JOURNAL OF
EMERGING INVESTIGATORS

Article

# Survival analysis in cardiovascular epidemiology: nexus between heart disease and mortality

**Anya Lachwani[1], Tim Gianitsos[2]**

[1] Saint Francis High School, Mountain View, California

[2] Inspirit AI, San Francisco, California

## SUMMARY

**In 2021, over 20 million people died from cardiovascular diseases, accounting for approximately one-third of all global deaths – an increase from the 1990's when approximately 12 million people died each year. Given the significant implications, it is imperative to explore further. Past studies identified variations in the median survival rate following heart failure between different sexes. While one study found correlations between the gender and age of the patient, another looked at age and blood pressure. However, it is noteworthy that these studies have primarily focused on no more than three factors when in fact multiple factors contribute to mortality following heart failure. We explored which of the many variables influences the outcome of patient mortality in the event of heart failure and how the influence changes as these features are removed or added. We hypothesized that age and high blood pressure would be most strongly correlated with mortality following an instance of heart failure due to the structural changes caused by both of these factors that may impede function. While aging prompts changes in the heart's structure including muscle cell deterioration, valve rigidity, and reduced chamber capacity, high blood pressure hampers arterial capabilities. Random forest (RF) was the most effective of the three models created, and the three most important factors influencing the outcome of patient mortality were determined to be time, serum creatinine, and ejection fraction. This study could be one step out of the many that are needed towards assisting with personalized medicine to improve the chances of a patient's survival following heart failure.**

## INTRODUCTION

In 2021, over 20 million people died from cardiovascular diseases (CVD), an increase from the 1990's where approximately 12 million people died each year (1). In the Framingham Heart Study, conducted on residents of Framingham, Massachusetts, the median survival rates for men and women following heart failure were 1.7 years and 3.2 years respectively (2). The mortality rate for men was 10% more than that for women in the 5-year study (2). Another study that assessed the causes of mortality in the event of heart failure in the Netherlands showed that mortality rates increased substantially with age (3). A third study conducted on older patients to determine the correlation of mortality and systolic blood pressure (SBP) found that those with a higher SBP were often associated with a lower risk of mortality over a month-long and year-long interval (4). All these studies looked at multiple metrics but were conducted without analyzing more than three variables in conjunction.

Generally, the treatment for heart failure (a type of CVD) includes medications and surgeries (5). These medications might include but are not limited to angiotensin-converting enzyme inhibitors to reduce blood pressure, sodium-glucose cotransporter-2(SGLT2) inhibitors often prescribed to those with type 2 diabetes, or angiotensin receptor/neprilysin inhibitors for people with reduced ejection fractions. (5). The purpose of this study was to evaluate the extent to which factors sex, age, SBP, anemia, creatine phosphokinase (CPK) level, ejection fraction, hypertension, platelets, serum creatinine level, serum sodium level, diabetes, smoking, and time play a role in the mortality of the patient following heart failure as well as examine the correlations between these factors.

To determine the most important factors we first used logistic regression, multi-layer perceptron (MLP) classifier, and RF models to find which model most accurately predicted the death event. We found the RF model to be most accurate at least for our dataset. We then used the RF model to further determine which features were most important to keep within normal limits and which could indicate the highest probability of survival for the patient. This could help doctors focus on prescribing medications to reach the optimal levels for the most important features for potential survival of the patient. While our findings are correlative rather than causative, our model provides a basis for future research in the causal relationships between the features and patient outcomes.

## RESULTS

We used the heart failure data set obtained from Kaggle (6). The data set contains information from 299 patients with multiple binary and non-binary factors. Binary factors include sex, whether the patient has hypertension, anemia, diabetes, and smokes as well as mortality outcome. Non-binary factors include age, CPK level, ejection fraction, platelet count, serum creatinine level, serum sodium level, and time. An elevated CPK level indicates that the patient has undergone stress to their heart (7). Ejection fraction indicates a patient's hearth strength, and a low ejection fraction signifies heart failure (8). Abnormal platelet count levels may signify risks associated with clotting and bleeding (9). High or low serum creatine levels indicate issues with kidney function (10). Serum sodium levels indicate the amount of sodium in the blood which, if not regulated, can lead to hyponatremia (11). Time referred to the amount of time between the heart failure and mortality outcome, increasing with every follow-up visit indicating the person is still alive. We first processed the data such that

https://doi.org/10.59720/24-024

the first 12 columns were the features that each patient was tested on and the last column was the mortality status of the patient. The mortality was given in the form of 0 (alive) or 1 (dead).

After the data was processed and split for training and testing, we used logistic regression, MLP classifier, and RF models to determine the patient death event. To assess the performance of each model, we computed a training accuracy value which indicated the accuracy of the model when analyzing the training data. We also test accuracy, which indicated the accuracy of the model on new data that was not part of the training data set. In its application to the unknown data, test accuracy was the most important metric to consider because when predicting a patient's mortality outcome, the patient's other factor levels would not always match the training data. Hence, the model's ability to adapt to the data given and predict based on that data is the most important metric. Training accuracy is crucial to determine whether the model selected can fit the data given to it well in addition to detecting overfitting where the model selected fits the data too well. A low training accuracy suggests the need for a different model, while a very high training accuracy necessitates checking the model's performance on test data to ensure it generalizes well and isn't overfitting.

After numerous changes to the hyperparameters using a GridSearch function, which finds the best model given the hyperparameters, the logistic regression model had a training accuracy of 78.2%, and a test accuracy of 65%. The accuracies for the MLP classifier model were the following: test accuracy of 58.33% and training accuracy of 70.3% The accuracies for the RF model were the following: accuracy on test data of 73.33% and an accuracy on the training data of 100%. These results suggest that RF was the best model, as it predicted the patient death event most accurately on unseen data (**Table 1**).

After assessing the performance of the three models, we sought to determine which features contributed the most to accurately predicting a patient's death or survival using RF importance. A higher RF importance value suggests that the feature provided more statistical disparity in determining the output. When ranking the features from high to low RF importance, the order was: time, serum creatinine, ejection fraction, age, creatinine phosphokinase, platelets, serum sodium, anemia, sex, diabetes, smoking, and high blood pressure (**Figure 1**). After the removing the most important feature (time) to assess the interactions between the other factors in the data, we observed that age became the most important factor, and anemia, creatinine phosphokinase, diabetes rose in importance. The new order of importance was: age, anemia, creatinine phosphokinase, diabetes, ejection fraction, high blood pressure, platelets, serum creatinine, serum sodium, sex, and smoking (**Figure 2**). The order was slightly different from when time was included. In addition, after removing time from the model, the model's training accuracy was 96.65% and the test accuracy was 66.67%. RF importance was computed during training and can be biased, so we also calculated permutation importance post-training, as this was unbiased. Overall, the feature rankings through RF importances and permutation importances were similar (**Figure 3**).

| | Logistic Regression Model | MLP Classifier Model | Random Forest Model |
|---|---|---|---|
| Training Accuracy | 78.2% | 70.3% | 100% |
| Test Accuracy | 65% | 58.33% | 73.33% |

**Table 1: Training and Test accuracies across different models.** The training accuracies of the models in descending order are as follows: Random Forest, Logistic Regression, MLP Classifier. The test accuracies of the models in descending order are Random Forest, Logistic Regression and MLP Classifier. This confirms that Random Forest was the best performing model and MLP Classifier was the worst performing model. Table 1: Training and Test accuracies across different models. The training accuracies of the models in descending order are as follows: Random Forest, Logistic Regression, MLP Classifier. The test accuracies of the models in descending order are Random Forest, Logistic Regression and MLP Classifier. This confirms that Random Forest was the best performing model and MLP Classifier was the worst performing model.

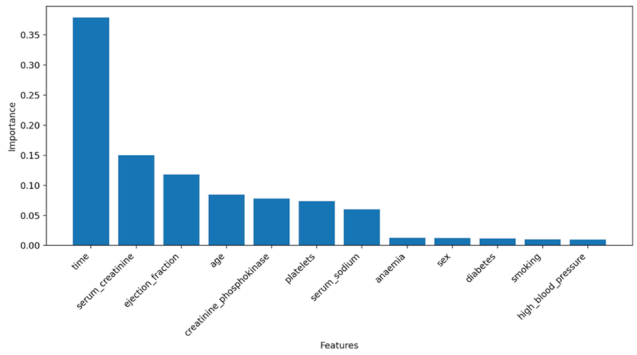


**Figure 1: Features in the order of importance.** Bar graph showing most important feature (time) to least important (high blood pressure) and their corresponding random forest importances which is an indication of predictive power ranking.

We then trained the model based on the most important feature, two most important features, three most important features, etc., until the model was trained on all the features. As the model was trained with increasing the number of features, the training accuracy was almost consistently at 100% once the number of important features being trained on was two or greater (**Figure 4**). It became apparent that when we used all twelve features, the test accuracy decreased.

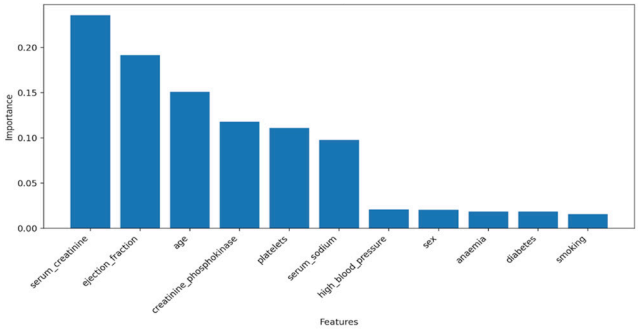


**Figure 2: Features in the order of importance after removing the most important feature (time).** Serum Creatinine, Ejection Fraction, and age were the three most important features and anemia, diabetes, and smoking were the three least important features.

https://doi.org/10.59720/24-024

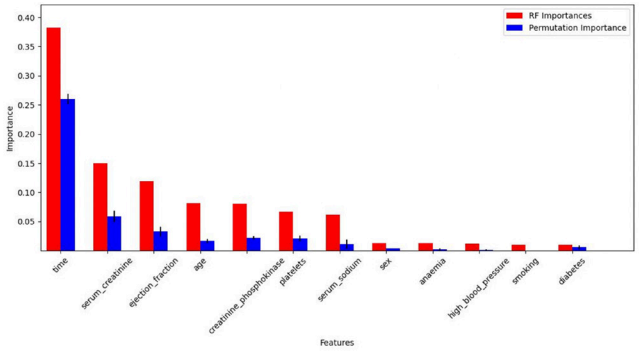


**Figure 3: Features in the order of importance of influencing mortality following heart failure.** The bar graph is showing a comparison between the rankings of feature importance determined by a random forest model and those determined by permutation importances. This comparison helps evaluate how similar or different the two methods are in identifying important features in the dataset.

When using the three most important features (time, serum creatinine and ejection fraction), the accuracy on the test data was at its highest: 79% (**Figure 5**).

To determine which were the noise inducing features, we decided to exclude one feature at a time and see the impact on accuracy. When the third feature was excluded, creatinine phosphokinase, the accuracy on the test data was the highest: 78%. The lowest accuracy on the test data came was when time (the most important feature obtained from permutation importance and RF importance) was excluded. When the second and third most important (serum creatinine and ejection fraction) were excluded, the test accuracy became 75% and 72%, respectively. Even while the importance of the feature diabetes was low, the test data accuracy dropped when excluding this feature (**Figure 6**). However, on the training data, even while excluding each individual feature one by one, the accuracy remained 100% (**Figure 7**).

For test data, it became apparent that when all 12 features were used, the test accuracy decreased. The highest accuracy seemed to be at 79% when using solely the three most important features: time, serum creatinine, and ejection fraction.

## DISCUSSION

We first started with using logistic regression because it

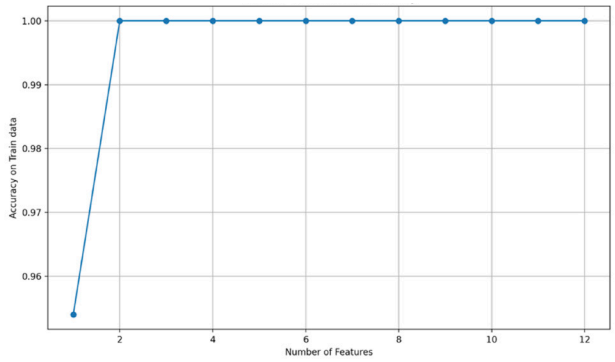


**Figure 4: The effect of number of features on the training accuracy using random forest model.** As we train on the 2, 4, and up to 12 most important features, the model's training accuracy is almost consistently at 100% once the number of important features being trained on is 2 or greater.

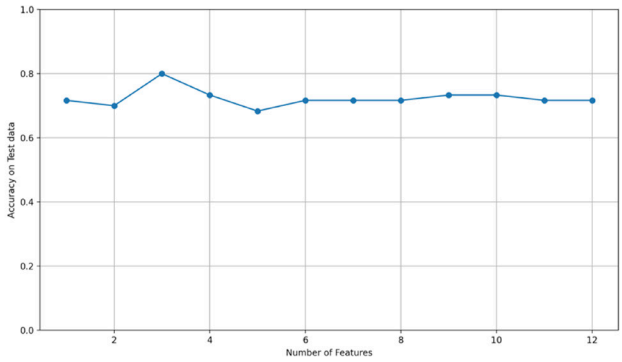


**Figure 5: The effect of the number of features on the accuracy of test data using random forest model.** In the graph above, it becomes apparent that if we use all 12 features, the test accuracy reduces. The highest accuracy seems to be at 79% when using the three most important features which are time, serum creatinine, and ejection fraction.

is a great baseline model (12,13) for binary predictions, and the objective was to predict whether the user would survive or die based on the levels of different features after an incident of heart failure. While the logistic regression model is fast and simple to train, it assumes linear relationships between the independent and dependent variables and is therefore prone to inaccurate predictions when there are missing values, outliers, and non-linear relationships.

MLP classifier, which extends the linear treatment of inputs to multiple layers, is known to be more complex and is apt for classification problems (14). Even with the best hyperparameters derived from using GridSearchCV() on different hyperparameters, the data did not fit well on the training data, as we obtained a mere ~70.3% training accuracy using this model. Similarly, the test accuracy of 58.33% also indicated that either the model did not fit the data properly or the model overfit on the data, yet the latter seems unlikely as the training accuracy was very low. The model exhibited both poor ability to fit the training data and poor generalization to unseen data. MLP is a type of feedforward artificial neural network, and neural networks require a lot of training data to perform well because they are generally used to estimate more parameters than a logistic regression, resulting in a larger model variance. Our dataset of 299 participants is rather small for MLP to work well and hence it did not perform

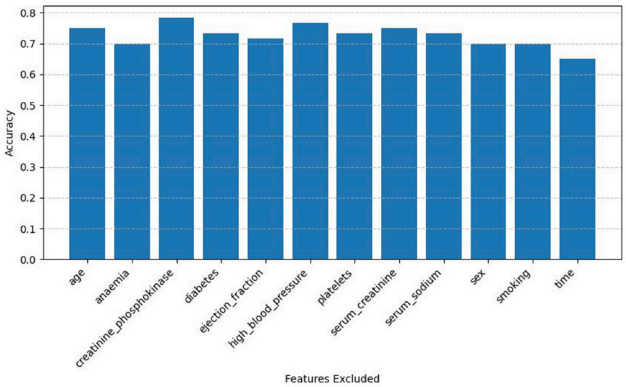


**Figure 6: Effect of removing a feature on accuracy of testing data using the random forest model.** The accuracy is the highest when creatinine phosphokinase is removed and drops the most when time is removed.

https://doi.org/10.59720/24-024

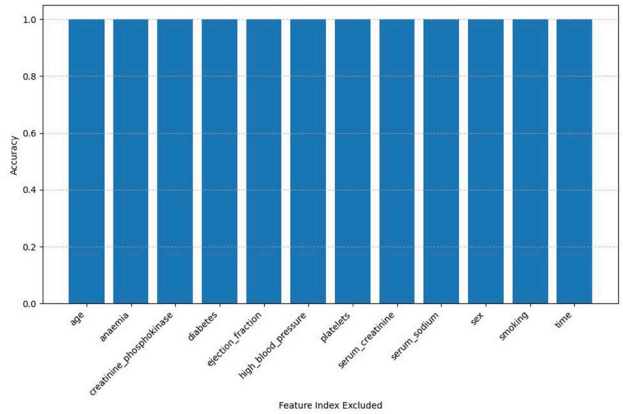


**Figure 7: Effect of removing a feature on Accuracy of Training data using random forest model.** Removing a feature, one at a time, has no impact on the training accuracy of the model.

as well as was expected.

Out of the three models we tested, RF was the best model since its training and test accuracies were the highest. However, the 73.33% on test accuracy suggests that the model did not excel in predicting the death event of an individual on unseen data due to potential overfitting. Random forests follow a hierarchical decision model. Like neural networks, some amount of real-world data is used only for training and creating decision trees. This data should be a sample of data that is a good representation of actual data. The training process then evaluates the relevance of input features based on the specific application and builds trees incrementally. These decision trees are uncomplicated to construct and are capable of handling various types of input including categorical and numerical data. They are also relatively easy to decode due to their structure. However, decision trees often lack dependability when applied to new data, that has never been seen before during training. One reason for this is their tendency to perfectly fit all samples in the training data – this is exactly what we observed with 100% accuracy on training data and 73.33% on test data.

In order to rank the importance of the features, it was necessary to use RF importances to build the decision trees based on feature importances during its training process. A high importance value suggests that the feature provided more statistical disparity in determining the output. When ranking the features from lowest to highest importance, the order observed was: time, serum creatinine, ejection fraction, age, creatinine phosphokinase, platelets, serum sodium, anemia, sex, diabetes, smoking, and high blood pressure (**Figure 1**). However, when the most important feature (time) was removed, the order changed (**Figure 2**). The order was different from when time was included, but that reinforced the differences that occurred when a discrete set of factors interacted with each other. In addition, with the removal of time, the model's training accuracy lowered.

Similarly, we used permutation importances to find the ranking of the most important features that contributed to the patient outcome and identified the following as the three most important features: time, serum creatinine, and ejection fraction. We showed that as long as we trained on the two most important features, the model's training accuracy was almost consistently at 100% whether trained on 2, 4, or up to 12 features (**Figure 4**). If we used all 12 features, the test accuracy decreased (**Figure 5**). The highest accuracy seemed to be at 79% when using the three most important features (time, serum creatinine, and ejection fraction). So, on test data if three features were used, they were enough to predict whether a person would survive or not, but additional features seemed to add noise to the prediction.

To determine which features contributed the most noise, we checked the accuracy of the model when each feature was removed in the order of the features given in the dataset. This showed that if we excluded the third feature, creatinine phosphokinase, and used the other 11 features, the accuracy on the test data is the highest at 78%. The lowest accuracy on the test data came with excluding twelfth feature, which was time (the most important feature obtained from permutation importance and RF importance). When excluding the serum creatinine and ejection fraction, the second and third most important factors, the test data accuracy became 75% and 72% respectively. With the exclusion of diabetes, the test data accuracy lowered, yet the importance of the feature diabetes was surprisingly low. This may be due to the shortage of data points in the data set. These results further demonstrate that time was the most important feature and creatinine phosphokinase was the most noise inducing feature (**Figure 6**). In the training data, when excluding each individual feature one by one, the accuracy remained 100% showing that it was overfitting and had created enough decision trees that even if only 1 feature was removed but other 11 features were present, the accuracy did not suffer (**Figure 7**).

Taken together, these results demonstrate that RF model can be used to predict patient's chances of survival based on the feature levels, but our study is not without limitations. One limitation of our study was our small sample size. This made finding a precise correlation challenging and the number of data points analyzed might not cover people with all different types of levels and backgrounds. In reality, everyone's body functions differently, so finding a precise single trend or single correlation between the features and survival is difficult. There are also many other features in the human body that may be pertinent to survival following heart failure like ferritin levels (15), calcium levels (16), potassium levels (17), etc. In this study, because there were only 12 feature levels, we did not account for all individuals' bodies as there might be other factors that outweigh the factors measured in this dataset. Similarly, other pre-existing conditions like asthma (which was not included in the data set) may also have an impact. As a continuation to this study, it would be beneficial to include more data or data sets with different features to examine how more features interact with one another and have an impact on the survival of the patient. Also, it would help to find datasets from more diverse sources.

Our findings could be used to help formulate strategies for treating patients who suffered heart failure to ensure best probability of survival. Measurements of the patient and the use of such models additionally can help tailor treatments towards a certain patient to help boost specific feature levels the patient lacks in. In addition to being used in healthcare spaces, this study may also lead to new studies that analyze how different factors in the human body contribute to one's survival following heart failure. Undeniably, there are many physiological parameters that may remain undiscovered or untested for correlation with mortality following heart

https://doi.org/10.59720/24-024

failure. By studying more features in the human body and its relationship with human mortality, it will be possible to more accurately deduce which particular features contribute most to a patient's death or survival following heart failure.

## MATERIALS AND METHODS

The data set used is the heart failure data set was obtained from Kaggle (6). Biomarker data from the heart failure prediction dataset on Kaggle were used to evaluate the prediction accuracy of three machine learning models. The heart failure prediction dataset consisted of serum and ejection fraction from 299 heart failure patients from the Faisalabad Institute of Cardiology and Allied Hospital in Faisalabad (Punjab, Pakistan) between April and December 2015 (1). The dataset included 105 women and 194 men, with ages spanning from 40 to 95 years.

The biomarkers retrieved from each participant included the sex, age, and whether the patient had hypertension along with other factors like anemia, CPK level, ejection fraction, platelets, serum creatinine levels, serum sodium levels, diabetes, smoking, time and mortality outcome.

The first step was to process the data. The first 12 columns were the features that each patient was tested on. The last column was the mortality status of the patient. The mortality was given in the form of 0 (alive) and 1 (dead). Because the original question was how the features influence the mortality of a patient who suffered from heart failure, the input data was the features and the respective values of the patients for each of those features. The output was the mortality status of the patient.

The next step was to train the data and create models. Prior to model fitting, the dataset was partitioned into training and testing sets, using an 80/20 scheme. The models used include logistic regression, MLP Classifier, and RF.

### Logistic Regression Model

For starting with logistic regression, we started with creating a logistic regression instance. The GridSearchCV() function was used to test a variety of values for each hyperparameter (C value, class weight, fit intercept, intercept scaling, max iterations, multi class, penalty, solver, tol, and warm start) to find the best fitting Logistic Regression curve for the data. We then applied the model to the training set and evaluated how the data performed on the training and testing set. We set the cross validation to five to ensure that the training data is separated into 5-folds: four for training and one for validation, testing the training data. We then found the accuracy on the training data, and the accuracy on the test data for the best set of hyperparameters found.

For logistic regression, the hyperparameter algorithm is the solver. We tested ['liblinear', 'lbfgs', 'saga', 'LogisticRegression'] and used max_iter(maximum iterations) of [100, 200, 300]. We used another hyperparameter called penalty which aims to reduce the impacts of overfitting with the values ['l1', 'l2', 'elasticnet']. Another hyperparameter was 'C' which also aims to prevent overfitting with the values [0.00001, 0.1, 3.0, 4.0]. We then used the fit_intercept parameter which tracks the dependent variable with the values of [True, False]. After that we used class weight which monitors imbalanced data with the values of [None, balanced], tol [0.0001, 0.001, 0.01], intercept scaling with [1, 2, 5], multi_class with ['ovr', 'multinomial'], and warm start with [False, True]. All these hyperparameters helped regulate the model in order to prevent overfitting.

### MLP Classifier Model

The next model used was MLP classifier, which extends the linear treatment of inputs to multiple layers and is known to be more complex and often suited for classification problems. Using the same train-test split of 80% to 20% and cross validation (a method for more thorough training) of setting five, we trained the MLP classifier model on its best hyperparameter values obtained through GridSearchCV. The hyperparameters adjusted were alpha, batch size, hidden layer size, learning rate init, and max iter. We then calculated the best score, accuracy on the data on which it was trained, and the accuracy on the data set it had never seen.

For MLP Classifier, we used the hyperparameter multiple hidden_layer_sizes with the values (100,), (50,), (100, 100, 100, 100), (50, 50, 50, 50), (50, 50, 50, 50, 50), (200,), (200, 200) where the numbers represent the number of nodes and the number of values within each parentheses represents the layers. For example, the third listed (100, 100, 100, 100) contains 4 layers, each with 100 nodes. The max_iter denotes the number of iterations through the entire training dataset. The values used were [100, 200, 300, 500, 1000]. The batch_size is the amount of data trained on at a given time. Values used were [15, 30, 50, 90]. The final hyperparameter used for the MLP classifier was learning_rate_init which is the learning rate and was tested on the values [0.001, 0.0001].

### Random Forest Model

Finally, we utilized the same train-test split, cross validation and the method on the following hyperparameters: n estimators, min samples leaf, max depth for the RF model. In order to change how the model should train the data, we changed the features n_estimators, min_samples_leaf, and max_depth with different values and had the model pick which values would be best for those hyper parameters to fit the data the best using the get_best_model function. The best performing model had the following values for the parameter: max depth: none, min_samples_leaf: 2, n_estimators: 100. For n-estimators that determine the amount of trees in the model, we had tested 20, 100, 200, 500, 750, 1000. For min_samples_leaf, we tested 1, 2, 4, 8. For max_depth, we tested solely none because the model should train to the depth it has to in order to achieve the best accuracy.

We then calculated the accuracy on the training set, and accuracy on the test set. After that we ranked the importances of each feature in relation to its impact on the death event of the patient. In order to do this, we ranked it in order of RF importances and permutation importances and plotted the findings. We also tested the order of the feature importances by training on the most important features one by one and omitting each feature one by one and finding the model's corresponding accuracies.

### Implementation

The Python packages used were sklearn, numpy, pickle, matplotlib. All code is available on GitHub: https://github.com/anyalachwani/MachineLearningProj.

**Received:** January 7, 2024
**Accepted:** June 5, 2024
**Published:** October 23, 2024

https://doi.org/10.59720/24-024

## REFERENCES


1. "Deaths from Cardiovascular Disease Surged 60% Globally over the Last 30 Years: Report." World Heart Federation, 9 Aug. 2023, www.world-heart-federation.org/news/deaths-from-cardiovascular-disease-surged-60-globally-over-the-last-30-years-report. Accessed 3 Jan. 2024.
2. Ho, K K, et al. "Survival after the Onset of Congestive heart failure in Framingham Heart Study Subjects." *Circulation*, vol. 88, no. 1, July 1993, pp. 107–115, https://doi.org/10.1161/01.CIR.88.1.107.
3. Buddeke, J., et al. "Mortality after Hospital Admission for heart failure: Improvement over Time, Equally Strong in Women as in Men." BMC Public Health, 10 Jan. 2020, https://doi.org/10.1186/s12889-019-7934-3.
4. Vidán, María T., et al. "The relationship between systolic blood pressure on admission and mortality in older patients with heart failure." *European Journal of heart failur*e, vol. 12, no. 2, 2010, pp. 148-155, https://doi.org/10.1093/eurjhf/hfp195.
5. "Treatment." NHS, www.nhs.uk/conditions/heart-failure/treatment. Accessed 3 Jan. 2024.
6. Ahmad, T., Munir, A., Bhatti, S. H., Aftab, M., and Ali Raza, M. "Survival Analysis of Heart Failure Patients: A Case Study." Dataset, PLOS, www.plos.figshare.com/articles/Survival_analysis_of_heart_failure_patients_A_case_study/5227684/1. Accessed 3 Jan. 2024.
7. "Creatine Phosphokinase Test." Mount Sinai Health Library, www.mountsinai.org/health-library/tests/creatine-phosphokinase-test. Accessed 17 Aug. 2024.
8. "Ejection Fraction." Cleveland Clinic, www.my.clevelandclinic.org/health/articles/16950-ejection-fraction. Accessed 17 Aug. 2024.
9. "Platelet Count." Cleveland Clinic, www.my.clevelandclinic.org/health/diagnostics/21782-platelet-count. Accessed 17 Aug. 2024.
10. "Crohn's Disease: Symptoms, Treatment, and More." Medical News Today, www.medicalnewstoday.com/articles/322380. Accessed 17 Aug. 2024.
11. "Hyponatremia." Cleveland Clinic, www.my.clevelandclinic.org/health/diseases/17762-hyponatremia. Accessed 17 Aug. 2024
12. Avicsebooks. "ML Part 4: Linear Classification." Medium, www.medium.com/@avicsebooks/ml-part-4-linear-classification-1d182c0b1eb3. Accessed 17 Aug. 2024.
13. "R Logistic Regression." Appsilon, www.appsilon.com/post/r-logistic-regression. Accessed 17 Aug. 2024.
14. Dutta, Sanjay. "Understanding Classification MLPS: An In-Depth Exploration." Medium, www.medium.com/@sanjay_dutta/understanding-classification-mlps-an-in-depth-exploration-22ff9eb15f9f. Accessed 17 Aug. 2024
15. Zhou, Zijing et al. "Serum ferritin and the risk of short-term mortality in critically ill patients with chronic heart failure: a retrospective cohort study." *Frontiers in physiology*, vol. 14, no. 1148891, 13 Jul. 2023, https://doi.org/10.3389/fphys.2023.1148891.
16. Jensen, Anne-Sofie Caroline, et al. "The Association Between Serum Calcium Levels and Short-Term Mortality in Patients with Chronic Heart Failure." *The American Journal of Medicine*, vol. 131, no. 12, Dec. 2018, pp. 1452-1460, https://doi.org/10.1016/j.amjmed.2018.10.006.
17. Linde, Cecilia et al. "Serum potassium and clinical outcomes in heart failure patients: results of risk calculations in 21 334 patients in the UK." *ESC heart failure*, vol. 6, no. 2 2019, pp. 280-290, https://doi.org/10.1002/ehf2.12402.