Prediction of Coronary Heart Disease using Machine Learning: An Experimental Analysis

Conference Paper · July 2019

DOI: 10.1145/3342999.3343015

4 authors, including:

Fadi Thabtah
University of Huddersfield
100 PUBLICATIONS **2,445** CITATIONS

SEE PROFILE

Rami Mustafa A Mohammad
Imam Abdul Rahman bin Faisal University
15 PUBLICATIONS **425** CITATIONS

SEE PROFILE

Prediction of Coronary Heart Disease using Machine Learning: An Experimental Analysis

Amanda H. Gonsalves
Digital Technologies, Manukau Institute of Technology
Manukau Station Road, Manukau, Auckland, New Zealand
+64 9 9754622
gons16@manukaumail.com

Fadi Thabtah
Digital Technologies, Manukau Institute of Technology
Manukau Station Road, Manukau, Auckland, New Zealand
+64 9 9754621
Fadi.fayez@manukau.ac.nz

Rami Mustafa A. Mohammad
Department of computer Information systems, College of Computer Science and Information Technology
Imam Abdulrahman Bin Faisal University, P.O. Box 1982, Dammam, Saudi Arabia
+966 13 333 1111
rmmohammad@iau.edu.sa

Gurpreet Singh
Digital Technologies, Manukau Institute of Technology
Manukau Station Road, Manukau, Auckland, New Zealand
+64 9 975 4717
Garry.singh@manukau.ac.nz

ABSTRACT
The field of medical analysis is often referred to be a valuable source of rich information. Coronary Heart Disease (CHD) is one of the major causes of death all around the world therefore early detection of CHD can help reduce these rates. The challenge lies in the complexity of the data and correlations when it comes to prediction using conventional techniques. The aim of this research is to use the historical medical data to predict CHD using Machine Learning (ML) technology. The scope of this research is limited to using three supervised learning techniques namely Naïve Bayes (NB), Support Vector Machine (SVM) and Decision Tree (DT), to discover correlations in CHD data that might help improving the prediction rate. Using the South African Heart Disease dataset of 462 instances, intelligent models are derived by the considered ML techniques using 10-fold cross validation. Empirical results using different performance evaluation measures report that probabilistic models derived by NB are promising in detecting CHD.

CCS Concepts
• **Information systems→ Information Systems Applications →Data Mining**

Keywords
Coronary Heart Disease; Data Mining; Machine Learning; Medical Informatics; Supervised Learning

ICDLT 2019, July 5–7, 2019, Xiamen, China

ACM ISBN 978-1-4503-7160-5/19/07…$15.00
https://doi.org/10.1145/3342999.3343015

1. INTRODUCTION
In this age of technology and digitalisation, data has proven to be the fuel for organisations and industries. The healthcare industry is not far behind in this respect. Nowadays, almost all hospitals and medical institutes have their patient's data stored in an electronic format. This includes their medical history, symptoms displayed, diagnosis, duration of illness, recurrences as well as any fatalities. As a result, the quantum of medical data being generated on the daily basis is constantly increasing [12]. However, this wealth of data is often left untapped due to the lack of effective analytical tools, methods and personnel to discover insights and hidden relationships in this data. If the data at hand is used to develop screening and diagnostic models, it will not only reduce the strain on medical personnel but also aid early detection and prompt treatment for patients thereby drastically enhancing the health system. Furthermore, it can also aid in devising a monitory and preventive program for those who might be susceptible to suffering from CHD, based on their medical and family history.

In recent years, researchers and experts working in the medical field have started realising the immense knowledge available in these medical datasets thereby inspiring medical analysis of data for instances of Dementia [13], Alzheimer [6], Tuberculosis screening [10], Autism [22,23,24], Cancer [18], etc. Amidst this vast array, one of the predominant diagnosis in the field of health analysis is CHD [27].

Coronary arteries play a vital role in delivering oxygen to the heart muscle. According to the Southern Cross Medical Care Society of New Zealand, constant build-up of fat or bad cholesterol within these artery walls leads to their narrowing down and eventual blockage thereby giving rise to CHD [20]. A mild-level of blockage might just lead to initial discomfort and alterations in the lifestyle of the person. However, when the flow of oxygen through the coronary arteries is severely hampered, it can prove to be fatal. The risk factors associated with CHD can be a combination of controllable factors like those influenced by one's lifestyle and uncontrollable factors like age, ethnicity, family medical history, etc [20]. Early detection of CHD symptoms can help the patient to control some of these risk

factors through lifestyle changes and/or medication thus preventing this disease from aggravating into a severe form and proving to be fatal.

In this era of Data Science, ML algorithms are constantly being used, across various fields, to gain meaningful insights and leverage the information mined to make decisions [1]. They have not only helped in optimising business in different domains but have also played a vital role in automating and simplifying various processes. ML is a discipline wherein predictive and descriptive models are learnt from data using intelligent techniques [2]. Thabtah et al., 2010 defines ML as an automated method used by systems to learn from data, identify useful patterns and minimize human interference in the decision-making process. ML algorithms can be broadly classified into two main types: Supervised Learning and Unsupervised Learning [15].

Supervised Learning involves training on a labelled dataset using techniques to generate specific knowledge using dependent and independent variable [4]. Here, the algorithm gets certain input variables along with the original output obtained and the algorithm draws comparison between the original and predicted output to find errors and thus modify the model correctly. On the other hand, Unsupervised Learning involves searching for patterns within the dataset without any restrictions on its variables [26]. Since the CHD involves predicting the type of blockage, this problem can be seen as a Supervised Learning Task within ML.

In this research, we look at the existing research on predicting CHD and try to answer the following research questions: Will machine learning improve CHD predictive accuracy ? and if so, which is the most effective ML algorithm for predicting CHD, for a South African Heart Disease dataset? For convenience, we restrict the scope of this essay to three Supervised Learning techniques; Decision Tree, Probabilistic NB, and Support Vector Machine (SVM). The reason for choosing these techniques is their applicability in different domains as well as the different learning methodologies they adopt. These techniques are briefly explained later on in Section IV.

This paper is organised as follows: Section II introduces the problem. Section III sheds the light on relevant literature. Section IV is devoted to the dataset description and the variables used for training phase. In this we also explain the experimental setting and conduct result analysis and finally conclusions are given in Section V.

2. PROBLEM STATEMENT AND METHODOLOGY

Medical diagnosis is an intrinsic and complicated task that demands being carried out with acute precision while taking into consideration various factors [11,21]. Moreover, prediction of CHD is a much complex challenge considering the level of expertise, and knowledge required for accurate result. According to a survey by WHO, medical professionals can correctly predict the heart disease with only 67% accuracy. In New Zealand alone, one out of twenty (that is around 180,000) adults suffer from heart disease and it claims a life every 90 minutes [20]. Considering that the heart disease casualties are expected to rise over the years, there is an immense research scope for predicting CHD.

ML techniques allow the use of intelligent methods across different datasets to reveal useful insights. This reprogrammable ability of ML in exploring, processing and interpreting datasets makes it favourable for decision makers in domains such as medical diagnosis. Since detecting CHD involves training a model based on historical dataset, ML seems to be an appropriate technology to deal with this problem.

In this context, a number of independent variables such as age, gender, medical history and symptoms among others will be used along with a dependent variable (CHD class) during the training phase to build a classification model. This model is then employed to forecast the dependent variable value in test dataset as accurately as possible.

Figure 1 shows the methodology adopted in this research to deal with predicting CHD. The historical medical data are obtained from South African Heart Disease - KEEL (Knowledge Extraction based on Evolutionary Learning) [28]. The initial dataset is pre-processed to eliminate any possible noise that may impact the results of the predictive analyses such as data conversion from .dat to .csv and missing values replacements if any.

The processed dataset is used to train the predictive models using multiple ML techniques, i.e. DT, Probabilistic NB and SVM [9,17,16]. Then, based on the classification models generated by these ML techniques the medical personnel and experts will be able to predict if a certain patient who shows the underlying traits of CHD does really suffer from CHD or not. Evaluation of the classification models are done using various evaluation metrics including error rate, accuracy, sensitivity and specificity among others. More details on the model evaluation and results analyses are discussed in Section IV.

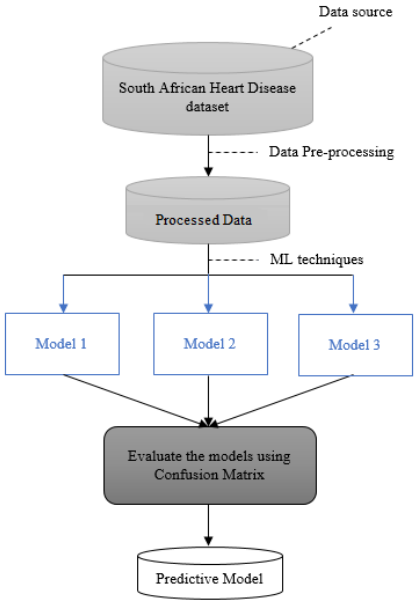

Figure 1 Proposed Methodology

3. LITERATURE REVIEW

Various research initiatives have been undertaken by experts, academic scholars and data science community in predicting and screening of medical data for various diseases. Multiple ML algorithms have been used in past researches to carry out these predictions. We will be reviewing relevant research works before we go ahead with our analysis on dataset.

In order to address the need of medical society to develop CHD prediction technique, [29] built a data mining model to predict CHD, using 100 CHD records and recording the survival rate information. The authors used Support Vector Machine (SVM), Artificial Neural Network (ANN) and Decision Trees (DT) on 502 instances using 10-fold cross validation technique and confusion matrix to measure the model performance. The accuracy obtained by his study was 92.1%,

91.0% and 89.6% for SVM, ANN and DT respectively. Thus, SVM proved to be a good classifier model.

Apte et al [3] carried out prediction of heart disease by using a dataset having 13 attributes, which included main attributes like sex, blood pressure and cholesterol. The authors added two more attributes: smoking and obesity. ANN, DT and NB classification techniques were used, and the results obtained pinpointed that ANN had the highest predictive accuracy on the utilized dataset.

Jenzi [8] established a relation between key patterns in the dataset of 14 attributes by using association rule mining. The authors built different classifier models using classification techniques like DT, NB and ANN. Microsoft .NET platform was used to build the graphical user interface (GUI) with the use of IKVM interface and Java libraries to form interconnections. The results of the models were depicted via the receiver operating characteristic (ROC) curves. The results showed that the area under ROC of ANN was slightly above 80% which was better than Naïve Bayes and DT algorithms.

Hazra et al [7] proposed a guide in using ML techniques to predict CHD. The authors initially reviewed CHD and then discussed relevant research studies on the application of ML to predict CHD. They then carried out a comparative study on various research papers concerning prediction of CHD and the methods devised in them.

Karthiga et al [11] conducted a research study to accurately predict the presence of heart disease using public dataset of 573 records. The authors adopted DT and NB classification technique and processed the dataset. Using MATLAB data analysis tool, the authors replaced all missing values and then generated results with respect to accuracy to determine models' performance. The reported results indicated that DT is more accurate than NB on the dataset considered.

Manimekalai [14] investigated the applicability of different ML techniques to predict CHD. Results derived showed that SVM classifier achieved 95% accuracy and was more superior than conventional ML methods.

4. DATA AND RESULTS ANALYSIS

4.1 DATA DESCRIPTION

The dataset for this research has been obtained from South African Heart Disease which is a subset of a larger dataset. It contains 462 instances (observations) and 10 attributes in all (shown in Table 1), of which 9 are independent factors and 1 variable, i.e. CHD is the dependent variable or labelled class. The dataset is a retrospective sample of males in a heart-disease high-risk region of the Western Cape in South Africa-KEEL [28] where the labelled class CHD has two predictive outcomes: positive (1) and negative (0).

Each high-risk patient was monitored in this study and the attributes obtained were as follows: systolic blood pressure (sbp), cumulative tobacco in kg (tobacco), bad cholesterol also known as low density lipoprotein cholesterol (ldl), adiposity, family history of heart disease (famhist), type-A behaviour (typea), obesity, current alcohol consumption (alcohol), and age at onset (age).

In order to get a clear understanding, we define few of the terms below [19]:

- Sbp: It is the blood pressure when the heart is contracting.
- Adiposity: It is measured as percent of body fat
- Type-A behaviour: It is characteristic of a person who is competitive, impatient and angry.
- Obesity: It is represented as Body Mass Index (BMI) which is calculated by dividing the weight of the person by the square of his height.

Figure 2 shows the first five instances of the dataset under study.

Table 1 Attributes Description of the dataset

Attribute	Domain	Data Type	Missing values?
Sbp	[101,218]	Decimal	No
Tobacco	[0.0,31.2]	Decimal	No
Ldl	[0.98,15.33]	Decimal	No
Adiposity	[6.74,42.49]	Decimal	No
Famhist	{Present, Absent}	Text	No
Typea	[13,78]	Decimal	No
Obesity	[14.7,46.58]	Decimal	No
Alcohol	[0.0,147.19]	Decimal	No
Age	[15,64]	Decimal	No
CHD (class)	[0,1]	Binary	No

	A	B	C	D	E	F	G	H	I	J
	Sbp	Tobacco	Ldl	Adiposity	Famhist	Typea	Obesity	Alcohol	Age	Chd
	160	12	5.73	23.11	Present	49	25.3	97.2	52	YES
	144	0.01	4.41	28.61	Absent	55	28.87	2.06	63	YES
	118	0.08	3.48	32.28	Present	52	29.14	3.81	46	NO
	170	7.5	6.41	38.03	Present	51	31.99	24.26	58	YES
	134	13.6	3.5	27.78	Present	60	25.99	57.34	49	YES

Figure 2 CHD Dataset Sample Instances

4.2 DATA PRE-PROCESSING

The South African Heart Disease dataset obtained from KEEL ((Knowledge Extraction based on Evolutionary Learning), 2004-2018) is available in .dat file format. However, the data analysis tools that we used, i.e. Waikato Environment for Knowledge Analysis (WEKA) [5] requires the file to be in.csv else .arff file format. As a result, we first converted the data from .dat to .csv. Moreover, in order to avoid confusion and gain clarity during analysis, we used the 'If…then' feature in Microsoft Excel to convert the data type of 'Class' from Integer to Factor. Thus, all the value '1' were replaced by 'Yes' and all value '0' were replaced by 'No'. Then we uploaded the processed data in WEKA.

4.3 EXPERIMENTAL SETTING

We used WEKA, an open source data analysis software version 3.8.3 to conduct the experiments. WEKA supports various ML tasks ranging from pre-processing, classification, regression, feature selection, clustering, association and visualisation. It is an easy and user-friendly software to use.

The estimation methodology used for testing the ML technique against the considered dataset was 10-fold cross validation. In this, the entire dataset is split into 10 folds and at each iteration, 9 folds are used for training the ML models and the remaining 1-fold is used for evaluation. This process is repeated 10 times and the error rate is recorded at each iteration. The final error of the model is the average of all errors produced during the 10 runs.

The ML techniques employed for the prediction of CHD are:

- Decision Tree – C4.5 (J48)
- Naïve Bayes Algorithm
- Support Vector Machine (SVM)

DT is based on the C4.5 algorithm [17]. The algorithm selects the best splitting attribute as a root by implementing the information gain ratio impurity method and sorting the data at every node. NB uses the theory of probability to carry out the classification of data. As a result, the most important assumption here is the independence of the predicting variable. In SVM, data is mapped to points in multi-dimensional space and labelled to different classes using maximum margin hyper-plane. The side of the margin that the new data falls on predicts the future outcome.

4.4 EVALUATION MEASURES

The performance of the classification models derived by the ML is measured using the confusion matrix. The confusion matrix is a contingency table that displays the number of instances assigned to each class thus allowing us to calculate the classification accuracy, sensitivity, specificity, true positives (TPs), true negatives (TNs), false positives (FPs), and false negatives (FNs) among others [25]. There can be two or more classes involved, however we have only two classes in the dataset thereby giving us a 2x2 confusion matrix for each classification model (Figure 3). For the experiment under consideration,

Class a = Yes (has CHD)
Class b = No (no CHD)

		Predicted Class	
Actual Class		a (has CHD)	b (no CHD)
	a (has CHD)	TP	FN
	b (no CHD)	FP	TN

Figure 1 Layout of 2x2 Confusion Matrix.

Let us understand the terms TP, FP, FN and TN used in confusion matrix.

- True Positive (TP): Number of patients that are predicted to have CHD and do actually have CHD.
- False Positive (FP): Number of patients that are predicted to have CHD and do not actually have CHD.
- False Negative (FN): Number of patients that are predicted to not have CHD but do actually have CHD.
- True Negative (TN): Number of patients that are predicted to not have CHD and do not actually have CHD.

One of the most common measure of performance comparison in classification analysis is Accuracy. It is the number of predictions made correctly out of the total predictions made by the model.

$$Accuracy = \frac{(TP+TN)}{(TP+FP+FN+TN)} \quad (1)$$

Another performance measure that goes hand in hand with accuracy is Error Rate. It is the total number of incorrect predictions made by the model with respect to the total number of predictions, after training the classifier with a given dataset.

$$Error\ Rate = 1 - Accuracy \quad (2)$$

The next measure of performance is Sensitivity. It measures the number of instances correctly predicted as positive by the classifier out of the total number of instances that are actually positive. Sensitivity is also known as Recall or True Positive Rate.

$$Sensitivity = \frac{TP}{(TP+FN)} \quad (3)$$

The last measure of performance that is used is Specificity. It measures the number of correct negative predictions by the classifier out of the total number of instances that are actually negative. It is also known as True Negative Rate.

$$Specificity = \frac{TN}{(TN+FP)} \quad (4)$$

4.5 RESULTS ANALYSIS

The classification accuracy results for the three classification algorithms i.e. DT, NB and SVM are summarized in Figure 4.a.

Based on the derived results, NB marginally outperforms SVM and DT with respect to accuracy. While all three models tend to show accuracy rates of more than 70%, accuracy alone can't be considered as the performance criterion for the underlying study. This is since the bi-variable response of the labelled class is unequal. Out of the original 462 instances, only 160 patients are said to have CHD whereas the remaining 302 individuals do not suffer from CHD. This discrepancy might sublimely influence the accuracy rate as the model can predict all values of the majority class and thus achieve an overall high accuracy while blinding out the mis-predictions occurring in the minority class. In order to avoid this imbalance affecting our performance measurement, we recorded other measures including sensitivity and specificity (Figure 4.b).

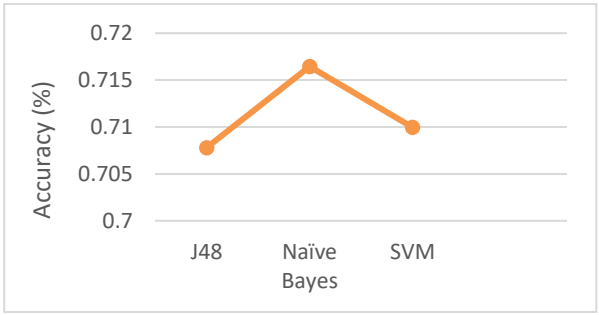

Figure 2a Accuracy of models built using ML techniques

We can see in Figure 4.b that there is only a slight difference between the performance of J48 and SVM models on the CHD dataset, with respect to Sensitivity and Specificity measures. According to the confusion matrix results (Figure 5), J48 and SVM misclassified only 54 and 53 instances respectively, as having CHD when in reality they didn't suffer from CHD. This in turn reduced the FPs and subsequently improved the Specificity of both the models to around 82%. This means that J48 and SVM are effective models at correctly predicting that a person doesn't have CHD. On the contrary, J48 and SVM both have Sensitivity of merely 49% which is too low. The confusion matrix indicates that, 81 instances which should be in 'Has CHD' class have been misclassified by J48 and SVM as belonging to 'No CHD' class, that is a shocking 50% FN rate. This makes the two models inappropriate, especially, since correctly predicting the presence of CHD when it is actually present is critical.

Figure 4.b shows that the sensitivity results of the models have a similar curve as their accuracy results. The final model NB is a more sensitive model at 63% as compared to the other models under consideration. The Sensitivity rate attained by NB model was better than others because of reduced FN rate of 36%. This happened

because the misclassification rate of the NB model reduced by approximately 27%. That is, it misclassified 59 instances, as belonging to 'No CHD' class when they actually were in the 'Has CHD' class. Thus, NB is more sensitive at predicting the presence of CHD correctly. On the other hand, Figure 4.b also shows that NB has a Specificity of 76.16% which is lower than that for J48 and SVM. This is because NB had approximately 25% higher FNs than other models. NB misclassified 72 instances as 'Has CHD' instead of 'No CHD. Thus, when it comes to correctly predicting that a person doesn't have CHD, NB has less Specificity rate as compared to those of J48 and SVM.

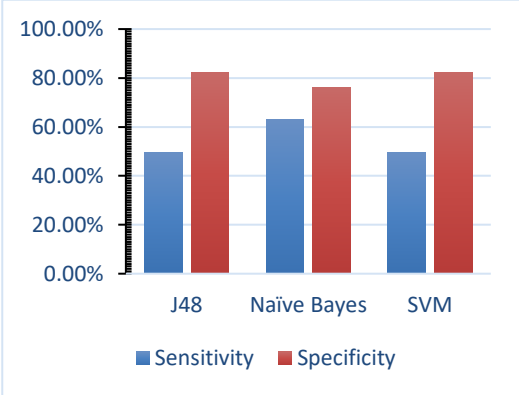

Figure 4b Sensitivity and Specificity rates of models built using ML techniques.

When we calculate the ratio of FP and FN, that is misclassified 'No CHD' and misclassified 'Has CHD', for all the three models NB proved to give more reliable results in comparison to J48 and SVM algorithms. Overall considering that it is better to incorrectly predict the presence of CHD than to fatally predict 'No CHD' when it is actually present; among the three models under study, NB Algorithm has proven to be the most effective algorithm for predicting CHD, for the South African Heart Disease dataset.

Generally, in medical research, the Sensitivity and Specificity rate of 80% or above is considered to be an acceptable rate. J48 and SVM did give a Specificity rate of around 82%, which is acceptable. However, at the same time its Sensitivity rate was less than 50% which is low. Among the models built for the dataset under consideration, NB did turn out to be the best model. But its Sensitivity and Specificity rate of less than 80%, i.e. 63% and 76% respectively, calls for further study. Increasing the number of instances under study and balancing the bi-variable responses of the labelled class might help to improve the performance of the ML models.

5. CONLCUSIONS & LIMITATIONS

Early screening and detection of diseases not only benefits the patients by accelerating their treatment but also helps the medical institutes and officials to better distribute their resources and devise ways to altogether prevent or at least reduce its occurrence. In some cases of fatal diseases, early detection leads to increased probability of cure. Many research approaches have been taken with respect to prediction and screening of diseases using medical analysis. There are various number of ML algorithms available to facilitate prediction of CHD. This research was an attempt to highlight a few of these available techniques of prediction and the performance measures associated with them. The aim of this research has been to find the most effective ML models in predicting the presence of CHD using the South African Heart Disease dataset.

In this research, we restricted our study to using SVM, DT and NB algorithms. These algorithms were decided based on literature review and attributes of the techniques. Different experiments have been conducted on real CHD dataset using the considered ML algorithms. The performance of the obtained models was analysed using evaluation measures of Accuracy, Sensitivity, Specificity, TPs, TNs, FPs and FNs. NB achieved the highest accuracy amongst the three models. SVM and DT J48 outperformed NB with a Specificity rate of 82% but proved to have an unacceptable Sensitivity rate of less than 50%. While NB Algorithm didn't reach the threshold of 80% Specificity and Sensitivity rate, it did turn out to be the best classifier for the considered dataset as its predictive rate is better that those of J48 and SVM algorithms at least on the considered dataset.

Based on the obtained ML results from the CHD dataset, future research needs to be carried out to improve the performance of the model, especially to increase the Sensitivity and Specificity rates. One such attempt could be to check if using unsupervised learning techniques before undertaking prediction, will enhance the model furthermore in terms of its prediction performance. Thereafter, the prediction model obtained through the research conducted can be used to develop a mobile application which will help people to track their health and thereby lead to early detection for CHD.

Possible shortcomings of this study that there was not enough instances and the CHD class was unbalanced. Therefore, treating these issues may enhance performance of ML algorithms.

Machine Learning Technique: DT J48				
		Predicted		
		Has CHD	No CHD	
Actual	Has CHD	TP= 79	FN= 81	160
	No CHD	FP= 54	TN= 248	302
		133	329	
Machine Learning Technique: Naïve Bayes (NB)				
		Predicted		
		Has CHD	No CHD	
Actual	Has CHD	TP= 101	FN= 59	160
	No CHD	FP= 72	TN= 230	302
		173	289	
Machine Learning Technique: SVM				
		Predicted		
		Has CHD	No CHD	
Actual	Has CHD	TP= 79	FN= 81	160
	No CHD	FP= 53	TN= 249	302
		133	329	

Figure 3 Confusion Matrix result for models built using ML techniques.

REFERENCES

[1] Abdelhamid N., Thabtah F., (2014) Associative Classification Approaches: Review and Comparison. Journal of Information and Knowledge Management (JIKM). Vol. 13, No. 3 (2014) 1450027.

[2] Abdelhamid N., Ayesh A., Thabtah F. (2012) An Experimental Study of Three Different Rule Ranking Formulas in Associative Classification Mining. Proceedings of the 7th IEEE International Conference for Internet Technology and Secured Transactions (ICITST-2012), pp. (795-800), UK.

[3] Apte, C. S. (2012). Improve study of Heart Disease prediction system using Data Mining Classification techniques.

International journal of computer application, 47(10), 44-48. doi:10.5120/7228-0076

[4] Hadi W., Thabtah F., Mousa S., ALHawari S., Kanaan G., Ababnih J. (2008). A Comprehensive Comparative Study using Vector Space Model with K-Nearest Neighbor on Text Categorization Data. Journal of Applied Sciences, volume 2:1-pp. 12-24. Science Alert.

[5] Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P., Witten I. (2009) The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.

[6] Hassan, S. A., & Khan, T. (2017). A Machine Learning Model to Predict the Onset of Alzheimer Disease using Potential Cerebrospinal Fluid (CSF) Biomarkers. International Journal of Advanced Computer Science and Applications, 8(12), 124-131.

[7] Hazra, A., Mandal, S. K., Gupta, A., & Mukherjee, A. (2017). Heart Disease Diagnosis and Prediction Using Machine Learning and Data Mining Techniques: A Review. Advances in Computational Sciences and Technology, 10(7), 2137-2159. Retrieved from https://www.researchgate.net/publication/319393368_Heart_Disease_Diagnosis_and_Prediction_Using_Machine_Learning_and_Data_Mining_Techniques_A_Review

[8] Jenzi, P. D. (2013). A Reliable Classifier Model Using Data Mining Approach for Heart Disease Prediction. International Journal of Advanced Research in Computer Science and Software Engineering, 3.

[9] John, G. H., & Langley, P. (1995). Estimating Continuous Distributions in Bayesian Classifiers. Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence (pp. 338-345). San Mateo : Morgan Kaufmann Publishers.

[10] Kalhori, S. R., & Zeng, X.-J. (2013). Evaluation and Comparison of Different Machine Learning Methods to Predict Outcome of Tuberculosis Treatment Course. Journal of Intelligent Learning Systems and Applications, 5(3), 184-193. doi:10.4236/jilsa.2013.53020

[11] Karthiga, A. S., Mary, M. S., & M.Yogasini. (2017). Early Prediction of Heart Disease Using Decision Tree Algorithm. International Journal of Advanced Research in Basic Engineering Sciences and Technology, 3(3).

[12] Kierkegaard, P. (2011). Electronic health record: Wiring Europe's healthcare. Computer Law & Security Review, 27(5), 503-515. Retrieved from https://www.sciencedirect.com/science/article/pii/S0267364911001257?via%3Dihub

[13] King, M. A. (2018). Dementia could be detected via routinely collected data, new research shows. Retrieved from University of Plymouth Website: https://www.plymouth.ac.uk/news/dementia-could-be-detected-via-routinely-collected-data-new-research-shows

[14] Manimekalai. K. (2016). Prediction of Heart Diseases using Data Mining Techniques. International Journal of Innovative Research in Computer and Communication Engineering, 4(2), 2161-2168. Retrieved from http://www.ijircce.com/upload/2016/february/73_27_Prediction.pdf

[15] Mohammed R., Thabtah F., McCluskey L., (2013) Intelligent Rule based Phishing Websites Classification. Journal of Information Security (2), 1-17. ISSN 17518709. IET.

[16] Platt, J. C., & Nitschke, R. v. (1998). Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines.

[17] Quinlan, J. R. (1993). C4.5: Programs for Machine Learning. San Mateo, California: Morgan Kaufmann Publishers.

[18] Reddy, P. V., & Suryachandra, P. (2016). Comparison of machine learning algorithms for breast cancer. 2016 International Conference on Inventive Computation Technologies (ICICT), (pp. 1-6). doi:10.1109/INVENTIVE.2016.7830090

[19] Sivakumar, S. (n.d.). Prediction of Coronary Heart Disease by learning from retrospective study. Retrieved from GitHub: http://srisai85.github.io/CHD/heart_attack.html

[20] Southern Cross. (2018). Coronary heart disease - causes, symptoms, prevention. Retrieved from Southern Cross: https://www.southerncross.co.nz/group/medical-library/coronary-heart-disease-causes-symptoms-prevention

[21] Thabtah F., Peebles D. (2019) A new machine learning model based on induction of rules for autism detection. Health Informatics Journal, 1460458218824711.

[22] Thabtah F. (2018a) An Accessible and Efficient Autism Screening Method for Behavioural Data and Predictive Analyses. Health Informatics Journal. 19:1460458218796636. doi: 10.1177/1460458218796636. 2018.

[23] Thabtah F. (2018b) Machine learning in autistic spectrum disorder behavioral research: A review and ways forward Informatics for Health and Social Care 43 (2), 1-20.

[24] Thabtah F, Kamalov F., Rajab K (2018) A new computational intelligence approach to detect autistic features for autism screening. International Journal of Medical Infromatics, Volume 117, pp. 112-124.

[25] Thabtah F. (2017) Autism Spectrum Disorder Tools: Machin Learning Adaptation and DSM-5 Fulfillment: An Investigative Study. Proceedings of the2017 International Conference on Medical and Health Informatics (ICMHI 2017), pp. 1-6. Taichung, Taiwan. ACM.

[26] Thabtah F., Hammoud S (2013) MR-ARM: A MapReduce Association Rule Mining. Journal of Parallel Processing Letter, 23 (3) 1-22, 1350012. World Scientific.

[27] World Health Organization. (2005). Preventing Chronic Diseases a vital investment. Switzerland: WHO Press.

[28] Knowledge Extraction based on Evolutionary Learning. (2004-2018). South African Heart data set. Retrieved from KEEL (Knowledge Extraction based on Evolutionary Learning): https://sci2s.ugr.es/keel/dataset.php?cod=184.

[29] Yanwei X, W. J. (2007). Combination data mining. Proceedings International Conference on Convergence Information Technology, (pp. 868-872).