Predicting Hospital Readmissions and Patient Outcomes Using MIMIC-III and Diabetes Data

Project Overview:

The goal of this project is to analyze hospital readmissions and patient outcomes by combining two datasets:

- **Diabetes 130-US hospitals dataset** (readmissions data).
- **MIMIC-III Clinical Database** (ICU data).

This will allow us to investigate which clinical factors affect readmissions, explore the relationships between various patient-level data, and build models to predict patient outcomes based on combined datasets.

Project Workflow:

Task 1: Merging and Analyzing the Data

1. **Data Collection and Cleaning**:
    - Prepare and clean data from both the **Diabetes 130-US hospitals** and **MIMIC-III ICU** datasets. Handle missing values, standardize formats, and ensure consistency across variables.
    - Extract key features from the ICU data (e.g., lab results, diagnoses, vital signs, medications), and combine it with the diabetes dataset based on relevant patient details.
2. **Exploratory Analysis**:
    - Analyze patient demographics (age, gender, race) and clinical features (ICU stay length, number of diagnoses, lab results). Investigate how these variables relate to patient outcomes such as readmission.
    - Create initial summary statistics for key variables to understand data distributions and potential trends.
3. **Data Visualization**:
    - Create visualizations such as histograms, box plots, and scatter plots to represent the distribution of key variables (e.g., ICU length of stay, number of lab procedures, readmission rate).

Task 2: Investigating Patterns and Relationships

1. **Exploring Variable Relationships**:
    - Investigate how different patient-level features, such as the number of medications, number of lab procedures, and ICU stay length, are associated with readmission outcomes.
    - Explore potential differences in outcomes based on patient characteristics, including age, gender, and diagnosis type.

2. **Checking Group Differences**:
    - Analyze whether different groups (e.g., age groups, medication categories, lab test counts) have significant differences in patient outcomes (e.g., readmission rates or length of stay).
    - Assess how clinical features (e.g., treatments, medications) differ among patients with different readmission outcomes.
3. **Exploring Distributions**:
    - Study the distribution of patient outcomes and clinical features (e.g., number of medications or lab results) to identify trends across different groups of patients.
    - Identify patterns that could indicate potential risk factors for patient readmission or longer hospital stays.

Task 3: Predicting Patient Outcomes

1. **Modeling the Data**:
    - Build a model to predict the likelihood of patient readmission based on the combined datasets. Use patient demographics and clinical data such as vital signs, lab procedures, and ICU stay information to predict outcomes.
    - Evaluate the model's performance using standard metrics and identify key factors influencing predictions.
2. **Segmenting Patients**:
    - Group patients into different segments based on their clinical data (e.g., number of medications, number of lab tests, ICU stay length). Use this segmentation to identify high-risk patient groups.
    - Analyze which patient segments are more likely to have specific outcomes such as readmission or extended ICU stays.
3. **Evaluating Outcomes**:
    - Assess the predictive power of the model by checking how well it identifies high-risk patients.
    - Identify the most important features that drive the model's performance and provide insights into which clinical factors have the most significant impact on patient outcomes.

Task 4: Deployment

- **Deploying the Predictive Model**:
    - Deploy the trained predictive model to a production environment (e.g., using Flask or Django) to enable real-time predictions for hospital readmissions.
    - Create a web interface for healthcare professionals to input patient data and receive a prediction regarding the likelihood of readmission.

Deliverables:

1. **Merged and Cleaned Dataset**:
    - A fully cleaned and merged dataset from the **Diabetes 130-US hospitals dataset** and **MIMIC-III ICU dataset**, prepared for analysis.
2. **Exploratory Analysis and Visualizations**:
    - Summary statistics and visualizations (histograms, box plots, scatter plots) showing key trends and relationships between variables such as ICU length of stay, number of lab procedures, medications, and readmissions.
3. **Analysis of Group Differences and Relationships**:
    - A detailed analysis of how patient demographics, clinical features, and treatments are associated with different outcomes. This includes insights into patient groups with significantly different outcomes.
4. **Predictive Models**:
    - A model to predict patient readmission based on the combined datasets, with metrics on the model's performance and a detailed analysis of important features.
5. **Segment Analysis**:
    - A segmentation of patients based on clinical features, showing which groups are at higher risk for readmission or extended ICU stays. Recommendations based on patient segmentation will be provided.
6. **Deployment of Model**:

- A deployed predictive model for hospital readmissions, accessible via a web interface for healthcare professionals.
- Documentation on how to use the interface, input patient data, and interpret the readmission predictions.

7. **Final Report and Presentation**:
    - A final report summarizing the key findings from the project, including exploratory analysis, relationships between variables, predictive modeling results, and actionable insights.
    - A presentation showcasing the analysis and model insights, ready for stakeholders.

Dataset Sources:

- **Diabetes 130-US hospitals dataset**: Download Link
- **MIMIC-III ICU dataset**: Dataset attached