Predicting Heart Failure From Various Health Factors

Angela Oku
Psychology
A15624633

Carmen Le
Economics
A15547577

Pallavi Saksena
Psychology
A16454660

Abstract

The dataset we've chosen lists different factors that can potentially cause heart failure, with columns that represent different factors that may contribute to heart failure (e.g. age, anemia, diabetes, high blood pressure, smoking, etc.). Some of the data is binary, such as sex, while others are continuous, such as number of platelets. For our project, we plan on using feature extraction to determine which variables have the most impact on death. By using various Machine Learning techniques such as a Logistic Regression, SVM classifier, Decision Tree, Random Forest, and Voting Classifier, our project aims to create robust models and see if our chosen features would be able to make good predictions. We used F1-scores and cross-validation to evaluate the performance of each model. We also wanted to determine which models were the best at classifying and predicting death from heart failure using certain factors. We will be looking at accuracy scores across the different models to establish which ones would work best, which could potentially be used in creating more efficient plans for prevention.

1 Introduction

Heart failure is a common occurrence that affects almost 6 million Americans and is the leading cause of hospitalization in people over 65 years old (National Heart, Lung, and Blood Institute). It occurs when the heart is weakened or damaged and there can be a plethora of causes from underlying conditions, which can make it difficult to pinpoint how someone could avoid this phenomenon. While some of these factors cannot be controlled, the risks of heart failure can be reduced by catching onto symptoms early and making the necessary lifestyle changes. Since heart failure is such a widespread issue, our group wanted to see whether we could use certain features from someone's medical information to predict whether or not they would experience death from heart failure. The motivation we had was due to our interest in exploring more about preventative measures for this medical condition. Determining which aspects are most indicative of heart failure can give us insight on what to focus on in order to gain control of these conditions and stop more potential deaths from happening.

2 Method

To find whether certain features of medical history information in the dataset could predict death from heart failure, we used several models for predictions: Logistic Regression, support vector machine (SVM), Decision Tree, Random Forest, and Voting Ensemble. Using a correlation matrix, we found which two attributes of medical history had the highest correlation with death from heart failure, which were the amount of serum creatinine in one's blood, labelled as "serum_creatinine," and the percentage of blood leaving the heart at each contraction, labelled as "ejection_fraction." Since serum creatinine is a chemical that is supposed to be filtered by blood and would have an abnormal amount from dysfunctional kidney function, and ejection fraction measures the performance of the heart in sufficiently pumps blood, these are attributes of bodily function that are reasonable indicators of health factors for death from heart failure ("Creatinine tests," 2021). From these attributes, we then used a Logistic Regression model to find the estimated probability that ejection fraction and serum creatinine predict death for a given patient. Since Logistic Regression can estimate probabilities for multiple variables, we chose this model to estimate the predictions of our two attributes of serum creatinine and ejection fraction for death from heart failure. In this model, we created training and testing data sets for training the model, and then used a confusion matrix to evaluate the accuracy of the Logistic Regression model. On its own, the strength of Logistic Regression in terms of accuracy is not as strong as a model of ensemble learning. To test the accuracy of this model, we also used cross-validation, which also showed the low accuracy of the Logistic Regression model compared to ensemble learning.

In addition to a Logistic Regression model, we also used a SVM. As a model that is flexible while avoiding overfitting and lower generalization, we chose to use a SVM so that we could find predictions for our attributes with its clear boundaries in its algorithm. In this model, we once again created testing and training data sets to train the model, and used a confusion matrix to evaluate the accuracy of the model, in addition to an f1 score. As with the Logistic Regression model, the SVM is not as strong as an ensemble learning model, and can be sensitive to outliers, however since we removed outliers when cleaning the data, this would not be an issue in our case. After the SVM, we used a Decision Tree. As with the SVM, we chose to do a Decision Tree for its flexibility, and also for an alternate form of visualization to interpret data through a decision tree. As with the other models, we created training and testing data sets, and performed an accuracy measure, in this case an f1 score. As with the Logistic Regression model and SVM, this model was also not strong in comparison to ensemble learning in terms of accuracy, but also had the lowest accuracy out of the three. Since Decision Trees have the limitation of higher risk for overfitting, this model having the lowest accuracy out of all of our models fits the general strength expected for a Decision Tree. Decision Trees can be strengthened with the use of Random Forests, however, as shown with higher accuracy with our use of a Random Forest.

As stated before, we used Ensemble learning models, Random Forest and Voting, in addition to the Logistic Regression model, SVM, and Decision Tree. As an Ensemble method using multiple Decision Trees, we chose to use Random Forest for its greater strength than our individual Decision Tree. As with our other models, we used training and testing data to train our model, as well as used an f1 score to measure its performance. As an Ensemble model, the Random Forest is a strong measure with the complexity of its algorithm and degree of testing and training data. As with the Random Forest, the final model we used, a Voting Ensemble method, also uses Ensemble learning. In this model, we used our Logistic Regression, Decision Tree, SVM, and Random Forest models as estimators. We chose this model for its strength and versatility to take in so many different models, and to provide a means to directly compare each of the models as well as use all of them to create a more powerful model with higher accuracy. Due to the power of the model in using so many methods with their own strengths, the Voting classifier is very strong, as reflected by it having the highest accuracy out of all the models we used.

3 Experiments

3.1 Dataset

The dataset we are using is the Heart Failure Prediction dataset, which is a CSV file that documents medical information for different patients who have experienced heart failure and aims to predict whether that person died or not. The 13 different columns are age, anaemia, creatinine phosphokinase levels, diabetes, ejection fraction percentage, high blood pressure, amount of platelets, serum creatinine levels, serum sodium levels, sex, smoking, time of follow-up period expressed in days, and whether or not a patient died during the follow-up period. In order to choose the features we wanted to use, we looked at the correlation between each column with the DEATH_EVENT column, and the 2 features with the most correlation (either positive or negative) would be used. We excluded the time column because the death of a patient would directly affect the length of the follow-up period, which we were afraid would bias the models.

To prepare the data to use for the models, we first showed the descriptive statistics for each column in the dataset and its shape. We checked for null values in each column and defined a function that would remove outliers using the interquartile range method. We found the first and third quartiles, and subtracted them from each other to find the interquartile range. We multiplied the interquartile range by 1.5, and either subtracted this number from the first quartile, where anything below it would be a lower outlier, or added this number to the third quartile, where anything above would be a higher outlier. After removing the outliers, we checked if there were any missing points in each column again and filled them in with the median values of each column. After looking at the correlation data and choosing the 2 best columns, we counted the number of unique values for each of them, and graphed their distribution.

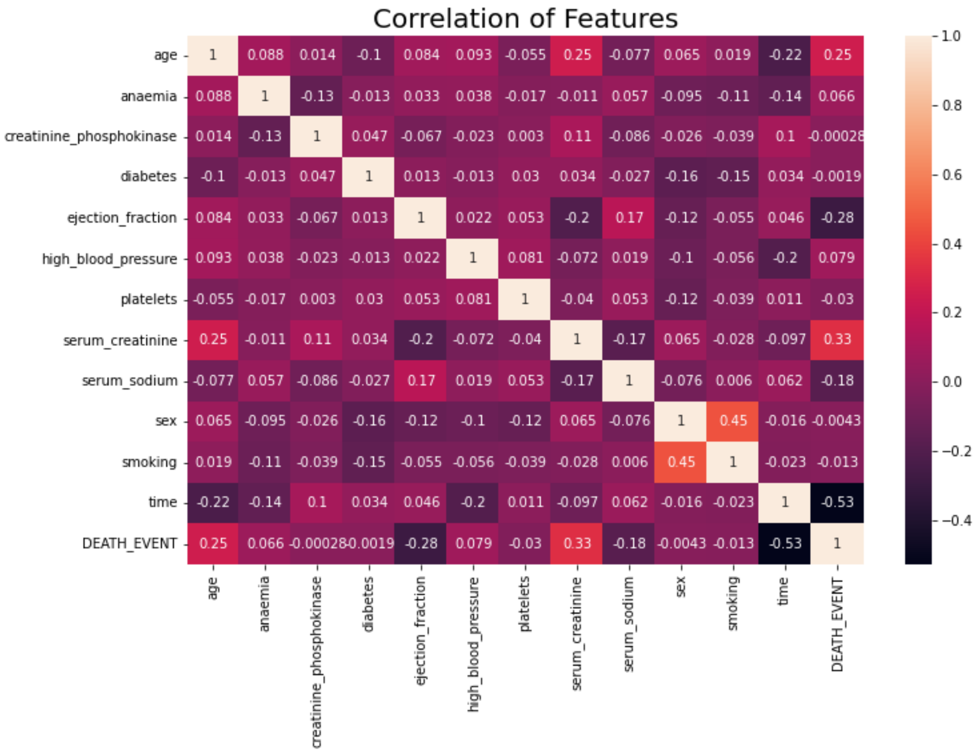

Figure 1: Correlation Matrix

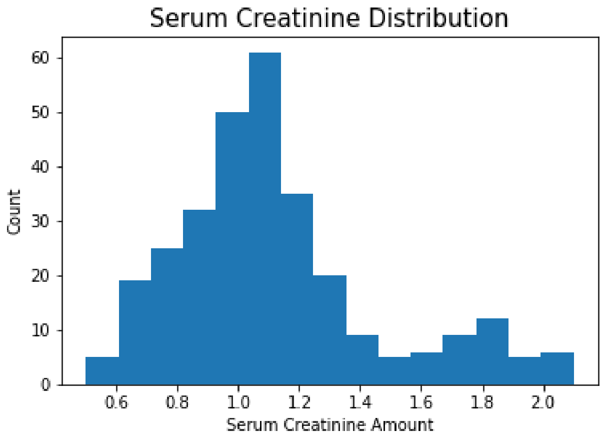

Figure 2: Distribution of Serum Creatinine Column

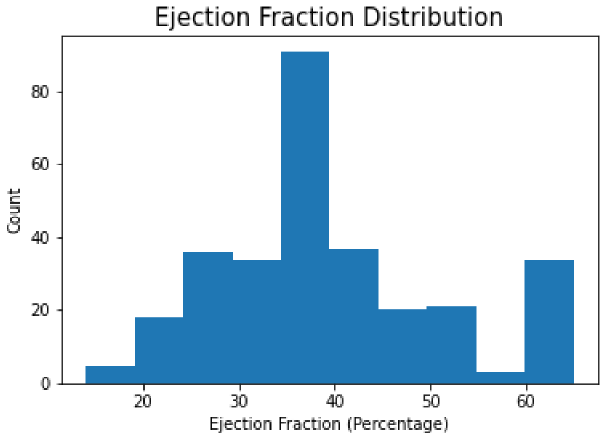

Figure 3: Distribution of Ejection Fraction Column

For each of the models, we would scale the dataset using StandardScaler and separate the data into training and testing sets. The size of the training sets would be 80% of the overall data.

3.2 Results

From the correlation matrix, we found that the 2 features with the strongest positive or negative correlation to death from heart failure were serum creatinine and ejection fraction. When using the Logistic Regression model, we found that the accuracy ended up being 83.3%, and the F1 score was 61.5%. The F1 score, which was the measure of the model's performance, favors classifiers that have similar precision and recall, which can be calculated by the values in the confusion matrix. (Gerón, 93). Our F1 score showed that the Logistic Regression model was a pretty decent classifier for death from heart failure.

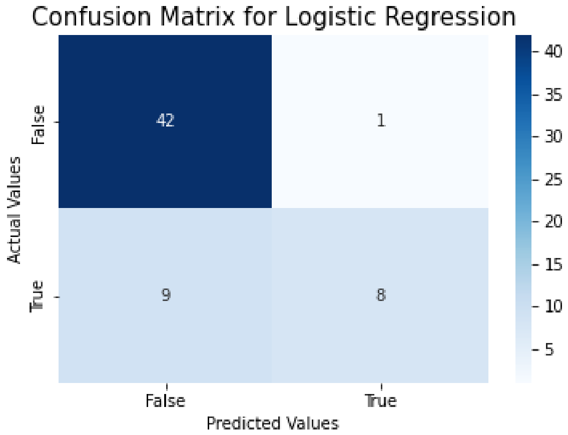

Figure 4: Confusion Matrix for Logistic Regression

For the SVM classifier, both accuracy and F1 score increased. For this model, we set the hyperparameters gamma = 5 and C = 25, and accuracy was 81.6% while F1 score was 70.2%. The confusion matrix looks better here, and shows that more instances were correctly predicted.

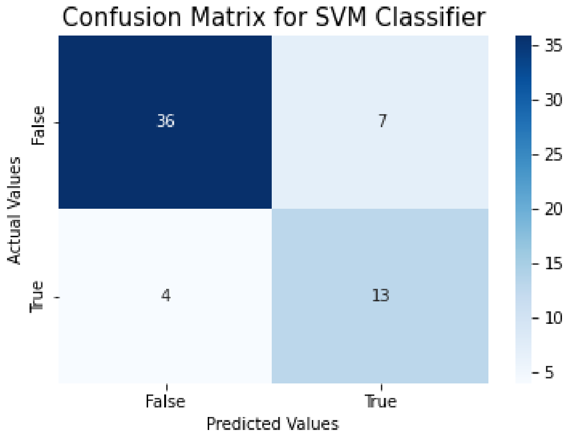

Figure 5: Confusion Matrix for SVM Classifier

When looking at the graph of the SVM classifier, it shows decision boundaries that are able to accurately predict and classify death from heart failure using our chosen features.

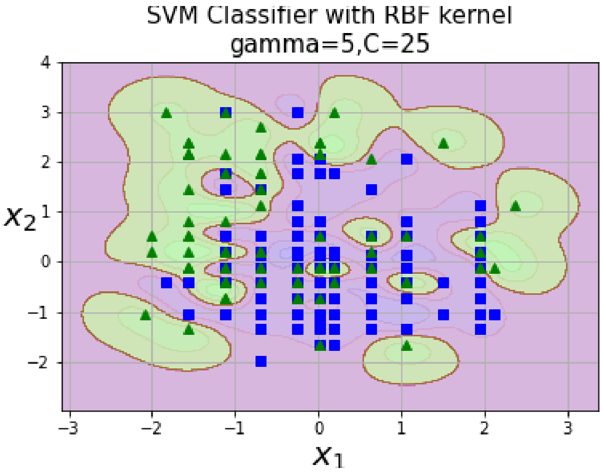

Figure 6: Plot for SVM Classifier

With the Decision Tree, we set the hyperparameters max_depth=3, which would limit the amount of nodes, and min_samples_leaf = 50, which set the minimum number of samples that a node should have before it can be split. For this model, accuracy was 78.3% and F1 score was 64.8%. This shows that the Decision Tree is also a pretty good classifier, but not the best.

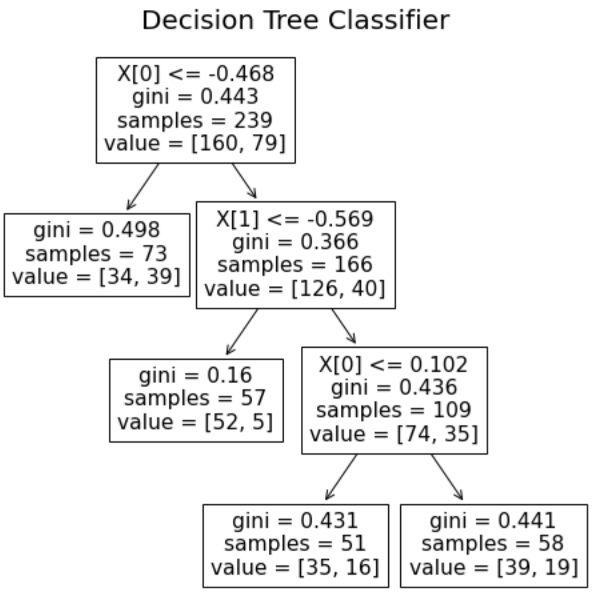

Figure 7: Decision Tree Plot

The classifiers that work better are the Ensemble Learning methods. With the Random Forest Classifier, which is a group of Decision Trees, the accuracy was 81.6% and F1 score was 70.2%. The voting classifier that aggregates the predictions from each classifier is shown to make the best predictions. The accuracy is 86.6%, and F1 score is 73.3%, which is the highest combination out of all the different models. In this image showing the accuracies of the various models, the Voting Classifier has the highest accuracy.

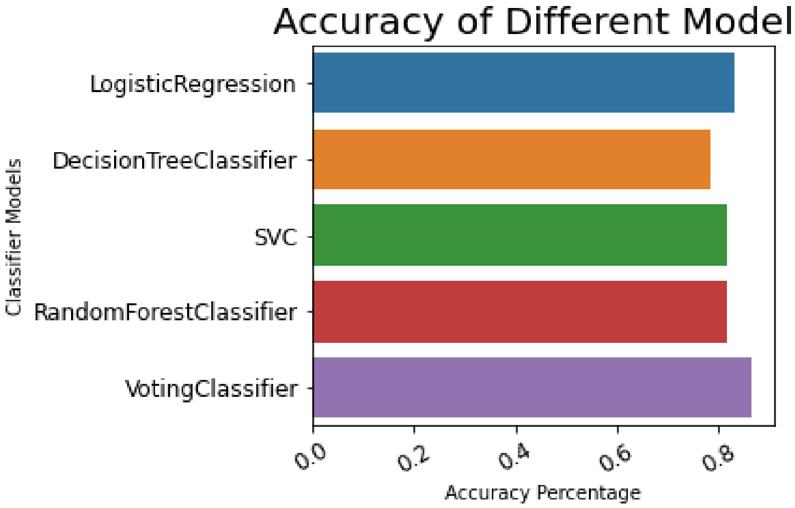

Figure 8: Accuracy of Different Models

The results from our findings show that when choosing features based on the amount of correlation with the target variable, the accuracy of the models are pretty high. Some models do better than others, but ultimately, the models using the Ensemble Learning technique do the best. This proves the idea that even if classifiers are weak learners, an ensemble can still be a strong learner when trying to make predictions. (Gerón, 190)

4 Discussion

4.1 Significance of Findings

We used multiple models and compared their performances in order to determine which ones would have better predictions, with higher accuracy scores and better performances. The Logistic Regression, SVM Classifier, and Decision Tree Classifier models were all able to classify the deaths fairly well, but they definitely could each be improved upon to raise their accuracy. However, we got much better results using Ensemble Learning techniques. For Ensemble Learning, first we used a Random Forest algorithm, which was more accurate than the single Decision Tree Classifier. Lastly, we used a Voting Classifier, which was the most accurate of all the models. Therefore, from all these findings, the Voting Classifier is the best one from these models that we would choose to apply in a real-world setting of predicting heart failure.

While some situations may not require very great accuracy, a situation involving heart failure is very important and necessary to study in order to prevent deaths from heart failure. This is why having well-trained, accurate classifier and predictor models are extremely important to have. As stated earlier, the models used in our experiments would need to be improved to have higher accuracy and F1 scores, since we would want to be able to use them when dealing with heart failure issues. Of course, a model will never be perfect, and our models are mostly above 80% accuracy, which seems pretty good. However, it would be much better if the accuracy score of these models was above 90% instead.

Especially in the healthcare system, people would benefit from a good model that can help them make better decisions for their health. As stated in the Introduction section, we wanted to examine if certain features about a person's medical health could predict them having heart failure later on. We did find this information, which is useful for determining preventative measures based on the most impactful factors, in order to reduce deaths from heart failure.

4.2 Speculation for Future Work

When visualizing the data, we found that serum creatinine and ejection fraction were the two factors with the highest correlation to heart failure. In terms of applied healthcare work, this finding suggests that doctors should focus on those two factors when looking at risk factors in their patients and for trying to prevent heart failure. In the "About this dataset" section on Kaggle, the uploader of this dataset describes that heart failure is caused by cardiovascular disease, which is the "number 1 cause of death globally". It is therefore especially important to be able to predict and prevent potential heart failure since this issue is so prevalent and affects many people.

Meanwhile, for the computational modeling aspect, we can use these two factors again to either improve the current models used in this project, or also to create and train new algorithms to see if those models are better. As shown in these experiments and as discussed earlier, an Ensemble Learning algorithm can become a strong learner even when it has many weak learners. Therefore it may be beneficial to focus on this type of algorithm more than any of the individual algorithms. It may also be of interest to look into other factors that were not included in this dataset, to see if anything else also has a high impact on heart failure.

One way to improve the models for the future is to have a larger dataset, since there were only 299 instances in this one. Machine learning algorithms require a lot more data than this, sometimes thousands or millions of instances, in order to really perform well (Gerón, 24). Having a much larger dataset with at least a few thousand instances would help us to better see and understand the true patterns in the data. Having more instances also allows us to generalize the results better, since they will be more representative of the population. In this case, we would be able to have a better idea of how different health factors, including serum creatinine and ejection fraction, affect a larger population outside of the sample used for the study. This is especially important because while the human body is fairly similar from person to person, no two people are exactly the same, and may be affected somewhat differently. Having more data lets us examine results for more diverse sets of people so we can see how heart failure may occur given different characteristics or combinations of factors for different people. Overall, it would allow for increased and improved preventative care to reduce heart failure for more people in the long run.

5 Contributions

Angela Oku wrote the Method section, Carmen Le wrote the Introduction and Experiments section, and Pallavi Saksena wrote the Discussion section.

References

Aurélien Gerón. Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow. O'Reilly Media, Inc. (2019).

Davide Chicco, Giuseppe Jurman: Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. BMC Medical Informatics and Decision Making 20, 16 (2020). https://www.kaggle.com/andrewmvd/heart-failure-clinical-data.

Mayo Foundation for Medical Education and Research. (2021, February 25). *Creatinine tests.* Mayo Clinic. Retrieved from https://www.mayoclinic.org/tests-procedures/creatinine-test/about/pac-20384646.

National Heart, Lung, and Blood Institute. (2021). *Heart Failure.* NHLBI. Retrieved from https://www.nhlbi.nih.gov/health-topics/heart-failure