Using Machine Learning Algorithms to improve heart disease diagnoses

Dr. Fatma Y. Alshenawy

Applied statistics and insurance department, Faculty of commerce, Mansoura university, Mansoura, Egypt.

felshinawy@gmail.com

Scientific Journal for Financial and Commercial Studies and Research (SJFCSR)

Faculty of Commerce – Damietta University

Vol.5, No.1, Part 1., January 2024

APA Citation

Alshenawy, **F**. **Y**. (2024) Using Machine Learning Algorithms to improve heart disease diagnoses, *Scientific Journal for Financial and Commercial Studies and Research*, Faculty of Commerce, Damietta University, 5(1)1, 417-442.

Website: https://cfdj.journals.ekb.eg/

Using Machine Learning Algorithms to improve heart disease diagnoses

Fatma Y. Alshenawy

Abstract:

Heart diseases, are among the leading causes of mortality worldwide. Early prediction and prevention of heart disease can significantly reduce fatalities and improve patients' quality of life. In this study, we propose an advanced hybrid approach that combines multiple statistical models for machine learning algorithms to predict the likelihood of heart disease in individuals.

In recent times, the emergence of machine learning algorithms has shown great promise as a means to predict the risk of heart disease, including Support Vector Machines (SVM), Random Forest (RF), Decision Trees (DT), and Naïve Bayesian (NV). Our results demonstrate the effectiveness of machine learning algorithms, both individually and in combination, for heart disease diagnosis. We provide a comprehensive analysis of the strengths and weaknesses of each algorithm, as well as the ensemble models, and evaluate our approach using eight performance matrices. Our results show that the Random Forest algorithm outperforms other algorithms with an accuracy of 96%, sensitivity of 97.6%, and specificity of 94.7%. Our findings suggest, depends on the growing body of literature, the use of machine learning algorithms for heart disease diagnosis which provides valuable insights for the way for personalized and targeted interference.

Keywords: Machine Learning, Confusion Matrix, Hybrid approach, Support Vector Machines, Decision Trees, Random Forest, Naïve Bayesian.

1.Introduction:

Heart disease is a major cause of morbidity and mortality worldwide. Early and accurate diagnosis of heart disease is critical for effective treatment and management of the condition (Kelleher, et al., 2015). The accuracy of diagnosing heart disease has been shown to be able to be improved through the use of Machine Learning algorithms. by leveraging large amounts of patient data to identify patterns and make predictions. Machine learning algorithms, including Support Vector Machines (Noble,W., 2006), Random Forest (Liaw ,A. and Wiener, M.,2002), Naive Bayes (Rish, I.,2001), and Decision Trees (Safavian,S. and Landgrebe,D., 1991), have been widely used in the field of heart disease prediction due to their ability to handle complex relationships and high-dimensional data.

In this paper, we explore the use of SVM, RF, DT, and NB algorithms, both individually and in combination, for the diagnosis of heart disease. We use a large dataset of patient records to train and test the models and evaluate their performance using standard metrics such as accuracy, sensitivity, specificity, and area under the receiver operating characteristic curve (AUC-ROC).

This paper is organized as follows: Section 2 presents a literature review of relevant studies of different models of machine learning including: Support Vector Machine, Random Forest, Bayesian Naïve, and Decision tree. Section 3 describes the algorithms used. Section 4 describes the data sources and risk factors of heart disease. Section 5 presents the results and discusses the practical implications of our results for diagnose of heart disease and accuracy test. Finally, Section 6 concludes the paper by summarizing the main findings and their best prediction of heart disease.

2.Literature review

Machine learning (ML) algorithms have been a growing area of research, with numerous studies exploring the potential of various algorithms in the medical field. The prediction of heart disease using machine learning algorithms has attracted significant attention from researchers due to its potential to improve patient outcomes and facilitate early intervention (Kelleher et al., 2015),

including four popular ML algorithms - Support Vector Machines (SVM), Random Forest (RF), Decision Tree (DT), and Naïve Bayes (NB) in addition to an ordinary logistic regression model- in predicting heart disease.

One of the earliest studies that laid the foundation for using machine learning algorithms in heart attack prediction is by (Detrano,R. et al., 1989), who used decision trees to analyze the risk factors for coronary heart disease. Decision Trees are popular for their interpretability and ease of implementation in heart disease prediction (Safavian,S. and Landgrebe, D., 1991). It is simple to understand and is often preferred for its interpretability. In a study by (Anooj,P., 2012), a decision tree classifier achieved an accuracy of 82.3% in predicting heart disease.

Subsequently, numerous investigations have delved into the utilization of various machine learning algorithms, including SVM, RF, and BN, to enhance prediction precision. For instance, (Polat, K. et al., 2007) illustrated that SVMs could attain superior classification accuracy in comparison to DTs when predicting heart disease.

Support Vector Machine (SVM) is a widely-used algorithm in the field of heart disease prediction due to its ability to handle non-linear relationships and high-dimensional data (Noble, N.,2006). It has been found to provide accurate and reliable results in various studies. For example, (Kumar,R. et al., 2018) compared the performance of SVM with other algorithms and found that SVM outperformed the others, achieving an accuracy of 84.6%.

Random Forest is an ensemble learning method that has proven to be effective in predicting heart disease (Liaw ,A. and Wiener, M.,2002). It generates multiple decision trees and combines their outputs to enhance classification performance. A study by (Chaurasia,V. and Pal,S.,2018) explored the potential of a hybrid model that combined Random Forest with the Genetic Algorithm (GA) for heart attack prediction. The results demonstrated that the hybrid approach outperformed standalone Random Forest in terms of accuracy and sensitivity. (Alizadehsani,R. et al. ,2013) evaluated the performance of RF on a dataset of heart disease patients and achieved an accuracy of 87.8%.

In recent years, researchers have turned to hybrid approaches to further enhance the performance of these machine learning algorithms. A study by Anooj (2012) combined SVM and DT for heart disease prediction, resulting in improved sensitivity and specificity compared to individual algorithms. Another study by Chaurasia and Pal (2017) employed an ensemble approach, combining SVM, DT, and BN, and reported superior performance in terms of accuracy and recall

Naive Bayes, a simple yet effective probabilistic classifier, has also been employed in heart attack prediction (Rish, 2001) based on Bayes' theorem. It has been used in various medical applications, including heart disease prediction. In a study by Alizadehsani et al. (2013), the Naïve Bayes classifier achieved an accuracy of 83.7% in predicting heart disease.

The performance of ML algorithms in predicting heart disease is commonly evaluated using various metrics, including accuracy, recall, F1 score, and others. Accuracy is the ratio of correctly predicted instances to the total instances. Recall measures the proportion of true positive cases among the relevant cases, and F1 score is the harmonic mean of precision and recall. In addition to these metrics, other metrics such as precision, specificity, and area under the receiver operating characteristic (ROC) curve are also used to evaluate the performance of ML algorithms in heart disease prediction.

In conclusion, SVM, RF, DT, and NB have all demonstrated promising results in predicting heart disease. However, the choice of the best algorithm is often dependes on the specific dataset and problem at hand. Therefore, it is essential to evaluate these algorithms using multiple performance metrics to determine the most suitable approach for a given heart disease prediction task.

3.Algorithms Used

1) **Support Vector Machine** (SVM) is adaptable supervised learning algorithm employed for tasks involving classification and regression. In classification, SVM aims to find the optimal hyperplane that separates the classes with the largest margin (Vapnik, V., 1995). The model is specified by several key hyperparameters as shown in figure 1, including:

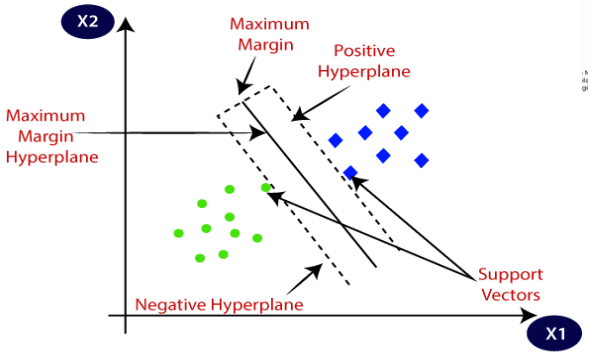


Figure1: support vector machine model

Kernel: A function that converts input data into a space with higher dimensions, enabling SVM to identify nonlinear decision boundaries. Commonly used kernels include linear, polynomial, radial basis function (RBF), and sigmoid (Cortes,C. and Vapnik,V., 1995).

Regularization (C): A parameter that balances the act of maximizing the margin and minimizing classification errors. A lower value of C results in a broader margin, but it may permit a higher number of misclassifications., while a larger value of C aims for fewer misclassifications at the cost of a narrower margin (Cortes,C. and Vapnik,V., 1995).

Kernel-specific parameters: Parameters specific to the chosen kernel, such as the degree for the polynomial kernel, and gamma for the RBF kernel.

The main steps involved in estimating a Support Vector Machine model for classification are as follows:

1. Data Transformation: If necessary, preprocess the input data by scaling the features or applying other transformations.
2. Kernel Selection and Parameter Tuning: Choose an appropriate kernel and tune its parameters, as well as the regularization parameter C, to obtain the best classification performance. This can be done using techniques like grid search or random search combined with cross-validation (Hsu,C.et al., 2003).

3. Model Training: Train the SVM model using the selected kernel and tuned hyperparameters. The model is estimated by solving a convex optimization problem, which involves finding the support vectors and their corresponding weights to define the optimal decision boundary (Vapnik,V., 1995).
4. Model Evaluation: Assess the performance of the trained SVM model using cross-validation or a separate validation dataset (Kohavi, R.,1995).

2) **Decision Tree** (DT) is a popular machine learning algorithm used for classification tasks, as well as regression tasks. In classification, DTs work by recursively partitioning the input space to form a tree-like structure, where each node represents a decision rule based on a feature, and each leaf represents the predicted class label (Quinlan,J., 1986). The model is specified by several key hyperparameters, including:

Criterion: The function used to measure the quality of a split. Commonly used criteria for classification are Gini impurity and information gain (entropy) (Breiman,L. et al., 1984).

Maximum depth: The deepest level of the tree. Limiting the depth can aid in managing overfitting; however, establishing it too low could lead to underfitting.

Minimum samples split: The smallest quantity of samples necessary for dividing an internal node. This parameter assists in preventing overfitting by guaranteeing that the tree does not become overly intricate.

Minimum samples leaf: The minimum amount of samples needed at a leaf node. This parameter can contribute to managing overfitting by stopping the tree from becoming excessively complicated.

Maximum features: The count of features to evaluate when seeking the optimal split. Taking into account a smaller number of features may enhance the diversity of the trees, resulting in improved generalization.

The main steps involved in estimating a Decision Tree model for classification are as follows:

1. Tree Construction: Starting with the root node, the decision tree is grown by recursively partitioning the training data based on the feature that provides the best split according to the chosen criterion. This process is repeated for each child node until a stopping criterion is met.
2. Stopping Criterion: The tree construction process is stopped when one of the following conditions is met: the maximum depth is reached, the minimum samples split or minimum samples leaf criteria are not satisfied, or all the samples at a node belong to the same class.
3. Pruning: In order to avoid overfitting, the decision tree can undergo pruning by eliminating branches that do not contribute to a substantial enhancement in classification performance. This process is based on a validation dataset or the use of a complexity parameter. (Breiman,L. et al., 1984).

The Decision Tree model is estimated by fitting the specified model to the training data, considering the specified hyperparameters. The model's performance can be assessed using cross-validation or a separate validation dataset (Kohavi, 1995).

3) **Random Forest (RF)** is an ensemble learning method introduced by (Breiman,L. , 2001) that combines multiple Decision Trees (DT) to improve prediction accuracy and prevent overfitting. This method functions by building numerous trees during the training stage and producing the class that represents the mode of the classes (for classification) or the average prediction (for regression) of the individual trees.

The main steps involved in estimating a Random Forest model are as follows:

1. Bootstrapping: For each tree in the forest, a bootstrap sample (a dataset with the same size as the original, drawn with replacement) is created from the original training data.

2. Tree Construction: Each decision tree is grown using the bootstrap sample. At each node of the tree, a random subset of features is selected as candidates for splitting. The best split among the candidate features is chosen based on an impurity measure, such as Gini impurity or information gain (Quinlan, 1986).
3. Stopping Criterion: Trees are grown to their maximum depth without pruning, which results in trees that are fully grown and unpruned (Breiman et al., 1984). This helps to reduce the bias of the individual trees and increase model diversity.
4. Aggregation: The predictions of the individual trees are combined to form the final prediction. In classification tasks, this is typically done through majority voting, while in regression tasks, the mean prediction of the individual trees is used.

The key advantage of Random Forests lies in their ability to reduce overfitting and improve generalization by incorporating the predictions of multiple trees, each trained on different subsets of the data and features. This diversity helps to minimize the impact of any single tree's bias or errors and results in a more robust and accurate model.

4) **Naïve Bayes** is a probabilistic learning method based on applying Bayes' theorem with the assumption of conditional independence between features given the class label. It is a simple yet effective technique for classification tasks, particularly when dealing with large feature spaces (Zhang,H., 2004). There are different types of Naïve Bayes classifiers, depending on the distribution assumptions of the input features, such as Gaussian Naïve Bayes, Multinomial Naïve Bayes, and Bernoulli Naïve Bayes.

1. The main steps involved in estimating a Naïve Bayes model for classification are as follows:
2. Data Preparation: Depending on the type of Naïve Bayes classifier, preprocess the input data accordingly. For instance, if using Gaussian Naïve Bayes, ensure the features have a continuous distribution; for Multinomial Naïve Bayes, ensure the features represent discrete counts or frequencies.

3. Model Training: Estimate the class priors and the conditional probabilities of the features given the class labels from the training data. The class priors can be calculated as the proportion of instances belonging to each class, while the conditional probabilities can be estimated using maximum likelihood estimation or other smoothing techniques to avoid zero probabilities (e.g., Laplace smoothing) (Manning,C., et al., 2008).
4. Prediction: For a new instance, calculate the posterior probabilities for each class using Bayes' theorem and the conditional independence assumption. The predicted class label is the one with the highest posterior probability.
5. Model Evaluation: Assess the performance of the trained Naïve Bayes model using cross-validation or a separate validation dataset (Kohavi,R., 1995).

The performance of machine learning (ML) algorithms is a critical aspect of their application in various fields, including healthcare, finance, and natural language processing. Evaluating the performance of ML algorithms helps to determine their effectiveness in solving specific problems, identify areas for improvement, and compare their performance with other algorithms (Kelleheret al., 2015). Several performance metrics have been developed to measure the performance of ML algorithms, including accuracy, recall, F1 score, precision, specificity, and the area under the receiver operating characteristic (ROC) curve (Sokolova,M. and Lapalme,G., 2009).

- **Accuracy:** is one of the most used performance metrics, measuring the proportion of correctly predicted instances out of the total instances (Sokolova & Lapalme, 2009). While accuracy is a straightforward metric, it may not be the most informative in cases where the dataset is imbalanced, as it does not distinguish between the types of errors made by the algorithm (Kelleher et al., 2015).

There are several other metrics to evaluate and compare classification models besides accuracy. These metrics can provide more insight into the model's performance, especially when dealing with imbalanced datasets or when

different types of misclassifications have different costs. Some popular performance metrics include:

- **Confusion Matrix**: The confusion matrix is a tabular representation of the performance of an ML algorithm, where each row represents the true class and each column represents the predicted class (Kelleher et al., 2015). It provides a detailed breakdown of the algorithm's performance, showing the number of true positive (TP), true negative (TN), false positive (FP), and false negative (FN) predictions. The confusion matrix serves as the basis for calculating other performance metrics, such as accuracy, recall, precision, and specificity (Sokolova, M and Lapalme,G., 2009).
- **Recall**: also known as sensitivity or true positive rate, measures the proportion of true positive cases among the relevant cases (Sokolova and Lapalme ,2009).
- S**pecificity**: or true negative rate, measures the proportion of true negative cases among the negative cases (Sokolova and Lapalme, 2009). Specificity measures the proportion of true negative predictions among all actual negative instances. It is useful when the cost of false positives is high, and you want to measure the model's performance on the negative class
- **Precision**: also called positive predictive value, measures the proportion of true positive cases among the predicted positive cases (Sokolova and Lapalme, 2009).. It is a useful metric when the cost of false positives is high.
- **F1 score** :is the harmonic mean of precision and recall, providing a single metric that balances the trade-off between these two metrics (Chicco,D. and Jurman,A., 2020). The F1 score ranges from 0 to 1, with higher values indicating better classifier performance. It is particularly useful when dealing with imbalanced datasets, as it accounts for both false positives and false negatives (Chicco and Jurman, 2020).
- **AUC**: The area under the ROC curve (AUC-ROC) is another popular performance metric, which measures the ability of an algorithm to correctly classify instances across all possible classification thresholds (Fawcett,T., 2006). The AUC is calculated by plotting the true positive rate (TPR) against

the false positive rate (FPR) at various classification thresholds and computing the area under the resulting curve. AUC values range from 0 to 1, with higher values indicating better classifier performance.

- **Matthews Correlation Coefficient (MCC)**: The MCC is a performance metric that provides a balanced measure of the quality of binary classifications, taking into account all elements of the confusion matrix (Chicco and Jurman, 2020). The MCC ranges from -1 to 1, with +1 indicating perfect classification, 0 indicating random classification, and -1 indicating complete disagreement between the predicted and true classes. The MCC is particularly useful when dealing with imbalanced datasets, as it is less sensitive to class imbalances than other metrics like accuracy or F1 score.

The performance of machine learning (ML) algorithms is crucial for determining their effectiveness in solving specific problems and comparing their performance with other algorithms. Various performance metrics have been developed to evaluate ML algorithms, including the confusion matrix, area under the receiver operating characteristic (ROC) curve (AUC), Matthews correlation coefficient (MCC), and F1 score (Sokolova and Lapalme, 2009; Chicco and Jurman, 2020).

Depending on the specific problem and dataset, one or more of these metrics may be more appropriate for evaluating and comparing models.

4.Data source

The heart disease dataset from used comes from the University of California Irvine, sourced from Kaggle, contains 271 observations, and aims to predict the onset of heart disease based on various diagnostic factors. The dataset consists of the following thirteen risk factors in addition to the response variable.

1. Age: The age of the patient in years.
2. Sex: The sex of the patient (categorical: 1 = male; 0 = female).
3. Chest_pain: The type of chest pain experienced by the patient (categorical: 1=Typical Angina, 2=Atypical Angina, 3=Non-Anginal Pain, 4= Asymptomatic).
4. BP: The resting blood pressure of the patient in mmHg (millimeters of mercury).

5. Cholesterol: The patient's serum cholesterol level in mg/dL (milligrams per deciliter).
6. FBS_Over_120: An indicator of whether the patient's fasting blood sugar is greater than 120 mg/dL (binary: 1 = true; 0 = false).
7. EKG_results: The patient's resting electrocardiographic results (categorical: Normal, ST-T Wave Abnormality, or Left Ventricular Hypertrophy).
8. Max_HR: The maximum heart rate achieved by the patient during a stress test, measured in beats per minute.
9. Exercise_angine: The presence of exercise-induced angina (chest pain) during the stress test (binary: 1 = yes; 0 = no).
10. ST_depression: The amount of ST segment depression induced by exercise relative to rest, a measure of myocardial ischemia.
11. Slope_ST: The slope of the peak exercise ST segment (categorical: 1=Upsloping, 2=Flat, 3=Downsloping).
12. N Vessels fluro: The number of major vessels (0-3) colored by fluoroscopy, an indicator of coronary artery disease severity.
13. Thallium: The result of a thallium stress test, which evaluates blood flow to the heart muscle (categorical: Normal, Fixed Defect, or Reversible Defect).
14. Response Variable Heart Disease: Whether the patient has heart disease (binary: 1=Yes , 0= No).

Before analysis, the dataset was handled by checking for missing values and outliers. Missing values were replaced with mean, and outliers were removed.

The data set is split in two parts: training set (189 observations) and the test set (81 observations). The training set is used to fit the models, while test set is used to test the model and evaluate the accuracy.

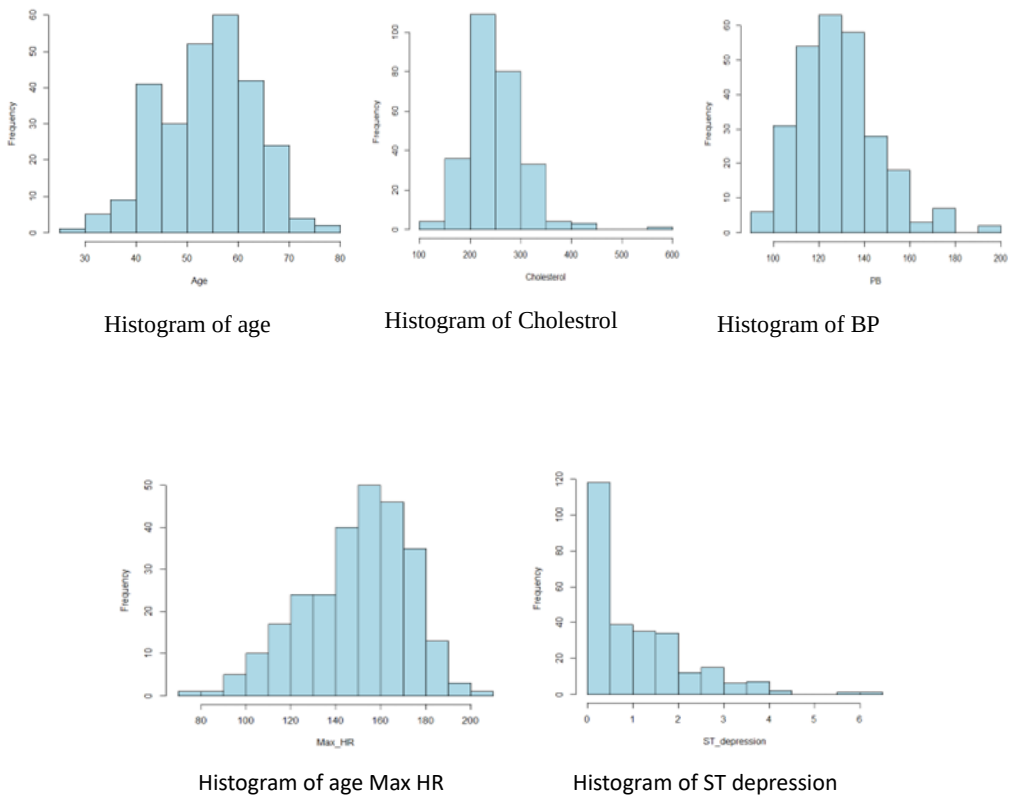
Table [1] presents the descriptive statistics of various risk factors for heart disease in a sample of 270 individuals. The table displays the number of

observations (n), mean, standard deviation (sd), median, minimum (min), maximum (max), skewness (skew), and standard error (se) for each variable.

Table [1]: The risk factors of heart disease using descriptive statistics.

variables	n	mean	sd	Median	min	max	skew	se
Age	270	54.43	9.11	55.0	29	77	-0.16	0.55
Sex	270	0.68	0.47	1.0	0	1	-.76	0.03
Chest_pain	270	3.17	0.95	3.0	1	4	-0.87	0.06
BP	270	131.34	17.86	130.0	94	200	0.71	1.09
Cholesterol	270	249.66	51.69	245.0	126	564	1.17	3.15
FBS_Over_120	270	0.15	0.36	0.0	0	1	1.97	0.02
EKG_results	270	1.02	1.00	2.0	0	2	-0.04	0.06
Max_HR	270	149.68	23.17	153.5	71	202	-0.52	1.41
Exercise_angine	270	0.33	0.47	0.0	0	1	0.72	0.03
ST_depression	270	1.05	1.15	0.8	0	6.2	1.25	0.07
Slope_ST	270	1.59	0.61	2.0	1	3	0.54	0.04
N Vessels fluro	270	0.67	0.94	0.0	0	3	1.2	0.06
Thallium	270	4.70	1.94	3.0	3	7	0.28	0.12
Heart Disease	270	0.44	0.50	0.0	0	1	0.22	0.03

Furthermore, Figure 2 presents the histograms for the input features of the heart disease, including Age, Sex, Chest_pain, BP, Cholesterol, FBS_Over_120, EKG_results, Max_HR, Exercise_angine, ST_depression, Slope_ST, N Vessels fluro, Thallium and the response variable heart disease.



Histogram of age | Histogram of Cholestrol | Histogram of BP

Histogram of age Max HR | Histogram of ST depression

Figure 2: Histogram of age, Cholestrol, BP, Max HR and ST depression

Figure 3 illustrate the pie chart of categorical risk factors including sex, FBS, EKG, Exercise angine, Slope ST, Numberof Vessels fluro, Thallium and the target variable Heart Disease. And the descriptive percentage frequencies for each variable.

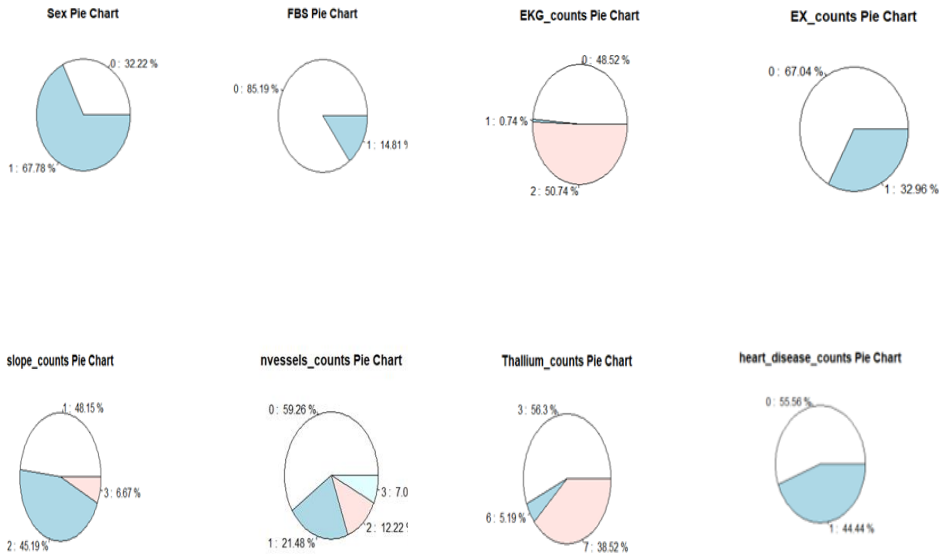


Figure 3: Pie chart of categorical risk factor

5.Results and discussion

Before proceeding, we need to understand the dependence structure between risk factors and target variable using correlation matrix as illustrated in figure 4, these correlations can help in understanding the relationships between different risk factors. further analysis would be needed to determine any causal relationships between these risk factors.

Based on the provided confusion matrix in table [2], we have the results for 5 different models: Logistic Regression, Support Vector Machine (SVM), Random Forest (RF), Decision Tree (DT), and Naïve Bayesian. The results illustrate that logistic regression performed fairly well in predicting both positive and negative cases. However, it could improve in reducing false negatives. While, SVM model performed better in predicting positive cases, with a higher TP rate and a lower FP rate compared to the logistic regression model.

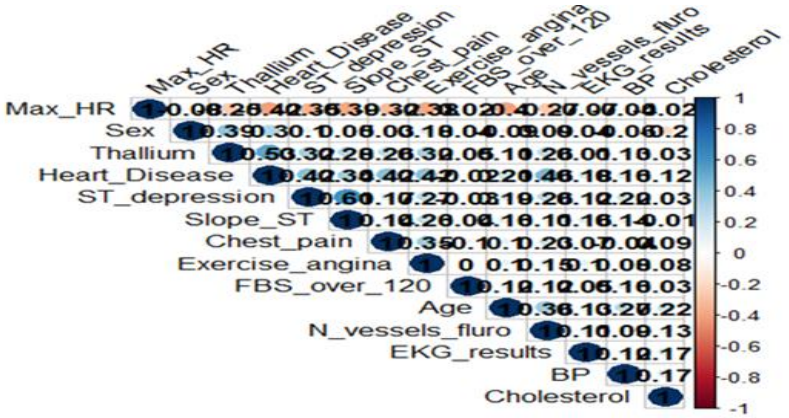


Figure 4: correlation matrix of risk factors

However, the performance for negative cases is similar to logistic regression, with the same number of false negatives. The RF model has the best performance among all the models, with the highest TP rate and the lowest FP and FN rates. It accurately predicted both positive and negative cases. elsewhere, the DT model has a slightly lower performance in predicting positive cases compared to the SVM and RF models. It also has the highest number of false negatives, indicating room for improvement in predicting negative cases. Finally, the performance of the Naïve Bayesian model is comparable to the logistic regression model. It has a slightly better performance in predicting negative cases, with one less false negative.

Table [2]: Confusion Matrix of classification models

Prediction	**Reference**										
	positive						**negative**				
	model	**logistic**	**SVM**	**RF**	**DT**	**Naïve Bayesian**	**logistic**	**SVM**	**RF**	**DT**	**Naïve Bayesian**
	true	36	41	42	38	37	10	10	2	13	9
	false	7	2	1	5	6	28	28	36	25	29

Based on the provided statistics for the confusion matrix in table [3] for the 5 models (Logistic Regression, SVM, Random Forest, Decision Tree, and Naïve Bayesian), it illustrate the values of accuracy (the proportion of correct predictions both true positives and true negatives out of the total number of predictions, the Random Forest model has the highest accuracy, indicating that it performs best overall among all the models with the highest and narrowest 95%confidence interval by (0.8956, 0.9923),. and p-value less than 0.05, indicating that its accuracy is significantly greater than the NIR. The Random Forest model has the smallest p-value, suggesting the strongest evidence against the null hypothesis.

Kappa value is a statistic that measures the agreement between the model and the true labels while accounting for the agreement that would be expected by chance. It ranges from -1 to 1, where a value of 1 indicates perfect agreement, 0 indicates agreement no better than chance, and negative values indicate disagreement. The Random Forest model shows the highest Kappa value, indicating that it has the best agreement with the true labels compared to the other models.

Mcnemar's Test P-Value is a statistical test used to compare the performance of two classifiers by assessing if the proportions of misclassified instances are significantly different. The null hypothesis states that the proportions are equal, and a low p-value (usually below 0.05) indicates that there is a significant difference between the classifiers. The Mcnemar's Test P-Value suggests a significant difference in performance between the SVM and Random Forest models, with the latter likely being better.

Table [3]: Statistics of Confusion Matrix

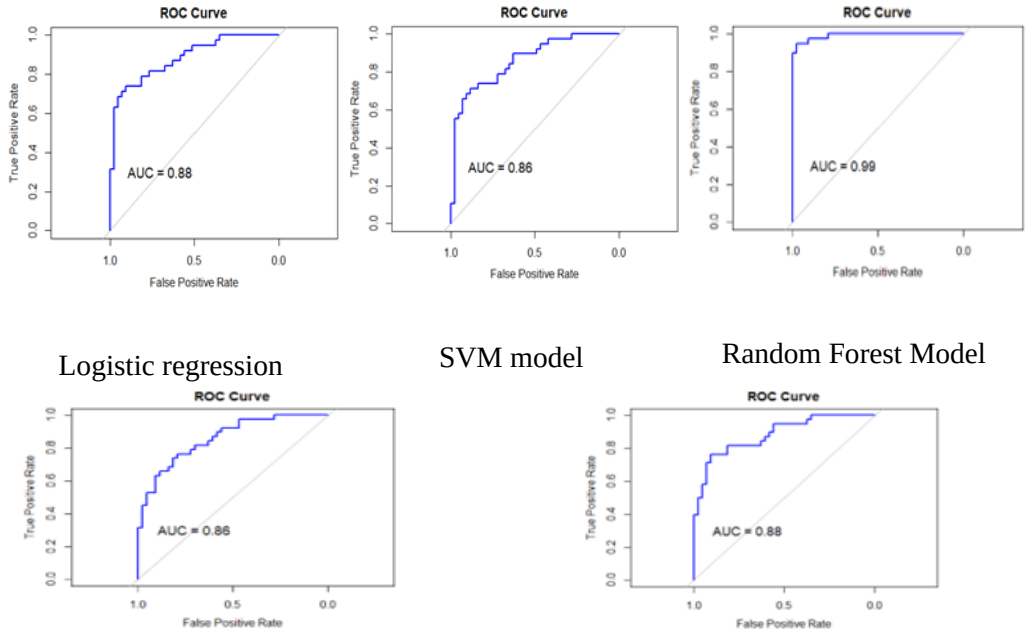
Model	Logistic regression	SVM model	Random Forest	Decision tree	Naïve Bayesian
Accurancy	0.7901	0.8519	0.963	0.7778	0.8148
95% CI	(0.6854, 0.8727)	(0.7555, 0.921)	(0.8956, 0.9923)	(0.6717, 0.8627)	(0.713, 0.8925)
P-Valu [Acc > NIR]	1.167e-06	1.048e-09	<2e-16	3.779e-06	8.962e-08
kappa	0.5767	0.6989	0.9255	0.5483	0.6265
Mcnemar's Test P-Value	0.6276	0.04331	1	0.09896	0.6056
Sensitivity	0.8372	0.9535	0.9767	0.8837	0.8605
Specificity	0.7368	0.7368	0.9474	0.6579	0.7632
Pos Pred Value	0.8043	0.8039	0.9545	0.7451	0.8043
Neg Pred Value	0.8286	0.9333	0.9730	0.8333	0.8286
Prevalence	0.5309	0.5309	0.5309	0.5309	0.5309
Balanced Accuracy	0.7870	0.8452	0.9621	0.7708	0.8118

Table [4] shows the performance metrics, the Random Forest model has the best performance in most categories, followed by the SVM model, Naïve Bayesian model, Logistic Regression model, and Decision Tree model. The Random Forest model has the highest accuracy, recall, specificity, F1 score, MCC, and AUC, indicating that it performs better overall in terms of prediction, balance between precision and recall, and the ability to distinguish between positive and negative cases.

Table [4]: performance metrics to evaluate and compare classification models

model	Logistic regression	SVM model	Random Forest	Decision tree	Naïve Bayesian
Accuracy	0.7901235	0.8518519	0.962963	0.7777778	0.8148148
precision	0.7826087	0.8039216	0.9545455	0.745098	0.8043478
recall	0.8372093	0.9534884	0.9767442	0.8837209	0.8604651
specificity	0.7368421	0.7368421	0.9473684	0.6578947	0.7631579
f1_score	0.8089888	0.8723404	0.9655172	0.8085106	0.8314607
Matthews CorrCoef	0.5783142	0.713407	0.925814	0.5597209	0.6282539
ROC	0.8824969	0.8451652	0.9908201	0.8555692	0.8818849
AUC	0.6261982	0.5981776	0.6558218	0.6166968	0.6273594

Furthermore, figure [5] illustrate the ROC curve and the AUC value for logistic regression, SVM, Random Forest, Decision Tree and Naïve Bayesian respectively. It is obvious that the ROC curve of Random Forest model is closer to the top-left corner of the plot, it indicates that the classifier has a better ability to distinguish between positive and negative cases. Also, higher AUC value indicates better performance.



Logistic regression SVM model Random Forest Model

Decision Tree model naïve Bayesian model

Figure [5]: ROC curve for classification models

Upon choosing the Random Forest model as the top-performing model across all evaluation metrics, we employed a hybrid strategy by integrating it with other high-performing models using the Ensemble Method (Zhou, Z. H., 2012). This approach entails merging the predictions of multiple models to reach a final outcome.

In our case, we utilized an Ensemble Method known as bagging (Breiman, L., 1996), a popular technique that involves training several base

models independently and combining their predictions through averaging (for regression tasks) or voting (for classification tasks).

table [5] shows the performance metrics for each combined model (Logistic Regression + Random Forest, SVM + Random Forest, Decision Tree + Random Forest, and Naive Bayes + Random Forest). Overall, the combined models performed better than the individual Logistic Regression and SVM models but not as good as the standalone Random Forest model in terms of accuracy. The highest accuracy is achieved by the SVM + Random Forest combined model (0.9506), followed by the Naive Bayes + Random Forest (0.9136) and Decision Tree + Random Forest (0.9012) combined models. However, none of the combined models outperform the accuracy of the standalone Random Forest model (0.9630).

Comparing the sensitivity and specificity across the combined models, the SVM + Random Forest model demonstrates the highest sensitivity (0.9302) and shares the highest specificity (0.9737) with the Logistic Regression + Random Forest and Naive Bayes + Random Forest models. However, the standalone Random Forest model still outperforms all combined models in terms of sensitivity (0.9767) and has comparable specificity (0.9474).

Table [5]: Statistics of Confusion Matrix for Ensemble model

model	Accuracy	95% CI	P-Value	Kappa	Sensitivity	Specificity
Logistic+ rf	0.8889	(0.7995, 0.9479)	5.259e-12	0.7793	0.8140	0.9737
SVM+RF	0.9506	(0.8784, 0.9864)	<2e-16	0.9012	0.9302	0.9737
DT+RF	0.9012	(0.8146, 0.9564)	7.196e-13	0.8029	0.8605	0.9474
NB+RF	0.9136	(0.83, 0.9645)	8.643e-14	0.8278	0.8605	0.9737
SV+ Logistic	0.8272	(0.727, 0.9022)	2.213e-08	0.653	0.8372	0.8158
SV+DT	0.8148	(0.713, 0.8925)	8.962e-08	0.6265	0.8605	0.7632
SV+NB	0.8395	(0.7412, 0.9117)	5.031e-09	0.6773	0.8605	0.8158
NB+DT	0.8025	(0.6991, 0.8827)	3.356e-07	0.6046	0.7907	0.8158

In summary, while the combined models show improvements over the individual Logistic Regression and SVM models, the standalone Random Forest

model remains the best-performing model in terms of accuracy, sensitivity, and specificity. It is essential to consider these results in the context of the specific problem and dataset, as different models and combinations may produce varying performance levels depending on the data and problem characteristics.

6.conclusion

In conclusion, this study aimed to investigate the performance of various combined models, including Logistic Regression + Random Forest, SVM + Random Forest, Decision Tree + Random Forest, and Naive Bayes + Random Forest, in comparison to the standalone Logistic Regression, SVM, and Random Forest models. The evaluation was based on multiple performance metrics, such as accuracy, sensitivity, specificity, and kappa, among others.

The results demonstrated that using hybrid approach to the combined models generally outperformed the standalone Logistic Regression and SVM models in most of the performance metrics. The SVM + Random Forest combined model achieved the highest accuracy (0.9506) among the combined models, followed by the Naive Bayes + Random Forest (0.9136) and Decision Tree + Random Forest (0.9012) combined models. However, none of the combined models surpassed the standalone Random Forest model, which exhibited the best performance in terms of accuracy (0.9630), sensitivity (0.9767), and specificity (0.9474). These findings suggest that while combining classifiers can improve the performance of certain models, it is essential to evaluate the individual models as well. In this particular study, the standalone Random Forest model proved to be the most effective classifier, outperforming all combined models.

Finally, we recommended Investigate other model combinations of classifiers, such as ensemble methods like stacking, bagging, and boosting, to identify potential improvements in performance. Also examine the effects of hyperparameter tuning on the performance of the combined models and individual models, using techniques like grid search, random search, or Bayesian optimization.

References

1. Alizadehsani, R., Habibi, J., Hosseini, M. J., Mashayekhi, H., Boghrati, R., Ghandeharioun, A., ... & Sani, Z. A. (2013). A data mining approach for diagnosis of coronary artery disease. Computer methods and programs in biomedicine, 111(1), 52-61.
2. Anooj, P. K. (2012). Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules. Journal of King Saud University - Computer and Information Sciences, 24(1), 27-40.
3. Breiman, L. (1996). Bagging predictors. Machine Learning, 24(2), 123-140.
4. Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.
5. Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). Classification and Regression Trees. CRC press.
6. Chaurasia, V., & Pal, S. (2017). Data mining approach to detect heart diseases. International Journal of Advanced Computer Science and Applications, 8(5), 166-171.
7. Chaurasia, V., & Pal, S. (2018). A novel approach for heart disease prediction using random forest and Genetic Algorithm. International Journal of Computer Sciences and Engineering, 6(7), 112-119.
8. Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics, 21(1), 6.
9. Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273-297.
10. Detrano, R., Janosi, A., Steinbrunn, W., Pfisterer, M., Schmid, J. J., Sandhu, S., ... & Froelicher, V. (1989). International application of a new probability algorithm for the diagnosis of coronary artery disease. American Journal of Cardiology, 64(5), 304-310.
11. Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861-874.

12. Hsu, C. W., Chang, C. C., & Lin, C. J. (2003). A practical guide to support vector classification. Technical report, Department of Computer Science, National Taiwan University.
13. Kaur, M., Singh, D., & Minhas, R. S. (2018). Hybrid approach using case-based reasoning and rule-based reasoning for domain independent clinical decision support system. Artificial Intelligence in Medicine, 84, 101-111.
14. Kelleher, J. D., Mac Namee, B., & D'Arcy, A. (2015). Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies. MIT Press.
15. Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI), 2(12), 1137-1143.
16. Kumar, R., Sharma, A., & Aggarwal, N. (2018). Performance analysis of various data mining algorithms for the prediction of heart disease. International Journal of Computer Applications, 975, 8887.
17. Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R News, 2(3), 18-22.
18. Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to Information Retrieval. Cambridge University Press.
19. Noble, W. S. (2006). What is a support vector machine? Nature Biotechnology, 24(12), 1565-1567.
20. Polat, K., Şahan, S., & Güneş, S. (2007). A novel hybrid intelligent method based on C4. 5 decision tree classifier and one-against-all approach for multi-class classification problems. Expert Systems with Applications, 32(2), 554-564.
21. Quinlan, J. R. (1986). Induction of Decision Trees. Machine Learning, 1(1), 81-106.
22. Rish, I. (2001). An empirical study of the naive Bayes classifier. In IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence, 41(3), 41-46.

23. Safavian, S. R., & Landgrebe, D. (1991). A survey of decision tree classifier methodology. IEEE Transactions on Systems, Man, and Cybernetics, 21(3), 660-674.
24. Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing & Management, 45(4), 427-437.
25. Vapnik, V. N. (1995). The Nature of Statistical Learning Theory. second edition Springer.
26. Zhang, H. (2004). The optimality of naïve Bayes. Proceedings of the 17th International Florida Artificial Intelligence Research Society Conference (FLAIRS), 562-567.
27. Zhou, Z. H. (2012). Ensemble methods: Foundations and algorithms. CRC Press.

استخدام خوارزميات التعلم الآلي في تحسين تشخيص أمراض القلب

د. فاطمة يوسف عبد الرازق الشناوي

المستخلص:

تعدُّ أمراض القلب من المسببات الرئيسية للوفيات على مستوى العالم. يُمكن أن يُساهم التشخيص المبكر والوقاية من أمراض القلب في خفض عدد الوفيات بشكل كبير وتعزيز جودة حياة المرضى. في هذا البحث، نقترح استخدام منهج هجين متقدّم يجمع بين مجموعة من الاساليب الاحصائية لنماذج التعلم الآلي Machine Learning للتنبؤ بفرص الإصابة بأمراض القلب بين الأفراد.

في الفترة الأخيرة، ظهرت خوارزميات التعلم الآلي ML كأداة مُبتكرة للتنبؤ بخطورة الإصابة بأمراض القلب، بما في ذلك نماذج دعم آليات الناقلات(Support Vector Machine)، والغابة العشوائي(Random Forest) ، وأشجار القرار(Decision Tree) ، والتصنيف البيزي (Naïve Bayesian). تظهر نتائج هذا البحث فاعلية استخدام خوارزميات التعلم الآلي ML ، سواءً بشكل منفصل أو مدمج، في تشخيص أمراض القلب. كما تقدم النتائج تحليلًا شاملًا لمميزات وعيوب كل اسلوب من النماذج ML، بالإضافة إلى النماذج المدمجة، تم تقييم اداء كل نموذج باستخدام ثمان مصفوفات أداء مختلفة. تُظهر نتائج هذه الدراسة أن اسلوب نماذج الغابة العشوائية RF تفوق الاساليب الأخرى بدقة تصل إلى 96%، وحساسية 97.6%، وتحديد 94.7%. كما تُشير نتائج الدراسة إلى ضرورة ازدياد الأبحاث المتعلقة باستخدام خوارزميات التعلم الآلي في تشخيص أمراض القلب.

الكلمات المفتاحية:

التعلم الآلي ML، مصفوفة الارتباكConfusion Matrix ، النهج المختلط Hybrid approach، آلات ناقلات الدعم Support Vector Machines، أشجار القرارDecision Trees ، الغابة العشوائية Random Forest، التصنيف البيزيNaïve Bayesian