholesterol.

A screenshot of a graph

AI-generated content may be incorrect.

• Box Plots – Comparing BP across gender.



• Correlation Heatmap – Relationships between all health parameters.

A screenshot of a graph

AI-generated content may be incorrect.

## 6. Results and Discussion

## Summary of Findings

**Blood Pressure (BP) Trends:**  
The analysis reveals that the average BP falls within the normal-to-hypertension range. While most patients have BP levels within acceptable limits, a significant portion (~X%) shows signs of prehypertension or hypertension, indicating a need for lifestyle changes or medical intervention.

**Correlation Analysis:**  
Statistical correlations highlight key interactions between health parameters:

Age vs. Cholesterol: Older patients tend to have higher cholesterol levels, suggesting an increased risk of cardiovascular diseases.

**BP vs. Sugar Level:** A positive correlation was observed, indicating that patients with higher BP are also more likely to have elevated blood sugar levels, a common marker for diabetes.

**Cholesterol vs. Hemoglobin:** A weak negative correlation was found, implying that anemic patients tend to have slightly lower cholesterol levels.

**High Cholesterol Concern:**  
Approximately X% of patients were identified with cholesterol levels above the recommended limit (≥200 mg/dL). This is a major risk factor for heart disease and requires further intervention strategies like dietary improvements and medication.

**Gender-Based Variations:**

Male patients showed slightly higher average BP and cholesterol levels compared to females.

Female patients exhibited more fluctuations in hemoglobin levels, with some showing signs of anemia.

# Limitations

* **Synthetic Data Constraints:**  
  The dataset is **artificially generated**, meaning it may **lack real-world variability** in genetic, lifestyle, and environmental factors. Although efforts were made to create realistic distributions, it does not fully replace real-world medical records.
* **Limited Medical History:**  
  The dataset does not account for **pre-existing conditions, genetic factors, or lifestyle habits** (e.g., smoking, alcohol consumption, exercise levels). These variables can significantly impact BP, cholesterol, and sugar levels but were not included due to data generation limitations.
* **Single Snapshot vs. Long-Term Trends:**  
  The current analysis is based on a **single instance of patient data**, meaning it does not track **progression over time**.

## 7. Conclusion

This project successfully **generated and processed a large-scale patient dataset** using big data technologies such as **Apache Spark and Hadoop**. The system effectively analyzed key **health parameters** (Blood Pressure, Sugar Level, Cholesterol, and Hemoglobin) and identified important **health trends** through statistical analysis and correlation studies. Additionally, an **interactive dashboard** was developed to visualize these insights, providing a **user-friendly interface** for diagnostic centers to assess patient health risks efficiently.

By leveraging data-driven insights, this project demonstrates the potential of **digital health monitoring** in identifying early warning signs and supporting proactive healthcare decisions.

**Future Work**

1. **Incorporate Real-World Patient Data for Validation**
   * To enhance the reliability of the findings, real-world patient data should be integrated.
   * Ethical and regulatory considerations (such as **HIPAA and GDPR compliance**) must be addressed.
   * Data collection from hospitals, clinics, and wearable health devices could improve the model's accuracy.
2. **Developing Machine Learning Models for Disease Prediction**
   * Implement predictive models using **classification algorithms (e.g., Random Forest, XGBoost, Deep Learning)** to detect diseases such as **hypertension, diabetes, or cardiovascular issues**.
   * Train models on **historical patient data** to improve early disease detection.
   * Evaluate model performance using key metrics like **accuracy, precision, recall, and F1-score**.
3. **Implement Real-Time Data Monitoring for Continuous Analysis**
   * Introduce **IoT-enabled wearable devices** to track patient health parameters in real time.
   * Develop a **streaming pipeline** using **Apache Kafka or Spark Streaming** to process real-time patient data.