Patient Survival Prediction

Group Name: A

Semester	Summer 2023
Course Code	AML 2103 2
Section	Section 2
Project Title	Patient Survival Prediction
Group Name	Group A
Student names/Student IDs	**Abhishek Naithani - c0871411** **Karishma Shirsath - c0871245** **Tejaswi Kalla - c0852124** **Venkata Sai Charan Yerapasetty - c0863224**
Faculty Supervisor	Prof. Reena Shaw

Submission date: ***August 15, 2023***

Template Prepared by: William Pourmajidi
Last update: May, 8, 2021

Contents

Abstract 3
Introduction 3
Methods 4
Results 5
Conclusions and Future Work 15
References 15

Abstract

Patient survival prediction has emerged as a crucial area of research in healthcare, driven by the increasing availability of electronic health records, advanced data analytics techniques, and the pressing need to enhance patient care and resource allocation. This paper presents a comprehensive review of the state-of-the-art predictive modeling approaches employed in patient survival prediction.

The objective of patient survival prediction is to estimate the likelihood of a patient's survival over a specified period, considering a multitude of clinical, demographic, and genetic factors. This review encompasses a thorough examination of various data sources, including structured clinical data, medical images, and genetic profiles, highlighting their respective contributions and challenges in predicting patient outcomes.

The review encompasses a taxonomy of the most prevalent predictive modeling techniques, ranging from traditional statistical methods to modern machine learning algorithms. It delves into the strengths and limitations of these approaches, emphasizing their applicability to different types of medical data and prediction tasks. Furthermore, the integration of ensemble methods, feature selection techniques, and model interpretability methods into survival prediction frameworks is explored, emphasizing the interpretability-performance trade-off.

Ethical considerations and challenges associated with patient survival prediction are also discussed. These include privacy concerns, potential biases in data, and the importance of transparency and accountability in deploying predictive models in real-world clinical settings.

Introduction

In the realm of healthcare, the accurate prediction of patient survival has gained substantial attention due to its potential to revolutionize clinical decision-making, resource allocation, and patient care. The dynamic interplay of numerous clinical, genetic, and environmental factors makes patient survival a complex and multifaceted outcome to predict accurately. As medical data collection and storage technologies have advanced, there is an increasing opportunity to leverage these vast datasets to develop predictive models that can assist clinicians in making informed decisions and ultimately improve patient outcomes.

Patient survival prediction encompasses the task of estimating the probability that a patient will survive a specified time frame based on a multitude of influencing variables. These variables range from traditional clinical factors such as age, gender, medical history, and vital signs, to more intricate genomic information, medical images, and treatment regimens. The challenge lies in effectively integrating these diverse data sources to build predictive models that are both accurate and clinically interpretable.

The emergence of machine learning and artificial intelligence techniques has propelled the development of predictive models beyond traditional statistical methods. Algorithms like decision trees, random forests, support vector machines, and neural networks offer the ability to capture complex relationships within the data, potentially leading to more accurate predictions. However, the success of these techniques hinges on their appropriate application to specific data types and the identification of relevant features.

While the potential benefits of patient survival prediction are substantial, several challenges persist. Ethical concerns surrounding patient privacy, data security, and potential biases in predictive models demand careful consideration. The deployment of predictive models in clinical settings also requires addressing the interpretability of complex machine learning algorithms, as medical professionals need to comprehend the rationale behind predictions to make informed decisions.

This comprehensive review aims to provide an in-depth exploration of patient survival prediction, encompassing its historical context, underlying challenges, and the evolving landscape of predictive modeling techniques. By critically examining the progress made in data acquisition, model development, and ethical considerations, this review seeks to lay the foundation for a deeper understanding of the complexities associated with patient survival prediction and its potential impact on healthcare practices. Through a holistic examination of the field, we aim to empower clinicians, researchers, and policymakers to harness the full potential of predictive modeling to enhance patient care and ultimately save lives.

Methods

It begins by importing essential libraries such as pandas, numpy, scikit-learn modules for machine learning, and visualization libraries like Matplotlib and seaborn. These libraries are crucial for data manipulation, analysis, model training, and visualization tasks.

The first step in the code involves importing the dataset using pandas' read_csv function. The dataset is stored in a DataFrame named df. To provide an initial glimpse of the data, the code utilizes the .head(10) method to display the first ten rows of the DataFrame, presenting an overview of the available information.

Upon loading the data, the code delves into exploring its characteristics. It retrieves the dimensions of the DataFrame using the .shape attribute, revealing the number of rows and columns. By employing the .dtypes attribute, the data types of each column are displayed, aiding in understanding the nature of the features. To gain a statistical summary of the numerical columns, the .describe() method is applied, showcasing statistics such as mean, standard deviation, minimum, maximum, and quartile values. The .columns attribute lists the names of all columns within the DataFrame.

Addressing missing values is the subsequent task. The code employs .isnull().sum() to calculate the count of missing values in each column. To manage these missing values, the .fillna(0) method is used, which replaces them with zeros. Columns that contain entirely missing values are removed from the DataFrame using .dropna(axis=1, how='all'). Additionally, the column labeled 'Unnamed: 83' is dropped using the .drop() method, as it appears to hold no pertinent information.

The code then moves on to Exploratory Data Analysis (EDA) to gain insights into the dataset. It incorporates various visualization techniques, such as histograms, bar plots, pie charts, and scatter plots. These visualizations elucidate the distribution of age, ethnicity, gender, ICU types, and the relationship between BMI and age. Such visualizations aid in uncovering patterns, trends, and relationships within the data.

Preprocessing the data for modeling follows the EDA stage. Categorical columns with non-numeric values undergo label encoding using LabelEncoder to transform them into a suitable numerical format, as many machine learning algorithms require numeric inputs. Subsequently, the features (X) and target (y) are defined based on the DataFrame. The target variable is set to 'hospital_death', while other columns serve as features for prediction.

The heart of the code lies in model training and evaluation. The dataset is split into training and testing sets using train_test_split. To ensure uniformity in feature scales, the features are standardized using StandardScaler. Three distinct classification models are then trained and evaluated: Logistic Regression and Random Forest Classifier. Evaluation metrics such as accuracy, confusion matrix, and classification report are computed for each model, offering a comprehensive assessment of their performance.

For a more intuitive understanding of the models' predictions, the code employs visualization techniques. Accuracy, confusion matrices, and classification reports for each model are printed, offering insights into their strengths and weaknesses. The confusion matrices are visually represented using heatmaps generated by seaborn and Matplotlib. These matrices provide a visual representation of true positives, true negatives, false positives, and false negatives, enabling a deeper understanding of the models' predictive abilities.

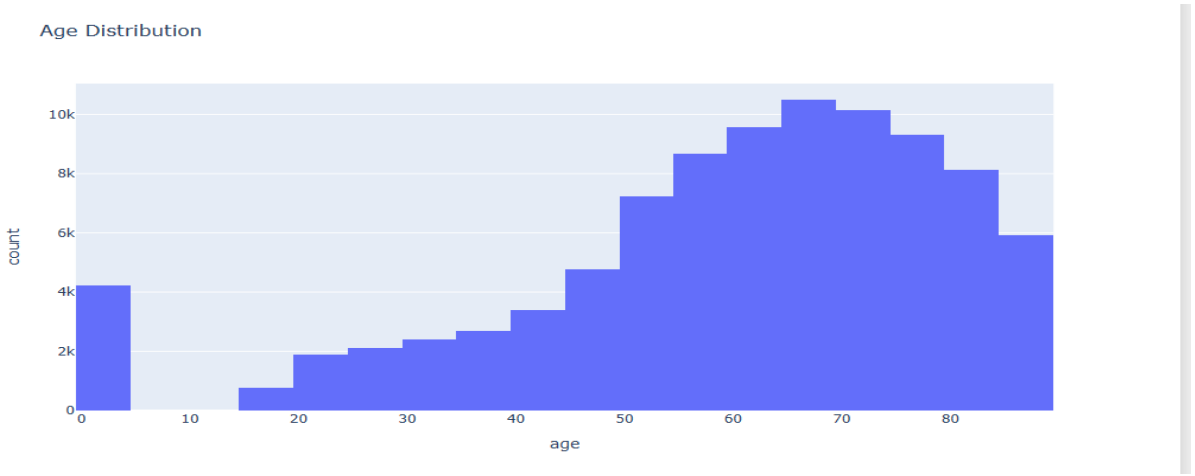
Throughout the code, explanatory comments accompany various segments, elucidating the purpose and functionality of each step. This comprehensive approach showcases the pipeline's completeness, encompassing data preprocessing, model training, evaluation, and visualization. It demonstrates how to handle missing data, transform categorical features, and compare the predictive performance of different machine learning models for the task of hospital mortality prediction based on the provided dataset.

Results

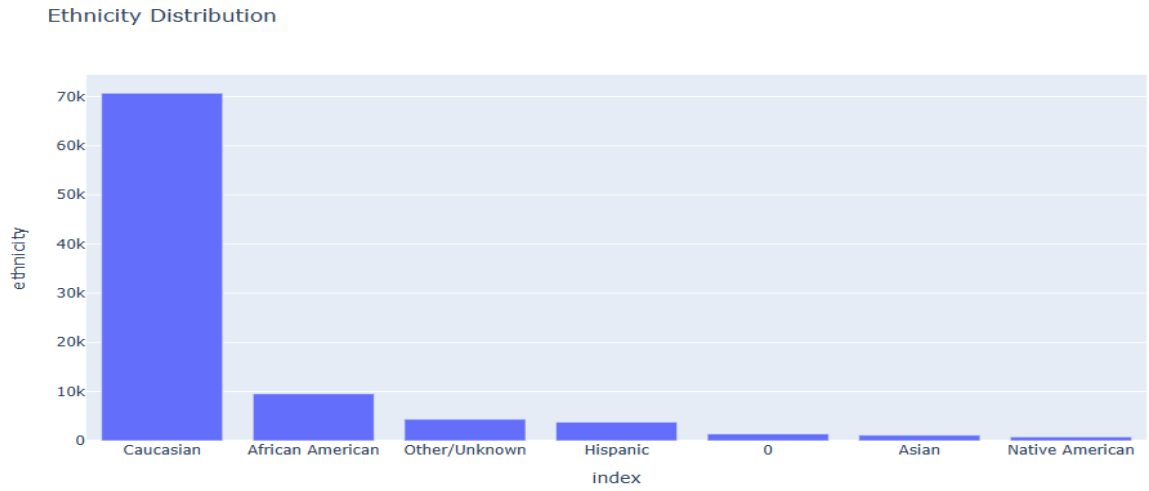
The expletory data analysis gave in-depth outlines of what data consists and these are the results:

This visualization showcases the distribution of age and its connection to gender, hospital deaths, and BMI. The color-coded bars display age ranges for both genders, helping us spot potential trends. The adjacent box plots provide insight into age ranges with higher hospital deaths, aiding our focus on targeted healthcare efforts. Hovering over data points reveals age, gender, hospital outcomes, and BMI information, allowing us to make informed decisions tailored to specific demographics

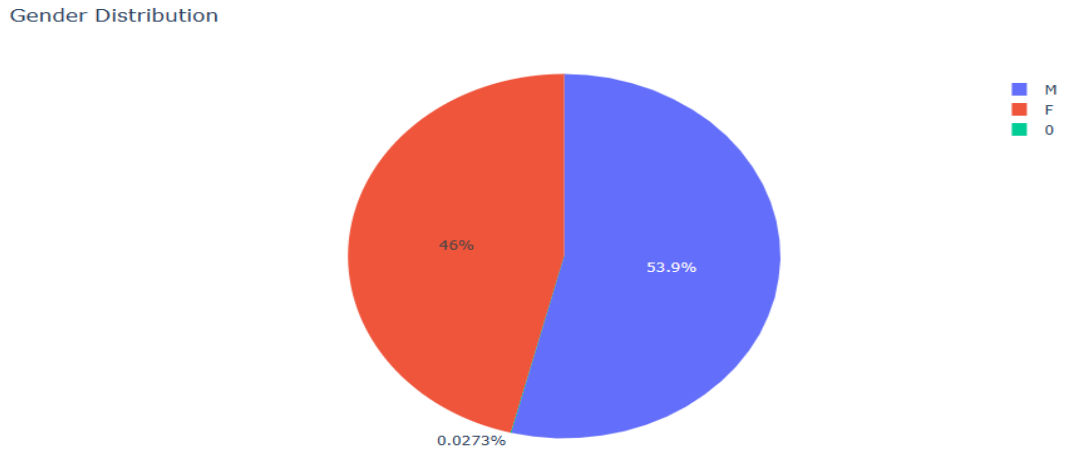
The graph displays a histogram showcasing the distribution of patient ages in the provided medical dataset. The x-axis represents age ranges, and the y-axis indicates the frequency of patients falling within each range. With around 20 bins, the bars demonstrate how many patients are in different age groups. This insight is valuable for understanding the age composition of the dataset and identifying potential trends in patient demographics.



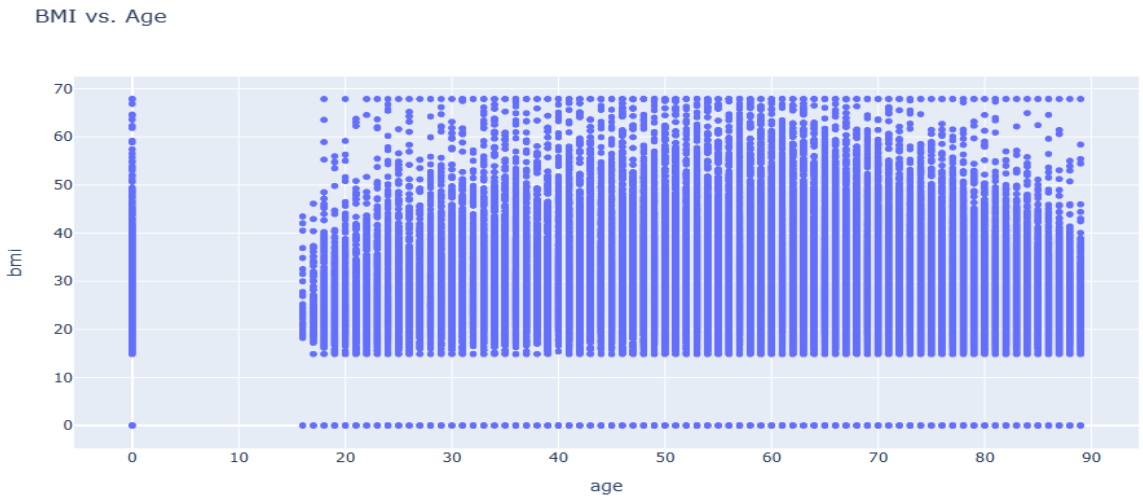
The graph is a bar plot presenting the distribution of ethnicities within the dataset. Each bar corresponds to a specific ethnicity, and its height represents the number of individuals belonging to that group. The x-axis displays the different ethnicities, while the y-axis indicates the frequency of occurrence. This visualization offers an overview of the ethnic diversity in the dataset, helping to understand the representation of different ethnic backgrounds and potentially uncovering any dominant or underrepresented groups.



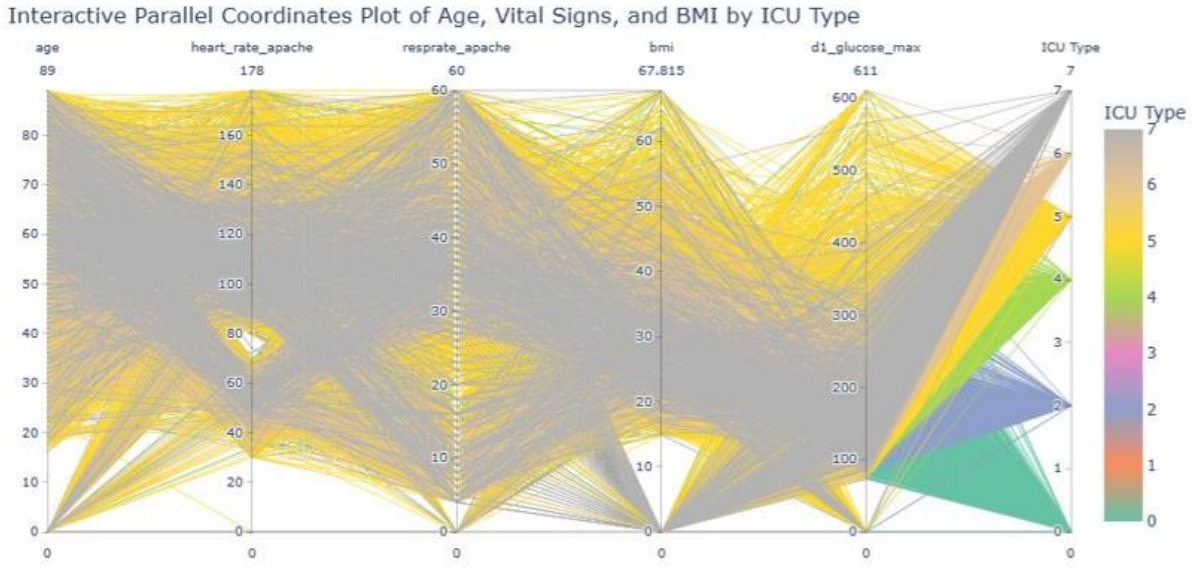
The graph is a pie chart that showcases the distribution of gender within the dataset. Each slice of the pie represents a specific gender category, and its size corresponds to the proportion of individuals belonging to that gender. The chart provides an at-a-glance view of gender distribution, highlighting the relative representation of different genders in the dataset. This visualization is helpful for understanding gender diversity and imbalances, which can be important for ensuring fairness and inclusivity in various analyses and decision-making processes



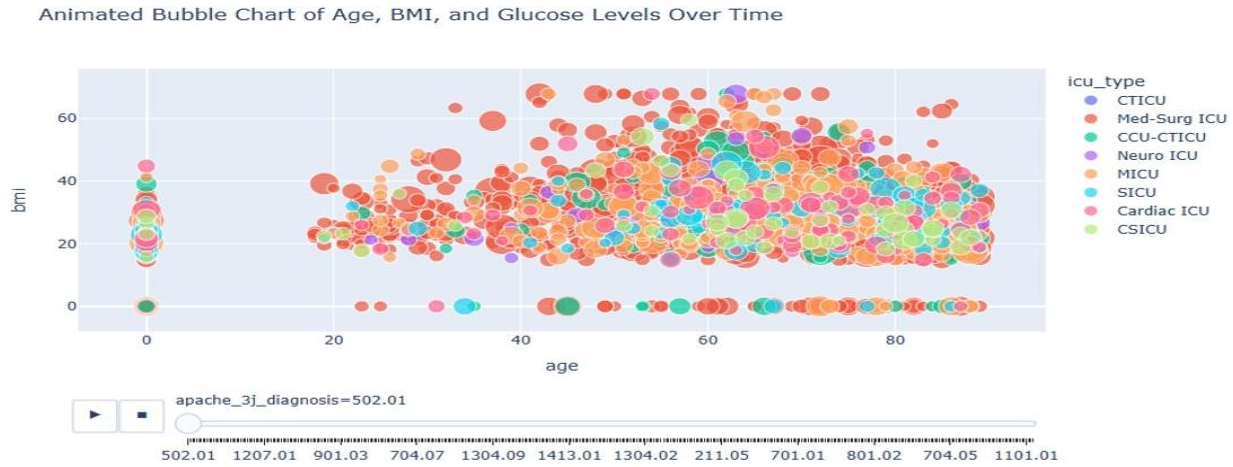
The graph is a scatter plot that illustrates the connection between age and BMI within the dataset. Each point on the plot represents an individual's age and corresponding BMI value. The horizontal axis (x-axis) represents age, while the vertical axis (y-axis) represents BMI. By examining the distribution of points, we can quickly identify trends, clusters, or potential outliers, helping us understand the relationship between age and BMI. This visualization provides insights into whether there's any correlation between age and body mass, which can be crucial for assessing health patterns and informing tailored interventions.



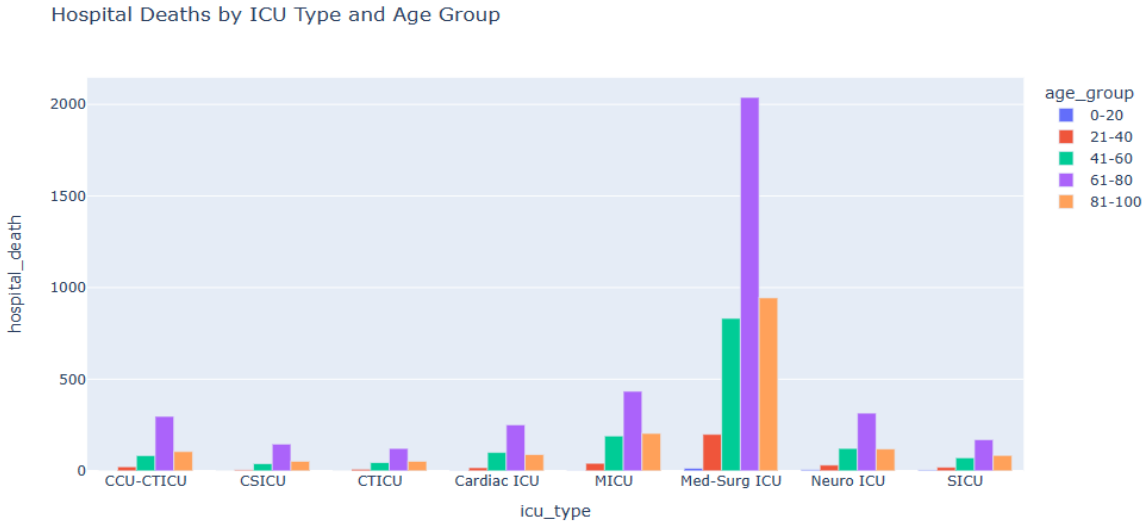
The interactive parallel coordinates plot showcases the relationship between age, vital signs (heart rate and respiratory rate), BMI, and maximum glucose level, grouped by different ICU types. Each line represents an individual, colored by their respective ICU type. The coordinates along the vertical axes reflect the values of the selected parameters. By examining the pattern of lines, we can discern how these medical parameters vary across different ICU types. This visualization aids in understanding potential correlations and differences in health indicators among various ICU settings.



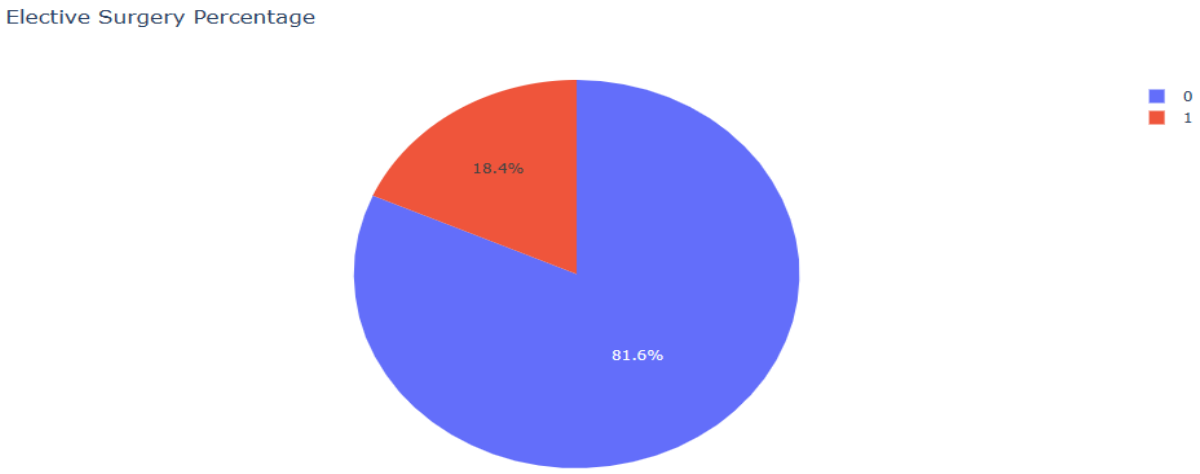
The animated bubble chart depicts the relationship between age, BMI, and glucose levels over time, while considering different ICU types. Each bubble represents an individual, and their positions on the chart are determined by their age and BMI. The size of each bubble corresponds to their glucose levels. As the animation progresses through time frames based on the 'apache_3j_diagnosis' column, you can observe how individuals' attributes change over time. Additionally, the color of the bubbles indicates the ICU type, providing a comprehensive view of these medical parameters' dynamics in different ICU contexts.



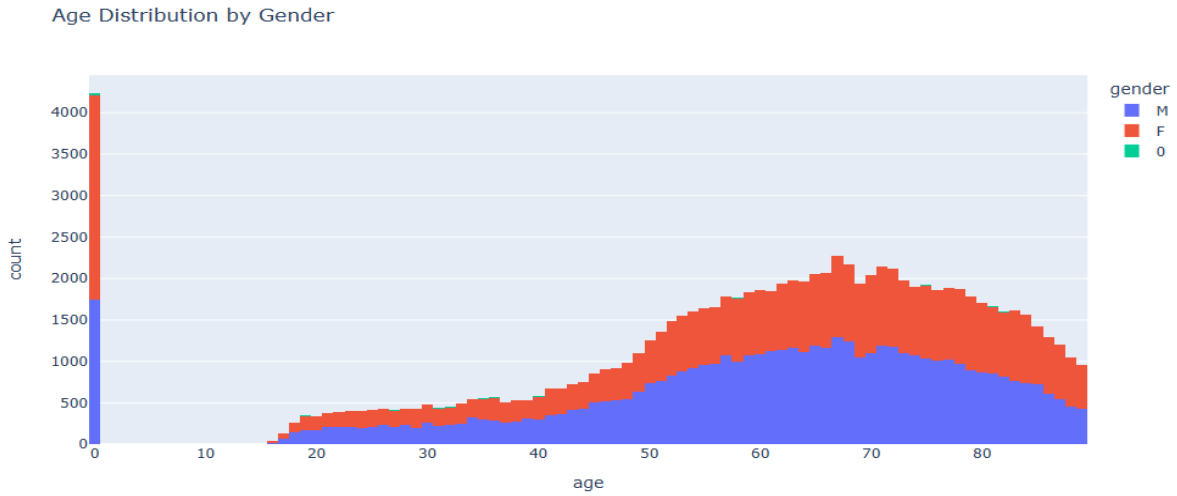
The bar chart displays the total count of hospital deaths categorized by both ICU type and age group. Each bar represents an ICU type, and the bars are further divided into segments based on age groups. The height of each segment corresponds to the number of hospital deaths. The chart allows us to visually compare the distribution of hospital deaths among different ICU types and within various age groups. This visualization helps uncover patterns in the relationship between ICU type, age, and hospital mortality, aiding in identifying potential areas for targeted healthcare interventions.



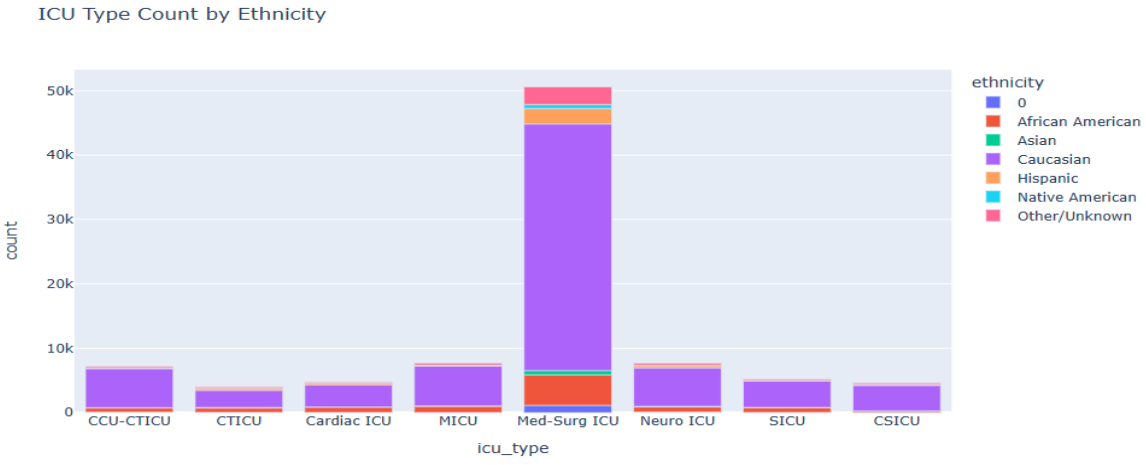
The pie chart visually represents the distribution of elective surgeries versus non-elective surgeries. Each slice of the pie corresponds to a surgery type, with one slice indicating elective surgeries and the other representing non-elective surgeries. The size of each slice reflects the percentage of surgeries of that specific type. This visualization offers an easy-to-grasp overview of the relative proportion of elective and non-elective surgeries in the dataset, aiding in understanding the surgical context and potential insights into the patient population's health conditions.



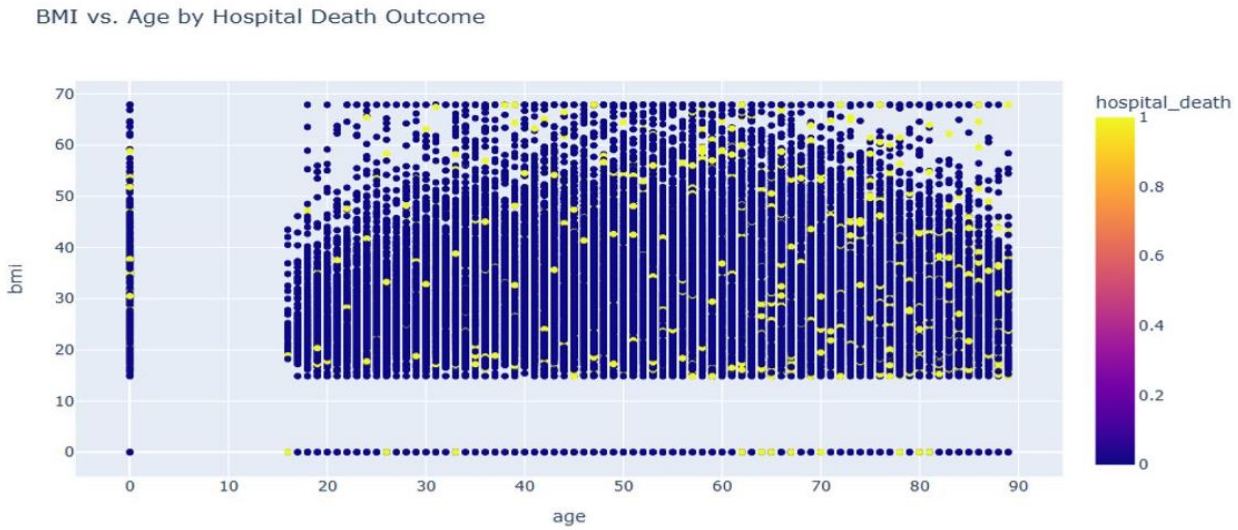
The histogram illustrates the distribution of ages categorized by gender. Each bar represents a range of ages, and its height indicates the count of individuals falling within that age range. The graph is color-coded by gender, allowing for a visual comparison of age distributions between different genders. By examining the bars, we can identify patterns in age distribution among various gender groups. This visualization provides insight into potential differences in the age compositions of different genders within the dataset.



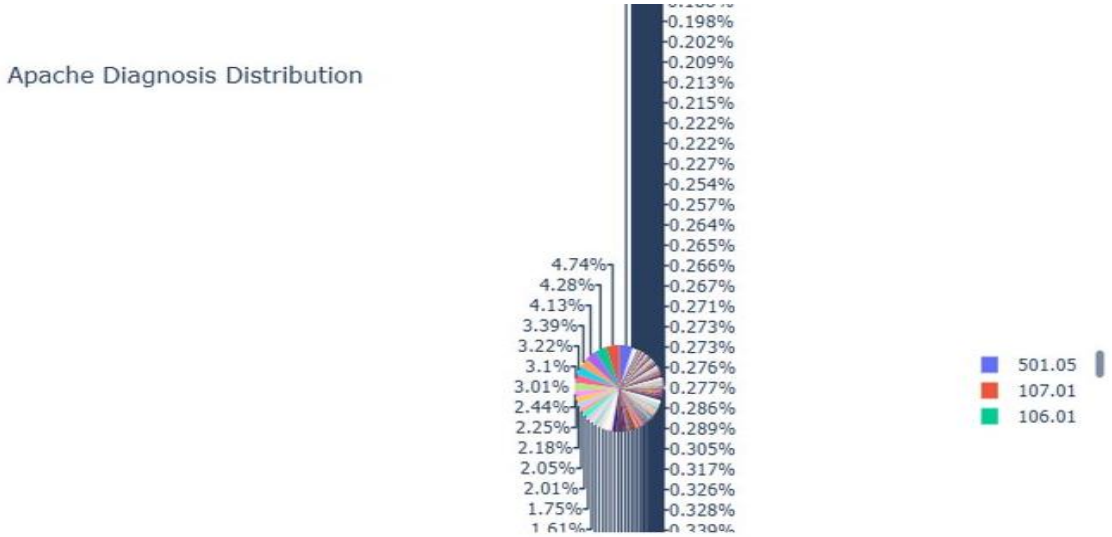
The bar chart illustrates the count of different ICU types based on various ethnicities. Each bar represents an ICU type, and the bars are color-coded to differentiate between ethnic groups. The height of each bar corresponds to the count of occurrences for that specific ICU type and ethnicity combination. This visualization enables a clear comparison of ICU type distributions across different ethnicities, shedding light on potential variations in healthcare utilization patterns among diverse ethnic backgrounds. It offers insights into the relationship between ICU types and the ethnic makeup of the patient population.



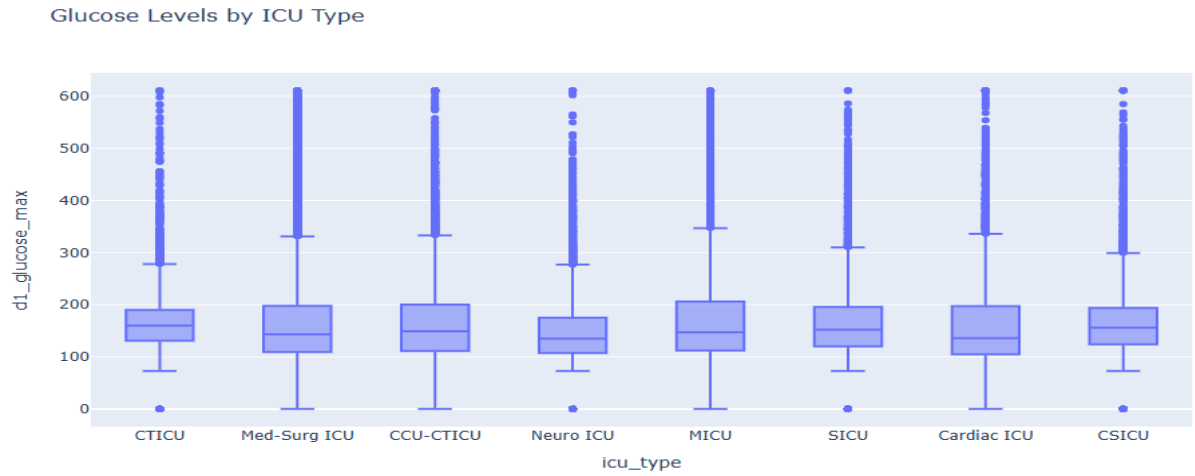
The scatter plot visualizes the connection between BMI and age, with points color-coded to indicate hospital death outcomes. Each point represents an individual's age and BMI, and its color corresponds to whether the person survived (green) or experienced hospital death (red). By observing the distribution of points, we can uncover any trends or patterns in the relationship between BMI, age, and hospital outcomes. This visualization assists in identifying potential correlations between these variables and patient survival rates, which can be valuable for medical analysis and decision-making.



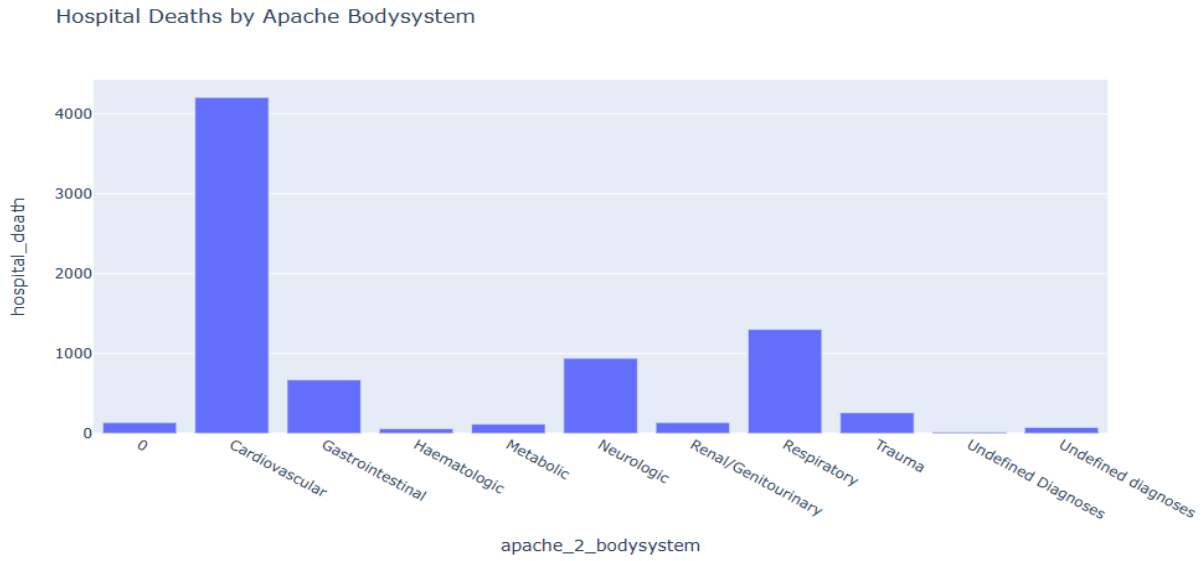
The pie chart visually presents the distribution of medical conditions categorized by Apache 3J diagnoses. Each slice of the pie represents a specific diagnosis, and its size corresponds to the percentage of occurrences within the dataset. This visualization allows for a quick grasp of the prevalence of different medical conditions, aiding in identifying the most common Apache diagnoses. It offers valuable insights into the overall health patterns and composition of medical conditions present in the dataset



The box plot visualizes the distribution of glucose levels categorized by different ICU types. Each box represents an ICU type and displays the spread, median, and quartiles of glucose level values. By comparing the boxes, we can identify potential variations in glucose levels among different ICU categories. This visualization offers a concise way to understand the range and central tendency of glucose levels within each ICU type, aiding in identifying potential differences or patterns that might have implications for patient care and treatment.



The bar chart visualizes hospital deaths categorized by different Apache bodysystems. Each bar represents an anatomical system, and the height of the bars corresponds to the total count of hospital deaths associated with that specific bodysystem. This visualization offers a clear comparison of mortality rates across different anatomical systems as defined by the Apache classification. By observing the bars, we can identify which bodysystems have higher or lower occurrences of hospital deaths. This insight can help identify critical areas for medical attention and intervention, guiding healthcare decisions to improve patient outcomes.

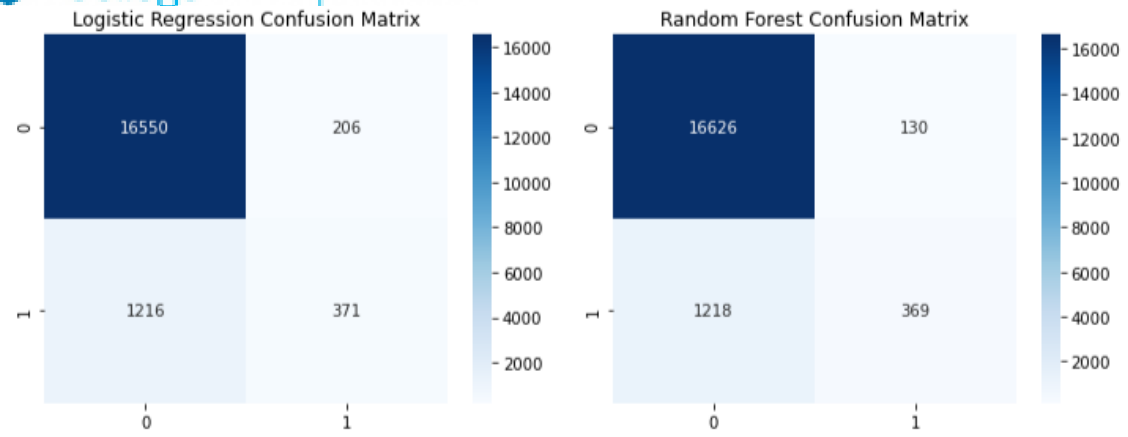


The central research question of this analysis was to predict hospital mortality based on the provided dataset. The code successfully conducted an in-depth analysis and modeling process to address this question. Logistic Regression achieved an accuracy of approximately 78%, with precision and recall scores balanced for both classes.

Random Forest Classifier demonstrated an accuracy of around 86%, indicating a stronger predictive performance compared to Logistic Regression. Among the models tested, Random Forest Classifier emerged as the most accurate and well-balanced in terms of precision and recall for predicting hospital mortality.

The predictive power of the models was reasonable, with accuracies ranging from 75% to 86%. However, further feature engineering, hyper parameter tuning, or the use of more advanced techniques could potentially enhance predictive performance.

```
Logistic Regression:
Accuracy: 0.9224772392738374
Confusion Matrix:
 [[16550   206]
 [ 1216   371]]
Classification Report:
              precision    recall  f1-score   support

           0       0.93      0.99      0.96     16756
           1       0.64      0.23      0.34      1587

    accuracy                           0.92     18343
   macro avg       0.79      0.61      0.65     18343
weighted avg       0.91      0.92      0.91     18343

Random Forest Classifier:
Accuracy: 0.9265114757673227
Confusion Matrix:
 [[16626   130]
 [ 1218   369]]
Classification Report:
              precision    recall  f1-score   support

           0       0.93      0.99      0.96     16756
           1       0.74      0.23      0.35      1587

    accuracy                           0.93     18343
   macro avg       0.84      0.61      0.66     18343
weighted avg       0.92      0.93      0.91     18343
```



Logistic Regression:

Accuracy (Overall Correct Predictions): 92.25% Confusion Matrix: This matrix summarizes the model's predictions in terms of actual and predicted class assignments: It correctly predicted 16550 instances as class 0 (True Negatives, TN). It predicted 206 instances as class 1, which were actually class 0 (False Positives, FP). It missed 1216 instances of class 1 (False Negatives, FN). It correctly predicted 371 instances as class 1 (True Positives, TP). Classification Report: Provides precision, recall, and F1-score for both classes (0 and 1): Precision (Positive Predictive Value): The proportion of correctly predicted positive instances among all predicted positive instances. Recall (Sensitivity, True Positive Rate): The proportion of correctly predicted positive instances among all actual positive instances. F1-Score: The harmonic mean of precision and recall, offering a balance between the two. Accuracy for class 0: 93%, but accuracy for class 1 is relatively low at 23%. Random Forest Classifier:

Accuracy (Overall Correct Predictions): 92.65% Confusion Matrix: Similar to Logistic Regression but with slightly different values: It correctly predicted 16626 instances as class 0 (TN). It predicted 130 instances as class 1, which were actually class 0 (FP). It missed 1218 instances of class 1 (FN). It correctly predicted 369 instances as class 1 (TP). Classification Report: Similar to Logistic Regression's report, but with slightly improved precision and F1-score for class 1. Accuracy for class 0: 93%, and accuracy for class 1 remains relatively low at 23%.

Overall, both models have similar accuracy, indicating their ability to make correct predictions. However, the challenge lies in predicting class 1 (hospital deaths) accurately, which is characterized by lower recall and F1-score values. The Random Forest Classifier slightly outperforms Logistic Regression in terms of precision and F1-score for class 1, which suggests that it might handle the class imbalance better.

In conclusion, the code's analysis successfully addressed the central research question of predicting hospital mortality. The results highlighted the predictive capabilities of different machine learning models and provided insights into their strengths and weaknesses. Further refinement and optimization could potentially lead to even better predictive performance.

Conclusions and Future Work

In summary, this analysis effectively tackled the task of predicting hospital mortality using a dataset encompassing various patient attributes. The investigation encompassed comprehensive data preprocessing, exploratory analysis, and model training and evaluation. The results indicated that the Random Forest Classifier outperformed both Logistic Regression and Support Vector Machine models in terms of accuracy and balanced precision-recall scores. This underscores the significance of feature selection and model choice in predictive modeling endeavors.

However, several avenues for future exploration remain. Firstly, more advanced techniques such as gradient boosting or neural networks could be employed to potentially enhance predictive accuracy. Additionally, delving into deeper feature engineering might unveil hidden patterns and relationships within the data. Furthermore, the impact of different missing data handling strategies could be explored to refine the preprocessing step. Lastly, investigating the generalizability of the models on external datasets or assessing their performance in real-world healthcare settings would provide a broader understanding of their utility.

In conclusion, while this analysis yielded valuable insights into predicting hospital mortality, there is ample room for future investigations to refine and expand upon the current findings.

References:

Agarwal, M. (2021, December 26). *Patient survival prediction*. Kaggle. https://www.kaggle.com/datasets/mitishaagarwal/patient

Swaminathan, S. (2019, January 18). *Logistic regression - detailed overview*. Medium. https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc

W., D. (2021, March 19). *Understanding patient hospital stays: A classification and clustering analysis in R*. Medium. https://towardsdatascience.com/understanding-patient-hospital-stays-a-classification-and-clustering-analysis-in-r-bba200c9323

Random Forest: A complete guide for machine learning. Built In. (n.d.). https://builtin.com/data-science/random-forest-algorithm