Data Description

Heart failure is a progressive condition that impacts the ability of the muscles of the heart to effectively pump blood throughout the body. This condition is induced from injury to the muscles due to events such as having a heart attack or suffering from chronic hypertension. As a result, the heart can either enter a weakened state or become stiff impacting cardiovascular functionality. This condition is progressively becoming more prevalent globally impacting millions currently and contributes significantly to cardiovascular disease death. Currently, 6.7 million Americans suffer from heart failure and it is projected to increase to 8.5 million Americans by 2030[1]. The dataset we obtained from Kaggle, consists of medical information from 299 patients suffering from heart failure collected from 2015. Our goal is to develop a classification model that will assist medical professionals in accurately predicting a heart failure patient's survival based on given medical record features. In addition, our analysis and assessment of significant features through machine learning algorithms will allow medical professionals to understand what type of features have a significant impact on heart failure patients.

Table 1 Dataset Information

#	Feature	Description	Data Type	Domain
1	Age	Patient's age	Integer	[40, 95]
2	Anemia	Presence of anemia	Boolean	0, 1
3	Creatine Phosphokinase	Level of enzyme in blood	Integer	[23, 7861]
4	Diabetes	Presence of diabetes	Boolean	0, 1
5	Ejection Fraction	Percentage of blood ejected per heartbeat	Integer	[14, 80]
6	High Blood Pressure	Presence of high blood preassure	Boolean	0, 1
7	Platelets	Platelet count in blood	Float	[25100, 850000]
8	Serum Sodium	Level of sodium in blood	Integer	[113, 148]
9	Serum Creatinine	Level creatinine in blood	Float	[0.5. 9.4]
10	Sex	Patient's gender	Boolean	0, 1
11	Smoker	Smoking status	Boolean	0, 1
12	Days Until Follow-Up	Follow-up period	Integer	[4. 285]
13	Patient Deceased	Patient died	Boolean	0, 1

To improve the accuracy of our classification models, we employ data preprocessing, exploratory data analysis, and feature selection techniques. Our dataset before preprocessing, shown in Table 1, consists of 13 feature columns and 299 rows of data. From the dataset, 12 columns are features representing medical information of the patient and 1 column which represents the target feature of whether a heart failure

patient survives until the follow up period. Grammatical errors in the original dataset were corrected and columns were renamed for clarity. Additionally, the target feature “DEATH_EVENT” was renamed to “patient_deceased” to be more clear and descriptive. Out of the 12 feature columns, the columns "anemia", "diabetes", "high_blood_pressure", "sex", "smoker" are categorical feature columns. These categorical columns are represented as a binary classification of whether a patient has this specific column condition, with 0 representing no and 1 representing yes. The column "sex" is represented by 0 being female and 1 being male. The remaining 7 columns, "age", "creatine_phosphokinase", "ejection_fraction", "platelets", "serum_creatinine", "serum_sodium", "days_until_follow_up" are numerical features. Creatine phosphokinase (CPK) is an enzyme that plays a role in energy metabolism, by generating ATP. Damage or death of heart muscle cells results in elevated CPK levels in the bloodstream due to leakage from the affected cells. Serum sodium and creatinine indicate levels of sodium and creatinine in the bloodstream. Lastly, ejection fraction is a measurement to assess how well the heart is pumping blood with each contraction.

Preprocessing of the dataset was performed before analysis of the dataset was started. Dataset validation included determining the number and percentage of missing data points among all the features. Additionally, we checked for the presence of duplicate rows of data. After conducting this validation facilitated by pandas, we observed that there were no instances of missing data points in any of the features, nor were there any duplicate rows of data.

Exploratory Data Analysis

Exploratory data analysis (EDA) was utilized to understand the underlying distribution of data among different features and the relationships among one another through the use of data visualization techniques. One key observation from EDA is the significant imbalance in the target class “patient_deceased”. Dividing the dataset into 2 subsets based on the target class label revealed that out of the total patients in the dataset, 203 patients survived while 96 patients did not survive. This imbalance could potentially impact the performance of any predictive models trained on this data, as they might be biased towards predicting the majority class.

Histograms and kernel density estimations (KDE) were generated among each numerical feature to assess whether the data under a specific feature follows a gaussian or skewed distribution. In addition, histograms were generated based on the subsets of data grouped by the target class label “patient_deceased”. Understanding this will allow us to determine if there are presence of outliers that will significantly impact our model accuracies. From the generated charts, we observed that almost all of the numerical feature columns are skewed distributions. The features age, creatine phosphokinase, ejection fraction, platelets, and serum creatinine exhibit a right-skewed distribution, indicating that most of their values lie to the left of the mean. Serum sodium follows a left skewed distribution meaning most of its values are to the right of the mean and days until follow up was observed to be a bimodal with 2 distinct peaks.

The analysis of categorical features was conducted through the use of bar charts. These charts facilitated the observation of any imbalances in the distribution of each categorical feature when grouped by the target feature, "patient_deceased". Out of the charts generated, we can observe there are only imbalances of heart failure patients who do not have high blood pressure, who are non-smokers and who are male patients in contrast to female patients. The remaining categorical features in the dataset do not exhibit a significant imbalance across the different classes.

Lastly, a heatmap was utilized to graphically represent the correlation matrix. Each cell of the heatmap represents the correlation coefficients between a pair of feature columns, providing insights about the strength and direction of their linear relationship. We can observe that there is a very weak association between each feature for a given patient. This suggests that the features largely vary independently of each other, and there is no strong linear relationship where a unit change in one feature would predict a similar change in another. Dimensionality reduction can not be performed utilizing a correlation matrix heatmap due to the weak linear relationship between pairs of features.

The key insights from our exploratory data analysis indicate that heart failure patients who do not survive generally exhibit lower levels of ejection fraction, higher levels of serum creatinine, lower levels of serum sodium, are older, and have elevated levels of creatine phosphokinase. Furthermore, these features display a broader range and a larger standard deviation from the mean, suggesting greater variability. Other

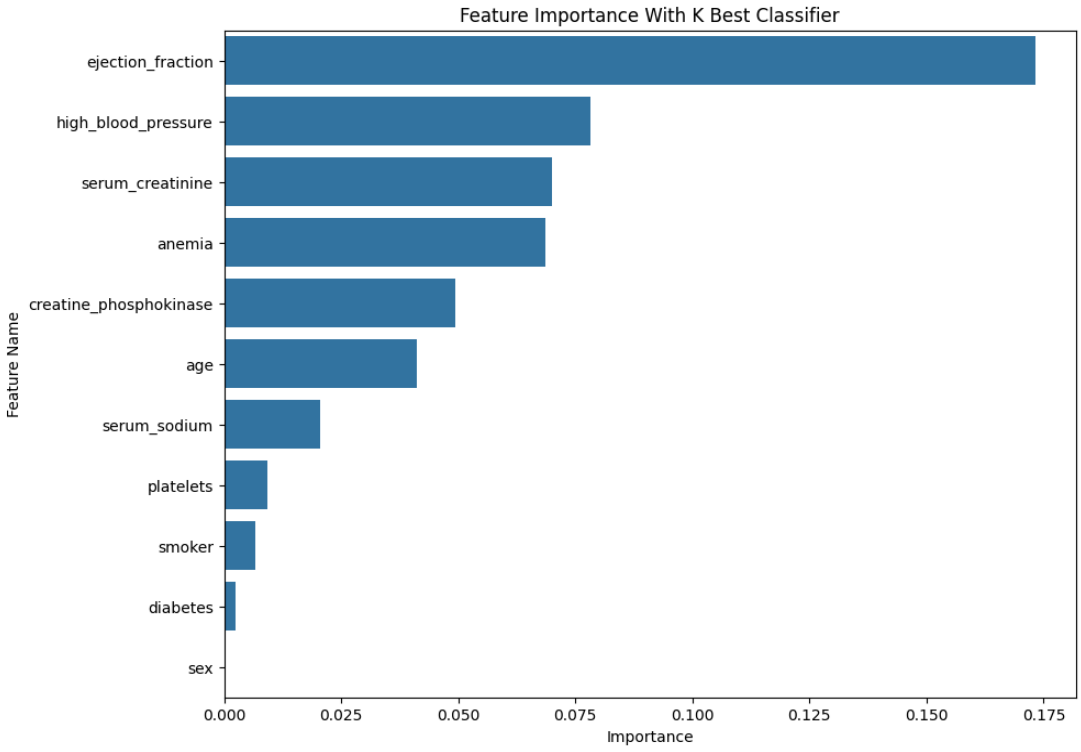
features appear to be less significant, as they demonstrate similar distributions and standard deviations across the board.

Outlier detection was performed through the use of statistical measures and box plots. Descriptive statistics were generated using pandas which summarizes the central tendency, dispersion and shape of each feature. Additionally, kurtosis was utilized which is a statistical measure that calculates the tailedness of a distribution in relation to its shape. The features “creatine_phosphokinase”, “platelets”, “serum_creatinine”, and “serum_sodium” were observed to have significantly high kurtosis values of 25.149, 6.209, 25.828, and 4.119 respectively. These features have a kurtosis value significantly greater than 3, indicating that the distribution has heavier tails when compared to a normal distribution. This suggests that outliers are more prevalent in these feature columns. Box plots were also generated in order to have a graphical representation of the presence of outliers in each feature column. As expected, there we could observe a significant number of outliers of the feature columns with high kurtosis values.

Through these statistical measures and data visualization techniques, we are able to conclude the presence of outliers. However, the main difficulty is determining whether the outliers present should be removed or retained. These outliers might hold significance in training accurate and generalized classification models. To overcome this, domain knowledge was required and we sought out research articles to observe typical ranges medical professionals tend to use for these medical feature columns. These research papers as referenced below allowed us to come to the conclusion that these outliers could be removed in order to prevent overfitting, misclassification and skewed class distributions when training classification models. The outlier detection technique used is determining the lower and upper bound using the formula Q1 - 1.5 * IQR and Q3 + 1.5 * IQR respectively. This technique was utilized because for skewed distributions, the mean and standard deviation are heavily influenced by outliers. Filtering was performed by iterating through all the numerical feature columns and removing values out of this range to create a new dataset. After filtering 223 rows of data remain from the original 299. Box plots generated from the new dataset show significant improvement in terms of skewness and presence of outliers from all the feature columns. We will take into account how the original dataset and dataset with no outliers fare in terms of accuracy and performance when training our models.

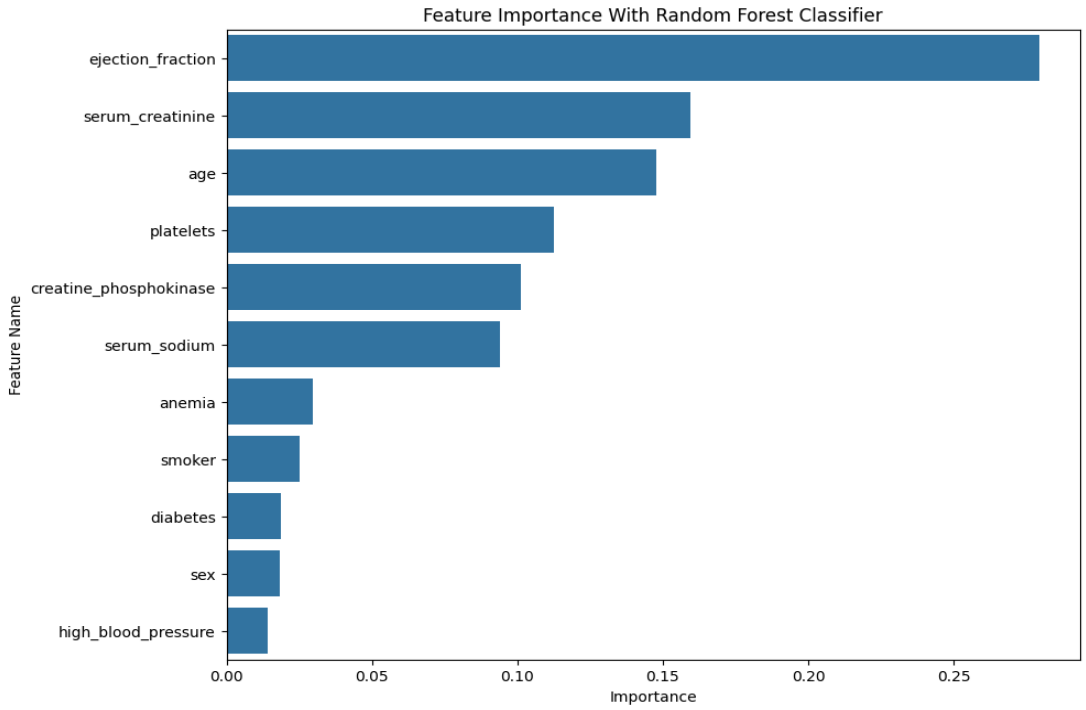
Feature Engineering and Selection

Feature selection was performed in order to determine if there were irrelevant features present that could be removed. This is performed in order to decrease the dimensionality of the dataset and in return improve performance, accuracy, avoid overfitting and reduce the computational time needed to train our models. Initially, the feature “days_until_follow_up” was removed because it represents the follow up period and has no reflection on the survival of a heart failure patient. We employed multiple feature engineering techniques to assist in validating the significance of various features. The first method used is SelectKBest. It is a filter-based feature selection method commonly used in machine learning that uses statistical measures such as the chi-squared test, the ANOVA F-test, and other measures to score and rank features based on their relationship with the output variable. In our case, mutual information classification was utilized which measures the dependency between a feature and its target predicted variable, “patient_deceased”. Higher values of mutual information scores indicate a strong dependency between two features and conversely scores close to 0 indicate that the features are independent of each other.



After ranking the features, SelectKBest selects the K = 7 best features with the highest scores to be included in the list of final features. Those features were "ejection_fraction", “high_blood_pressure”, “serum_creatinine”, “anemia”, “creatine_phosphokinase”, and “age” as the most important features.

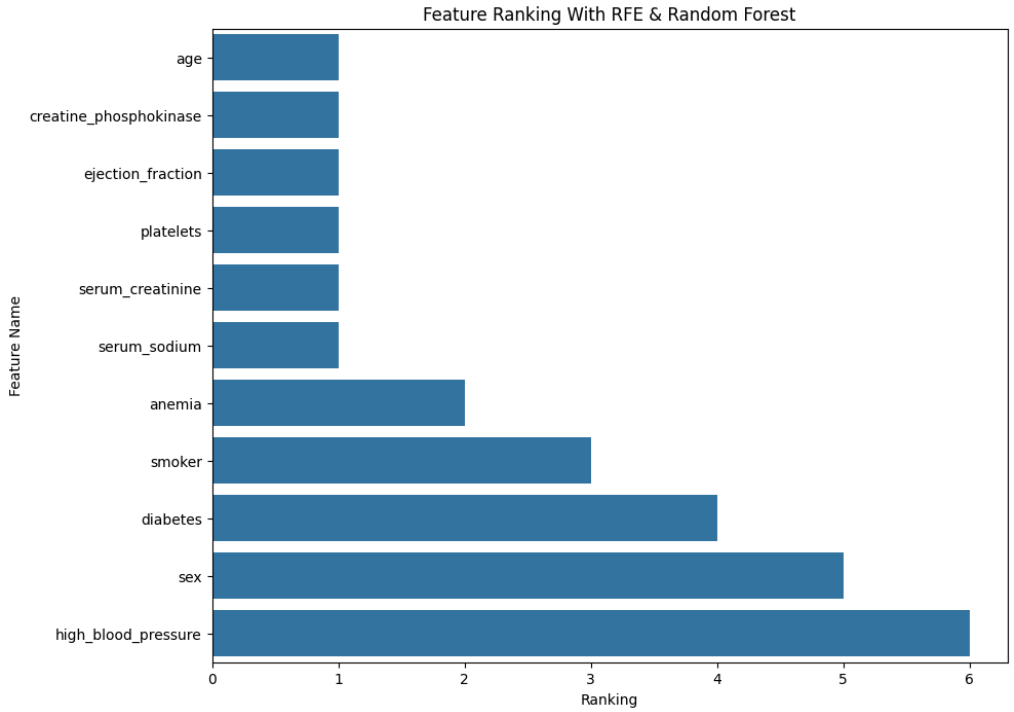
Besides using SelectKBest for feature selection, we also used other methods in choosing the most important features. Those methods are Recursive Feature Elimination (RFE) and Random Forest Classifier. Random forest classifier fits a number of decision tree classifiers on sub samples of the dataset through bootstrapping. This combines many decision trees into a single model. Random forest performs averaging to improve accuracy and prevent overfitting resulting in more stable predictions. Intrinsically, random forest ranks the importance of features by computing scores based on which features decrease overall impurity. Random forest was selected due to its robustness to overfitting and ability to manage non linear relationships.



Using the random forest classifier machine learning model we were able to observe that the features "ejection_fraction”, “serum_creatinine”, “age”, “platelets”, “creatine_phosphokinase”, and “serum_sodium” are the most important features.

Recursive Feature Elimination is a wrapper method for feature selection. It is utilized by training a machine learning algorithm, in our case, Random Forest Classifier,

to obtain the importance of our features, removes the features with the least importance, and then recursively re-trains the model with the remaining features until the desired number of features remains.



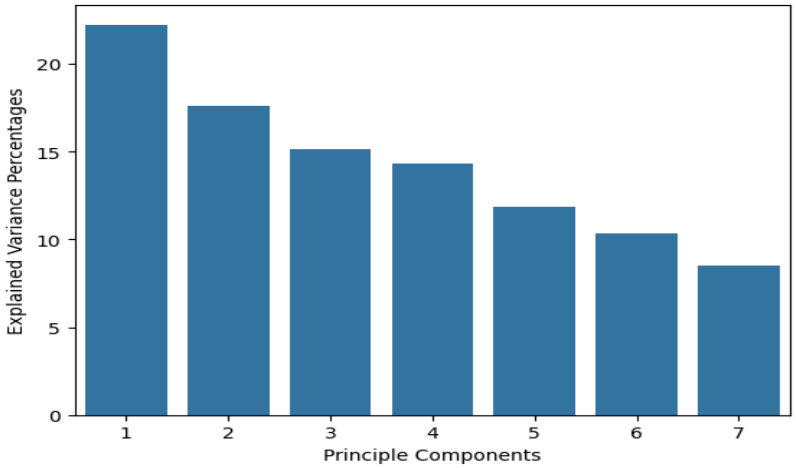
RFE ranked “age”, “creatine_phosphokinase”, “ejection_fraction”, “platelets”, “serum_creatinine”, and “serum_sodium” as all having the same ranking of 1 signifying as the most important.

With the use of 3 feature selection techniques we have come to the conclusion that the features “smoker”, “diabetes”, “sex”, and “high_blood_pressure” are irrelevant in producing classification models to predict the survival of heart failure patients. As a result we have reduced the dimensionality of the dataset from 12 medical features to 7 medical features and 1 target class feature to train our models.

Table 2 Summary of Feature Selection

#	Feature	SelectKBest	RFE	Random Forest Classifier
1	Age	0.051509	1	0.147632
2	Anemia	0	2	0.029527
3	Creatine Phosphokinase	0	1	0.101106
5	Ejection Fraction	0.157745	1	0.279329
7	Platelets	0.01288	1	0.112421
8	Serum Sodium	0.016935	1	0.093862
9	Serum Creatinine	0.087941	1	0.15968

Principal Component Analysis (PCA) is a feature extraction technique that is used to reduce the dimensionality of data while preserving as much information as possible. The result of PCA is a set of principal components which are uncorrelated and capture most of the variance in the data.



After standardizing our final dataset features to have a mean of 0 and variance of 1, we can observe how much variance is captured by each principal component. The cumulative sum of all the principal components is equal to 1, signifying it captures all the variance of our dataset. This dataset created through PCA can be used to train our classification model. We will compare the accuracies of the models with respect to the PCA dataset and final dataset created after feature selection.

References

1. Bozkurt, Biykem, et al. “Heart Failure Epidemiology and Outcomes Statistics: A Report of the Heart Failure Society of America.” *Journal of Cardiac Failure*, vol. 29, no. 10, 2023, p. 40. *Heart Failure Epidemiology and Outcomes Statistics: A Report of the Heart Failure Society of America*, https://onlinejcf.com/article/S1071-9164(23)00264-6/fulltext#:~:text=Approximately%206.7%20million%20Americans%20over,develop%20HF%20in%20the%20 lifetime.
2. Jurman, Giuseppe. “Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone - BMC Medical Informatics and Decision Making.” *BMC Medical Informatics and Decision Making*, 3 February 2020, https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-020-1023-5. Accessed 23 February 2024.
3. Srivastava, Amit. “What Is Principal Component Analysis (PCA) & How It Works?” *Analytics Vidhya*, https://www.analyticsvidhya.com/blog/2016/03/pca-practical-guide-principal-component-analysis-python/. Accessed 23 February 2024.
4. Wannamethee, S. G., Shaper, A. G., & Perry, I. J. (1997). Serum creatinine concentration and risk of cardiovascular disease. *Stroke*, *28*(3), 557–563. https://doi.org/10.1161/01.str.28.3.557
5. Gabr, R.E., El-Sharkawy, AM.M., Schär, M. et al. Cardiac work is related to creatine kinase energy supply in human heart failure: a cardiovascular magnetic resonance spectroscopy study. J Cardiovasc Magn Reson 20, 81 (2018). https://doi.org/10.1186/s12968-018-0491-6