1. **Project Title**: Diabetic Patients Readmission in Hospitals Prediction

2. **Group Member**:

I am doing this project Alone.

- Akshay Kumar Mahto, [**650848332**].

3. **Problem Statement:**

Eventually, this will improve health outcomes and reduce healthcare spending because healthcare providers can manage their patients better. For this project, we want to build a machine learning model for predicting 30-day readmission among diabetic patients using patient demographic information, laboratory test results, and treatment data. With risk predictions, timely interventions will be provided to prevent such unnecessary readmission, hence raising the level of quality in care. The problem of readmission is very well-suited to machine learning due to a host of influencing variables.

It is important to know if a patient will be readmitted in some hospital. The reason is that you can change the treatment, in order to avoid a readmission.In this database, you have 3 different outputs:

1. No readmission;
2. A readmission in less than 30 days (this situation is not good, because maybe your treatment was not appropriate);
3. A readmission in more than 30 days (this one is not so good as well the last one, however, the reason can be the state of the patient.

The main problem this project has tried to solve is the prediction of diabetic patient readmission to the hospital within 30 days following discharge. Hospital readmission is one of the major areas of concern for any health institution because it always comes with a cost implication and can be a possible pointer for less-than-adequate care for the patient. Diabetes being a chronic disease requires extra caution in management; therefore, an unplanned readmission mostly points to a complication which could have been avoided by proper attention in time.

The focus of this project, through the application of machine learning techniques, will be to ascertain patterns and factors resulting in high readmission rates. Such patterns, if understood, will also help healthcare providers take necessary preventive measures, manage resources effectively, and enhance patient outcomes.

Significance:

1.Healthcare Impact: One of the main priorities at hospitals is to decrease the rate of readmission. This is because of heavy monetary penalties from regulators and for the betterment of healthcare for patients. Machine learning can identify high-risk patients, thus allowing early interventions and management of better care.

2.Cost reduction: Prevention of readmission may drastically lower healthcare expenditures for patients, hospitals, and insurance providers. Accurate identification of patients at risk of readmission may enable hospitals to target at-risk patients more efficiently with the purpose of decreasing readmissions.

3.Data complexity: The data associated with a patient has many dimensions, like demography, history, treatment, and laboratory results. The volume and complexity of data need conclusions that are difficult to arrive at by mere manual reasoning; hence, adaptation to machine learning algorithms could analyze and give meaningful patterns.

Expected Challenges:

1.Class Imbalance: The count of readmitted patients is usually much lesser than those not readmitted. Hence, most of the machine learning algorithms might get biased to the negative class and thus give poor results. So, balancing the dataset using some techniques like resampling or adjustment of decision thresholds will be required.

2.Missing Data: The medical and treatment information of some entries in the dataset may be incomplete. There will be a need for efficient imputation strategies so that the models can produce valid predictions without prejudice to lost information.

3.High Dimensionality: this dataset consists of more than a couple of thousand variables that include medication, lab tests, and many more. This all requires feature selection or dimensionality reduction to then focus on the most important factors associated with readmissions.

4. Patient variability: different patients have different responses to various treatments. This again causes a problem for generalizing across diverse populations. Models will need to account for this variability in order to make proper predictions.

Suitability to Machine Learning: This problem becomes a good candidate for machine learning due to the volume of data, a large number of features, and nonlinear relations between variables that may impact readmission risk. It is challenging or impossible for conventional statistical methods to model such complex interactions, while machine learning algorithms will be able to identify patterns and trends in the data not easily found by traditional methods. Moreover, the more data machine learning models receive, the better they will be. Thus, this may allow healthcare providers to work their predictive capabilities in a progressively positive direction.

4. **Dataset:**

a) **Name**: Diabetes 130-US hospitals for years 1999-2008 dataset.

b). **Source**: UCI Machine Learning Repository.

c). Size: The dataset consists of over 100,000 records of diabetic patients, including various attributes related to patient demographics, medical history, lab test results, and treatments.

d). Key Features:

- **Demographics**: Age, gender, ethnicity.
- **Medical Information**: Number of medications, types of diagnoses, results of lab tests.
- **Hospitalization Details**: Length of stay, number of previous admissions, discharge type.
- **Treatment Data**: Medications prescribed, number of lab procedures, and insulin levels.

These features are critical for determining the factors that influence hospital readmissions for diabetic patients. For example, treatment data and previous hospital admissions are likely significant predictors of future readmission.

Relevance to Problem:

This dataset is directly relevant to the prediction problem of hospital readmission among diabetic patients. The richness in this dataset is multidimensional, enabling it to be used in the analysis of several factors that contribute toward the readmission of patients, relating demographic factors and treatment plans with previous hospitalization. It allows performing comprehensive analytics on data to find the patterns leading toward readmission. This makes it a very good choice to build machine learning models for the prediction of readmission risk with the goal of improving patient outcomes.

5. **Methodology**:

Machine Learning Techniques:

1. **Logistic Regression**:
   - ○ A simple and interpretable classification algorithm, ideal for binary outcomes like predicting whether a patient will be readmitted or not. It provides insights into how individual features affect the likelihood of readmission.
2. **Random Forest**:
   - ○ A robust ensemble learning technique that reduces overfitting and improves accuracy by aggregating predictions from multiple decision trees. It is well-suited for handling the complex interactions between patient attributes and provides feature importance scores.
3. **XGBoost**:
   - ○ A high-performance gradient-boosting algorithm that works well with structured data, making it an excellent choice for predictive tasks like this. XGBoost handles missing data and outliers effectively and is known for delivering high predictive accuracy.

4. **Neural Networks:** A neural network is a set of algorithms that attempts to recognize underlying relationships in a batch of data using a methos that mimics how the human brain works. Neural networks, in the context, refer to systems of neurons that can be organic or artificial in nature.

Justification for Approaches:

- Logistic Regression: It is chosen for its simplicity and interpretability. In the healthcare context, understanding the contribution of each feature to the prediction is crucial for clinicians.
- Random Forest : It is used due to its ability to capture non-linear relationships and interactions between features, which is important for complex healthcare datasets.
- XGBoost : It is included for its strong performance in handling imbalanced datasets and its capability to optimize predictive accuracy through gradient boosting.
- Neural Networks: A neural network is a set of algorithms that attempts to recognize underlying relationships in a batch of data using a methos that mimics how the human brain works. Neural networks, in the context, refer to systems of neurons that can be organic or artificial in nature.

These approaches complement each other, providing a balance between interpretability and prediction performance.

Preprocessing and Feature Engineering:

1. **Handling Missing Data**:
    - ○ Missing values will be imputed using appropriate techniques like mean/mode imputation for numerical and categorical variables. Alternatively, XGBoost's built-in mechanism to handle missing values will be utilized.
2. **One-Hot Encoding**:
    - ○ Categorical variables (e.g., discharge type, ethnicity, and diagnosis codes) will be converted into numerical format using one-hot encoding to ensure they are machine-readable.
3. **Feature Scaling**:
    - ○ Continuous variables like age, length of stay, and lab test results will be scaled using standardization (z-score normalization) to bring them onto the same scale, improving model performance, especially for algorithms sensitive to scale (e.g., Logistic Regression).
4. **Class Imbalance**:
    - ○ Since hospital readmissions are expected to be less frequent, techniques such as oversampling (SMOTE) or adjusting decision thresholds will be used to address the class imbalance.

Exploratory Data Analysis (EDA) and Visualization:

- **Correlation Analysis**: A correlation matrix will be used to understand relationships between numeric features like age, length of stay, and number of medications.
- **Data Visualization**: Histograms, bar charts, and box plots will be created to explore the distribution of key features like age, gender, number of previous admissions, and treatment type. Visualizations will help identify trends and anomalies in the dataset.
- **Feature Importance**: Random Forest and XGBoost will provide feature importance metrics, which will be analyzed to identify the most significant factors contributing to patient readmission.

6. **Evaluation Metrics:**

1. **Accuracy**:
    - ○ Measures the percentage of correctly predicted instances (both readmitted and not readmitted) out of the total predictions.
    - ○ **Suitability**: While accuracy gives a general sense of model performance, it might not be sufficient due to potential class imbalance (fewer readmitted cases).
2. **Precision**:
    - ○ Indicates the proportion of true positive predictions (correctly identified readmissions) out of all predicted positives (patients predicted to be readmitted).

   - **Suitability**: Precision is important in this context because predicting a readmission incorrectly (false positives) may lead to unnecessary interventions or resource allocation.
3. **Recall**:
   - Measures the proportion of true positive predictions (correctly identified readmissions) out of all actual positives (patients who were actually readmitted).
   - **Suitability**: Recall is critical because the primary goal is to minimize the number of missed readmissions (false negatives), which could result in inadequate follow-up care for high-risk patients.
4. **F1-Score**:
   - The harmonic mean of precision and recall, providing a balanced measure that accounts for both false positives and false negatives.
   - **Suitability**: Given the potential for class imbalance in the dataset, the F1-score offers a more balanced assessment of model performance than accuracy alone.
5. **Area Under the ROC Curve (AUC-ROC)**:
   - The ROC curve plots the true positive rate (recall) against the false positive rate. The AUC measures the model's ability to distinguish between the two classes.
   - **Suitability**: AUC-ROC is a useful metric for evaluating model performance across different classification thresholds, making it ideal for imbalanced datasets where precision and recall are both important.

Why These Metrics Are Suitable:

- **Class Imbalance**: Since hospital readmission rates are likely lower compared to non-readmission cases, relying solely on accuracy might lead to misleading conclusions. Precision, recall, and F1-score provide deeper insights into the performance of the model, especially for the minority class (readmissions).
- **Healthcare Implications**: Precision ensures that we avoid false positives (patients wrongly predicted to be readmitted), while recall ensures that we capture as many true positives (actual readmissions) as possible. Balancing these with the F1-score and AUC-ROC helps in making informed decisions regarding patient care.

7. **Timeline:**

Week 1-2: Data Preprocessing and Exploratory Data Analysis (EDA)

- Clean the dataset by handling missing values, outliers, and inconsistencies.
- Perform exploratory data analysis (EDA) to understand the dataset structure, distribution of key variables, and relationships between features.
- Apply feature engineering techniques such as one-hot encoding for categorical variables and feature scaling for continuous variables.
- Visualize important features using correlation matrices, histograms, and box plots to guide model selection.

Week 3: Model Building and Tuning

- Train machine learning models such as Logistic Regression, Random Forest, and XGBoost using the preprocessed data.
- Tune hyperparameters for each model using techniques like cross-validation to optimize performance.
- Address class imbalance by applying oversampling techniques (e.g., SMOTE) or adjusting class weights.

Week 4: Model Evaluation and Reporting

- Evaluate the performance of the models using accuracy, precision, recall, F1-score, and AUC-ROC.
- Compare the models and select the best-performing one based on the evaluation metrics.
- Document findings and prepare a final report summarizing the methodology, results, and insights from the project.

—--

https://github.com/zeglam/Diabetes-Patient-Readmission-Classification/blob/master/README.md

https://www.youtube.com/watch?v=h6dOhTRWSaU

https://github.com/Mohith-Kota/Diabetic-patients-Readmission-in-Hospitals-Prediction

https://archive.ics.uci.edu/dataset/296/diabetes+130-us+hospitals+for+years+1999-2008

https://www.kaggle.com/datasets/brandao/diabetes