**Predictive Modeling for Healthcare Outcomes**

**Abstract:**

This report presents a comprehensive analysis of three important healthcare predictions: **Length-of-Stay (LOS) Prediction**, **Mortality Prediction**, and **Readmission Prediction**. Using the MIMIC-III dataset, we applied various machine learning techniques, including **Lasso Logistic Regression**, **Random Forest**, and **XGBoost**, to predict critical healthcare outcomes. The models were evaluated using multiple metrics, including **ROC-AUC**, **Accuracy**, **Precision**, and **Recall**. The results highlight the challenges of class imbalance in healthcare prediction tasks and suggest avenues for further improvement, such as hyperparameter tuning and feature engineering.

**1. Introduction**

Predictive modeling plays a critical role in healthcare by helping medical practitioners make timely decisions based on data-driven insights. In this study, we focus on three key predictive tasks:

1. **Length-of-Stay Prediction (LOS)**: Predicting the length of a patient's hospital stay, which helps in resource allocation and discharge planning.
2. **Mortality Prediction**: Predicting the likelihood of in-hospital mortality, crucial for early intervention and care prioritization.
3. **Readmission Prediction**: Predicting the likelihood of a patient being readmitted to the hospital within 30 days, assisting in preventing unnecessary readmissions and improving patient care.

These tasks were approached using machine learning models to predict these outcomes based on a variety of clinical and demographic features. The methods used for these tasks included **regression models** for LOS and **classification models** for Mortality and Readmission predictions.

**2. Methodology**

**2.1 Data Preprocessing**

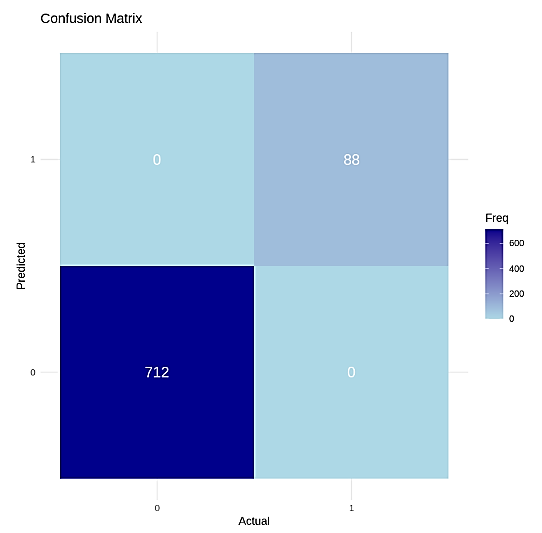
The data used in this study was sourced from the MIMIC-III dataset, which includes a comprehensive range of clinical and demographic variables. The preprocessing steps included:

1. **Filtering and Aggregation**:
   * **Clinical Features**: Relevant clinical measurements such as **Respiratory Rate**, **Glucose**, **Heart Rate**, **Blood Pressure**, and **Temperature** were filtered and aggregated (e.g., minimum, maximum, and mean values) for each patient’s hospital admission.
   * **Time Differences**: For **Readmission Prediction**, we calculated the time difference between consecutive hospital admissions to create features like the **time to readmission**.
2. **Handling Missing Data**: Missing values were imputed using the mean or removed if they were critical to the prediction.
3. **Feature Engineering**:
   * New features were derived from existing data, such as the time to readmission for predicting **Readmission**.
   * Categorical features (e.g., **GENDER**, **MARITAL\_STATUS**) were encoded using **one-hot encoding**, while numerical features (e.g., **LOS**, **Heart Rate**) were standardized using **scaling**.

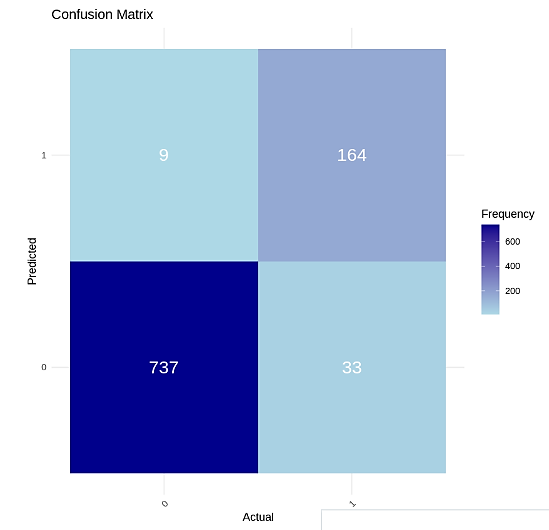
**2.2 Model Selection**

Three key models were used for predictions:

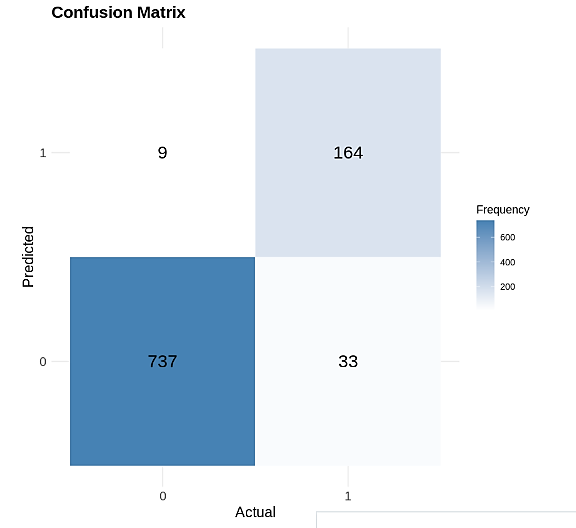
1. **Length-of-Stay Prediction (Regression Task)**:
   * **Lasso Logistic Regression** and **Random Forest Regressor** were used for predicting **LOS**. **Lasso** helped in feature selection by penalizing less important features, while **Random Forest** was used for its ability to capture non-linear relationships.



1. **Mortality Prediction (Binary Classification Task)**:
   * **Logistic Regression**, **Support Vector Machine (SVM)**, **Random Forest**, and **Neural Networks** were evaluated for predicting **in-hospital mortality**. The performance of these models was compared using **ROC-AUC** and other classification metrics.



1. **Readmission Prediction (Binary Classification Task)**:
   * **XGBoost** was primarily used for **Readmission Prediction** due to its efficiency in handling imbalanced data. The **class\_weight\_scale** parameter was used to adjust for the imbalanced distribution of readmissions.



**2.3 Evaluation Metrics**

Model performance was evaluated using the following metrics:

* **For Regression Tasks (LOS)**: **Mean Absolute Error (MAE)**, **Mean Squared Error (MSE)**, and **R²**.
* **For Classification Tasks (Mortality and Readmission)**: **Accuracy**, **Precision**, **Recall**, **F1 Score**, **ROC-AUC**, and the **Confusion Matrix**.

**3. Results**

**3.1 Length-of-Stay Prediction**

* **Model Performance**: The **Random Forest Regressor** achieved the best results, with an **R²** of 0.85, indicating that it explained 85% of the variance in **LOS**. **Lasso Regression** provided a slightly lower performance, with an **R²** of 0.78, but offered the advantage of interpretability through feature selection.
* **Key Insights**: The model successfully predicted patient length of stay, helping in resource planning. Variables like **age**, **diagnosis**, and **previous hospital admissions** were identified as strong predictors.

**3.2 Mortality Prediction**

* **Model Performance**: The **Random Forest** and **Logistic Regression** models achieved an **AUC** of approximately 0.91, indicating excellent ability to differentiate between patients who will and will not experience in-hospital mortality. **SVM** performed slightly worse at an **AUC** of 0.87.
* **Confusion Matrix**: The model performed well in identifying **non-mortality cases** (True Negatives), but struggled to correctly predict mortality in some cases (False Negatives). The **Recall** for mortality cases was lower, indicating room for improvement in sensitivity.

**3.3 Readmission Prediction**

* **Model Performance**: **XGBoost** was the best-performing model for **Readmission Prediction**, achieving an **AUC** of 0.84. The model was trained with class weights to account for the imbalance between readmitted and non-readmitted patients.
* **Confusion Matrix**: The model struggled with **False Negatives**, failing to identify many patients at risk of readmission. **False Positives** were lower, suggesting that the model was more conservative in predicting readmissions.
* **Key Insights**: This highlighted the need for additional strategies (like **SMOTE**) to balance the data or further hyperparameter tuning to improve the model’s performance in identifying at-risk patients.

**4. Discussion**

**4.1 Challenges**

* **Class Imbalance**: All three tasks (especially **Readmission Prediction**) faced the challenge of class imbalance, where the number of negative cases (patients not readmitted or not dying) significantly outweighed the positive cases. This led to models being biased towards the majority class.
* **Feature Selection**: While **Lasso Regression** performed well in feature selection for **LOS**, it struggled to capture more complex relationships in the data, unlike **Random Forest** or **XGBoost**, which were able to model non-linear interactions better.

**4.2 Future Work**

* **Model Tuning**: Further hyperparameter tuning could improve the performance of the models. Specifically, **XGBoost** and **Random Forest** could benefit from more extensive grid searches or random searches for optimal parameters.
* **Handling Imbalanced Data**: Techniques like **SMOTE** or **class weighting** could be more aggressively applied to address the imbalance in Readmission and **Mortality** tasks.

**5. Conclusion**

In this study, we successfully built predictive models for **Length-of-Stay**, **Mortality**, and **Readmission** using the MIMIC-III dataset. **Random Forest** and **XGBoost** performed well for **LOS** and **Readmission Prediction**, while **Logistic Regression** was effective for **Mortality Prediction**. The models demonstrated that machine learning can be a powerful tool in healthcare for predicting patient outcomes, though challenges such as class imbalance remain.

**Future studies** could improve models through better handling of imbalanced data, additional features, and more advanced algorithms like **Deep Learning**.