**ED.title**

A Machine Learning Approach: Using Predictive Analytics to Identify and Analyze High Risks Patients with Heart Disease

Using Machine Learning Approach to Identify and Analyze High Risks Patients with Heart Disease

**ED.Abstruct**

**Background:** Cardiovascular disease is one of the most threatening diseases to human health today. **Exploring the performance of different models in predicting cardiovascular diseases will help medical practitioners to make more accurate medical diagnoses using non-invasive means to save lives.** In this paper, a comparative analysis of different classification prediction models was applied on the prediction of heart disease cases using heart disease data from UCI machine learning Repository. This data source contains 14 dimensions of data for 303 patients. The classifiers applied in this study were decision trees, random forests, support vector machines(SVM) and logistic regression. To examine the performance of each classifier, criteria such as accuracy, sensitivity, and specificity were used, and a 10-fold cross-validation method was used to measure the unbiased estimates of these prediction models.

**Results:**

**Conclusion**: According to our results, SVM is able to make predictive judgments for suspected cardiovascular disease cases to the maximum extent possible.

**ED.keywords**

Keywords: heart disease,classification,decision tree, random forest, support vector machine.

**ED. Introduction**

Heart disease is the leading cause of death before cancer and traffic accidents **[ED.I.04]**. Data from the CDC **[ED.I.05]** reveal that heart disease causes about 655,000 deaths each year, which accounts for 25% of all deaths in the United States. According to the National Cancer Institute's definition of heart disease, heart disease refers to a category of diseases that affect the heart or blood vessels. In fact, the risk of cardiovascular disease is associated with other factors as well; smoking, high blood pressure, high cholesterol, unhealthy diet, lack of exercise and obesity all contribute to the potential risk of heart disease. However, identifying heart disease can be difficult due to the uncertainty of clinical practice. Due to this limitation, scientists have turned to modern methods, such as data mining and machine learning, to predict disease. In this regard, it is particularly important to compare the performance of various techniques and algorithms and determine the best approach.

This study builds classification models to make predictions on typical data from heart disease patients and compares the performance of different algorithms to gain insight into the ability of different models to reveal hidden patterns in medical data. Understanding the ability of models can empower non-invasive means in cardiovascular disease diagnosis and assist medical professionals to make more accurate judgments. This can not only help save lives through early diagnosis of heart problems, but also save money by avoiding costly invasive treatments. Existing classifier algorithms, such as decision trees**[ED.I.03]**, logistic regression, random forests, and support vector machine**[ED.I.02]**, and hybrid data mining **[ED.I.01]**, have been used to explore different types of cardiac problems. Medical data mining has great potential in exploring hidden patterns in datasets in the clinical domain.

### ED.Related Works

The extensive use of classifiers to model and diagnose cardiovascular diseases has promoted this work. This section will describe the results of a brief literature survey. In reference [RC.2], the authors used an open heart disease dataset from the Cleveland Clinic and classified 303 patients by different methods of decision trees(CART, ID3, DT) and concluded that the CART classifier had the best performance with an accuracy of 83.49%. A similar study was conducted in [RC.1] using the same dataset, where researchers trained a deeper neural network on the feature scaling data and improved the accuracy significantly to 96%. In reference [EC.3], the authors used feature-selected patient data to predict heart disease by building an H-model and obtained an accuracy of 93.3%. And in the context of the proposed hybrid prediction model [RC.4], Luxmi and Sangeet tested the performance of eight different classifiers(SVM, neural network, decision tree, generalized linear model, Lasso, Bayesian regularized neural network, classification and regression tree) for the prediction of the UCI heart disease database showing that for a single model, support vector machine, the logistic regression classifier and random forest performed better, with accuracies of 86%, 84% and 83%, respectively. In a similar study in [RC.6], the authors compared the performance of different models using the same dataset, and this time, logistic regression classifier and support vector machine proved to be the best methods. Another study [RC.5] compared the effect of different feature selection methods on the prediction effectiveness of the models. The experimental results proved that using the combination of CFS and PSO was effective in improving the prediction of the model, and the combination improved the correct rate of the MLP algorithm by almost 7%. The following table summarizes and describes in detail all the methods previously analyzed. The obvious conclusion is that the hybrid algorithm-based methods provide more accurate results than those using a single algorithm. Also, feature selection can improve the performance of the model. The pre-analysis of the existing literature helped the study to adopt the best algorithm based on the results obtained in previous papers. For this purpose, the study will optimize the input data through feature selection and train the test set on support vector machines, decision trees, and logistic regression models.

**ED. Data& Methods**

The following section is a briefly discussion of the method and materials applied in this research.

**ED.M.Data Source**

The dataset used in study is obtained from **UCI machine learning Repository**[ED.M.D.01], contains 303 instances and 14 attributes, which is also the most commonly used by researchers. After processing the missing values, six samples will be removed and the sample used for this study is composed of 13 characteristics from 297 patients. The output field, which is defined as angiographic disease status, has a value range of 1 to 4. To simplify the prediction, the new target will appear as a binary value, value 0 for cases without risk of heart disease, and other value(value 1,2,3,4) means the presence of cardiovascular disease. Complete information and descriptions of the 297 instances of the 13 features in the dataset are given in Table [ED.M.Data Source].

**Table[ED.M.Data Source]: Attributes of UCI heart disease dataset**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Index | Attribute | label | Description | Domain range of values |
| 1 | Patient's age | age | age in years | 29-77 |
| 2 | Patient's gender | sex | 1 = male  0 = female | 0,1 |
| 3 | Chest pain type | cp | 1 = typical angina  2 = atypical angina  3 = non-anginal pain  4 = asymptomatic | 1,2,3,4 |
| 4 | Resting blood pressure | trestbps | in mm Hg on admission to the hospital | 94-200 |
| 5 | Cholesterol measurement | chol | in mg/dl | 126-564 |
| 6 | Fasting blood sugar | fbs | 1 = fbs> 120 mg/dl  0 = fbs≤120mg/dl | 0,1 |
| 7 | A blood disorder called thalassemia | thal | 3 = normal  6 = fixed defect  7 = reversable defect | 3,6,7 |
| 8 | Resting electrocardiographic results | restecg | 0 = normal  1 = having ST-T wave abnormality  2 = probable or definite left ventricular hypertrophy | 0,1,2 |
| 9 | maximum heart rate | Thalach |  | 71-202 |
| 10 | Exercise induced angina | exang | 1 = yes  0 = no | 0,1 |
| 11 | ST depression induced by exercise relative to rest. | Oldpeak |  | 0-6.2 |
| 12 | Slope of peak exercise ST segment | slope | 0 = downsloping  1 = flat  2 = upsloping | 0,1,2 |
| 13 | Number of major vessels(0-3) colored by flourosopy | ca |  | 0,1,2,3 |

**ED.Proposed model**

The following section describes how the proposed model has been developed to predict the occurrence of cardiovascular disease and how the performance of different feature selection methods and machine learning algorithms are tested. Feature selection algorithms, such as ANOVA and LASSO were used to pick up important features, and the the performance of the famous classifiers applied in related works, such as KNN, decision tree, Naive Bayes, random forest, SVM and logistic regression were tested. Cross-validation method will be applied in the process of training data. To evaluate the results of the statistical analysis, this study also uses different performace metrics . The proposed model consists of four functional modules, in the order of data perpoessing, feature selection, machine learning, and evaluation. Figure ed.p.1 shows the main workflow of the whole process.

Diagram

Description automatically generated

Figure 1: The main workflow of predicting heart disease

**ED.P. Data preproessing**

Low quality data will lead to low quality mining results, besides that, the noise in the clinical data set increases the computational effort and computation time. Thus pre-processing of data is necessary to build effective machine learning classifiers. This process involves dealing with missing values and standard scalar.

**ED.P.platform**

With respect to [ED.M.1], python provides a good platform for data analysis and machine learning due to its object-oriented, open source nature, and scikit-learn package enables a comprehensive list of machine learning methods and is a handy toolkit for statisticians. Scholars have found the python environment to be concise and accurate, and have used it to predict heart disease in medical data[ED.M.2].

In this paper, several packages,especially scikit-learn, is used in python environment to applying a variety of statistical models for data analysis, data visualization, feature selection and training and testing using machine learning methods.

**ED.P.Feature Selection**

Medical experts face several problems when using algorithms to make diagnoses on clinical datasets: because clinical datasets are often complex, unintuitive, and contain subjective data, some features in the dataset may be redundant or irrelevant, which can lead to degradation of the classifier's performance[ED.P.FE.02]. At this point, feature selection will become our concern. Feature selection is an integral part of data analysis and mining. By selecting features in the data that contribute more to the final result, overfitting can be reduced in a way that reduces data noise. At the same time, clinical datasets are usually high-dimensional, which limits the medical experts from manually removing features that contribute very little to the results, leading us to turn to automatic means of feature selection. According to related studies comparing different means of feature selection[ED.P.FE.01], SelectKBest paired with a classifier has the highest performance in the processing of high-dimensional data, yielding 97% accuracy in 0.11 seconds.

This study uses three feature methods for selecting the most informative features.

##### ED.P.F. Random Forest Feature selection

The main idea of feature importance assessment using random forests is that by calculating the contribution made by each feature on each CART tree in a random forest and averaging the sum, a comparison of the contribution size of different features can be made. The measures of contribution include Gini index and oob error(out-of-bag error). The Gini index will be used as the evaluation criterion in this study.

##### ED.P.F. ANOVA Method for Feature selection

The basic idea of ANOVA is that it allows the contribution of different sources of variation to the total variation to be assessed, and thus to objectively determine the magnitude of the influence of controllable factors on the study results[ED.PF.ANOVA.01]. The method can select a subset of features with significant influence on the system state from many sample features with strong correlation and redundancy, as a new sample of features reflecting the system state.

##### ED.P.F. Least Absolute Shrinkage and Selection Operator

Least absolute shrinkage and selection operator, also Lasso or LASSO, is a commonly used feature selection method in machine learning. It makes the training solving parameter process take into account the magnitude of the coefficients by adding a penalty term to the loss function(i.e., optimization objective), and by setting the scaling factor(penalty coefficient), it will make the coefficients of the less influential features decay to zero and only retain the important features. Lasso feature selection method suffers from low stability, i.e., it may lead to large model differences even when there are small changes in the data. Also, it cannot handle data with multicollinearity: when there is a high correlation between the data, it tends to select one from each group and ignore the others [ED.PF.LASSO.01]

#### ED.P.classifiers

Machine learning classifiers are good at identifying hidden patterns and regularities within data, and in this study a number of machine learning classifiers were used to predict cardiovascular disease incidence events. In the following, their theoretical background will be briefly described.

##### ED. P.C.Logistic Regression

Logistic regression is a generalized linear regression model used to deal with classification problems. The probability of a patient having cardiovascular disease is 1 when the event occurs and 0 when healthy. p is the predicted probability of disease occurrence and takes values in the range [0,1], then 1-p is the probability that the patient has no risk of heart disease. Suppose that the independent variables are x1,x2,…,xn, then the logistic regression formula can be expressed as

,while βi(I = 0, 1, 2, …, k) is the regression coefficient.

##### ED. P.C.SVM

The basic idea of support vector machine is to map the sample data to a higher dimensional space, and to build a hyperplane in such a higher dimensional space so that the distance between the hyperplane and different class sample sets is maximized for the purpose of classification. In this study, the svc function in the sklearn package is used to implement the model prediction work of support vector machine, and the penalty parameter is 0.5 when the kernel function is radial basis kernel function(RBF) and 1 when the kernel function is linear, as confirmed by the hyperparameter auto search module GridSearchCV.

##### ED. P.C.Naive Bayes

Naive Bayes is a simple but very powerful linear classifier. It uses the training data set to find the conditional probability value of each given class vector based on data point xi in the feature data set, and after calculating the probability conditional value of each vector, it calculates the class of the new vector based on its conditional probability. The Naive Bayes approach has a stable classification efficiency in prediction.

##### ED. P.C.Decision Tree

Decision trees usually have three steps: feature selection, decision tree generation, and decision tree pruning. When building a decision tree, a feature of the instance will be tested starting from the root node, and the instance will be assigned to its child nodes according to the test result, at which time each child node corresponds to a value taken for that feature, and so on recursively, the instance will be tested and assigned until it reaches the leaf node, and finally the instance will be assigned to the class of the leaf node.

##### ED. P.C.Random Forest

A random forest is an algorithm that integrates multiple decision trees through the idea of integrated learning. To explain it intuitively, each decision tree is a classifier, then for one input sample, N trees will have N classification results. And the random forest integrates all the classification votes, designating the category with the most votes as the final output. In the process of building the random forest model, after parameter tuning, the maximum depth of the tree is 4 and the minimum number of samples of leaf nodes is 2, with calculating Gini as the criterion.

##### ED. P.C.K-Nearest Neighbor

The principle of KNN is that when predicting a new value x's, it determines which class x belongs to based on what class it is from the nearest K points. Compared to other algorithms, the model training time is fast and it is easy to obtain higher accuracy. When comparing the performance of different knn algorithms, the model has good performance when k is equal to 6. The following section will not show the rest of the knn model training results.

#### ED.P.Cross Validation

To avoid data bias from a single partitioned dataset, k-fold cross-validation is applied. The following picture shows it's basic idea. By dividing the whole data into 10 copies and taking one copy at a time randomly without duplication as the test set, while using the other 9 copies as the training set to train the model, ten biases will be obtained. Their mean value will be the evaluation result, which will be the closest to the real performance of the model[ED.P.CV.01].

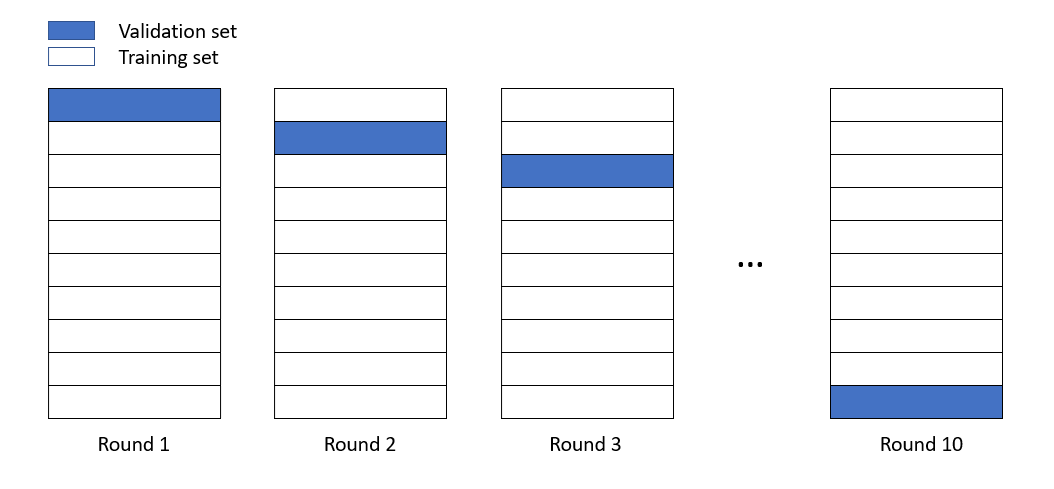


Figure 1: The main idea of 10-fold cross vaildation

#### ED.P.Performance Mertics

This study will evaluate the performance of different models using confusion matrices. The confusion matrix can help us to quickly visualize the model performance to help us to further adjust the parameters or choose a better model. Confusion matrix consists of a N × N matrix(N category), where each row represents the true category to which the data belongs, and the total number of data in each row represents the number of data instances in that category. The values in each column indicate the number of instances of the real data predicted to be in that category. The following figure shows its basic form.

Table : Confusion matrix

|  |  |  |
| --- | --- | --- |
|  | Predicted: At risk of heart disease | Predicted: Healthy |
| Actual: Cases diagnosed with heart disease based on angiographic findings(target = 1) | TP | FN |
| Actual: healthy(target = 0) | FP | TN |

In this case, if a patient(target = 1)is correctly labeled as person at the risk of heart disease, this case will be classified as TP. For those subjects who were correctly classified as healthy(target = 0) by the model, they are classified as TN. At the same time, FP, which is also called type I error, means that the model has made a wrong decision to label a healthy person as cardiac patients. FN shows the cases when cardiac patients are incorrectly predicted to be healthy by predictive models(Type II error).

On the basis of the different classifications, several evaluation criteria are calculcated:

1. *Accuracy:* The percentage of correctly predicted cases out of all cases, which is calculated as:
2. *Sensitivity(Sn):* Sensitivity, also be called recall, measures how apt the model is to detecting events in the positive class. In this case, Sn quantifies how many of the actual heart disease patients are correctly predicted as at the risk of heart disease. is calculated as:
3. *Specificity(Sp):* Specificity measures how exact the assignment to the positive category. In this case, Sp quantifies how many of the healthy ones are correctly predicted in the actual healthy category. is calculated as:
4. *Matthews’ correlation coefficient(MCC):* The Matthews correlation coefficient(MCC) or phi coefficient is a simple and efficient measurment for binary classification problems[ED.P.PM.01]. It has a range between 1 and -1, a coefficient of +1 will be considered the perfect prediction, while the output -1 means a total mismatch.
5. *Area Under The Curve(AUC):* AUC is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve. The higher the AUC, the better the performance of the model at distinguishing between the positive and negative classes.
6. *Processing time:* this crition measures the time efficiency of the model. The processing time of the whole training part will be recorded using the time package in python.

**ED. Experimental Results**

### ED.E.R1. Result of Selected Features by Random Forest Feature Selection Algorithm

Figure 2 shows the ranking of important features by RF FS Algorithm

### ED.E.R2. Result of Selected Features by ANOVA Feature Selection Algorithm.

Result of Selected Features by LASSO Features Selection Algorithm.

### ED.E.R3. Result of Selected Features by LASSO Feature Selection Algorithm.

### ED.E.R4. Results of K-Fold Cross-Validation for Classifiers Performance on Full Features(n = 13).

### ED.E.R5. Results of K-Fold Cross-Validation(k = 10) Classifier Performance on Selected Features(n = 6) by Random Forest Feature Selection Algorithm

### ED.E.R6. Results with K-Fold Cross-Validation of Classifiers Performance on Selected Features(n = 6) by ANOVA Feature Selection Algorithm.

### ED.E.R7. Results with K-Fold Cross-Validation of Classifiers Performance on Selected Features(n = 6) by LASSO Feature Selection Algorithm.

**ED.Conclusion**

Different data mining techniques can be used to identify and prevent cardiovascular disease in patients, with the potential to be put into new artificial intelligence devices and to assist professionals in their judgment. This paper compares the performance of different classifiers for predicting cardiovascular disease in patients: decision trees, random forests, support vector machines, and logistic regression. These techniques are compared on the basis of sensitivity, specificity, accuracy, error rate, true positive rate and false positive rate. Our study shows that the support vector machine model is the best classifier for cardiovascular disease prediction.

**ED.Future work**

In the future, we intend to improve the performance of these basic classification techniques by creating metamodels, which will be used to predict cardiovascular disease in patients.

**WI. Appendix----------**

**Reference**

|  |  |  |
| --- | --- | --- |
| **INDEX** | **Location** | **Reference** |
| 1 | ED.M.D.01 | [UCI Machine Learning Repository: Heart Disease Data Set](https://archive.ics.uci.edu/ml/datasets/heart+Disease) |
| 2 | ED.M.1 | 1. Hao J, Ho TK. Machine Learning Made Easy: A Review of Scikit-learn Package in Python Programming Language. Journal of Educational and Behavioral Statistics. 2019;44(3):348-361. doi:10.3102/1076998619832248 |
| 3 | ED.M.2 | A. Singh and R. Kumar, "Heart Disease Prediction Using Machine Learning Algorithms," *2020 International Conference on Electrical and Electronics Engineering (ICE3)*, 2020, pp. 452-457, doi: 10.1109/ICE348803.2020.9122958. |
| 4 | ED.P.EDA.01 | What is Exploratory Data Analysis? by Prasad Patil Towards Data Science[online]<https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15> |
| 5 | [ED.P.FE.02] | M. Shouman, T. Turner and R. Stocker, "Using data mining techniques in heart disease diagnosis and treatment," *2012 Japan-Egypt Conference on Electronics, Communications and Computers*, 2012, pp. 173-177, doi: 10.1109/JEC-ECC.2012.6186978. |
| 6 | [ED.P.FE.01] | Powell, A., Bates, D., van Wyk, C. & Darren de Abreu, A. A cross-comparison of feature selection algorithms on multiple cyber security data-sets. CEUR Workshop Proc. 2540, 196–207 (2019). |
| 7 | ED.P.CV.01 | Kahramanli, H. & Allahverdi, N. Design of a hybrid system for the diabetes and heart disease, Expert systems with applications, pp. 82-89, 2008. |
| 8 | ED.I.01 | M. Saini, N. Baliyan and V. Bassi, "Prediction of heart disease severity with hybrid data mining," *2017 2nd International Conference on Telecommunication and Networks (TEL-NET)*, 2017, pp. 1-6, doi: 10.1109/TEL-NET.2017.8343565. |
| 9 | ED.I.02 | J. S. Sonawane and D. R. Patil, "Prediction of heart disease using multilayer perceptron neural network," International Conference on Information Communication and Embedded Systems (ICICES2014), 2014, pp. 1-6, doi: 10.1109/ICICES.2014.7033860. |
| 10 | ED.I.03 | Randa El-Bialy, Mostafa A. Salamay, Omar H. Karam, M. Essam Khalifa,Feature Analysis of Coronary Artery Heart Disease Data Sets,Procedia Computer Science,Volume 65, 2015, Pages 459-468, ISSN 1877-0509, <https://doi.org/10.1016/j.procs.2015.09.132>. |
| 11 | ED.I.04 | M. Kirmani, "Cardiovascular disease prediction using data mining techniques," Oriental Journal of Computer Science and Technology, vol. 10, pp. 520-528, 2017. |
| 12 | ED.I.05 | Heart Disease | cdc.gov. [online]<https://www.cdc.gov/heartdisease/index.htm> |
| 13 | ED.P.CV.01 | Kahramanli, H. & Allahverdi, N. Design of a hybrid system for the diabetes and heart disease, Expert systems with applications, pp. 82-89, 2008. |
| 14 | RC.1 | Darmawahyuni, A. Coronary Heart Disease Interpretation Based on Deep Neural Network. Comput. Eng. Appl. J. **2019**, 8. |
| 15 | RC.2 | Chaurasia, Vikas and Pal, Saurabh, Early Prediction of Heart Diseases Using Data Mining Techniques (2013). Caribbean Journal of Science and Technology, Vol. 1, 208-217, **2013** , Available at SSRN: <https://ssrn.com/abstract=2991237> |
| 16 | RC.3 | L. ALI, A. RAHMAN, A. KHAN, M. ZHOU, A. JAVEED, J. A. KHAN: An Automated Diagnostic System for Heart Disease Prediction Based on Chi2 Statistical Model and Optimally Configured Deep Neural Network, IEEE Access, 7 (2019), 34938-34945. |
| 17 | RC.4 | A. UL HAQ, J. PING LI, M. H. MEMON , S. NAZIR, R. SUN: A Hybrid Intelligent System Framework for the Prediction of Heart Disease Using Machine Learning Algorithms , Mobile Information Systems, 2018 (2018), 1–21 |
| 18 | RC.5 | L. VERMA, S. SRIVASTAVA, P. C. NEGI: A Hybrid Data MiningModel to Predict Coronary Artery DiseaseCases Using Non-Invasive Clinical Data , Journal of Medical Systems, 40 (2016), 1–7 |
| 19 | RC.6 | M. A. JABBAR, B. L. DEEKSHATULU, P. CHANDRA:Prediction of heart disease using random forest and feature subset selection , Advances in intelligent systems and computing, 424 (2016), 187–196 |
| 20 | ED.PF.LASSO.01 | Fonti, V., and Belitser, E. (2017). Feature selection using LASSO.  Friston, K. J., Williams, S., Howard, R., Frackowiak, R. S., and Turner, R. (1996). Movement-related effects in fMRI time-series. Magn. Reson. Med. 35, 346–355. doi: 10.1002/mrm.1910350312 |
| 21 | ED.PF.ANOVA.01 | George A Morgan. SPSS for introductory statistics [M].Lawrence Erlbaum Associates, 2004. |
| 22 | [ED.P.PM.01] | Boughorbel S., Jarray F., El-Anbari M. (2017). Optimal classifier for imbalanced data using Matthews correlation coefficient metric. PLoS ONE 12:e0177678. 10.1371/journal.pone.0177678 |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |