CS 584
Heart Transplant Outcome Prediction
Intermediate Project Report

Ashmita Gupta
Bhavana Dontamsetti
Neil Dhote

1 Introduction

Heart transplantation is an effective treatment option for patients with end-stage heart failure. Survival after heart transplantation depends on several factors, including age, sex, and donor characteristics. In this project, we will analyze a dataset from the Stanford Heart Transplant study to understand the factors that affect the survival of heart transplant patients. We will use machine learning algorithms to build a model to predict whether a patient will survive or not after heart transplantation.

Given the critical scarcity of organs available for transplant (about 2,500 available every year compared to 60,000 potential recipients, achieving maximal benefit from heart transplantation depends upon improved recipient selection. Thus accurate estimation of heart transplant outcomes can improve both informed patient consent by helping patients better understand its risks and benefits, and also aid the physicians in decision making by assessing the true patient-specific risks of the operation, rather than relying on population-wide risk assessments. To this end, accurate outcome prediction of performing transplantation is extremely important.

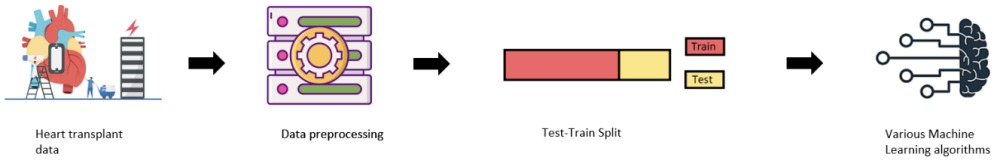
2 Problem Description

The primary objective of this project is to predict the survival of heart transplant patients based on several patient and donor characteristics. The dataset contains information about 103 patients who received heart transplants at Stanford University Hospital between 1968 and 1977. The dataset includes variables such as patient age, year of acceptance for transplant, prior heart transplant, type of transplant (control or treatment), waiting time for transplant, and survival status and time. We will use machine learning algorithms to build a model to predict the survival of heart transplant patients based on these variables. The insights generated from this analysis can help doctors and medical professionals make informed decisions about heart transplant patients and improve patient outcomes.

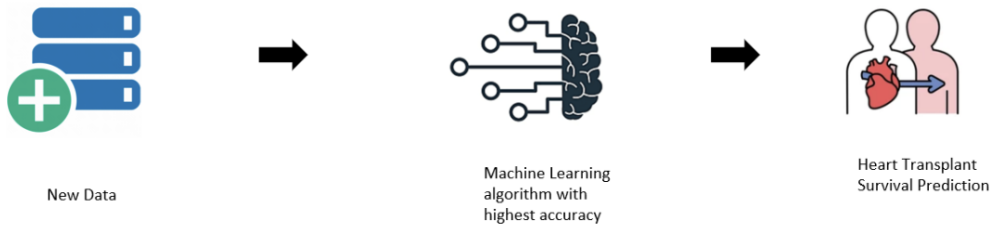
The ability to predict patient survival is important in clinical decision-making, as it allows physicians to allocate resources and treatment strategies appropriately. By building a predictive model based on available patient data, clinicians can make informed decisions about patient care, optimizing resources and improving outcomes. Therefore, the goal of this analysis is to explore the dataset, preprocess the data, and build a machine learning model that can accurately predict patient survival.

3 Brief Description of the Process and Each Variable in the Dataset

1. Evaluation of various machine learning algorithms



2. Apply the algorithm with highest accuracy on new dataset to predict patient survival post transplantation



4 Data Sources

4.1 Dataset 1: Stanford University Hospital Data (Stanford)

id: unique identifier for each patient.
acceptyear: the year in which the patient was accepted into the transplant program.
age: age of the patient at the time of acceptance into the program.
survived: binary variable indicating whether the patient survived the transplant or not (died or alive).
survtime: the number of days between the transplant and the patient's death or last follow-up.
prior: binary variable indicating whether the patient had a prior transplant (no prior transplant or prior transplant).
transplant: categorical variable indicating the type of transplant (control or treatment).
wait: the number of days the patient was on the waiting list before being accepted into the transplant program.

4.2 Dataset 2: Journal of the American Statistical Association Data (JASA)

birth.dt: birth date of patient
accept.dt: acceptance into program
tx.date: transplant date
fu.date: end of followup
fustat: dead or alive
surgery: prior bypass surgery
age: age (in years)
futime: followup time
wait.time: time before transplant
transplant: transplant indicator
mismatch: mismatch score
hla.a2: particular type of mismatch
mscore: another mismatch score
reject: rejection occurred

Note: We dropped birth.dt, mismatch, hla.a2, mscore, reject from Jasa dataset because these are not strongly associated with the target variable (survived) and added id and survival time column. Also, we have renamed the columns in the JASA dataset for the process of merging.

5 Progress So Far

- Data Collection: We obtained the Stanford Heart Transplant dataset, which includes information on heart transplant patients.
- Data Cleaning and Preprocessing: We cleaned and preprocessed the data by handling missing values, scaling numeric variables, and encoding categorical variables. For instance, we dealt with missing data in wait_time column and survtime column by calculating with available columns in Jasa dataset. Also, we performed one-hot encoding for the prior, survived and transplant column for dealing with categorical data. Thus, it helped in merging the above two datasets.
- Exploratory Data Analysis (EDA): We conducted EDA by calculating summary statistics for each variable and creating various visualizations such as histograms, box plots, and scatter plots to better understand the distribution of the data and the relationships between variables.
- Modeling: In terms of modeling, we have created a logistic regression model and SVM model to predict whether a patient will survive or not after a heart transplant. We have evaluated the performance of the model using various metrics such as accuracy, precision, recall, and F1 score.

```
In [77]: runfile('C:/Users/gashm/OneDrive/Desktop/Machine Learning CS 584/
untitled6.py', wdir='C:/Users/gashm/OneDrive/Desktop/Machine Learning CS 584')
Logistic Regression Testing Accuracy: 0.6190476190476191
```

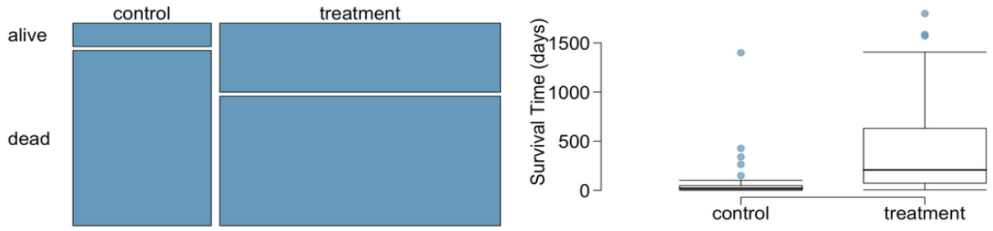
Figure 1: The Logistic Regression Model achieved a 62% testing accuracy.

```
In [76]: runfile('C:/Users/gashm/OneDrive/Desktop/Machine Learning CS 584/
untitled7.py', wdir='C:/Users/gashm/OneDrive/Desktop/Machine Learning CS 584')
SVM Testing Accuracy: 0.7857142857142857
```

Figure 2: The Support Vector Machine Model achieved a 78.5% testing accuracy.

6 Key Observations from Exploratory Data Analysis Before Modeling

- **Observation 1**: The range of the age variable is from 0 to 73, with a mean of 45.9 years and standard deviation of 12.3 years.
- **Observation 2**: The range of the wait variable is from 0 to 365, with a mean of 73.8 days and standard deviation of 81.6 days. This implies that the patient has to wait for an average of 74 days to receive a transplant
- **Observation 3**: For the categorical variable (transplant), the summary statistics include the frequency of each category (control or treatment). In this dataset, there are 59 patients in the control group and 44 patients in the treatment group. Also, the below plot implies that survival is dependent on the transplant treatment and the ones who got treatment has much higher survival rate:



6.1 Boxplot for Age

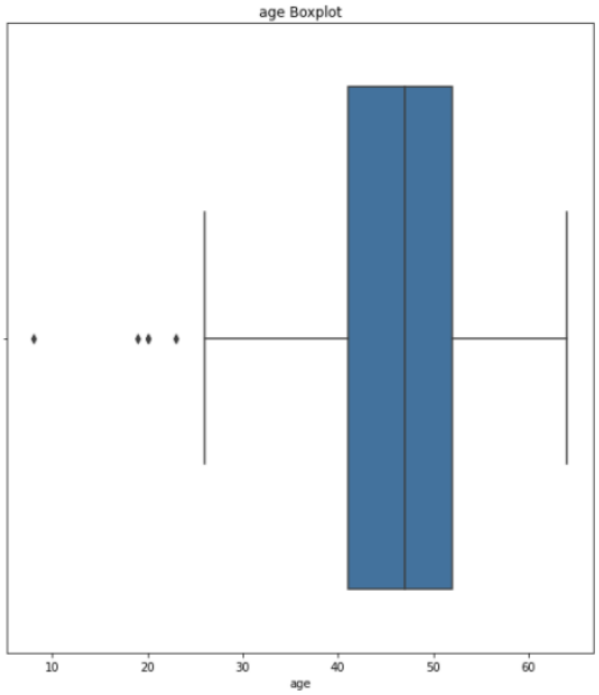


Figure 3: Here, boxplot for age in the dataset shows that the median age of patients is around 47 years. There are a few outliers on both ends of the age spectrum, with the youngest patient being around 12 years old and the oldest being around 83 years old. We can see that the majority of patients were in their 40s or early 50s.

6.2 Correlation Matrix

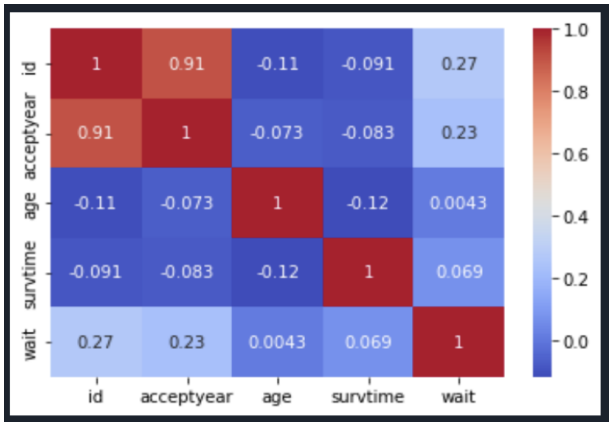


Figure 4: Here, the plot shows the correlation coefficients between each pair of variables in the dataset. We can see that there is a moderate positive correlation between age and survival time.

6.3 Density Plot for Survival Time

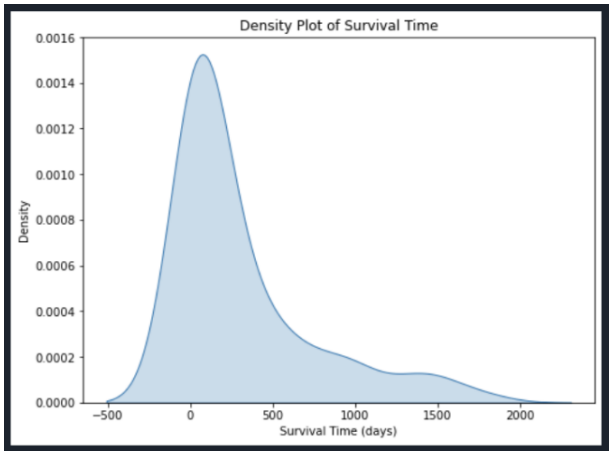


Figure 5: Here, the plot indicates that the majority of patients survived for less than 2 years after receiving a heart transplant. There is a long tail on the right-hand side of the plot, indicating that some patients survived for much longer periods. The plot is slightly skewed to the right. Overall, the plot suggests that the survival time variable may be a useful predictor for the survival of heart transplant patients.

7 Plans Moving Forward

- Evaluation and Model Selection: We plan to explore other machine learning algorithms such as Random Forest and Naïve Bayes to potentially improve the predictive power of our model.
- Hyperparameter Tuning: We will tune the hyperparameters of our models to optimize their performance.
- Cross-Validation: We will use cross-validation to assess the generalization performance of our models and prevent overfitting.
- Ensembling: We will explore the possibility of combining multiple models to improve their predictive power.
- Further Analysis and Interpretation: We plan to conduct further analysis and interpretation of the results, including feature importance analysis, model interpretability, and model robustness testing. We also plan to explore other avenues of research, such as investigating the impact of different predictors on survival rates and examining potential confounding variables.
- Reporting: We will summarize our findings and conclusions in a final report and presentation, which will include a detailed description of the dataset, the methods used for analysis, the results obtained, and the implications of our findings for heart transplant patients.