Introduction / Problem Statement

Heart failure is a serious medical condition that affects millions of people worldwide, and early detection of patients at risk of dying from heart failure can help healthcare providers intervene before the situation becomes critical. The creation of a model to predict mortality caused by heart failure can help healthcare providers to identify high-risk patients and develop personalized interventions to improve patient outcomes. The use of machine learning algorithms, such as decision tree, neural network, and logistic regression in this case, can improve the accuracy and reliability of such models and facilitate the development of more effective interventions. Overall, the benefits of creating a model to predict mortality caused by heart failure are potentially significant, both for individual patients and for the healthcare system as a whole.

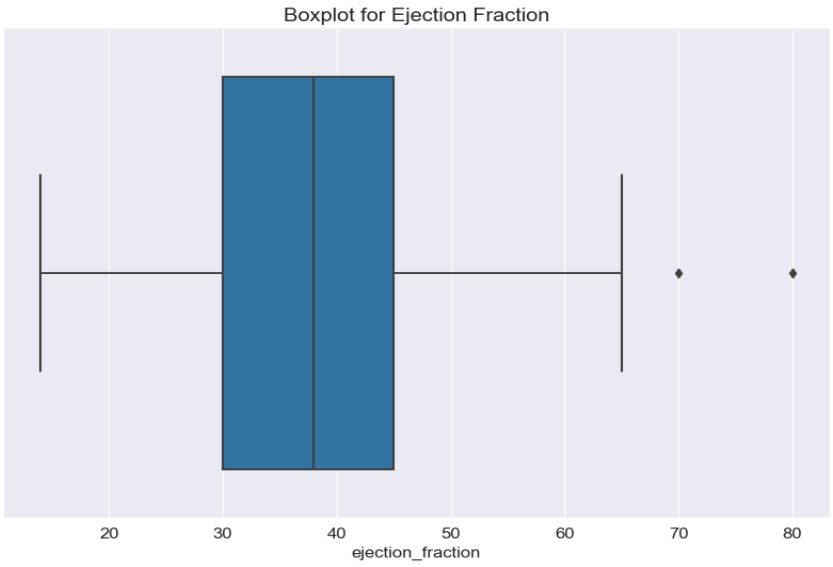
Here are some of the features in the dataset:

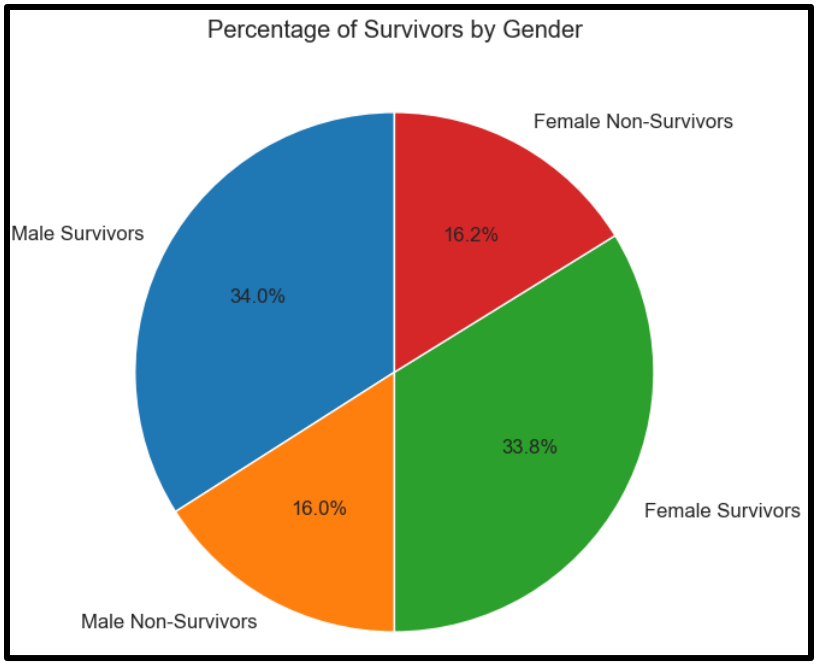
age: the age of the patient in years.
anaemia: a boolean feature that indicates if the patient has a decrease in red blood cells or hemoglobin.
high blood pressure: a boolean feature that indicates if the patient has hypertension.
creatinine phosphokinase (CPK): the level of the CPK enzyme in the blood in mcg/L.
diabetes: a boolean feature that indicates if the patient has diabetes.
ejection fraction: the percentage of blood leaving the heart at each contraction.
platelets: the number of platelets in the blood in kiloplatelets/mL.
sex: a binary feature that indicates if the patient is a woman or a man.
serum creatinine: the level of serum creatinine in the blood in mg/dL.
serum sodium: the level of serum sodium in the blood in mEq/L.
smoking: a boolean feature that indicates if the patient smokes or not.
time: the follow-up period in days.
[target] death event: a boolean feature that indicates if the patient deceased during the follow-up period.

Methods / Implementation

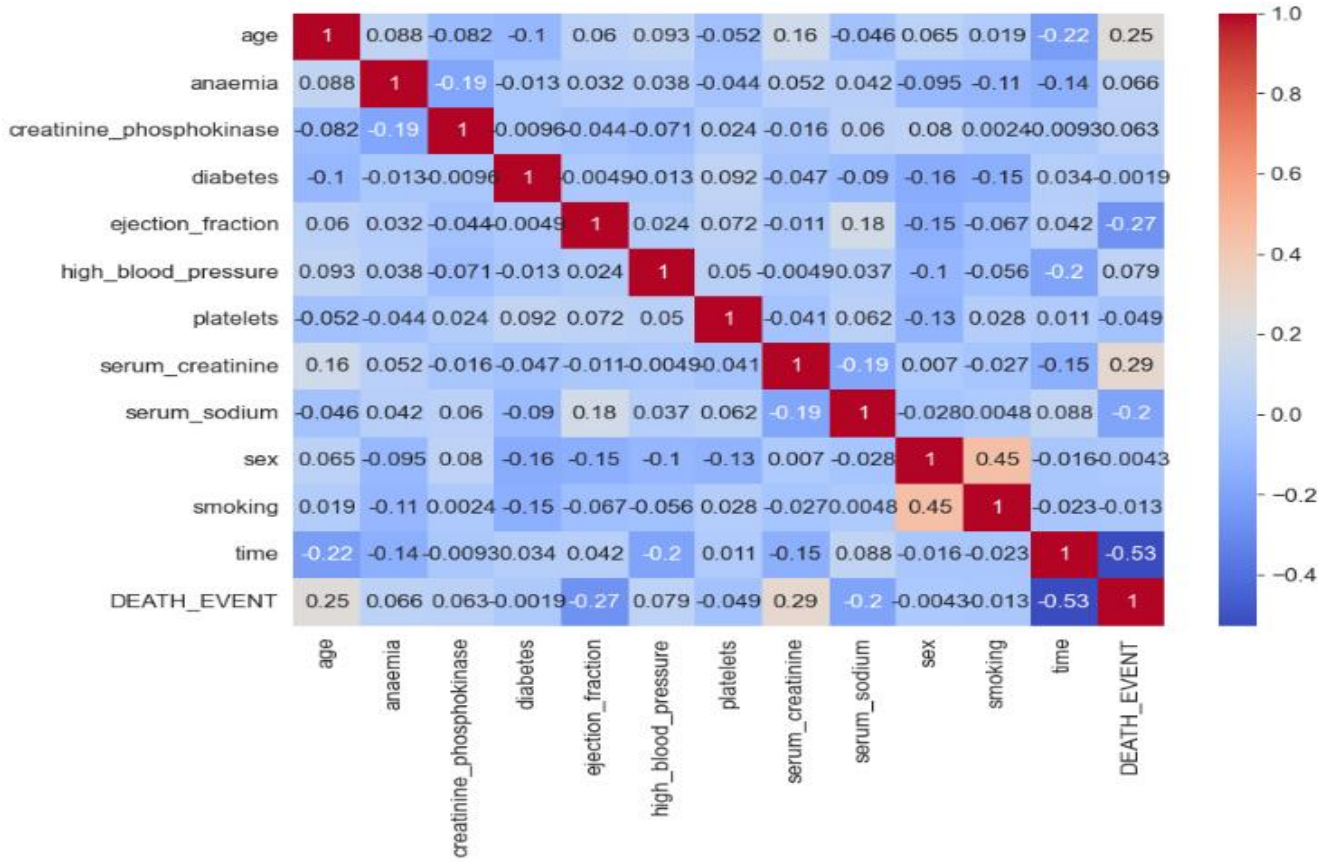
We began by analyzing the relationships between certain features to see if they were correlated with each other.

The box in the plot represents the interquartile range (IQR) from 30 to around 45, which is the middle 50% of ejection fraction. The median ejection fraction of around 38 is represented by the straight line in the box. The minimum ejection fraction was around 13, while the maximum ejection fraction was around 65. We can also see that the boxplot shows a fairly symmetric distribution of ejection_fraction.





The plot shows that the percentage of female survivors is slightly higher than the percentage of male survivors. However, the difference is not very significant. Overall, the majority of patients in the dataset survived.



Looking at the heatmap, we can see that there is some multicollinearity in our dataset. For example,

time is strongly negatively correlated with death event. Ejection fraction is negatively correlated with death event, while serum creatinine is positively correlated with death event. Additionally, there is a positive correlation between age and death event.

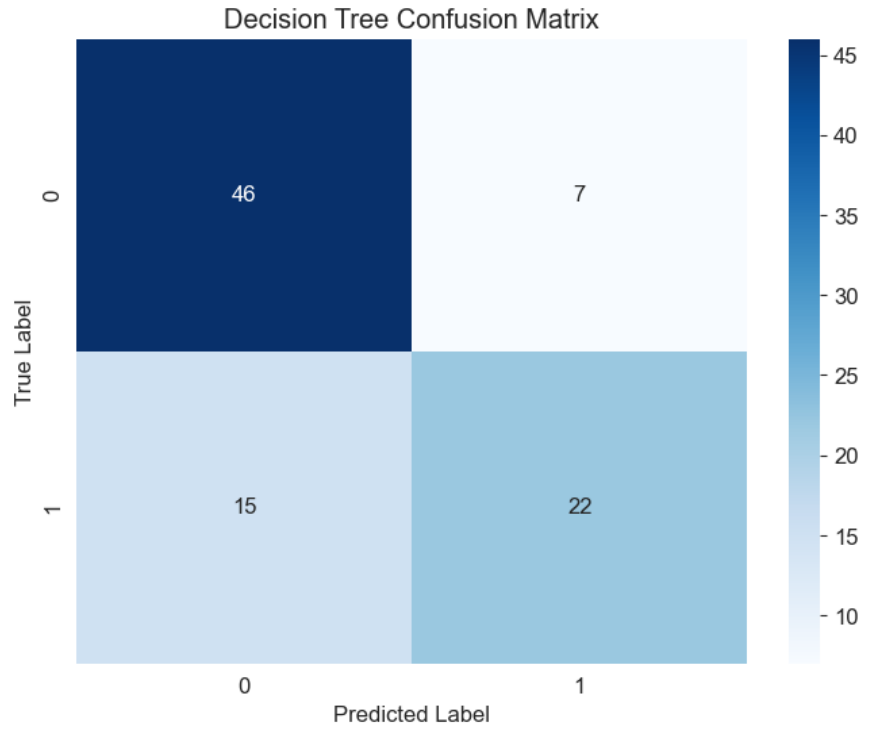
Results

The decision tree model showed a better performance with an accuracy of 0.76. The logistic regression model showed a worse performance with an accuracy score of 0.72. However, both models showed comparable performance. To achieve better performance, we could use cross-validation with different cv values to optimize the logistic regression model and experiment with different tree depth values.

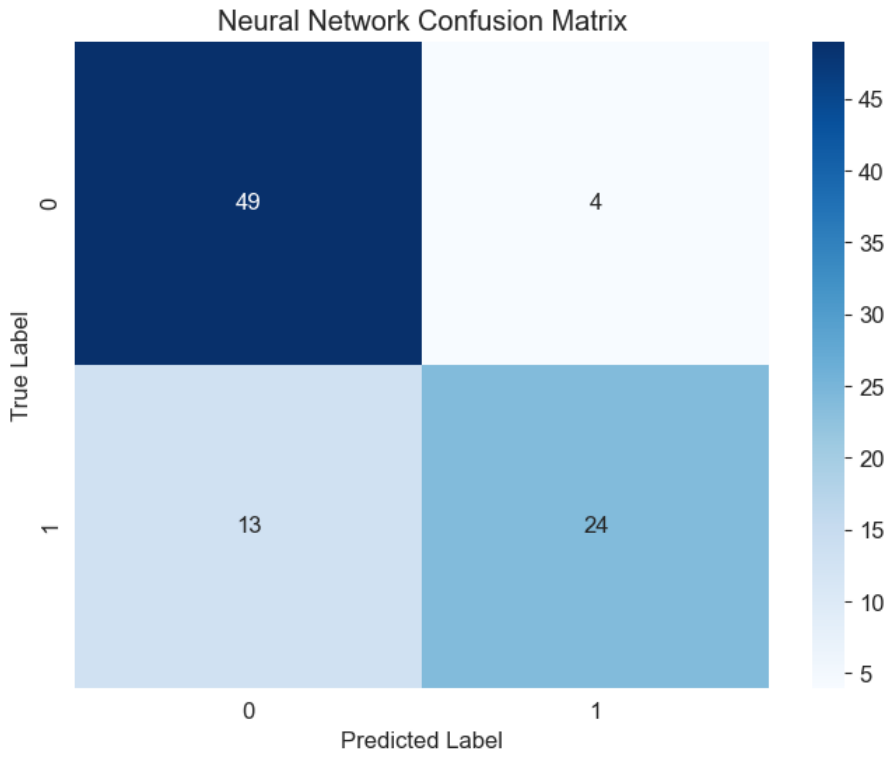
Discussion / Explanation

All three models can moderately predict whether a patient with heart failure will survive or not based on certain features. This can be useful for healthcare professionals in making decisions about the treatment and care of their patients. When selecting a model, it is important to consider the nature of the problem, the size of the dataset, the type of data, and the desired outcome. In this case, we have a binary classification problem with a relatively small dataset of 299 observations and 13 features. We aim to predict whether a patient will experience death due to heart failure. Based on these considerations, we chose to use logistic regression, neural networks, and decision trees. Logistic regression and neural networks are commonly used for binary classification problems, and they are suitable for datasets of this size. Additionally, they are models that can handle both categorical and continuous data, which is important when dealing with complex datasets. Neural networks are more flexible than logistic regression and decision trees in terms of model complexity, as they can capture both linear and nonlinear relationships between variables. Neural networks are also capable of learning complex patterns in the data, making them suitable for more complex problems. While decision trees can handle both categorical and continuous data, they tend to overfit the training data, resulting in poor performance on unseen data. On the other hand, logistic regression can be more robust and generalize better. Neural networks have the potential to generalize well, but they can also suffer from overfitting if the model is too complex or the dataset is too small. In terms of interpretability, decision trees create a clear decision path with a series of if-else statements, whereas logistic regression outputs a probability value. Neural networks, on the other hand, are less interpretable due to their complexity.

Here are the confusion matrices for each model:



The decision tree makes more sense to use (over a logistic regression model) to predict if a patient will die because it has less false negative predictions (type 2 error). Clearly, we want to minimize type 2 errors because that error represents the model predicting that the patient will survive even though the patient dies.



The neural network outperformed both the decision tree and the logistic regression model. It had the truest predictions and the least false predictions while also having a higher accuracy. Thus, it makes the most sense to use to predict death from heart failure.

The logistic regression model did end up predicting more cases where the patient survives when compared to the decision tree and neural network. However, it is more important to minimize type 2 errors in this particular study to help more patients survive.