Key Learnings from Healthcare Dataset

1. Dataset Overview

The healthcare dataset consists of various records related to patients, including information such as Hospital_code, Hospital_type_code, City_Code_Hospital, Hospital_region_code, Department, Ward_Type, Ward_Facility_Code, Bed Grade, patientid, Age, Type of Admission, Severity of Illness, Visitors with Patient, Admission_Deposit, and whether the patient was discharged or not.

2. Key Learnings

• The dataset includes 18 features that provide insights into hospital demographics, patient characteristics, and admission details.

• Patient Age, Admission Type, Severity of Illness, and Admission Deposit are critical features in predicting discharge outcomes.

• Hospitals are categorized by type and region, which may affect patient handling and discharge likelihood.

• Departments and Ward types can indicate specialization and resource allocation, impacting patient outcomes.

3. Insights from Visualizations

• Most admissions come under 'Trauma' and 'Emergency' types, indicating critical care demands.

• Discharge rates vary based on severity, with patients having 'Extreme' illness showing lower discharge percentages.

• Younger patients (0–10 years) and older patients (70+) show distinct trends in discharge rates and admission deposits.

4. Using PySpark for Analysis

PySpark is a powerful tool for handling large-scale healthcare datasets in a distributed computing environment. By leveraging PySpark’s DataFrame API and MLlib, we can efficiently:

• Clean and preprocess massive datasets (handle missing values, normalize fields, convert types, etc.).

- Perform EDA (Exploratory Data Analysis) using SQL queries on structured data.

- Build machine learning models such as logistic regression or random forests to predict patient discharge outcomes.

- Use parallel processing to significantly reduce computation time compared to pandas-based workflows.