Journal of Population Therapeutics & Clinical Pharmacology

RESEARCH ARTICLE
DOI: 10.53555/jptcp.v30i19.3661

A COMPARATIVE STUDY OF PREDICTIVE SUPERVISED-MACHINE LEARNING ALGORITHMS ON CARDIOVASCULAR DISEASES (CVD)

Alina Faisal[1]*, Dr. Syed E. Ahmed[2], Mahnaz Makhdum[3], Dr. Farah Naz Makhdum[4]

[1]*Lahore University of Management Sciences; alinafaisal151@gmail.com
[2]Department of Mathematics and Statistics, Brock University.
[3]Department of Statistics, Gulberg College.
[3]NCBA&E.
[4]Lahore College for Women University.

***Correspondence:** Alina Faisal
*Email: alinafaisal151@gmail.com, ORCID: https://orcid.org/my-orcid?orcid=0009-0000-9495-8057 Tel.: (00923244442011)

Abstract
Supervised machine learning (S-ML) applications in the medical niche assist in attenuating the fatality rates as it is an arduous challenge for cardiologists to predict the patterns from the clinical data. The objective of this comparative research study is to determine the best S-ML algorithm amongst the full model, submodel I and II for the prediction of the incidence of death events due to heart failure. S-ML classification algorithms including logistic regression, ridge classifier, random forest, decision tree, and SVM with and without hyperparameter tuning using GridSearch CV were applied to a Kaggle dataset for extensive performance analysis. Feature importance techniques including random forest feature selection, and SHAP approach using the XGBoost library were used to create two submodels. The conclusion drawn from the predicted results suggested decision tree as the best algorithm due to its highest accuracy (78%, 77%, 74%) and least root-mean-square error (RMSE) (0.471, 0.483, 0.506) among the S-ML algorithms implemented on all the models. To the best of our knowledge, the implementation of linear and James-Stein shrinkage estimator strategies is the first empirical analysis of a CVD dataset. It showed submodel II as the best fit, and BIC scores showed submodel I as a better-performing model.

Keywords: heart failure; supervised machine learning; decision tree; CVD; feature importance; shrinkage estimators

I. Introduction

Globally, healthcare professionals are confronted with a staggering increase in mortality due to cardiovascular diseases (CVDs) by far the most frequent cause. CVDs are one of the serious causes of fatalities and are responsible for about 43% of all fatalities [1] (p. 88). The umbrella term CVD comprises various diseases related to the heart and blood vessels, mainly the circulatory system. CVD refers to certain conditions which include heart disease (atherosclerosis), heart failure, ischemic stroke, hemorrhagic stroke, arrhythmia, rheumatic heart disease and other conditions [2]. There are several behavioral risk factors that contribute to CVDs including an unhealthy diet, excessive stress,

smoking, alcohol intake, hypertension, high cholesterol levels, raised blood glucose and lipids, obesity and several other factors [3]. According to WHO, at least three-quarters of the world's deaths occur in low and middle-income countries due to CVDs [3]. The treatment of CVDs has incurred considerable socio-economic costs for the governments and the patients. The rate of rehospitalizations for CVDs has also accelerated relatively to other age-related diseases. Therefore, CVDs can be identified and avoided by prediction models through early intervention [4] (p. 48).

Recent research shows that roughly 18 million deaths are caused due to CVD each year and they are predominant in Asia [5]. Consequently, it is essential to utilize supervised machine learning algorithms which are a potent tool for extensive biomedical applications. Predictive supervised machine learning algorithms on CVDs play a critical role in the medical research world for better and accurate classification, which encourages cardiologists to provide satisfactory treatment to patients. S-ML algorithms analyze and predict the convoluted relationships and outcomes bet ween the features and the output response label, thus promising great clinical results [6] (p. 311). A comparison of models based on CVDs has been proposed to achieve the greatest prediction accuracy and reliability possible for several algorithms, but it is unknown whether it adds predictive information to standard parameters or not. These will require an intelligent analysis in order to promptly predict the expected outcomes of CVD patients.

S-ML measures which include regression algorithms, random forest, decision trees, support vector Machine (SVM), k-nearest neighbor (KNN), neural network, and naive bayes algorithm are of paramount importance in quantifying the degree to which the constructed model is successfully achieving its designated result and capable of accurately predicting CVDs. Moreover, machine learning combines statistics, optimization and probabilistic techniques that measure performance and permit computer programs to interpret patterns from noisy and complex data sets and algorithms. It is a data analysis technique utilizing significant factors in model development, including loss functions (L1-absolute error loss, L2-squared error loss, hinge loss), lasso and ridge regression through feature selection, tuning parameters through cross-validation, performance metrics (area under the receiver operating characteristic curve (AUC) and area under the precision-recall curve), confusion matrix and several other evaluations techniques. Prevention of CVDs is necessary for individuals at relatively high risk and prediction through machine learning algorithms may provide outcome measures of interest, which is an important aspect of medical research for accurate CVD prediction.

In the study, a comparative understanding of various models in prediction of the CVDs has been performed. The study aims to improve the prediction of CVDs through the optimization of Supervised ML techniques on the heart failure dataset, a type of CVD, obtained from the Kaggle ML repository. The preliminary phase of the work provides insight into the dataset used and will have an overview of the classification models implemented. The objective of the research is to enhance the prediction rate of heart failure. The study thoroughly investigated the performance of ML classification methods on the full model. Then it extracted significant features through the random forest feature extraction method and the SHAP method, creating two sub models using classification machine learning algorithms for heart failure prediction for better prediction results. Overall, the main purpose of the research study is to provide significant insights to healthcare workers carrying out clinical trials for improved treatment of heart disease patients.

Heart failure-related diseases have remained at the cutting edge of global deaths. Presently, the healthcare industry has undergone vast improvement in the conventional methods of predicting heart disease. Novel approaches have been widely adopted by deploying machine learning algorithms that have demonstrated efficacy and competence in the cardiology field. The machine-learning methods used to overcome the conventional methods include non-parametric regression techniques catering to a large number of predictors and automated variable selection methods [7] (pp. 323-329). With the aid of computationally powerful approaches, the interpretation and analysis of the data are more accelerated [8]. The conventional statistical approaches may not be adequate to make predictions for studying unstructured data. Moreover, the accumulation of large amounts of medicinal cardiological data has provided researchers with an unparalleled chance to predict outcomes by implementing and

developing new algorithms [1] (p. 88). Various prior researchers and studies have used ML algorithms to forecast the prediction of CVD along with feature selection using cardiological clinical datasets.
Kavitha et al. [9] applied a hybrid form of heart disease prediction model of random forest and decision trees leading to an accuracy of 88.7%. In another comparative study by Chua et. al. [10] ML models, namely logistic regression, naïve bayes (NB), k-nearest neighbor (k-NN), decision tree (DT) and support vector machine (SVM), were applied to analyze the results of the strategies for CVD prediction. Based on the performance results from [10], logistic regression was the better fit for the prediction relative to the other algorithms. According to a study by Drożdż et al. [11] machine learning techniques were used to decide the significant risk factors subjected to metabolic fatty liver disease in cardiovascular disease patients. Different machine learning approaches including multiple logistic regression classifiers and univariate feature ranking were used to develop a model for the recognition of the highest risk of CVD. The results concluded 85.11% of vulnerable patients and 79.17% of non-critical patients in the study [11].
Moreover, several articles comprised comparative analysis such as Amin et al. [12] (pp. 82-93) in their study identified the important features contributing to the objective of the research and employed seven ML algorithms along a hybrid model, Vote, including Naive Bayes and logistic regression. The best accuracy achieved was 87.41% using the significant features. In a study carried out by Dritsas et al. [13] in 2022, the authors aimed to investigate the potential CVD risk factors in individuals older than 50 years and they utilized a publicly available CVD dataset. To add more, the accuracy and recall of S-ML models were compared in individuals and showed the logistic regression classifier as the best fitting and relevant amongst naive bayes, SVM and random forest with 72.1% accuracy, recall and 78.4% area under curve [13]. In a research study conducted by Yousefi [14], several S-ML algorithms were utilized to predict heart disease on a dataset from the UCI repository evaluating based on cross-validation and feature importance scores. The results of the study [14] concluded decision trees and logistic regression as the better algorithms for the prediction of heart disease.
Feature selection techniques have also been used in machine-learning approaches for the improvement of prediction accuracy among heart patients. In a study by Ahmed et al. [15] (pp. 714-722) the researchers used feature selection algorithms which would reduce the number of features for model simplification and better prediction results. The study validated the classification models with the selected features to increase heart disease prediction performance. Another research study using the machine learning framework in predicting CVD risk was conducted by Saleh and Srinivas [16]. XGBoost gave the highest prediction accuracy amongst the four ML classifiers used to predict adulthood CVD risk. The Shapley feature importance approach was also used to find the significant features for CVD risk prediction. Additionally, Amanda et. al. [17] used three S-ML methods namely SVM, naïve bayes and decision tree, to discover correlations in coronary heart disease on the South African medical dataset that assisted in improving the prediction rate. The results reported that the probabilistic models derived by naïve bayes were promising in detecting heart disease. Hasan Mehedi et al. [18] in their research applied two feature importance techniques, the recursive method on naïve bayes, and the minimum redundancy maximum relevance method, to predict the death probability in heart failure patients. The results of the study [18] suggested ejection fraction and serum creatinine as the two factors for prediction of heart failure and decision tree classifier achieved the highest overall accuracy of 80% with both features.
The prior research demonstrated and highlighted the capability of ML algorithms in predicting CVDs for early detection of risk. However, in the proposed model in our research study, different feature-importance techniques are utilized for heart disease prediction due to death events and the results are compared through the performance metrics, root mean square error scores, BIC and linear shrinkage estimator strategy and Stein shrinkage for a more optimized conclusion.
The work in the research article is organized in the following way: Section II describes the dataset, explains the methodology of the full model, submodel I and submodel II using the S-ML classification algorithms in detail; Section III presents the experimental results with analysis. Section IV concludes the study and proposes future research directions with the challenges and limitations.

1.1 Objectives of the Study

- To determine the most efficient S-ML algorithm which predicts the incidence of death event due to heart failure with the highest predictive accuracy and least root mean square error amongst the full model, sub models I and II.
- To determine the submodel which gives the least Bayesian Information Criterion (BIC) scores as compared to the full model while predicting the death incidence due to heart failure.
- To identify which submodel performs better in terms of linear and James-Stein shrinkage estimator strategies.

II. Materials and Methods

The proposed study aims to utilize supervised machine learning algorithms to predict efficient clinical results based on data related to cardiovascular diseases. Some supervised machine-learning algorithms were implemented on a dataset to bring the objective into effect. The architecture of the methodology includes several phases: data pre-processing (data cleaning), assessment, feature selection, model training-testing, evaluating, and comparing the results of the algorithms.

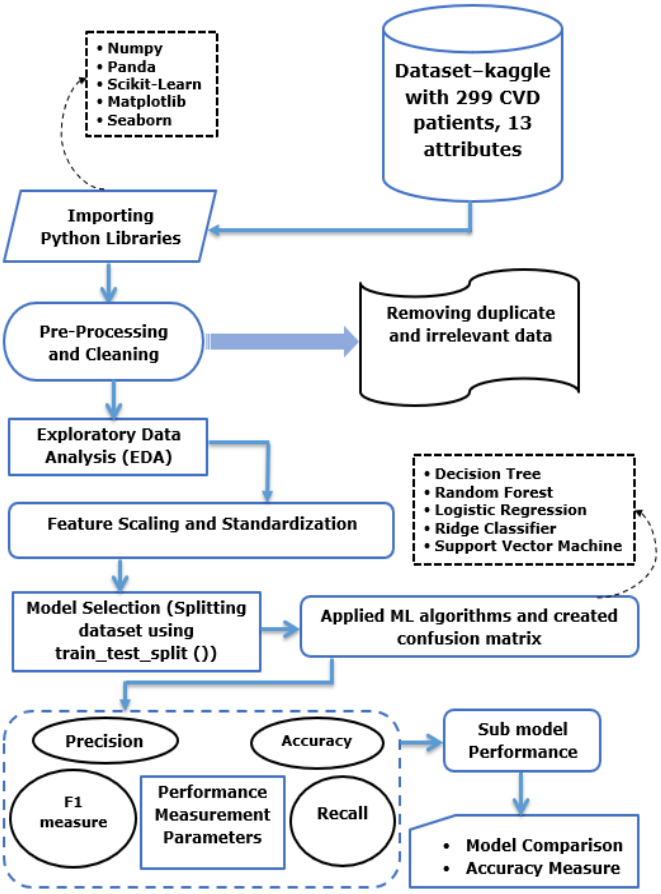


Figure 1. Model for CVD (heart failure due to death events) Prediction

2.1. Dataset Description

Kaggle dataset has been used in the study named as the "heart_failure_clinical_records_dataset" to investigate the effect of predictors on the likelihood of a death event [19]. The data contains variables: age, anaemia, creatinine_phosphokinase (cpk enzyme), diabetes, ejection_fraction, high_blood_pressure, platelets, serum_creatinine, serum_sodium, sex, smoking, time, and death_event. The sample contains 299 records of the middle-aged and elderly population and comprises 13 features, including 6 categorical (anaemia, diabetes, high_blood_pressure, sex, smoking, death_event) and 7 numerical variables (age, creatinine_phosphokinase (cpk enzyme), ejection_fraction, platelets, serum_creatinine, serum_sodium, time). The patient's medical parameters provided by the dataset are used to

recognize the patients diagnosed with heart disease, a type of CVD. The dataset is used to determine and categorize patients with a possibility of developing cardiac conditions.

2.2. Data Preprocessing

The dataset in the study is used to determine the likelihood of the death event and its association with the risk factors. In exploratory data analysis (EDA), the time column was dropped as it was irrelevant to our objective and research questions. Google Colab with Python version 3.10.12 and data wrangling techniques were used for exploratory data analysis.

2.3. Description of Features

1. Age (numerical and independent variable) - in years.
2. Anaemia (categorical and independent variable) - 1 if the patient suffers from anaemia and 0 if the patient does not. Anaemia is used to worsen cardiac conditions.
3. Creatinine Phosphokinase/CPK (numerical and independent variable) - CPK enzyme levels in the blood (mcg/L) which are elevated in cardiac conditions.
4. Diabetes (categorical and independent variable) - 1 if the patient suffers from diabetes and 0 if the patient does not. Diabetic patients are more likely to have heart disease.
5. Ejection_fraction (numerical and independent variable) - the lower the percentage of ejection fraction, the higher the risk of cardiac arrest.
6. High blood pressure/hypertension (categorical and independent variable) - 1 if a patient suffers from high blood pressure and 0 if not; hypertension increases the risk of cardiac diseases.
7. Platelets (numerical and independent variable) - Elevated platelet count (kilo platelets/mL) can block blood flow, increasing the risk of clotting and leading to strokes or heart failure.
8. Serum_creatinine (numerical and independent variable) – High serum creatinine levels (mg/dL) in the blood raise the risk of developing cardiac conditions.
9. Serum_sodium (numerical and independent variable) - Heart failure patients have low serum sodium levels (mEq/L).
10. Sex (categorical and independent variable) - 1 if the patient is male and 0 if female.
11. Smoking (categorical and independent variable) - 1 if the patient is a smoker and 0 if not.
12. Death_event (categorical and dependent variable) - 1 if the patient dies after heart failure, 0 if not.

During the data-cleaning phase, removing the duplicated data and the inaccurate records is essential, as it affects the model's generality. However, there were no missing or null data in the dataset, meaning there were no errors in the data entry process.

2.4. Exploratory Data Analysis

Seaborn count plots were used to plot the death event against the categorical features to determine the distinctions between whether the death event occurred or not. According to the count plot, smoking is not correlated to the death event due to more deaths among non-smokers. In the high blood pressure (high_bp) count plot, the proportion of patients with high blood pressure is significant when the death event occurs. Similarly, diabetes is also not correlated to the death event. However, the count plot for the anaemic patients indicates that the proportion of death versus non-death is greater in non-anaemic patients than in anaemic patients, thus indicating that anaemia may increase the likelihood of death events.

Moreover, the box plots were also used to compare the distributions of numerical features with the death event. Majority of the patients have lower ejection fractions compared to the typical ranges of 50 to 75 per cent, according to the American Health Association. The box plot also shows that the creatinine phosphokinase (cpk enzyme) levels were inflated from the normal range.

The histogram of the age distribution was plotted concerning the death event count of the patients, indicating that the death event was most likely in patients aged 55 to 60 and 68 to 75 approximately. It was followed by the kernel density plots to visualize the distribution of numerical features with the death event, each showing clear and precise distinctions between the death event and the numerical

risk factor. Furthermore, a correlation heatmap of the numerical features was also constructed using the seaborn library of Python, showing generally weak positive and negative correlations as depicted in figure 2.

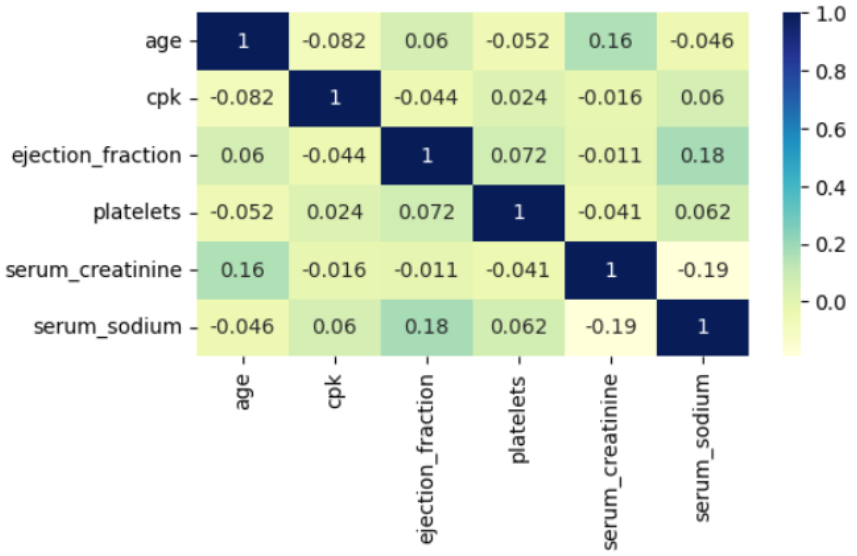


Figure 2. Correlation heatmap of the non-categorical features

2.5. Prediction through S-ML Algorithms

Various S-ML algorithms were implemented, and performance analysis was carried out to accurately predict the death event from risk factors causing heart failure. After data pre-processing, the data were shuffled to avoid overfitting the data and to generalize the model, which would then reduce the variance. It was followed by splitting the dataset into train and test data by importing the model_selection function of the sklearn library and applying sklearn's train_test_split () function. The data was split into 70% training and 30% testing data. The model was fit on the training dataset and validated on the testing dataset. The data were normalized and scaled through the standardization process, which was applied using StandardScalar() and fit_transform() functions of the sci-kit-learn library. Normalization and feature scaling was applied for a standard scale followed by the dataset's numeric columns without changing the values' ranges.

For the evaluation of the algorithms, accuracy, precision, recall, and F1-score were compared. These performance metrics were calculated by:

Precision= TP/(TP+FP)

Recall= TP/(TP+FN)

F1-score= (2*Precision*Recall)/(Precision+Recall)

Accuracy= (TP+TN)/(TP+FP+TN+FN)

TP stands for true positives, FP stands for false positives, TN stands for true negatives and FN stands for false negatives.

III. Results

3.1. S-ML Algorithms on Full Model

3.1.1. Logistic Regression

Logistic Regression is a statistical technique used in the medical domain for the risk analysis of various chronic diseases. It is a standard classification statistical approach to predict future trends/results and is most popular due to its high classification performance [20] (pp. 132-148). For the study, it was used to calculate the probability of the occurrence of the death event. The accuracy of our test data was 0.73, and the cross-validation score was 73.19 percent.

Moreover, the receiver operating characteristic (ROC) curve was also plotted to inform about the death event's practicality. It plots the sensitivity, which is the true positive rate against the false positive rate (Specificity subtracted from 1). The ROC curve or the probability curve indicates the

diagnostic ability of a binary classifier as the threshold is varied. AUC, the area under the ROC curve, which gives the probability was calculated to be 0.802. Hence, the greater the area, the greater the probability.

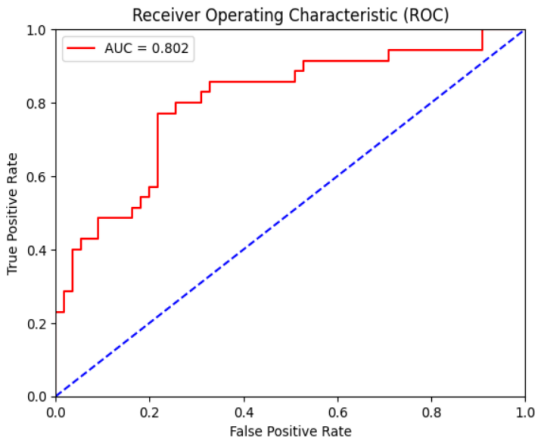


Figure 3. ROC curve

The coefficients of all the features were calculated along with their p-values, among which ejection fraction, serum-creatinine and age had the highest coefficient values indicating that they were the significant risk factors for predicting the death event due to heart failure.

3.1.2. Ridge Classifier - Regularized Linear Regression Approach

The ridge regression or L2 regularization uses a ridge classifier, a penalized linear regression model, to predict numerical data. It is a form of regularization used to prevent overfitting. It imposes a penalty on the size of the classifier to reduce the standard error by including some bias in the regression estimates. The alpha parameter which is the penalty term controls the amount of shrinkage. The ridge classification takes the square of the coefficients. With a greater alpha value, the shrinkage is more and the higher is the smoothness constraint.

In Figure 4 the ridge coefficients are plotted as the function of regularization as the dataset did not include an exorbitantly high number of features to predict heart failure. Toward the end of the plot, alpha (tuning or penalty term) tends to zero but does not reach zero showing the effect of collinearity.

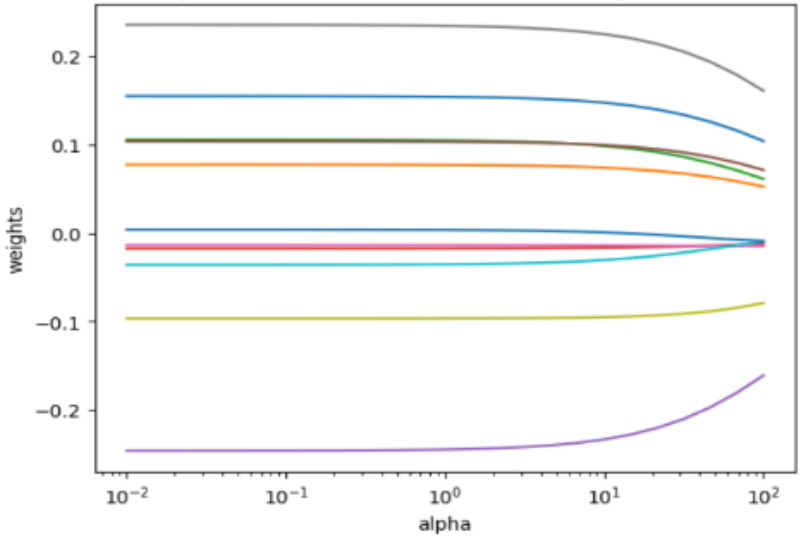


Figure 4. Ridge coefficients as a function of the regularization

The maximum accuracy of predicting death events due to heart failure by implementing the ridge regression classifier was 70%. Similar to logistic regression, the coefficients of all features were calculated, showing ejection fraction, serum-creatinine and age having the highest coefficient values. It

was interpreted that these three features were the significant risk factors for predicting death events due to heart failure.

3.1.3. Decision Tree

A decision tree classifier represents a process incorporating a tree-like structure built by top to down divide and conquer method [21] (pp. 1561-1572). These trees have a promising effect in terms of occurrence implications, optimization strategy and performance. They are used to manage large data and the enlarged data requires little data preparation. Decision trees are widely used due to their efficiency, reliability, and clarity.

The non-parametric S-ML method used the decision tree classifier to optimize the decision tree performance by using criterion entropy. Impurity was used to split the decision tree nodes and such homogeneity measure of the node labels was applied using sci-kit learn's information gain and gini index. However, the information gain criterion parameter was utilized to measure the impurity. The node was split in such a way as to give the maximum information gain. Through the selection measure attribute of the default tree, 72% accuracy was achieved, and the other performance metrics were calculated as well.

Pre-pruning of the decision tree classifier was carried out using different max depths. Max depth 3 gave the best accuracy of 78% on the test data set. Other metrics calculated include precision, recall and f1-score. Maximum depth 3 decision tree was the best classifier of all the decision trees, and it controlled the overall general complexity of the tree.

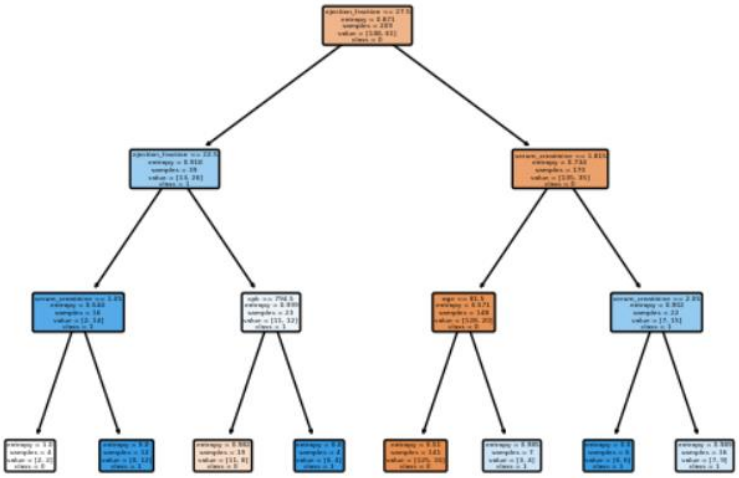


Figure 5. Decision Tree of maximum depth 3

The importance of the input features was calculated, and the results indicated ejection fraction, platelets and serum creatinine with the highest importance scores for predicting the death event due to heart failure through the decision trees algorithm.

3.1.3. Random Forest an Ensemble Method

Random forests create multiple decision trees by utilizing several features and data. Predictions are made by calculating the prediction for each decision tree and then selecting the most significant one. They split on a random subset of features and only consider a small number of features from the dataset to minimize the variance [22] (p. 8352). The high performance of the decision trees is leveraged through the random forest model. Along with its quick prediction property, all the decision trees in the random forest are averaged to produce results which have low bias and moderate variances.

For the model optimization, the random model was framed to judge the best number of estimators and the maximum depth. The highest accuracy was 74% with max depth 6 and 114 estimators. Other performance metrics and the confusion matrix were also calculated to compare the results.

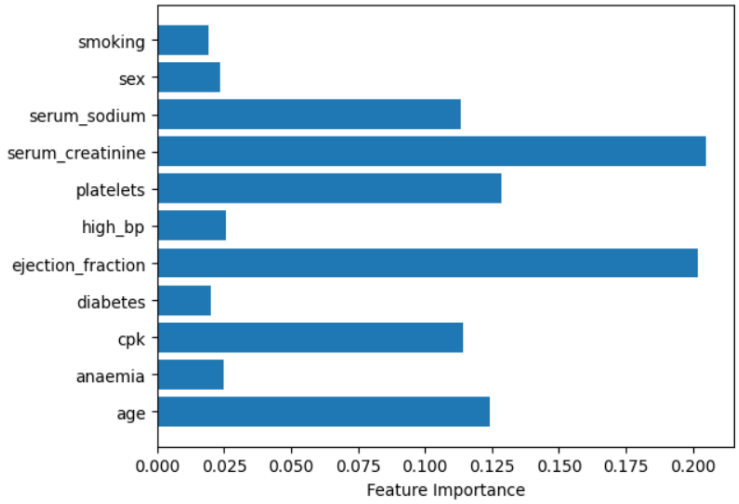


Figure 6. Random Forest feature importance scores

Similarly, to the decision tree algorithm, the random forest model also predicts serum creatinine, ejection fraction, platelets, and age as significant features due to their highest importance scores in predicting the label outputs occurrence due to heart failure, a cardiovascular disease as shown in figure 6.

3.1.4. Support Vector Machine (SVM)

SVM a supervised algorithm generates and finds the marginal hyperplane to minimize the dataset error and divide the classes in a multidimensional space. The decision plane, identified as the hyperplane, separates the data points and the points closest to the hyperplane are known as the support vectors which are essential for the construction of the classifier. The SVM classifier using the hyperplane separates the data points with the greatest margin. Firstly, the SVM model was implemented without hyperparameter tuning and the accuracy came out to be 71%. Secondly, the SVM model was implemented with hyperparameter tuning for suboptimal results without minimizing the loss function. The accuracy and precision scores were increased as compared to the results of SVM implementation without hyperparameter tuning. Gridsearch CV was applied for hyperparameter tuning and the best parameter and kernel in the model was 'C': 10, 'gamma': 0.01, and 'kernel': 'rbf'. Ejection fraction, serum creatinine and age were the significant features for predicting the death event due to heart failure (a type of CVD) as calculated by applying the feature importance.

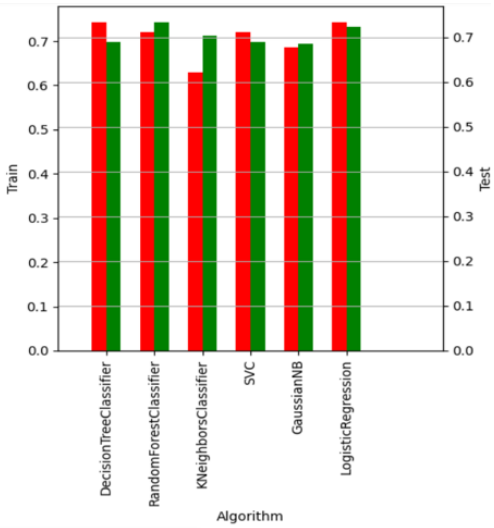


Figure 7. Model for the comparison of test/train accuracy

The green bars in figure 7 define the training score and the red bars depict the test scores. Overall, the results suggested that decision tree classifiers with max-depth 3 gave the highest accuracy on the test data followed by the random forest algorithm with max-depth 6. There was little variation in the performance of logistic regression, random forest, and decision tree classifier algorithms in the train/test accuracy.

Table 1. Full Model Performance Metrics

Model	Accuracy	F1-Score	Recall	Precision	Alpha Level	Max-depth	Estimators
Logistic Regression	0.733	0.520	0.371	0.867	-	-	-
Ridge Classifier	0.700	0.426	0.286	0.833	1.470	-	-
Decision Tree	0.711	0.594	0.543	0.655	-	-	-
Decision Tree	0.778	0.643	0.514	0.857	-	3	-
Random Forest	0.744	0.705	0.697	0.754		6	114
SVM w/o Hyper-parameter tuning	0.711	0.536	0.429	0.714	-	-	-
SVM with Hyper-parameter tuning	0.733	0.520	0.371	0.867	-	-	-

The Bayesian information criterion (BIC) score, a model selection statistical approach for the evaluation of the models, was calculated for the full model. A penalty term for the number of parameters in the full model was introduced and the BIC score was 142.789. See Appendix A for calculations.

3.2. Submodel I

In submodel I, the random forest feature importance method, and the significant coefficients of all the features along with their p-values in logistic regression were considered to extract the essential risk factors. The four features with the highest feature importance scores included serum creatinine, ejection fraction, age and platelets. A new model known as the submodel I was created by dropping the non-significant risk factors of the death event due to heart failure from the dataset. The submodel I was extracted from the full model's training and testing dataset. The insignificant features were dropped from both the training and the testing data of the full model (70:30) to keep the results consistent. Random sampling was applied to both the training and testing data of submodel I. X_train1, X_test1, y_train1 and y_test1 were used for the implementation of S-ML algorithms and the results were contrasted with the full model. Feature Scaling was applied in a similar way as on the full model. Similar machine learning algorithms were implemented as on the full model and the results were analyzed. The accuracy values of the submodel I were relatively greater than the accuracy values of the full model except for SVM with hyper-parameter tuning algorithm and decision tree with max-depth 3 which had less accuracy value in the submodel I than the full model. Overall, the decision tree with max-depth 3 had the highest accuracy followed by random forest with max-depth 7 among the machine learning algorithms implemented in submodel I for the prediction of death events due to heart failure, a type of CVD.

Table 2. Submodel I Performance Metrics

Model	Accuracy	F1-Score	Re-call	Precision	Al-pha Level	Max-depth	Estimators
Logistic Regres-sion	0.744	0.566	0.429	0.833	-	-	-
Ridge Classifier	0.733	0.520	0.371	0.867	1.470	-	-
Decision Tree	0.733	0.636	0.600	0.677	-	-	-
Decision Tree	0.767	0.687	0.657	0.719	-	3	-
Random Forest	0.744	0.715	0.708	0.740		7	102
SVM w/o Hy-per-parameter tuning	0.733	0.520	0.371	0.867	-	-	-
SVM with Hy-per-parameter tuning	0.644	0.448	0.371	0.565	-	-	-

The coefficients and the p-values were calculated through logistic regression indicating that age, ejection fraction and serum creatinine were the significant factors as their p-values were less than .05. These risk factors were the significant features contributing to the death event due to heart failure, a type of CVD. Moreover, the BIC score calculated for submodel I was 118.988. See Appendix A for calculations.

3.3. Submodel II

However, for submodel II, the SHAP (SHapley Additive explanations) was adopted for feature selection. The explain ability, agnostic tool was used for the model prediction as it computes the global interpretation. It is an "open-Source project maintained by Scott Lundberg, a Microsoft researcher, and it's available under MIT License" [23].

An explainable AI, post hoc approach was applied using the XGBoost, an efficient gradient boosting library. The SHAP library was installed using pip. It was applied after the training phase of the model. The trained model was then inserted into shap.Explainer(model). The SHAP value of each feature was plotted in the bar graph to determine the importance of the features. Using the SHAP bar plot, we made the global feature importance plot as shown in figure 8 according to which serum creatinine, creatinine phosphokinase (cpk enzyme), age and platelets were the most significant features. See Appendix A for the method.

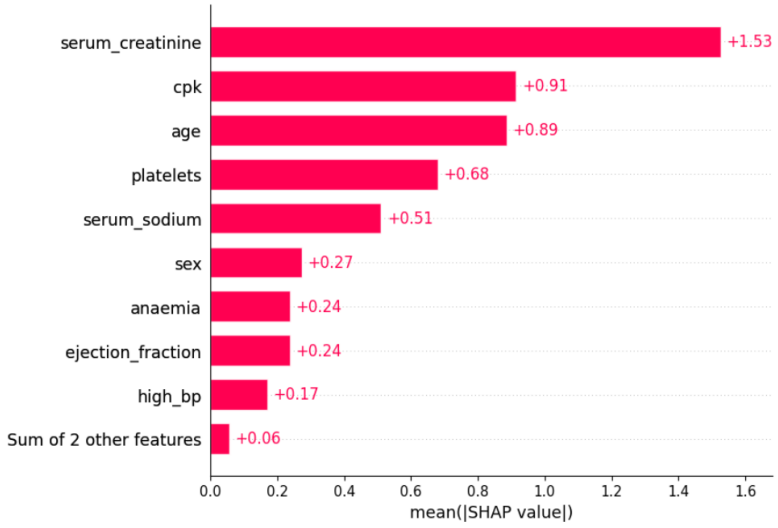
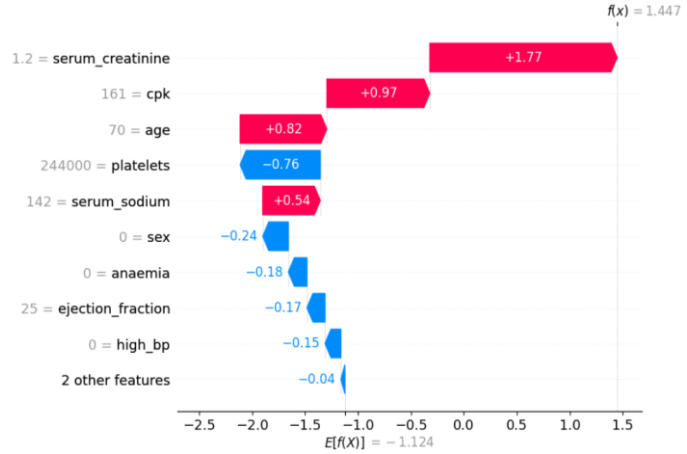


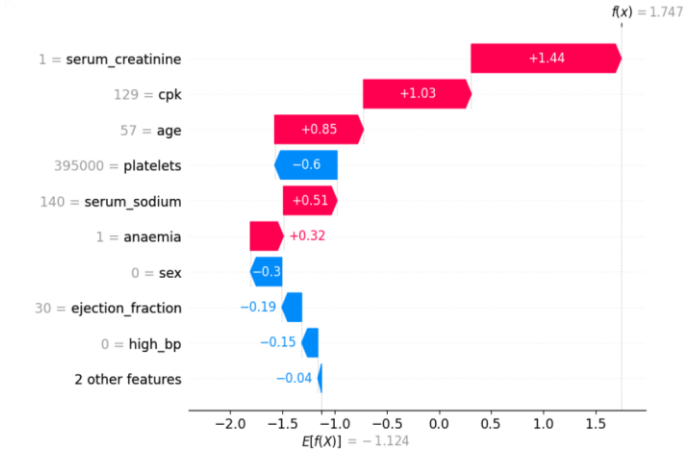
Figure 8. Global Feature Importance

3.3.1. SHAP Waterfall Plots

Figure 9 (a) shows the contribution of all risk factors to the first prediction and figure 9 (b) to the second prediction. E[f(X)] is the prediction in both plots. The red bar highlights the increase in the value of the prediction of every feature whereas the blue bar shows the decrease in the value of prediction. To generate the model, each feature is replaced by its average value using SHAP.



(a)



(b)

Figure 9. (a) Waterfall plot 1; (b) Waterfall plot 2

The overall impact of each feature shows that serum creatinine, creatinine phosphokinase (cpk enzyme), age and platelets had the highest feature importance scores using the SHAP feature selection method. These four features were used to create a Submodel II, by dropping the non-significant features from the dataset. The submodel II was extracted from the full model which was split into a 70:30 ratio using the sklearn libraries. The insignificant features were dropped from both the training and the testing data of the full model (70:30) to keep the results consistent. Random sampling was applied to both training and testing data of the submodel II. X_train2, X_test2, y_train2 and y_test2 were used for the implementation of S-ML algorithms and the results were contrasted with the full model. The data were normalized again using feature scaling. Generally, the accuracy values of the submodel II were lesser than the accuracy values of the full model except random forest algorithm which had the same accuracy value as of the full model. Overall, the decision tree with max depth 3 and random forest with max depth 7 had the highest accuracy among the S-ML algorithms implemented in submodel II for the prediction of death events due to heart failure, a type of CVD.

Table 3. Submodel II Performance Metrics

Model	Accuracy	F1-Score	Recall	Precision	Alpha Level	Max-depth	Estima-tors
Logistic Regres-sion	0.667	0.286	0.171	0.857	-	-	-
Ridge Classifier	0.656	0.244	0.143	0.833	1.470	-	-
Decision Tree	0.711	0.581	0.514	0.667	-	-	-
Decision Tree	0.744	0.610	0.514	0.750	-	3	-
Random Forest	0.744	0.678	0.677	0.820		7	114
SVM w/o Hyper-parameter tuning	0.644	0.158	0.086	1.00	-	-	-
SVM with Hy-per-parameter tuning	0.667	0.318	0.200	0.778	-	-	-

The coefficients and the p-values were calculated through logistic regression indicating that age and serum creatinine were the significant factors as their p-values were less than .05. These risk factors were the significant features contributing to the death event due to heart failure, a type of CVD. Additionally, the BIC score calculated for submodel II was 125.015. See Appendix A for calculations.

3.4. Linear and James-Stein Shrinkage Estimation Strategies

Shrinkage estimation method utilizes the full model and the submodel estimates. They are combined in such a way that shrinks the least square estimates toward the submodel estimates. The regression coefficients estimator vector ($\hat{\beta}_{full}$) is shrunk toward the structured submodel estimator ($\hat{\beta}_{sub}$). In this research study, the performance of each combination of the submodel and the full model was evaluated through the estimation of the prediction errors dependent on cross-validation. The model was fitted using the training data, which was used to make predictions on the test data. The prediction errors were calculated using the MSE scores and the linear shrinkage technique was applied to both submodel I and submodel II [24]. The linear shrinkage estimator is defined in the following equation:

$\hat{\beta}^{LS} = c\,\hat{\beta}_{sub} + (1-c)\hat{\beta}_{full}$, where

$\hat{\beta}_{sub}$ is the submodel estimator, $\hat{\beta}_{full}$ is the estimator based on full model, and, $\hat{\beta}^{LS}$ denotes linear shrinkage estimator; finally, c (0,1) is the shrinkage constant whose optimal value may be determined through cross-validation method, among others and MSE is the mean squared error.

The optimal parameter score calculated for submodel I was 0.766 and for submodel II was 0.732. The MSE calculated from the linear shrinkage on the submodel I was 1.264 and for the submodel II it was 0.923. According to the results, submodel II had less MSE score, therefore it was a better performing model.

A biased James-Stein estimator was introduced, and it used both the full model and the submodel. Similarly, the JS estimator was applied as it shrinks the estimates [25] (pp. 1-107). The likelihood-ratio test (LRT) was used, and chi-square value was extracted from the test statistic, $\mathcal{L}$ [26] (pp. 396-420). According to Wilk's theorem [27], LRT uses chi-squared distribution, which is the sum of the squared values of independent normal random variables. The Stein-type estimator results for both submodels were calculated using the following equation:

$\hat{\beta}^{JS} = \hat{\beta}_{sub} + \left(\hat{\beta}_{full} - \hat{\beta}_{sub}\right)\left(1 - (p-2)\frac{1}{\mathcal{L}}\right),\ p \geq 3$

$\mathcal{L}$ = -2 [log (null results) – log (alt results)], where

p is the degree of freedom (no. of insignificant features in each submodel), $\mathcal{L}$ is likelihood-ratio test statistic, null results are the null hypothesis results, and alt results include the values from the alternative hypothesis as discussed in Appendix B.
For submodel I, there were 8 insignificant features and for submodel II there were 9 insignificant features. The MSE value calculated from JS shrinkage method on submodel I was 1.034 whereas the MSE value for submodel II was 0.761 suggesting submodel II as a better fitting model too. See Appendix B for the explanation regarding calculations for both shrinkage estimator strategies.

3.5. Root Mean Square Error (RMSE)

RMSE which is the variance of the residuals, fits the model to the data and measures the closeness of the predicted value of the model to its actual value. It is an absolute measure of fit and estimates the accuracy of the predicted value of the forecasting model versus the actual values. The measure of the standard deviation of the residuals is used to predict the best algorithm among the algorithms trained on the dataset with the lowest RMSE value giving the least error and indicating the best fit. The RMSE values were calculated for each full model and submodel I and II. The results reported in table 4 highlight that the decision tree with max depth 3 has the least RMSE value in the full model, submodel I and submodel II. Overall, the results suggest that the decision tree with max depth 3 is the best prediction model for the death event due to heart failure among all algorithms.

Table 4. Root Mean Square Error (RMSE)

Model	Full Model	Submodel I	Submodel II
Logistic Regression	0.516	0.506	0.577
Ridge Classifier	0.548	0.516	0.587
Decision Tree	0.537	0.516	0.537
Decision Tree max-depth 3	0.471	0.483	0.506
Random Forest	0.537	0.527	0.527
SVM with hyperparameter tuning	0.516	0.548	0.596
SVM w/o hyperparameter tuning	0.537	0.596	0.578
Linear shrinkage estimator			
James-Stein shrinkage estimator	-	1.124	0.961
	-	1.017	0.872

4. Discussion

A comparative CVD (heart failure prediction) research study was conducted using the most common S-ML algorithms in healthcare: logistic regression, ridge classifier, random forest, decision tree, SVM with and without hyperparameter tuning. GridSearch CV was applied for the hyper-parameter tuning of the algorithms to get optimized results. The performance metrics used in the research study included accuracy, F1-score, precision, recall and RMSE scores. Two types of feature importance techniques were applied on the train-test split (70:30) of the full model: random forest feature importance and the SHAP approach. The study compromised two submodels formed through the two feature importance methods. The results predicted the likelihood of death event due to heart failure, a type of cardiovascular disease. The results of the research study using RMSE scores and the performance metrics of the classification S-ML algorithms implemented on the full model and both the submodels suggested the decision tree with max-depth 3 to be an optimal algorithm, as it had the highest accuracy (78%, 77%, 74%) and the least RMSE score (0.471, 0.483, 0.506) among all the supervised algorithms implemented for the three models. Linear shrinkage estimation strategy, James-Stein estimator and BIC scores were compared for both submodels. The results of the shrinkage estimator suggested that submodel II was a better-performing model due to its less MSE score as compared to submodel I. However, among the BIC scores calculated on all the models, submodel I had the least BIC score; therefore, it was a better-fitting and an optimal model. The change in the BIC value for submodel I

was $\Delta BIC = 23.801$ indicating that the evidence favouring the submodels vs. the full model was very strong and significant.

4.1. Limitations and Challenges

The value of machine learning algorithms is growing because of the symbiotic rise of CVDs, as these methods are better in the prediction of death events due to heart failure and for taking efficient clinical decisions. Albeit S-ML can be a valuable tool for CVD prediction, it should not merely be depended upon due to its ethical and transparency issues. Despite its advantages, the generalization of the prediction results has specific restrictions. One of the significant issues confronted by the myriad researchers includes dataset production, as seen in the research study which does not include the socio-economic status of the patients contributing to heart failure. Further research is solicited for improving the quality and validation of the findings. Moreover, the study was grounded on a single dataset of a certain type of CVD and comprised a limited set of factors for heart failure. The sample size for the predictive analysis was also small, and the pooled results could be potentially biased too. Due to a small sample size, the count plot of death events with the categorical feature of explicitly smoking, it was determined that there were more deaths in patients who were non-smokers hence contradictory to the past research. In addition, a significant challenge includes the probability of error in diagnosis and the prediction process using ML algorithms which can be catered to by doing clinical trials. Considering the constraints and challenges, it is recommended to conduct further research on various types of CVDs for an extensive understanding of the potential of S-ML algorithms in the advancement of CVD prediction through intelligent applications.

4.2. Future research

For future research studies, unstructured data such as images, websites, text files, and documents may also be included to develop concrete and better practices in the clinical systems. The subsequent level in the development of ML methods in the CVDs context is to apply them more widely to different types of CVDs and on divergent populations with CVD-related diseases. The urgent need of the time is to adopt advanced tools and techniques in CVD healthcare to solve problems to make more accurate decisions as the dataset is growing in volume. Thus, a potential handicap may be massive CVD data and special types of devices are needed to process such quality and quantity of data. All the methods and models have different purposes so one may outperform others. Additionally, it can be beneficial to compare the performance of various types of shrinkage estimators including the James-Stein shrinkage estimator on each submodel. The evaluation of the shrinkage estimators can be done on simulation data first before implementing it on a real dataset. Furthermore, regularization can be performed to improve the performance, avoid overfitting, and decrease the time required for training the model. For optimization and fine-tuning, the lib linear solver can be used with regularized regression, including the L1 penalty. The penalty applied to the coefficients would then further multiply all the regressors and shrink the coefficients of the input variables that do not contribute much to the predictor variable. Ideally, these techniques will be more effective on a large dataset with either more features than data points or vice versa. Lasso regression can be used to decrease the complexity of the model. Various decision tree algorithms including extreme gradient boost and its predecessors including adaptive boosting can be applied which exceeds the performance of other decision tree approaches, hence giving better accuracy results. Future research directions should have generalizability and robustness of the results by the algorithms for an improvised understanding of the study's findings. Conclusively, it is pertinent to validate the accuracies of the ML algorithms with high dimensional data on CVD population for future research prospects. The research study can be expanded to actual systems utilizing deep learning approaches where test reports of patients with CVD can be uploaded as images for future predictive and clinical analysis.

Data Availability Statement: A real-world Kaggle dataset was used for the study. The name of the dataset is Heart Failure Prediction dataset. The link to the dataset: https://www.kaggle.com/datasets/andrewmvd/heart-failure-clinical-data?resource=download

Acknowledgments: I sincerely thank my grandfather, Makhdum Tariq Salim and my mother Gulnaz Makhdum for their encouragement, unwavering support and wisdom which has been a constant source of inspiration for me. I would also like to extend my sincere gratitude to Professor Ejaz Ahmed for his invaluable guidance and mentorship during the research and for writing this article.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results. This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Appendix A

1. The BIC scores were calculated using the following code for each submodel as follows:

```
X_test = sm.add_constant(X_test)

# Create the model
model = sm.OLS(y_test, X_test)

# Fit the model
results = model.fit()
bic = results.bic
print ('BIC: ', bic)
```

2. For SHAP method the following code was applied as follows:

```
import xgboost
import shap
model = xgboost.XGBClassifier().fit(X_train, y_train)
predict = model.predict(X_test)
explainer = shap.Explainer(model)
shap_values = explainer(test_df[cols])
shap.plots.bar(shap_values)
shap.plots.waterfall(shap_values[1])
shap.plots.waterfall(shap_values[0])
```

Appendix B

The shrinkage estimator values were calculated using the equations mentioned in section 3.4. The optimal parameter or shrinking factor was calculated using the cross-validation scores for both the submodels through the python code as follows:

```
from sklearn.model_selection import cross_val_score
accuracies=cross_val_score(estimator=model,X=X_train1, y=y_train1, cv=10)
print ("Accuracy:", format(accuracies.mean()*100))
```

The value of the linear shrinkage estimator was calculated using the equation in section 3.4. The python code used for the calculations was:

```
Beta_hat=Beta_hat_full+Beta_hat_sub1
Beta_hat_T=np.transpose(Beta_hat)
print (Beta_hat_T)
```

Beta_hat_full was calculated by the multiplication of (1-c) with the full model coefficients. This was applied to both submodel I and submodel II.

```
print(model.coef_, model.intercept_)
Beta_hat_full=0.234*model.coef_
print(Beta_hat_full)
```

Similarly, Beta_hat_sub1 was calculated in the same way. However, the value of c (optimal parameter) was value used for the multiplication with the submodel I coefficients. The MSE scores for both submodel I and II were calculated to compare the results.

```
y_predicted=np.dot(X_test,Beta_hat_T)
y_predd=np.asarray(y_predicted)
y_test_list=np.array(y_test)
mse=np.mean((y_predd-y_test_list)**2)
print(mse)
```

Furthermore, for the James-Stein estimation the equation mentioned in section 3.4. was used. This method was applied to control the size of the bias. The degrees of freedom for both the submodels were calculated by subtracting the full model features from the significant features used in each submodel. $\mathcal{L}$ is the test statistic to test H_0: $\alpha = 0$ and H_A: $\alpha \neq 0$ [26]. Under H_0, the chi-squared distribution is followed by the test statistic. After the data were split into training and testing data and further into dependent and independent variables the chi-squared value was used from the likelihood ratio test applied in python using the following code:

```
X_test1 = sm.add_constant(X_test1)
X_test = sm.add_constant(X_test)

##null model
null_model=sm.OLS(y_test1, X_test1)
null_results= null_model.fit()

##alternatemodel
alt_model=sm.OLS(y_test1, X_test)
alt_results= alt_model.fit()
```

Following this, likelihood ratio test of null model to alternative model was implemented using python likelihood ratio chi-squared test statistic library:

```
likelihood_ratio = -2*(null_results.llf - alt_results.llf)
print(likelihood_ratio)
```

The likelihood ratio was then used in place of L in equation 3.4. The p used in equation 3.4. was calculated by subtracting the number of predictor variables used in full model and each submodel. The MSE values were calculated through the similar method used for linear shrinkage estimator. The whole procedure was repeated for submodel II.

References

1. Bhatt, C. M.; Patel, P.; Ghetia, T.; Mazzeo, P.L. Effective Heart Disease Prediction Using Machine Learning Techniques. *Algorithms* **2023**, *16(2)*, 88. DOI: https://doi.org/10.3390/a16020088.
2. Heart Attack and Stroke Symptoms. Available online: https://www.heart.org/en/health-topics/consumer-healthcare/what-is-cardiovascular-disease (accessed on 10 June 2023).
3. World Health Organization. Available online: https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1 (accessed on 15 June 2023).

4. Kerexeta, J.; Larburu, N.; Escolar, V.; Lozano-Bahamonde, A.; Macía, I.; Iraola, A. B.; Graña, M. Prediction and Analysis of Heart Failure Decompensation Events Based on Telemonitored Data and Artificial Intelligence Methods. *Journal of Cardiovascular Development and Disease* **2023**. *10(2)*, 48. DOI: https://doi.org/10.3390/jcdd10020048.
5. Arunachalam, S.K.; Rekha, R. A novel approach for cardiovascular disease prediction using machine learning algorithms. *Concurrency and Computation Practice and Experience* **2022**. *34 (19).* DOI: https://doi.org/10.1002/cpe.7027.
6. Shou, B.L.; Chatterjee, D.; Russel, J. W.; Zhou, A.L.; Florissi, I, S.; Lewis, T.; Verma, A.; Benharash, P.; and Choi. C. W. Pre-operative Machine Learning for Heart Transplant Patients Bridged with Temporary Mechanical Circulatory Support. *Journal of Cardiovascular Development and Disease* **2022**, *9(9)*, 311. DOI: https://doi.org/10.3390/jcdd9090311.
7. Chen, X.; Ishwaran, H. Random forests for genomic data analysis. *Genomics* **2012**, *99(6)*, 323-329. DOI: https://doi.org/10.1016/j.ygeno.2012.04.003.
8. Faizal, A.S.M.; Thevarajah, T.M.; Khor, S. M.; Chang, S. A review of risk prediction models in cardiovascular disease: conventional approach vs. artificial intelligent approach. *Computer Methods and Programs in Biomedicine* **2021**, *207*. DOI: https://doi.org/10.1016/j.cmpb.2021.106190.
9. Kavitha, M.; Gnaneswar, G.; Dinesh, R.; Sai, Y.R.; Suraj, R.S. Heart Disease Prediction using Hybrid machine Learning Model. In Proceedings of 6th International Conference on Inventive Computation Technologies (ICICT), Coimbatore, India, Date of Conference (20-22 January 2021). DOI: 10.1109/ICICT50816.2021.9358597.
10. Chua, S.; Sia, V.; Nohuddin, P. N. E. Comparing Machine Learning Models for Heart Disease Prediction. In Proceedings of IEEE International Conference on Artificial Intelligence in Engineering and Technology (IICAIET), Kota Kinabalu, Malaysia, Date of Conference (13-15 September 2022). DOI: 10.1109/IICAIET55139.2022.9936861.
11. Drożdż, K.; Nabrdalik, K.; Kwiendacz, H.; Hendel, M.; Olejarz, A.; Tomasik, A.; Bartman, W.; Nalepa, J.; Gumprecht, J.; Lip, G. Y. H. Risk factors for cardiovascular disease in patients with metabolic-associated fatty liver disease: a machine learning approach. *Cardiovascular Diabetology* **2022**, *21*. DOI: https://doi.org/10.1186/s12933-022-01672-9.
12. Amin, M.S.; Chiam, Y.K.; Varathan, K. D. Identification of significant features and data mining techniques in predicting heart disease. *Telematics and Informatics* **2019**, *36*, 82-93. DOI: https://doi.org/10.1016/j.tele.2018.11.007.
13. Dritsas, E.; Alexiou, S.; Moustakas, K. Cardiovascular Disease Risk Prediction with Supervised Machine Learning Techniques. In Proceedings of 8th International Conference on Information and Communication Technologies for Ageing Well and e-Health, Greece, Date of Conference (April 2022). DOI: 315-321. 10.5220/0011088300003188.
14. Yousefi S. Comparison of the performance of machine learning algorithms in predicting heart disease. Frontier Health Informatics **2021**, *10*. DOI: https://doi.org/10.30699/fhi.v10i1.349.
15. Ahmed, H.; Younis, E. M. G.; Hendawi, A.; Ali, A.A. Heart disease identification from patients' social posts, machine learning solution on Spark. *Future Generation Computer Systems* **2020**, *111*, 714-722. DOI: https://doi.org/10.1016/j.future.2019.09.056.
16. Salah, H.; Srinivas, S. Explainable machine learning framework for predicting long-term cardiovascular disease risk among adolescents. *Scientific Reports* **2022**, *12*. DOI: https://doi.org/10.1038/s41598-022-25933-5.
17. Gonsalves, A. H.; Thabtah, F.; Mohammad, R.M.A.; Singh, G. Prediction of Coronary Heart Disease using Machine Learning: An Experimental Analysis. In Proceedings of the 3rd International Conference on Deep Learning Technologies, Association for Computing Machinery, New York, USA, Date of Conference (July 2019). DOI: https://doi.org/10.1145/3342999.3343015.
18. Hasan, M.A.M.; Shin, J.; Das, U.; Srizon, A. Y. Identifying Prognostic Features for Predicting Heart Failure by Using Machine Learning Algorithm. In Proceedings of the 11th International Conference on Biomedical Engineering and Technology, Association for Computing Machinery USA, Date of Conference (March 2021). DOI: https://doi.org/10.1145/3460238.3460245.

19. Kaggle. Available online: https://www.kaggle.com/datasets/andrewmvd/heart-failure-clinical-data?resource=download&select=heart_failure_clinical_records_dataset.csv (accessed on 7 June 2023).
20. Ramesh, TR.; Lilhore, U.K.; M, P.; Simaiya, S.; Kaur, A.; Hamdi, M. PREDICTIVE ANALYSIS OF HEART DISEASES WITH MACHINE LEARNING APPROACHES. *Malaysian Journal of Computer Science* **2022**, 132-148. DOI: https://doi.org/10.22452/mjcs.sp2022no1.10.
21. Trabay, D.; Gharibi, W.; Abd-Elhafiez, W. M. Effective Models for Predicting Heart Disease Using Machine Learning Techniques – A Comparative Study. *Information Sciences Letters* **2023**, *12*, 1561-1572. DOI: 10.18576/isl/120505.
22. Reddy, K.V.V.; Elamvazuthi, I.; Aziz, A.A.; Paramasivam, S.; Chua, H.N.; Pranavanand, S. Heart Disease Risk Prediction Using Machine Learning Classifiers with Attribute Evaluators. *Applied Sciences* **2021**, *11*, 8352. DOI: https://doi.org/10.3390/app11188352.
23. Silva, M. P. Feature Selection using SHAP: An Explainable AI approach. Undergraduate Thesis, University of Brasília, Brazil, 2021. DOI: https://doi.org/10.1111/bjop.1226.
24. Ahmed, S.E. *Penalty, Shrinkage and Pretest Strategies*, 1st ed.; Springer Cham: Brock University, St. Catherines, Canada, 2014; pp. 115. DOI: https://doi.org/10.1007/978-3-319-03149-1.
25. Ahmed, S.E.; Ahmed, F.; Yusbasi, B. *Post-Shrinkage Strategies in Statistical and Machine Learning for High Dimensional Data*, 1st ed.; CRC Press: Boca Raton, Abingdon, Oxon, USA, England, 2023; pp. 1-107.
26. Yuzbasi, B.; Asar, Y.; Ahmed, S.E. Liu-type shrinkage estimations in linear models. *Statistics* **2022**, *56(2)* 396-420. DOI: https://doi.org/10.1080/02331888.2022.2055030.
27. Medium. Available online: https://towardsdatascience.com/the-likelihood-ratio-test-463455b34de9 (accessed on 28 August 2023).