Adil Khan

Data Science, Bellevue University

DSC 680: Applied Data Science

Dr. Catie Williams

Sep 26, 2022

Topic

Heart Failure Predictions Based on Patients' Health Attributes

Research Question/Abstract:

According to estimates, cardiovascular diseases account for 17.9 million annual deaths worldwide, or 31% of all fatalities. Heart failure arises when the heart is unable to pump enough blood to meet the body's needs. Machine learning, particularly when used with medical data, may be a helpful tool for both forecasting the prognosis of each patient displaying heart failure symptoms as well as for identifying the most important clinical features (or risk factors) that may culminate in heart failure. (Latha, 2019). 2019 (Lawler).

Machine learning may help scientists with feature evaluation as well as clinical prediction. A new tool for clinicians to utilise in assessing whether a patient with heart failure will survive or not may be created by using machine learning and data science to clinical practise in the healthcare sector. In example, while attempting to evaluate whether a patient would survive after experiencing heart failure, clinicians frequently place an emphasis on serum creatinine and ejection fraction (Chicco, 2020). In order to build a prediction model that will enable us to forecast a patient's chance of developing heart failure based on the existing health data, the objective of this research is to examine and grasp the data.

Dataset:

The dataset is publicly available on Kaggle website and can be access through the link provided, https://www.kaggle.com/andrewmvd/heart-failure-clinical-data. There are 13 characteristics in the dataset that provide light on clinical, physical, and lifestyle data related to patients. The medical records of 299 patients with heart failure were gathered from April to December 2015 at the Allied Hospital and Faisalabad Institute of Cardiology in Punjab, Pakistan (Ahmad, 2017). Between the ages of 40 and 95, the sample's 194 men and 105 women represented a wide age range. The dataset includes categorical factors including anaemia, hypertension, diabetes, sex, smoking, and mortality events. In data that has been boolean-typed, these category characteristics are represented. If a patient's hematocrit level was below 36%, they were deemed anaemic (Chicco, 2020). Age, creatinine phosphokinase (CPK), ejection fraction, platelets, serum creatinine, serum sodium, and time make up the remaining characteristics. Continuous variables are used to represent these characteristics. Creatinine phosphokinase (CPK) is released into the circulation when muscle tissue is injured. Consequently, heart failure may be a sign of elevated CPK levels in the blood (Chicco, 2020). High levels of serum creatinine may be caused by renal failure; serum creatinine is a byproduct of creatinine produced during muscle catalysis (Stephens, 2019). In the dataset, the death event variable identifies whether a patient passed away or lived before the conclusion of the follow-up period, which on average lasted 130 days. (Ahmad, 2017).

Ethical Considerations

The leading cause of mortality worldwide, cardiovascular diseases (CVDs), claim 17.9 million lives annually, or 31% of all fatalities worldwide. The majority of cardiovascular illnesses may be avoided by employing population-wide measures to target behavioural risk factors such cigarette use, poor eating and obesity, inactivity, and problematic alcohol consumption.

Early detection and management of people with cardiovascular disease or who are at high cardiovascular risk (due to the presence of one or more risk factors like hypertension, diabetes, hyperlipidemia, or already established disease) are essential, and a machine learning model can be very helpful in this regard.

Methods

The strategy was divided into three stages in order to create a predictive model. The activities that must be completed before moving on to the next step are included in each phase.

- **Phase 1** – exploratory data analysis. The initial stage in every data science analysis task is this phase. Since there are 13 variables total in the dataset, several of them may or may not be correlated, particularly with the feature in question, the death event variable. We must thus picture and comprehend how they are distributed. We must also make careful to look for outliers and missing numbers.

• **Phase 2** – this is the feature selection phase. Once we are aware of how each variable relates to and correlates with our main variable, death event. Then, to base the prediction model on, we may choose the attributes that have the most effect and connection with the primary variable we have chosen. To determine which machine learning model is best for our prediction model, we will test a variety of them. Decision trees, logistic regression, support vector machines, k-nearest neighbours, random forests, and other machine learning models will all be employed in this model.

• **Phase 3** – The characteristics will be applied to the construction of the prediction models in this step after being chosen. The dataset's existing data will be used to execute and train the model.

Results

Phase 1 - Exploratory Data Analysis (EDA)

Python and Jupyter Notebook were used for all of the coding and analysis. Any data science analysis started with preprocessing the dataset to check for any null or missing values. Since there were no missing data, we can now look for correlations or other links between the various attributes. The first analysis was performed to determine the age distribution of each sample in the dataset.

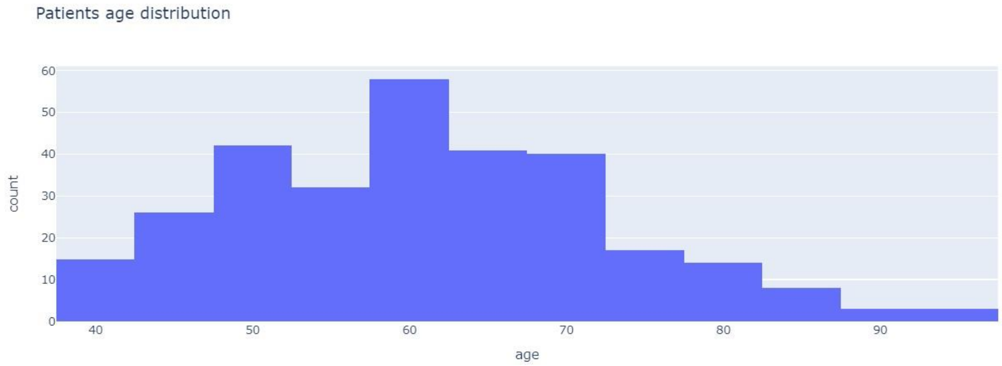


Figure 1: Patients' age distribution

According to Figure 1, the majority of patients are in the 50–70 age range, with the bulk of patients being about 60 years old and a minority being at least 90 years old.

In order to observe the link between age and gender in our sample set, we added another variable to the age distribution. The association between age and gender in the sample is depicted as a box-and-whisker plot in Figure 2.

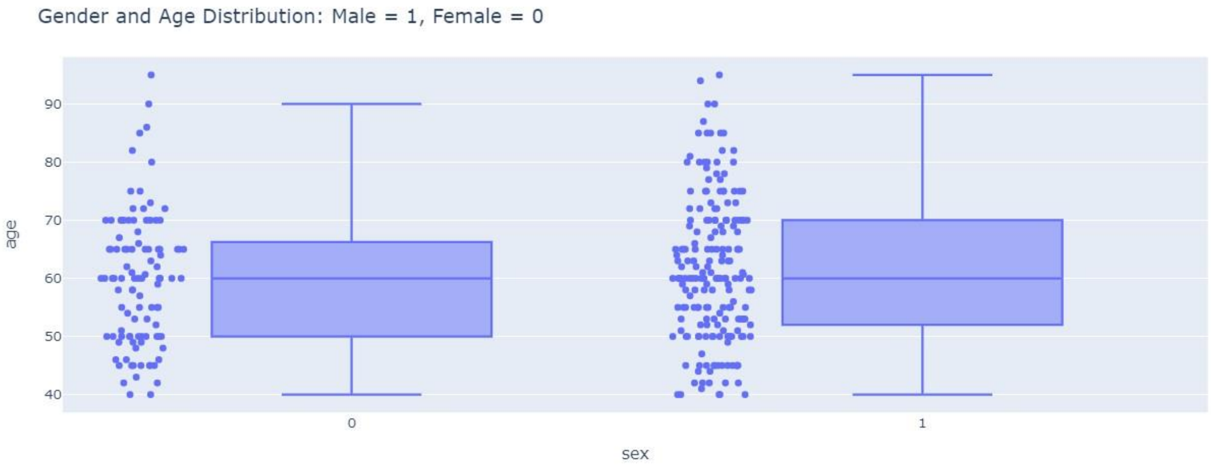


Figure 2: Gender and Age distribution with male being 1 and female being 0

According to Figure 2's gender and age distribution, male patients have a far wider range of ages than female patients. Male age distribution ranges consistently from 40 to 90, while female age distribution is mostly centred in the 40 to 70 age range.

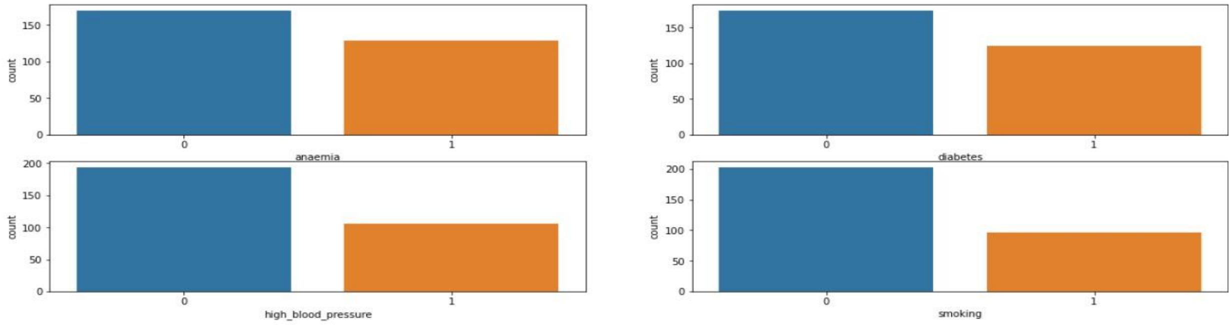


Figure 3: Distributions of all categorical variables in the dataset

Figure 3 shows that around 43% of the sample has anaemia, 41% has diabetes, 35% has high blood pressure, 32% smokes, and some individuals may have several illnesses. As we dug further into the information to examine associations between various variables, we created a histogram to compare the death event vs the quantity of follow-up visits, as seen in Figure 4.

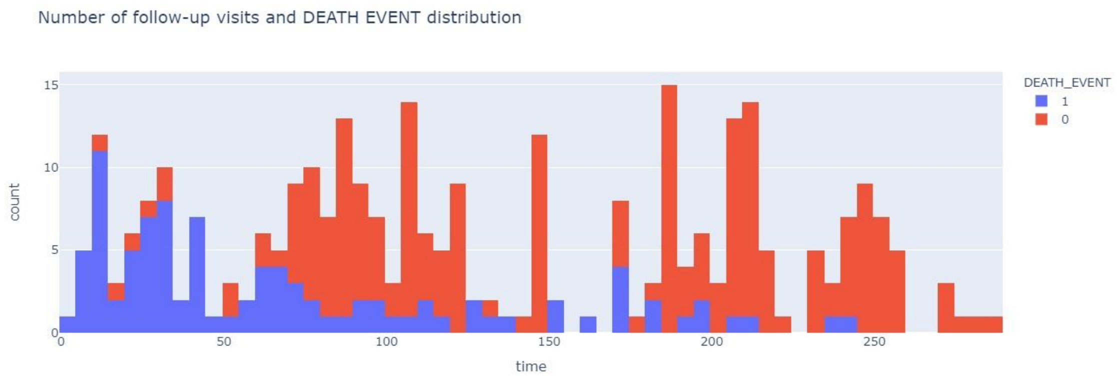


Figure 4: Number of follow-up visits and death_event occurrence distribution

The likelihood that a patient would survive a heart failing condition or pass away appears to be correlated with the frequency and quantity of follow-up visits. A patient has a greater likelihood of survival the more frequently they attend their scheduled follow-up appointments, and vice versa.

Before, it was believed that if a muscle tissue is damaged, creatinine phosphokinase will flow into the blood, which may also be one of the symptoms of heart failure (Stewart, 2020). Therefore, we sought to identify any associations between creatinine phosphokinase and fatal events, as shown in Figure 5.

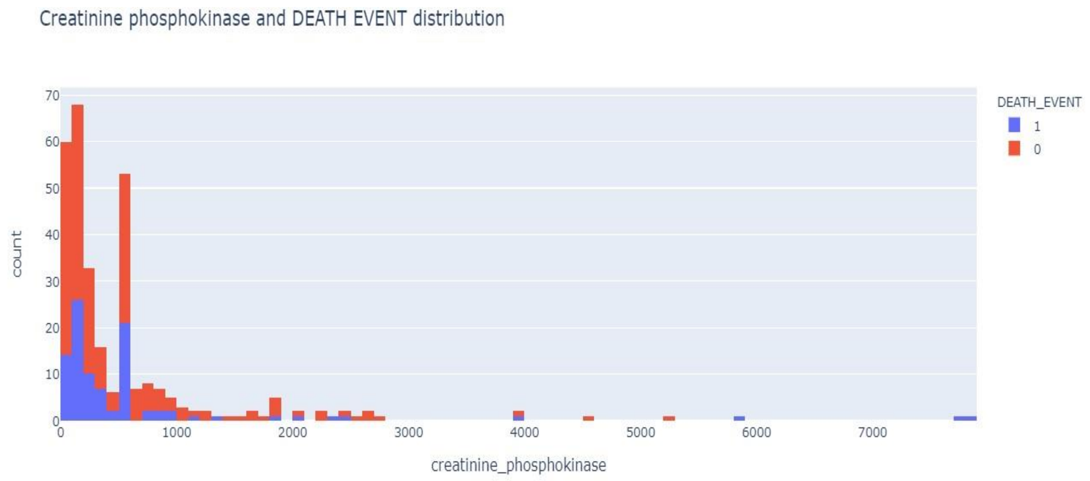


Figure 5: Relationship and distribution between creatinine phosphokinase level and death event

The distribution pattern for the enzyme creatinine phosphokinase, or CPK, appears to be rather comparable in samples that have survived and those that have not. Although the CPK level had a few outliers, they were small and were ignored in the study.

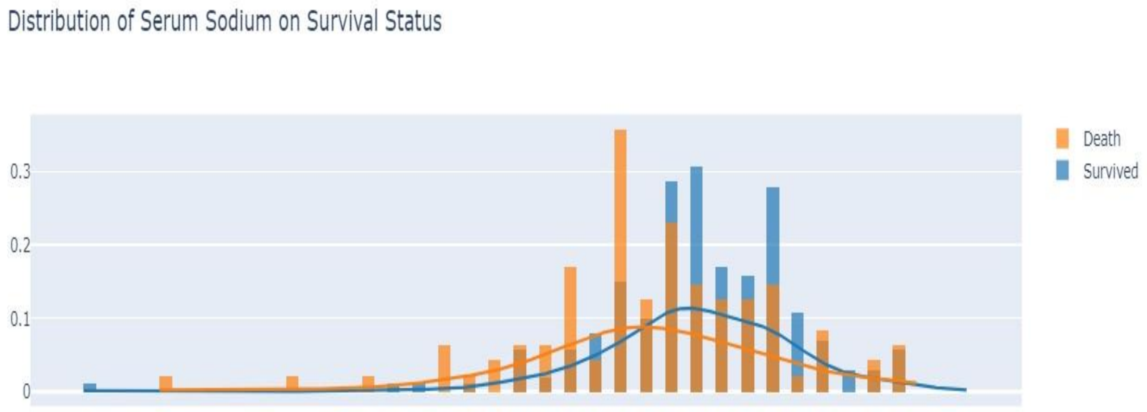


Figure 6: distribution of serum sodium on survival status

If we look at Figure 6's distribution trend for serum sodium on the survival status of either survived or died from heart failure, those who survived heart failure had a little higher distribution curve of serum sodium in their bloodstream than those who did not. Those who lived had a median level of 137 serum sodium in their circulation, compared to 134 for those who did not.

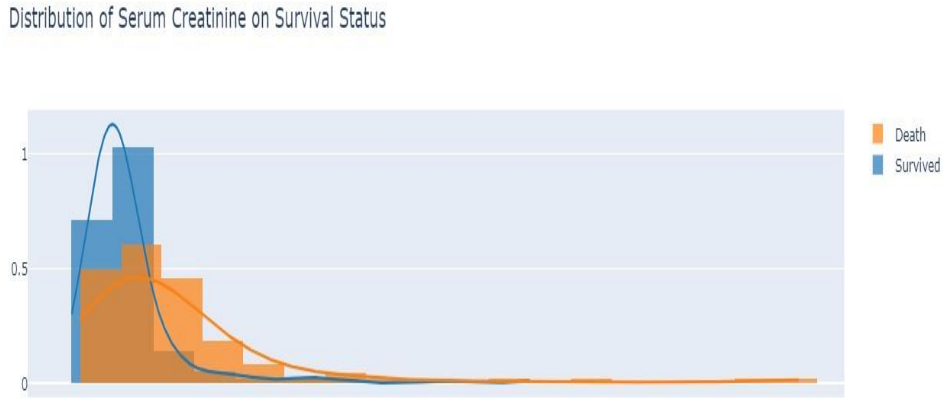


Figure 7: distribution of serum creatinine on survival status

According to the pattern seen in Figure 7, individuals who survived with heart failure disease had lower serum creatinine levels than those who did not. For people who survived with heart failure, a median serum creatinine level of 1 is considered normal. On the other hand, individuals who passed away from heart failure had a median serum creatinine level of 1.5. Since serum creatinine is produced as a byproduct of creatinine, renal dysfunction may be to blame for elevated serum creatinine levels (Stephens, 2019). This may help to explain why those who died of heart failure had such elevated serum creatinine levels.

Distribution of Ejection Fraction on Survival Status

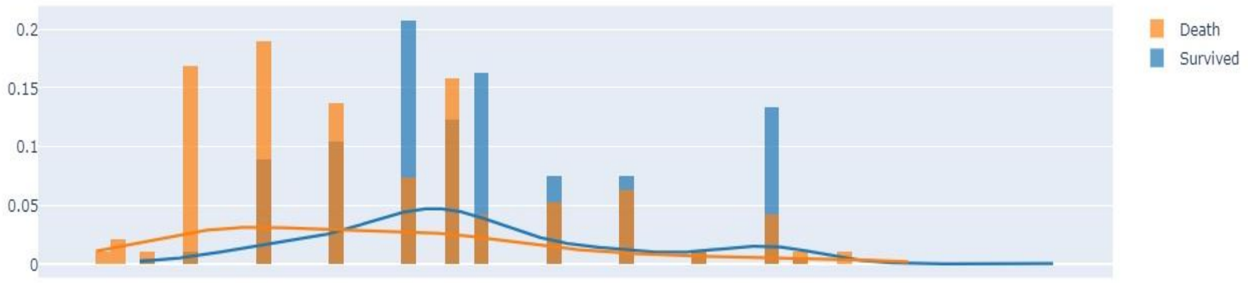


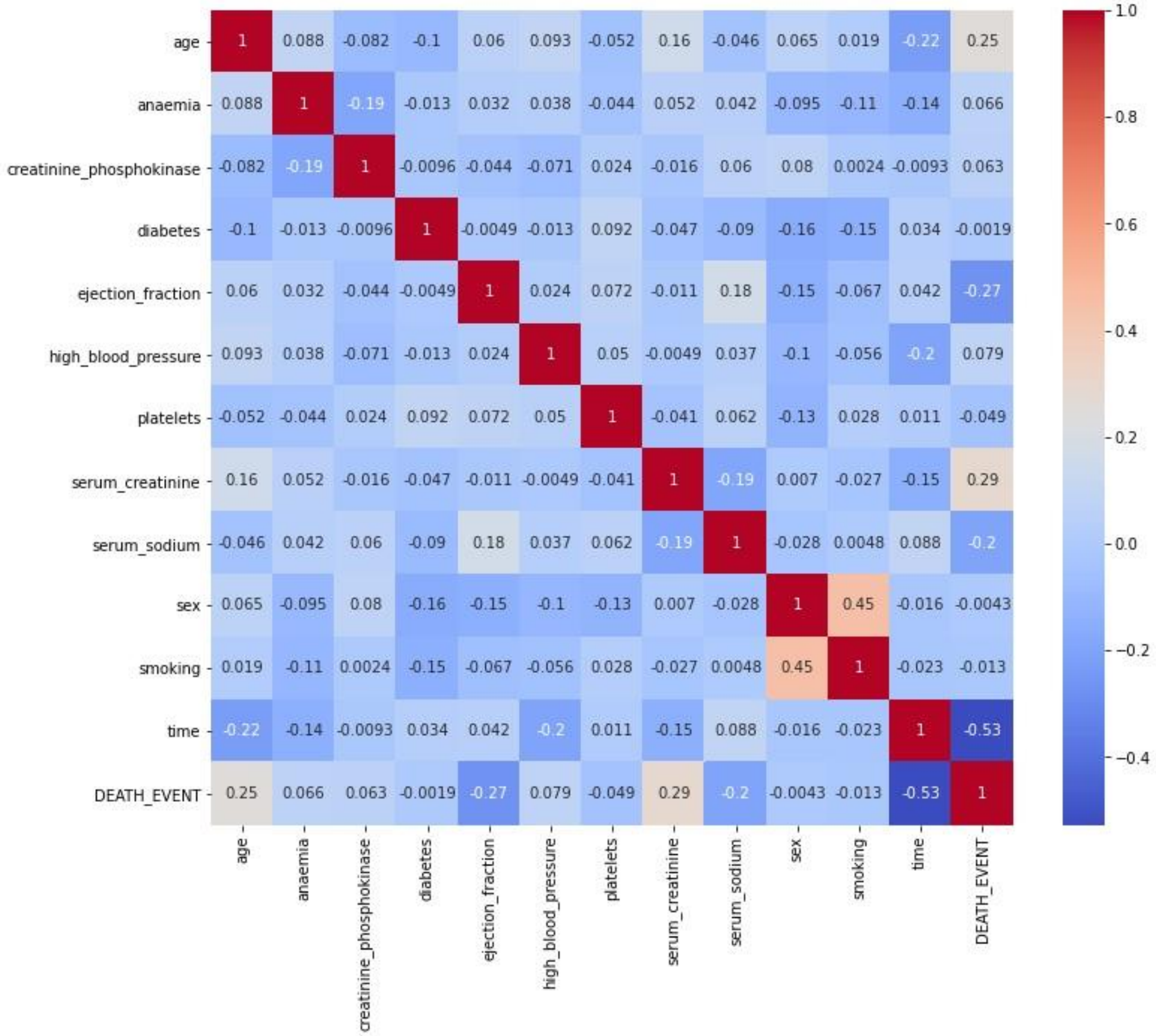
Figure 8: The distribution of ejection fraction between death and survived group of patients with heart failure.

Regarding ejection fraction, one of the traits in the group of patients who survived and lived after being diagnosed with heart failure was a high percentage of ejection fraction in blood. But those who passed away from cardiac failure had substantially lower blood levels of ejection fraction (Tripoliti, 2017).

In order to determine if platelet counts had any bearing on mortality occurrences, we also wished to examine this trend and another variable. However, there were no variations in the amount of platelets between patients who survived and those who passed away from heart failure. Figure 9 of the chart is located in the Appendix A section.

Phase 2 – Features selection

Using a heatmap correlation, we did a feature selection process to determine which factors had the most impact on the death event variable depicted in Figure 10 after analysing the general correlations and patterns between various features.



	age	anaemia	creatinine_phosphokinase	diabetes	ejection_fraction	high_blood_pressure	platelets	serum_creatinine	serum_sodium	sex	smoking	time	DEATH_EVENT
age	1	0.088	-0.082	-0.1	0.06	0.093	-0.052	0.16	-0.046	0.065	0.019	-0.22	0.25
anaemia	0.088	1	-0.19	-0.013	0.032	0.038	-0.044	0.052	0.042	-0.095	-0.11	-0.14	0.066
creatinine_phosphokinase	-0.082	-0.19	1	-0.0096	-0.044	-0.071	0.024	-0.016	0.06	0.08	0.0024	-0.0093	0.063
diabetes	-0.1	-0.013	-0.0096	1	-0.0049	-0.013	0.092	-0.047	-0.09	-0.16	-0.15	0.034	-0.0019
ejection_fraction	0.06	0.032	-0.044	-0.0049	1	0.024	0.072	-0.011	0.18	-0.15	-0.067	0.042	-0.27
high_blood_pressure	0.093	0.038	-0.071	-0.013	0.024	1	0.05	-0.0049	0.037	-0.1	-0.056	-0.2	0.079
platelets	-0.052	-0.044	0.024	0.092	0.072	0.05	1	-0.041	0.062	-0.13	0.028	0.011	-0.049
serum_creatinine	0.16	0.052	-0.016	-0.047	-0.011	-0.0049	-0.041	1	-0.19	0.007	-0.027	-0.15	0.29
serum_sodium	-0.046	0.042	0.06	-0.09	0.18	0.037	0.062	-0.19	1	-0.028	0.0048	0.088	-0.2
sex	0.065	-0.095	0.08	-0.16	-0.15	-0.1	-0.13	0.007	-0.028	1	0.45	-0.016	-0.0043
smoking	0.019	-0.11	0.0024	-0.15	-0.067	-0.056	0.028	-0.027	0.0048	0.45	1	-0.023	-0.013
time	-0.22	-0.14	-0.0093	0.034	0.042	-0.2	0.011	-0.15	0.088	-0.016	-0.023	1	-0.53
DEATH_EVENT	0.25	0.066	0.063	-0.0019	-0.27	0.079	-0.049	0.29	-0.2	-0.0043	-0.013	-0.53	1

Figure 10: Heat map correlation of all variables in the dataset.

Platelets and Creatinine Phosphokinase enzymes should be dropped from our model because, as shown in Figure 10, they do not appear to have the same impact on survival as other factors like smoking, sex, high blood pressure, diabetes, and anaemia. As a result, we should not choose these factors to base our model on.

Time, serum creatinine, ejection fraction, age, and serum sodium are the variables I will pick to include in my prediction models based on the association heatmap.

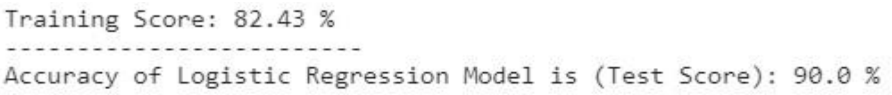
Phase 3 – Models Selection and Evaluation

To choose the optimal model for the predictions in this step, we employed an iterative modelling strategy. Five distinct models were combined: decision tree, random forest, support vector machine, logistic regression, and k-nearest neighbours. Following are results and accuracy ratings for each model:

- **Model 1: Logistic Regression**

Because it utilises a logistic function to plot a binary output model, logistic regression is more like a classification model than a regression model, while it shares certain similarities with linear regression.

Figure 11 shows the accuracy score that was obtained.



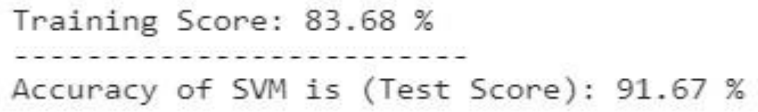
```
Training Score: 82.43 %
-------------------------
Accuracy of Logistic Regression Model is (Test Score): 90.0 %
```

Figure 11: Logistic Regression accuracy table

We were able to use logistic regression to achieve an accuracy score of 90% with a training score of about 82%.

- **Model 2: Support Vector Machine**

The support vector machine is a tool that may be used for regression and classification. The SVM accuracy score is shown in Figure 12



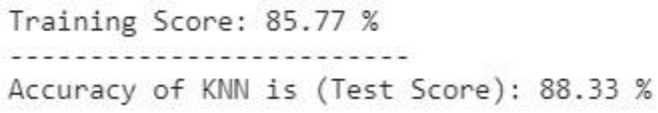
```
Training Score: 83.68 %
-------------------------
Accuracy of SVM is (Test Score): 91.67 %
```

Figure 12: Support Vector Machine accuracy table

The accuracy score for this prediction model using SVM was 91.67%, while the training score was 83.68%.

- **Model 3: K-Nearest Neighbors (KNN)**

A non-parametric model known as KNN is employed in both classification and regression. It is sometimes referred to as a model of lazy learning with local approximation (Tripoliti, 2017). When using KNN, we find k neighbours and make a forecast. We picked a k of 5 for the forecast since anything higher will cause our accuracy model to become less accurate.



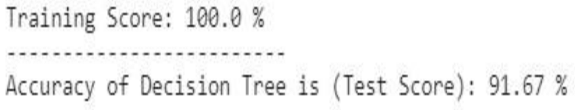
```
Training Score: 85.77 %
-------------------------
Accuracy of KNN is (Test Score): 88.33 %
```

Figure 13: Accuracy score for KNN predictive model

Figure 13 shows that for KNN, the accuracy we obtained was 88.33% and the training score was 85.77%.

- **Model 4: Decision Tree**

Regression and classification issues are solved using decision trees. For dependent variables with continuous values, decision trees are employed, whereas classification trees are used for dependent variables with discrete values. The independent variables are used to create a decision tree, with each node having a condition over a feature. Based on the criteria, the nodes choose which node to travel to next. An output is anticipated once the leaf node is reached (Latha, 2019).



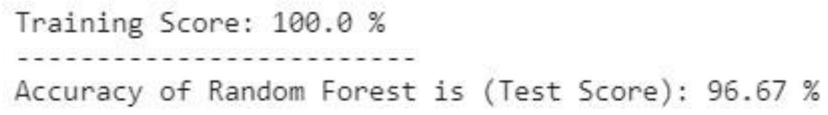
```
Training Score: 100.0 %
------------------------
Accuracy of Decision Tree is (Test Score): 91.67 %
```

Figure 14: Accuracy score for Decision Tree predictive model

Figure 14 shows that our training score was flawless and the decision tree prediction model's accuracy was 91.67%, which is quite intriguing.

- **Model 5: Random Forest**

In an ensemble model called Random Forest, different decision trees are blended to create a more robust and precise model. With binary, categorical, and continuous characteristics, Random Forest develops a reliable, accurate model that can handle a wide range of input data (Latha, 2019).

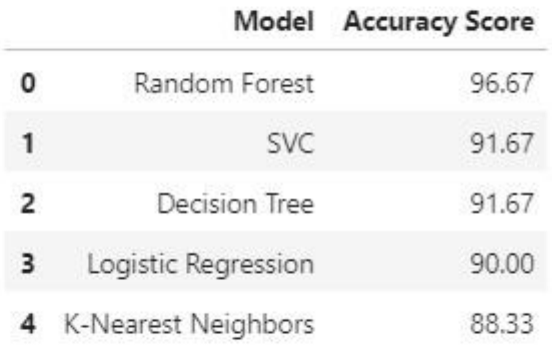


```
Training Score: 100.0 %
-------------------------
Accuracy of Random Forest is (Test Score): 96.67 %
```

Figure 15: Random Forest predictive model accuracy score

Based on Figure 15, Random Forest had a 96.67% accuracy score with a flawless training score on our dataset.

Table 1 displays the overall summary of the accuracy scores obtained by the five models, ranking them from top to lowest in terms of accuracy scores.



	Model	Accuracy Score
0	Random Forest	96.67
1	SVC	91.67
2	Decision Tree	91.67
3	Logistic Regression	90.00
4	K-Nearest Neighbors	88.33

Table 1: Accuracy score summary table of all predictive models

Figure 16 shows the accuracy score comparison among all models in a bar plot

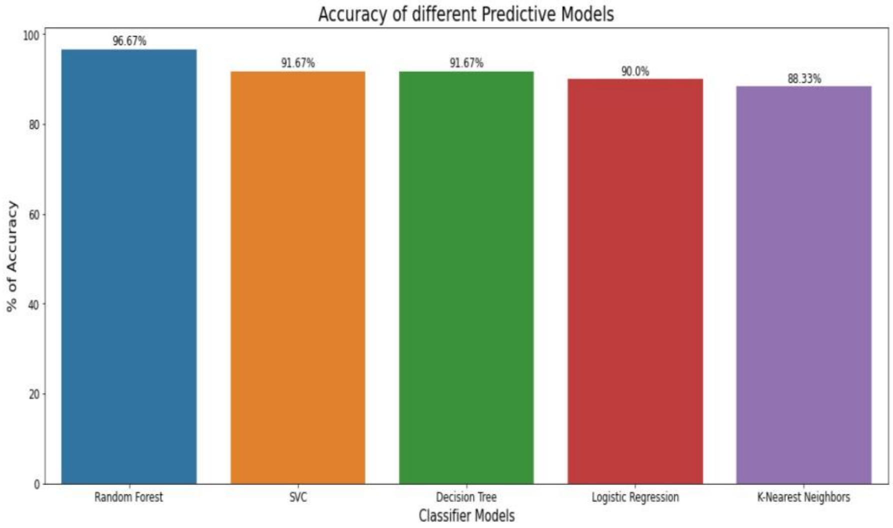


Figure 16: Comparison of different predictive models on accuracy score

Discussion and Conclusion

The dataset and its variables were better understood as the exploratory data analysis phase was carried out, and their linkages were depicted. Utilizing data analysis tools like histograms, heatmaps, and a data profiling stage facilitates the compilation of all EDA analysis. We found no missing values, indicating that the dataset was very clean and well-organized. In the second and third stages of our method, important characteristics were found using correlation heatmaps, and predictive models were created using these relevant and significant features. Out of the 13 variables, only 5 were ultimately determined to be significant and have a significant impact on our predictive model.

Random Forest had the greatest accuracy score (96.67% for the prediction) out of the five models we created using training data from our dataset and testing. Out of all the models we examined, the KNN model had the worst accuracy (88.33%). If we can modify the hyperparameters using cross-validation on our dataset, there is greater possibility for advancement. Additionally, I think the n = 299 data sample may be a little too small. Our predictive model might be a lot more precise with a larger sample size, and we might even be able to comprehend and choose new features to include in our model. In general, depending on the provided health indicators, I think we can reliably forecast survival from heart failure disease using the Random Forest Model.

Appendix A

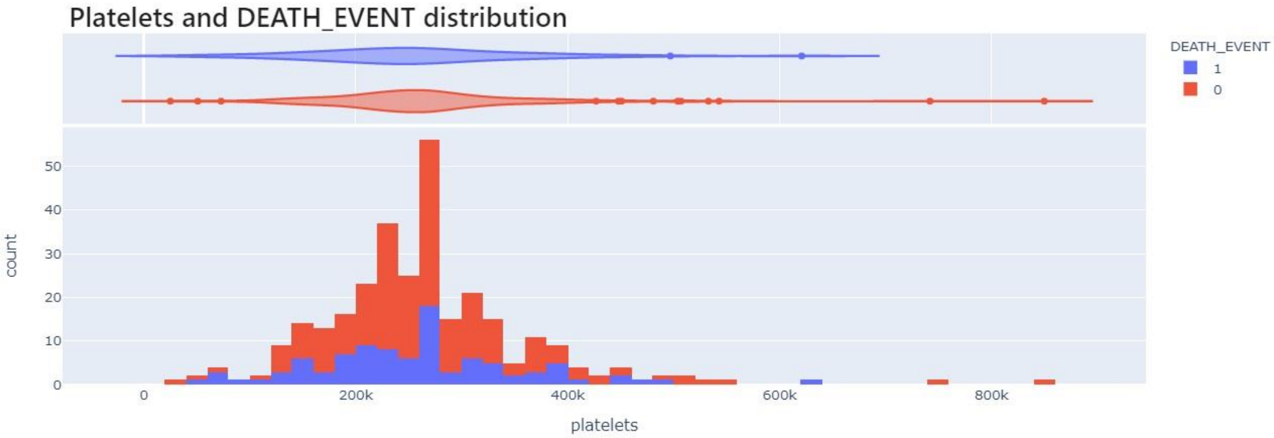


Figure 9: Platelets Count on Survival Status

References

1. Ahmad T, Munir A, Bhatti SH, Aftab M, Raza MA. Survival analysis of heart failure patients: a case study. PLoS ONE. 2017; 12(7):0181001.
2. Chicco, D., Jurman, G. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. BMC Med Inform Decis Mak 20, 16 (2020). https://doi.org/10.1186/s12911-020-1023-5
3. Dalen, J. E., Alpert, J. S., Goldberg, R. J., & Weinstein, R. S. (2014). The Epidemic of the 20th Century: Coronary Heart Disease. The American Journal of Medicine, 127(9), 807–812. https://doi.org/10.1016/j.amjmed.2014.04.015
4. Faggella, D. (2020, March 4). 7 Applications of Machine Learning in Pharma and Medicine. Emerj. https://emerj.com/ai-sector-overviews/machine-learning-in-pharma-medicine/

5. HealthITAnalytics. (2018, September 18). Using Big Data, Machine Learning to Reduce Chronic Disease Spending. https://healthitanalytics.com/news/using-big-data-machine-learning-to-reducechronic-disease-spending
6. Latha, C. B. C., & Jeeva, S. C. (2019). Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. Informatics in Medicine Unlocked, 16, 100203. https://doi.org/10.1016/j.imu.2019.100203.
7. Lawler, R. (2019, July 15). How doctors are using machine learning to improve health outcomes. Samsung NEXT. https://samsungnext.com/whats-next/how-doctors-are-using-machine-learningto-improve-health-outcomes/
8. Stephens C. What is a creatinine blood test? https://www.healthline.com/health/creatinine-blood. Accessed 25 Jan 2019.
9. Stephens, W. (2019, June 19). Machine Learning Can Predict Heart Attack or Death More Accurately Than Humans. AJMC. https://www.ajmc.com/view/machine-learning-can-predict-heart-attackor-death-more-accurately-than-humans
10. Stewart, J., Addy, K., Campbell, S., & Wilkinson, P. (2020). Primary prevention of cardiovascular disease: Updated review of contemporary guidance and literature. JRSM Cardiovascular Disease, 9, 204800402094932. https://doi.org/10.1177/2048004020949326.
11. Tripoliti, E. E., Papadopoulos, T. G., Karanasiou, G. S., Naka, K. K., & Fotiadis, D. I. (2017). Heart Failure: Diagnosis, Severity Estimation and Prediction of Adverse Events Through Machine Learning Techniques. Computational and Structural Biotechnology Journal, 15, 26–47. https://doi.org/10.1016/j.csbj.2016.11.001.
12. Varghese, D. (2019b, May 10). Comparative Study on Classic Machine learning Algorithms. Medium. https://towardsdatascience.com/comparative-study-on-classic-machine-learning-algorithms24f9ff6ab222
13. Varghese, D. (2019, February 18). Comparative Study on Classic Machine learning Algorithms , Part2. Medium. https://medium.com/@dannymvarghese/comparative-study-on-classic-machinelearning-algorithms-part-2-5ab58b683ec0
14. Yang, L., Wu, H., Jin, X. et al. Study of cardiovascular disease prediction model based on random forest in eastern China. Sci Rep 10, 5245 (2020). https://doi.org/10.1038/s41598-020-62133-5