**Health Monitoring System - Project Report**

**1. Introduction**

The **Health Monitoring System** is a data-driven approach to analyzing and visualizing patient health records. The project includes a dataset of **10,000 patients**, PySpark-based data processing, and various statistical and visual insights. The system leverages **Big Data technologies**, including **Hadoop** and **PySpark**, for scalable data processing and analysis. The dataset was generated using a Python script that simulates real-world patient health metrics based on statistical distributions.

**2. Dataset Overview**

The dataset contains the following attributes:

* **Demographics**: Patient\_ID, Name, Age, Gender
* **Physical Health**: Weight\_kg, Height\_cm, BMI
* **Vital Signs & Lab Results**: BP\_Systolic, BP\_Diastolic, Sugar\_Level, Cholesterol, Hemoglobin
* **Size**: 10,000 records with no missing values.
* **Data Generation**: The dataset was created using the Faker library in Python, ensuring a diverse and realistic set of patient records.

**3. PySpark & Big Data Analysis**

**3.1 Data Processing**

* **Loaded Data**: Read CSV into a PySpark DataFrame.
* **Schema & Summary**: Displayed column types and descriptive statistics.
* **Hadoop Integration**: Processed large-scale data using Hadoop's distributed storage.

**3.2 Statistical Aggregations**

* **Average Metrics**:
  + BMI: Analyzed distribution across different age groups.
  + Sugar Levels & Cholesterol: Examined trends and deviations.
* **Blood Pressure Cases**:
  + Counted patients with high blood pressure to assess overall health risk.

**3.3 Filtering & Risk Analysis**

* Identified high-risk patients based on BMI, blood pressure, and sugar levels.
* Applied PySpark functions to classify abnormal health metrics for further analysis.

**4. Data Visualization & Insights**

**4.1 BMI Analysis**

* **Underweight** (1142 patients) – 11.42%
  + These patients may be at risk of malnutrition or other deficiencies.
* **Normal Weight** (2635 patients) – 26.35%
  + A small percentage of patients fall within a healthy BMI range.
* **Overweight** (6223 patients) – 62.23%
  + The majority of patients are overweight, which could indicate potential health risks related to obesity.

**4.2 Blood Pressure Analysis**

* **High BP (Hypertension):** 4494 patients (44.94%)
  + Nearly half of the patients have high blood pressure, increasing their risk of cardiovascular diseases.
* **Low BP:** 0 patients
  + No cases of dangerously low blood pressure were found.
* **Normal BP:** 3326 patients (33.26%)
  + Only one-third of the patients maintain a normal blood pressure level.

**4.3 Diabetes Risk (Blood Sugar Analysis)**

* **Diabetes Risk:** 6097 patients (60.97%)
  + A significant proportion of the dataset exhibits high blood sugar levels, indicating a potential risk of diabetes.
* **Low Sugar:** 0 patients
  + No instances of hypoglycemia were recorded.
* **Normal Sugar Level:** 3903 patients (39.03%)
  + A minority of patients maintain normal blood sugar levels.

**4.4 Cholesterol Analysis**

* **High Cholesterol:** 5005 patients (50.05%)
  + Half of the patients have elevated cholesterol levels, which could lead to heart disease.
* **Low Cholesterol:** 0 patients
  + No instances of abnormally low cholesterol levels.
* **Normal Cholesterol:** 4995 patients (49.95%)
  + The dataset is evenly divided between normal and high cholesterol levels.

**5. Key Findings & Insights**

* **Blood Pressure Trends**: A substantial portion of the population is at risk of hypertension, necessitating lifestyle interventions.
* **Obesity & Sugar Levels**: A strong correlation exists between high BMI and elevated blood sugar levels.
* **Cholesterol Levels**: A significant percentage of the dataset has cholesterol-related health risks.
* **Big Data Scalability**: Using **Hadoop & PySpark**, the dataset was efficiently processed, demonstrating the capability of handling large-scale health data.

**6. Conclusion**

This project successfully demonstrates a **data-driven approach to health monitoring**. By leveraging **Big Data technologies like Hadoop**, **PySpark for scalable data analysis**, and **Seaborn/Matplotlib for visualization**, meaningful insights into patient health trends were extracted. Future work could involve **predictive modeling** for health risk assessment, integrating real-time data processing with **Hadoop & Spark Streaming** to enhance efficiency and accuracy.

**Author**: Abhishek Kumar Mishra  
**GitHub Repository**: [Health Monitoring System](https://github.com/arvbrahman/Health-Monitoring-System)