**RESEARCH PROPOSAL**

**PROJECT SUMMARY**

This project aims to use the Medical Information Mart for Intensive Care III (MIMIC-III) dataset and tools such as R, Python, SQL, and Tableau to study hypertension and its clinical outcomes. With the enhancement in understanding of hypertension and the utilization of data-driven hypertension strategies, we can decrease the morbidity and associated mortality, improve in public health outcomes, reduce healthcare costs, lead to personalized treatment plans that are more effective for individual patients, and provide the evidence base necessary for policies for managing hypertension at the population level.

This project has three aims: 1) Exploratory Data Analysis and Feature Engineering; 2) Predictive Analysis of Mortality and LOS; 3) Shiny App and Tableau Dashboard Development

**SPECIFIC AIMS**

1. Exploratory Data Analysis and Feature Engineering
2. Predictive Analysis of Mortality and LOS
3. Shiny App and Tableau Dashboard Development

**RESEARCH STRATEGY**

* SIGNIFICANCE

According to the CDC, hypertension is a significant health concern in the United States. Nearly half of U.S. adults (48.1%) suffer from hypertension. In 2021, it was a primary or contributing factor in nearly 691,095 deaths. Also, the economic impact of high blood pressure is substantial, costing the U.S. an average of $131 billion annually from 2003 to 2014. Given that most Americans will develop hypertension during their lifetime, it is crucial to implement early preventive strategies and timely management.

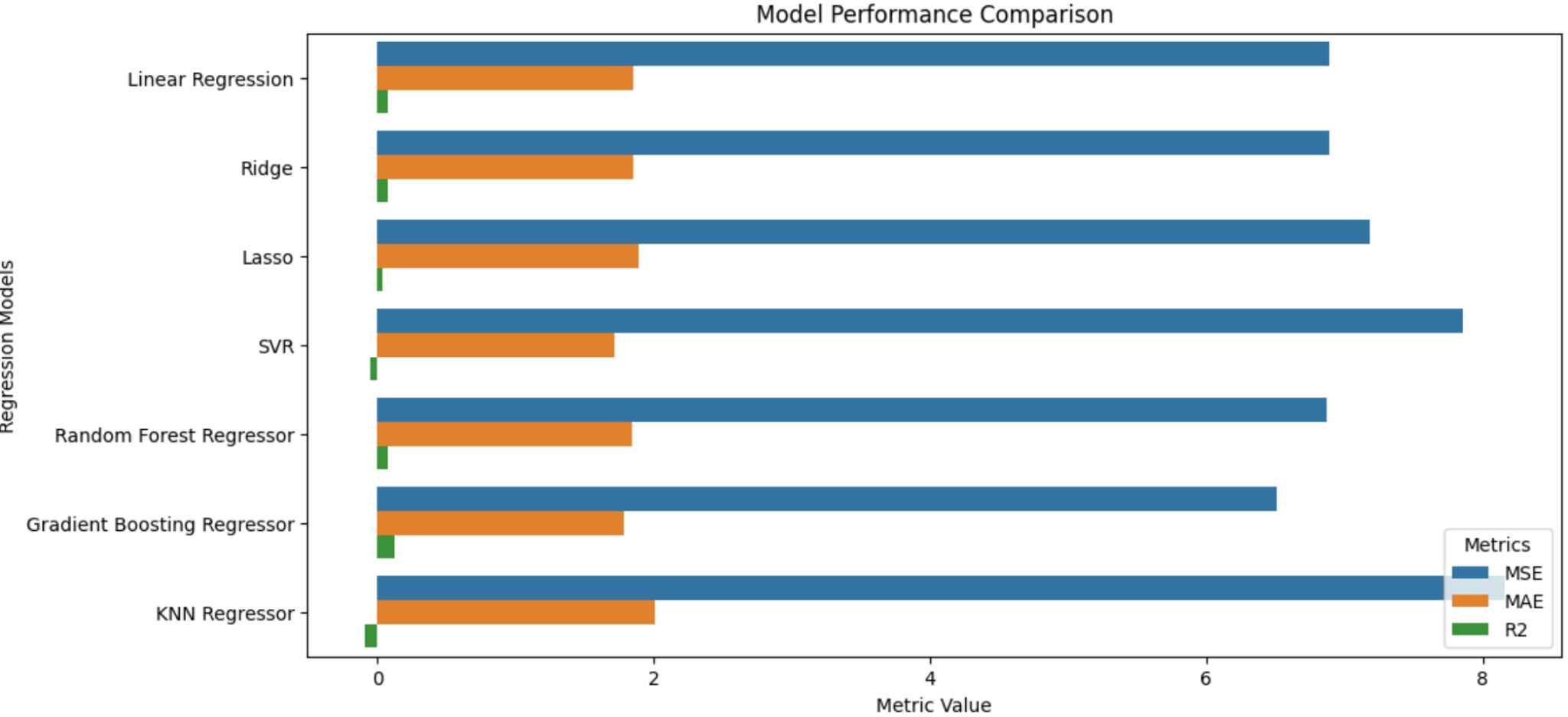
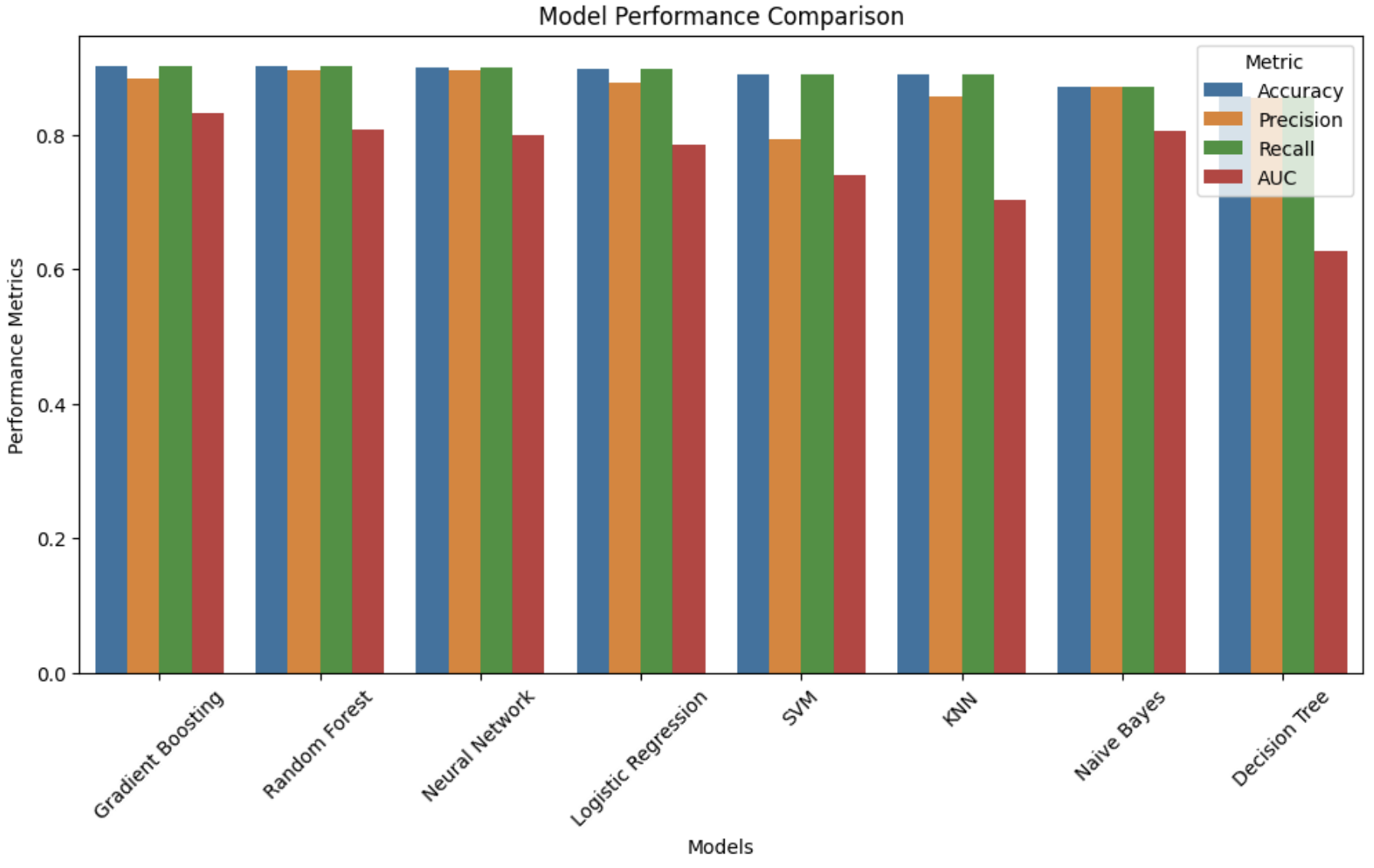
Our project presents an important advancement in the management and understanding of hypertension. By investigating the correlations between patient demographics, lab results and clinical outcomes like length of stay (LOS) and mortality, this study provides evidence support for hypertension management, which is vital for optimizing treatment plans and improving patient care efficiency.

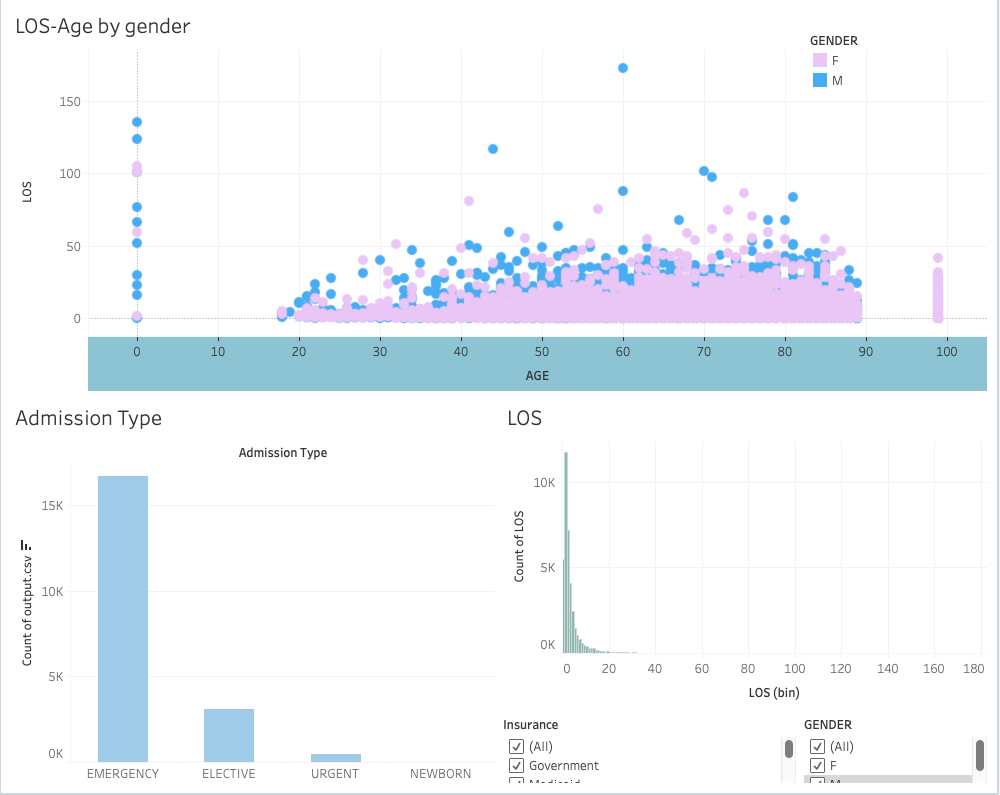
* INNOVATION

The innovation of our project lies in the development of multiple predictive models, specifically for hypertension. By using Tableau and R Shiny, we create interactive tools that can transform our findings into practical application. The Tableau dashboard could serve as important support for decision-making and monitoring. The R shiny tool allows healthcare providers to input patient-specific data and receive real-time predictions on LOS and mortality risks. Overall, our project not only advances medical research but also directly contributes to improved healthcare outcomes and hypertension management through data-driven strategies.

* SPECIFIC AIM 1: Exploratory Data Analysis(EDA) and Feature Selection
  + Hypothesis: EDA enables a basic understanding of the data characteristics that prepare for future data modeling. Results from EDA are converted to dataframe for easy handling and visualization. We will also investigate the correlation between various variables and mortality and LOS via a random forest model.
  + Rationale: By EDA, we can better understand patterns within the data, detect outliers and errors, select, and exclude variables in the dataset. It ensures the dataset for following modeling is applicable and valid. By using a variety of tests, procedures, and plots, we can determine the best distribution and structure of the data.
  + Experimental Approach: By setting up an SQLite database and establishing a connection to Google BigQuery, we have access to retrieve the datasets in MIMIC-III database and use SQL queries to perform data manipulation and cleaning. Python modules such as matplotlib, seaborn, pandas, and numpy are used to create visualizations and summary tables. ‘Wordcloud’ module is applied to produce a word cloud for variables in the dataset, and random forest regressor and classifier of scikit-learn are utilized to perform feature selection.
  + Interpretation of Results: Important variables that are selected by the feature selection process will be extracted for further analyses.
  + Potential Problems and Alternative Approaches: Problem: Drawing incorrect conclusions from EDA if the methods are not well-suited to the data's characteristics. Approaches: Supplement EDA with more robust statistical tests.
* SPECIFIC AIM 2: Predictive Analysis of Mortality and LOS
  + Hypothesis: Hypertension NOS is theoretically related to mortality and LOS. The data we acquired from the MIMIC database are samples across the United States and have the same distribution as the real-word data.
  + Rationale: By implementing the random forest model in the data selection stage, we are able to select variables according to their importance in the random forest model. We will use several classification models including tree-based mode, support vector machine, ensembling models and neural networks to predict the mortality rate. We will also use regression models including LASSO, ridge, KNN and advanced regressors to predict the LOS.
  + Experimental Approach: We will first conduct EDA and data pre-processing. As Rationale mentioned, we will respectively establish classification and regression models. To evaluate the model performance, we will use AUC, precision, recall and accuracy for the classification models and R square, Mean Absolute Error (MAE) and Mean Square Error (MSE) for the regression models.
  + Interpretation of Results: We will use bar plots to compare the model performance using different evaluation methods.
  + Potential Problems and Alternative Approaches: The best model may have less transparency or interpretation. Therefore, we should take the application potential into consideration.
* SPECIFIC AIM 3: Shiny App and Tableau Dashboard Development
  + Hypothesis: Interactive predictive tools via R Shiny and Tableau dashboard could become important guidance for physicians decision-making, hypertension management, and hospital resource management.
  + Rationale: Interactive tools can foster a better and deeper understanding of features and their impact on clinical outcome, provide insightful visuals from patient data, enable real-time predictions that will enhance decision-making processes in clinical settings.
  + Experimental Approach: With R Shiny, we aim to develop two interactive tools to dynamically apply our predictive models: The Length of Stay (LOS) Prediction Tool and the Mortality Prediction Tool enable users to input data for selected variables to receive an estimated LOS. This estimate is derived from the insights gained through our model. With Tableau, a dashboard will be designed for the extracted hypertension data.
  + Interpretation of Results: These tools will offer insights into hypertension and guide decision-making of hypertension management. Our results have the potential to improve patient care, contributing meaningfully to the related field.

**PROGRESS REPORT**

* Aim 1 - 100%
* Aim 2 - 70%
* Aim 3
  + Tableau - 50%



* + R Shiny - 0%