**PREDICTING RISK OF PATIENT READMISSION**

**Team Members:**

* **Tatiana Celine**
* **Kamal Ali**
* **William Muthama**
* **Gilead Gad**
* **Simon Muema**
* **Graffin Kiprotich**

GitHub Repository: [Predictive Model for Patient Readmission Risk](https://github.com/Tatianaceline/Predictive-Model-for-Patient-Readmission-Risk)

**Project Objective**

The primary objective is to develop a predictive model that can accurately identify patients at high risk of readmission. Such a model will help healthcare providers proactively manage patient care and reduce costs.

**CRISP-DM Methodology**

The CRISP-DM method was used in this project to ensure a systematic approach to predicting patient readmission risk. By following CRISP-DM, the project team was able to align the model development with the healthcare objective of identifying high-risk patients, making sure each phase; from understanding business needs, to preparing and analyzing data, directly contributed to this goal. This structured approach provided a clear path from initial data exploration to model evaluation, helping the team address the complexities of medical data while focusing on building a reliable predictive model that healthcare providers can use effectively.

**1. Business Understanding**

Understanding the problem of hospital readmissions is critical, as it affects patient outcomes and hospital costs. This project aims to:

* Identify key predictors of readmission risk.
* Assess the correlation of patient characteristics and hospital metrics with readmission likelihood.
* Develop a model that flags high-risk patients, aiding hospital staff in improving patient management strategies.

**Key Questions:**

* What patient characteristics are highly correlated with readmission?
* How significant is the length of stay in predicting readmission risk?
* Can actionable insights be derived from the predictive model?

**2. Data Understanding**

The dataset used includes detailed patient information across demographics, hospital visits, medical treatments, and readmission indicators. Key columns considered include:

* **Demographics:** age, gender, weight.
* **Hospital Metrics:** time in hospital, number of procedures, number of medications, outpatient/emergency/inpatient visits.
* **Medical Metrics:** number of lab procedures, primary care specialty, diabetes medication status.
* **Target Variable:** readmission indicator.

This phase involved an initial review to determine available data and assess each feature's relevance to predicting readmission.

**3. Data Preparation**

**3.1 Data Import and Library Setup**

Libraries were imported to support data analysis, cleaning, and modeling processes. Essential libraries include those for data handling, visualization, and machine learning.

**3.2 Data Cleaning**

Steps taken for data cleaning include:

* **Dropping Unlisted Columns:** Non-essential columns were removed to focus on relevant data.
* **Handling Missing Values:** Missing data was identified and handled appropriately to prevent bias in modeling.
* **Handling Duplicated:** Repeated values in the unique identifier column were removed.

**4. Exploratory Data Analysis (EDA)**

The EDA phase provided insight into the data distribution, relationships, and outliers:

* **Data Distribution Analysis:** Reviewed the spread of continuous variables like age, weight, and time in hospital.
* **Feature Correlation:** Analyzed correlations to identify which features significantly impact readmission risk.
* **Outlier Detection and Treatment:** Identified outliers that could skew model performance, considering medical context for appropriate handling.

**5. Further Data Cleaning Based on EDA**

Adjustments were made to the data based on findings from EDA, ensuring the dataset was well-prepared for modeling. This phase included refining feature selection and removing any remaining inconsistencies.

**6. Modeling and Evaluation**

Multiple models were developed, evaluated, and compared to find the best predictive approach for readmission risk. The process involved:

* **Model Selection:** Various machine learning algorithms were explored, including logistic regression, decision trees, and ensemble methods.
* **Hyper-parameter Tuning:** Conducted to optimize each model's performance.
* **Cross-Validation:** Used to validate model robustness and avoid overfitting.
* **Evaluation Metrics:** Performance measured using accuracy, precision, recall, and F1-score, focusing on models with high recall to minimize missed high-risk patients.

**7. Predictive Modeling and Final Evaluation**

The final model selection was based on achieving a balanced performance that maximized readmission detection accuracy while minimizing false positives. Testing on unseen data provided a realistic evaluation of model effectiveness.

**8. Conclusion and Future Work**

The project concluded with a successful model for predicting patient readmission. Key takeaways include:

* The importance of specific patient demographics and hospital metrics in determining readmission likelihood.
* The need for continuous data updates to enhance model adaptability.

Future recommendations include integrating real-time patient data and exploring advanced deep learning models for potentially improved performance.