

A MACHINE LEARNING MODEL FOR THE PREDICTION OF 1-YEAR MORTALITY AFTER HEART TRANSPLANTATION IN ADULTS WITH CONGENITAL HEART DISEASE

Poster Contributions

For exact presentation time, refer to the online ACC.22 Program Planner at <https://www.abstractsonline.com/pp8/#!/10461>

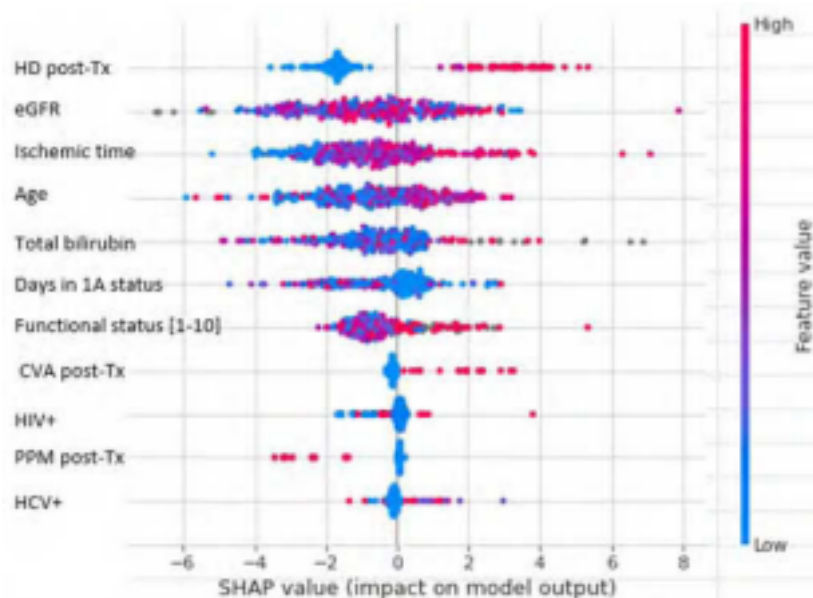
Session Title: Heart Failure and Cardiomyopathies Flatboard Poster Selections: Mechanical Support and Cardiac Transplantation Abstract Category: 10. Heart Failure and Cardiomyopathies: Mechanical Support and Cardiac Transplantation

Authors: [Maria Emfietzoglou](#), Athanasios Siouras, Jef Van den Eynde, Serafeim Moustakidis, Ilias Doulamis, George Giannakoulas, Dimitrios Vasileios Avgerinos, Alexandros Briasoulis, Polydoros Kampaktsis, Columbia University Irving Medical Center, New York, NY, USA, University of Oxford, Oxford

Background: Machine learning (ML) can be used to assist clinical decision-making. We developed a ML model for the prediction of 1-year mortality after heart transplantation (HT) in adults with congenital heart disease.

Methods: The United Network for Organ Sharing (UNOS) database was queried from 2000-2020 for adults with congenital heart disease who underwent isolated HT and had at least 1-year of follow-up. The cohort was randomly split into derivation (70%) and validation (30%) datasets used to train and test a CatBoost decision tree model, respectively. Recipient and donor characteristics were used. The primary outcome was 1-year mortality. Explainability analysis with Shapley Additive exPlanations (SHAP) was performed.

Results: A total of 1,032 recipients were included in the study (35 ± 13 years, 61% males). At 1 year after HT, there were 205 deaths (19.9%). After feature selection, area under the curve and predictive accuracy for the final ML model were 0.82 and 77% respectively. The impact of each model variable for each individual prediction in the validation dataset is represented by its SHAP value.



Conclusion: A ML model developed using data from the UNOS database showed satisfactory predictive accuracy for 1-year mortality after HT in adults with congenital heart disease. Explainability analysis helps interpret the results in a clinical manner.

Computers in Biology and Medicine 137 (2021) 104813

Contents lists available at [ScienceDirect](https://www.sciencedirect.com)

Computers in Biology and
Medicine



Interpretable prediction of 3-year all-cause mortality in patients with heart failure caused by coronary heart disease based on machine learning and SHAP

Ke Wang^{a,b,c}, Jing Tian^d, Chu Zheng^{a,c}, Hong Yang^{a,c}, Jia Ren^a, Yanling Liu^{a,c}, Qinghua Han^{d,**}, Yanbo Zhang^{a,c,*}

^a Department of Health Statistics, School of Public Health, Shanxi Medical University, Taiyuan, People's Republic of China

^b Department of Epidemiology and Biostatistics, Xuzhou Medical University, Xuzhou, People's Republic of China

^c Shanxi Provincial Key Laboratory of Major Diseases Risk Assessment, Shanxi Medical University, Taiyuan, People's Republic of China ^d

Department of Cardiology, The First Affiliated Hospital of Shanxi Medical University, Taiyuan, People's Republic of China

and stratify the risk of 3-year all cause mortality in patients with HF caused by CHD. ML combined with SHAP could provide an explicit explanation of individualized risk prediction and give physicians an intuitive understanding of the influence of key features in the model.

ARTICLE INFO

Keywords:

Interpretable model
Heart failure
Machine learning
SHAP value

1. Introduction

ABSTRACT

Background: This study sought to evaluate the performance of machine learning (ML) models and establish an explainable ML model with good prediction of 3-year all-cause mortality in patients with heart failure (HF) caused by coronary heart disease (CHD).

Methods: We established six ML models using follow-up data to predict 3-year all-cause mortality. Through comprehensive evaluation, the best performing model was used to predict and stratify patients. The log-rank test was used to assess the difference between Kaplan–Meier curves. The association between ML risk and 3-year all cause mortality was also assessed using multivariable Cox regression. Finally, an explainable approach based on ML and the SHapley Additive exPlanations (SHAP) method was deployed to calculate 3-year all-cause mortality risk and to generate individual explanations of the model's decisions.

Results: The best performing extreme gradient boosting (XGBoost) model was selected to predict and stratify patients. Subjects with a higher ML score had a high hazard of suffering events (hazard ratio [HR]: 10.351; $P < 0.001$), and this relationship persisted with a multivariable analysis (adjusted HR: 5.343; $P < 0.001$). Age, N terminal pro-B-type natriuretic peptide, occupation, New York Heart Association classification, and nitrate drug use were important factors for both genders.

Conclusions: The ML-based risk stratification tool was able to accurately assess

occupation of medical resources and high mortality, HF poses a heavy economic and social burden. Seeking effective measures to improve

Heart failure (HF), which is characterized by cardiac systolic or diastolic dysfunction, is a major public health problem worldwide [1] and has become one of the deadliest cardiovascular diseases of the 21st century [2]. There are many causes of HF, but coronary heart disease (CHD), as the world's leading cause of heart failure, is still associated with high morbidity and mortality [3]. Patients with HF caused by CHD often have poor prognosis and high mortality due to poor physical function and a prolonged disease duration [4]. Owing to the long-term

patient prognosis and reduce mortality has become an important goal of HF management. Therefore, obtaining accurate mortality risk predictions for patients with HF caused by CHD and understanding what drives these predictions is vitally important to determine targeted interventions in clinical settings.

Machine learning (ML) algorithms provide researchers with powerful tools. It uses statistical methods in large datasets to infer relationships between patient attributes and outcomes and allows for objective

* Corresponding author. Department of Health Statistics, School of public health, Shanxi Medical University, Yingze district 56 New South Road, Taiyuan, China.

** Corresponding author.

E-mail address: sxmuzyb@126.com (Y. Zhang).

integration of data to predict outcomes. ML has been used in many medical-related fields, such as diagnosis, outcome prediction, treatment, and medical image interpretation [5,6], and it has also been used to predict adverse outcomes in patients with HF by integrating clinical and other data in recent studies [7–9]. However, there is still a lack of research on ML for the prognosis of HF caused by CHD, especially me

dium and long-term mortality risk prediction. Moreover, despite the promising performance of ML in previous studies, evidence on its application in a real-world clinical setting and explainable risk prediction models to assist disease prognosis are limited [10,11]. Because of the “black-box” nature of ML algorithms, it is difficult to explain why certain predictions should be made about patients; that is, what specific characteristics of the patient lead to a given prediction. The lack of interpretability has so far limited the use of more powerful ML approaches in medical decision support [12], and the lack of intuitional understanding of ML models is also one of the major obstacles to implementation of ML in the medical field [13].

To solve these disadvantages, this study combined the advanced ML algorithm with a framework based on SHapley Additive exPlanations (SHAP) [14]. In addition to improving the accuracy of predicting 3-year mortality risk in patients with HF caused by CHD, it provides intuitive explanations that lead patients to predict risk, thereby helping clinicians better understand the decision-making process for assessing disease severity and maximizing opportunities for early intervention. This is an important step forward for ML in medicine [12] and will help develop interpretable and personalized risk prediction models.

1.1. Related work

ML technology does not require assumptions about input variables and their relationship with output. The advantage of this completely data-driven learning without relying on rule-based programming makes ML a reasonable and feasible approach [15]. Among various data-driven methods, the performance of computational models that predict health outcomes has been improved by applying more sophisticated approaches, investigating techniques in the areas of statistics and ML [16]. An increasing number of studies are applying ML to predict cardiovascular disease [17], and various risk models can be used to assess the risk of patients across the HF spectrum [18,19]. Nowadays, people’s interest in using the interpretation and tree ensemble models has grown to the development of mortality prediction models, such as random forest (RF) and Gradient Boosting Decision Tree [20,21]. Although tree ensemble models are more accurate and can also provide a ranking of feature importance, they cannot tell users whether these important factors are protective or dangerous, while logistic regression (LR) can. The “black-box” characteristics of ML algorithms make it difficult to understand and correct errors when they occur [22]. Meanwhile, improving collaboration between humans and artificial intelligence is critical for applications where explaining ML model predictions can enhance human performance [23]. A balance between model accuracy and interpretation is often difficult to achieve, and the probabilities of risk that the model outputs are not easily understood by most physicians.

2. Methods

2.1. Study population

This was a prospective, multi-center, cohort study to predict the 3-year risk of all-cause mortality in patients with HF caused by CHD. Patients were enrolled in a regional cardiovascular hospital and the cardiology department of a medical university hospital in Shanxi Province, China from January 2014 to June 2019 according to the inclusion and exclusion criteria.

The inclusion criteria were as follows: (1) aged ≥ 18 years; (2) diagnosed with HF according to the guideline for the diagnosis and

(NYHA) classification II–IV disease; (4) diagnosis of CHD [4]; and (5) underwent HF treatment while hospitalized. Patients who had an acute cardiovascular event within 2 months prior to admission or were unable or refused to participate in the program for any reason were excluded.

2.2. Data collection

Patient information was collected according to the case report form of chronic HF (CHF-CRF) developed by this research group based on the content of the case records and HF guidelines [25]. The CHF-CRF included patients’ demographics, medical history, physical status and vitals, currently used medical therapy, echocardiography results, electrocardiography results, and laboratory parameters. All patients were followed up by a trained specialist over the telephone every 6 months after discharge to record survival information for patients with HF. Based on inclusion and exclusion criteria, we collected a total of 5188 patients with HF caused by CHD, and finally identified 1562 patients with a follow-up duration of >3 years or who died.

The cohort used in this study was from a prospective cohort study of CHF registered by our research group in the Chinese Clinical Trial Registry (ChiCTR2100043337).

2.3. Study outcomes

The primary endpoint of the study was all-cause mortality throughout 3 years of follow up. All-cause mortality was defined as death due to any cause.

2.4. Feature selection and data preprocessing

Our structured database initially contained hundreds of clinical variables (so-called “features” in ML). Features with a missing percent age of not more than 30% were retained and filled in with the method of missForest [26]. Because the range of different features widely varied and some of the used algorithms required quantitative data normalization, Min-Max normalization was used, and multi-category variables were processed by One-Hot [27]. After the single-factor preliminary screening, the recursive feature elimination (RFE) based on RF with five-fold cross-validation (CV) was used to screen the overall features. The main idea of RFE is to build a model, select the best feature, pick out the selected feature, and then repeat this process for the remaining features until all the features are traversed.

2.5. Model development

We developed six ML models using follow-up data to predict 3-year all-cause mortality. In addition to five commonly used models [28], including LR, k-nearest neighbors (KNN), support vector machines (SVM), naive Bayesian (NB), and multi-layer perceptron (MLP), we introduced extreme gradient boosting (XGBoost). XGBoost is an optimized implementation of gradient boosting. It is based on the ensemble of weak learners and has the characteristics of high bias and low variance. XGBoost uses a second-order Taylor series to approximate the value of the loss function, and further reduces the possibility of over-fitting through regularization [29]. According to whether the endpoint occurred, stratified random sampling was used to divide 1562 patients into a training set and a test set in a 4:1 ratio. The training set was pretreated using the synthesizing minority oversampling technology combined with edited nearest neighbors (SMOTE + ENN) technique [30] to balance them between positive and negative categories. The synthetic minority oversampling technique combined with the editing nearest neighbor (SMOTE + ENN) technique [30] was used to preprocess the training set to achieve a balance between positive and negative categories. A Grid Search method with five-fold CV was used to optimize the hyper-parameters of ML models (details in [Supplementary Table 1](#)). Finally, the performance of each model was evaluated and compared in

the test set. To obtain a more robust performance estimate, avoid reporting biased results and limit over-fitting, we repeated the persis
Computers in Biology and Medicine 137 (2021) 104813

2.7. Statistical analysis

tence method 100 times with different random seeds and calculated the average performance in these 100 repetitions [31] (Fig. 1). Through comprehensive evaluation of multiple evaluation indicators, the best performing model among the six models was selected for further risk prediction and stratification. Furthermore, the optimal model was developed for men and women separately to assess gender-based dif
ferences in the prognostic importance of covariates.

2.6. Model interpretation and feature importance

ML models are often considered as black boxes because it is difficult to interpret why an algorithm provides accurate predictions for a particular patient cohort; therefore, we introduced the SHAP value in this study. SHAP is a unified framework proposed by Lundberg and Lee [14] to interpret ML predictions, and it is a new approach to explain various black-box ML models. It has previously been validated in terms of its interpretability performance [11,32]. SHAP can perform local and global interpretability simultaneously, and it has a solid theoretical foundation compared with other methods [12]. We leveraged SHAP to provide an explanation for our predictive model, which includes related risk factors that lead to death in patients with HF caused by CHD. To determine the main predictors of

all-cause mortality in the patient population, we calculated the importance of ranking features from the final model.

All analyses and calculations were performed using Python Version 3.6.5 (imblearn, sklearn, xgboost, lifelines, and shap packages) and R version 4.0.2 (survival and survminer packages).

Multiple evaluation indices, including sensitivity, specificity, F1-score, and area under the receiver operating characteristic curve (AUC) were used to comprehensively evaluate the discrimination of ML models. The Brier score [9] was used to evaluate model calibration. The evaluation indices of the six models were compared to one-way analysis of variance and multiple comparisons of least-significant difference. The highest Youden's index was used to define an optimal cut-off value and to separate patients with a low and high ML risk. The log-rank test was then used to assess the difference between Kaplan–Meier curves. The association between ML risk and 3-year all-cause mortality was also assessed using multivariable Cox regression. The statistical significance was based on a two-tailed P value of ≤ 0.05 .

3. Results

3.1. Patient characteristics

A total of 1562 patients with HF caused by CHD were followed for at least 3 years or died within 3 years, including 1023 male patients (65.49%) with an average age of 65.27 ± 11.14 years and 539 female patients (34.51%) with an average age of 70.80 ± 9.64 years. The average age of all patients was 67.18 ± 10.96 years. During the 3-year

Fig. 1. Analysis flow for the development and evaluation of models.

Through single factor and the five-fold CV RFE-RF feature selection, the optimal number of features was 45 (Fig. 2, Table 1) (details in Supplementary Table 2).

3.3. ML to predict outcomes

Over the 3-year follow-up period, the XGBoost model achieved a mean AUC of 0.8207 (95% confidence interval [CI]: 0.8143–0.8272) and an F1-score of 0.4476 (95% CI: 0.4407–0.4546) for mortality. These values were significantly higher compared to the respective values in the other five models ($P < 0.001$). The mean sensitivity of 0.7520 (95% CI: 0.7471–0.7569) and specificity of 0.7493 (95% CI: 0.7356–0.7630) with XGBoost were also relatively high. Furthermore, the mean Brier score of XGBoost (0.1960; 95% CI: 0.1926–0.1995) was second only to SVM (0.1448; 95% CI: 0.1422–0.1473) among the six models (Table 2). Therefore, XGBoost was selected for further prediction in this study.

3.4. Categorization of prediction score and risk stratification

The XGBoost model was used to predict and stratify the 3-year risk of all-cause mortality in individuals with HF caused by CHD in the test set. Patients were divided into high-risk and low-risk groups with the maximal Youden's index as the optimal cut-off value (0.5339) (Fig. 3A). At this cut-off value, the prediction scores were associated with a sensitivity and specificity of 0.7857 and 0.7638, respectively. As depicted by Kaplan–Meier curves, a gradual decline in survival was observed for high-risk patients over 3 years, indicating that subjects with higher prediction scores are more likely to experience death (log-rank test: $P < 0.001$; Fig. 3B).

3.5. Visualization of feature importance

In order to visually explain the selected variables, we used SHAP to illustrate how these variables affect the 3-year mortality rate in the model. Fig. 4A shows the top 20 risk factors evaluated by the average absolute SHAP value. Fig. 4B displays the top 20 most important features in our model. The feature ranking (y-axis) indicates the importance of the predictive model. The SHAP value (x-axis) is a unified index that

Computers in Biology and Medicine 137 (2021) 104813

responds to the influence of a certain feature in the model. In each feature important row, the attributions of all patients to the outcome were drawn with dots of different colors, where the red dots represent the high risk value and the blue dots represent the low risk value. Older age, a higher N-terminal pro-B-type natriuretic peptide (NT-proBNP) concentration, NYHA classification, left and right diameter of the right atrium (RA1), serum creatinine (CR) concentration, and a lower left ventricular ejection fraction (EF), red blood cell (RBC) count, weight, and body mass index (BMI) were associated with a higher predicted probability of 3-year all-cause mortality. Furthermore, mental work, pulmonary aortic valve regurgitation-1 (PVSIAI-1), pulmonary disease (PULMONARY), lung

Fig. 2. Results of feature screening by RFE with 5-fold CV. >0.800 , the ML risk score significantly outperformed other currently

infection (INFECTION), and a history of treatment for central nervous system disease (HISTORYOF0) also increased the risk of all-cause mortality.

3.6. Cox regression analysis

In the unadjusted analysis, a high ML risk was significantly associated with 3-year all-cause mortality (unadjusted hazard ratio [HR]: 10.351; 95% CI: 4.949–21.650; $P < 0.001$), with a corresponding concordance index of 0.761 (95% CI: 0.698–0.824). After adjusting for the five most influential factors (age, NT-proBNP concentration, NYHA classification, RA1, and occupation), the association between a high ML risk and death persisted (adjusted HR: 5.343; 95% CI: 2.402–11.881; $P < 0.001$), with a concordance index of 0.834 (95% CI: 0.773–0.895). The results of the multivariable Cox analysis are shown in [Fig. 5](#).

3.7. Gender-based analysis

In the sex-specific sub-analysis, age, NT-proBNP concentration, occupation, NYHA classification, nitrate drug use, PVSIAI-1, RBC count, HISTORYOF0, direct bilirubin (DBIL), neutrophil ratio, and blood urea nitrogen appeared as important predictors in both men and women. However, some factors, such as RA1, weight, and CR concentration, were only important predictors in men; they were not in the top 20 predictors in women. Similarly, some factors, such as arrhythmia, BMI, and potassium concentration, were only important predictors in women; they were not in the top 20 predictors in men (Fig. 6A and B). In both sexes, a high ML risk score was associated with a significantly higher 3- year all-cause mortality (Fig. 6C and D).

3.8. Interpretation of personalized predictions

SHAP values show the contribution of each feature to the final prediction and can effectively clarify and explain model predictions for individual patients. Moreover, a new visualization method [11] was used to make the results more intuitive. We provide two typical examples to illustrate the interpretability of the model: an 81-year-old man who died during the follow-up period and a 45-year-old woman who survived to the end of the follow-up period (Fig. 7). The arrows show the influence of each factor on prediction. The blue and red arrows indicate whether the factor reduced (blue) or increased (red) the risk of death. The combined effects of all factors provided the final SHAP value, which corresponded to the prediction score. For the representative man, there was a high SHAP value (5.41) and prediction score (0.9955); for the representative woman, there was a low SHAP value (-3.14) and prediction score (0.0414).

4. Discussion

In this study, we developed and tested an interpretable ML-based risk stratification tool to predict all-cause mortality in patients with HF caused by CHD during a 3-year follow-up period. Among the six ML classifiers, XGBoost demonstrated the best performance; therefore, this model was used to create the ML risk score. With an average AUC of

K. Wang et al.

Table 1

Variable No event Event *P* value Variable No event Event *P* value

AGE (years) 66.0 (59.0–75.0) 77.0 (59.0–82.0) <0.001 PUMONARY <0.001 WEIGHT (kg) 69.0 (60.0–75.0) 65.0 (55.0–75.0) <0.001 No 1221 (90.3%) 150 (71.4%) BMI 25.0 (22.9–27.3) 23.8 (21.1–26.1) <0.001 Yes 131 (9.7%) 60 (28.6%) RBC(10¹²/L) 4.4 (4.1–4.8) 4.1 (3.7–4.6) <0.001 CANCER 0.013 RDW (%) 13.6 (13.1–14.3) 14.3 (13.5–15.4) <0.001 No 1340 (99.1%) 204 (97.1%) HGB (g/L) 137.0 (126.0–148.0) 130.0 (114.0–143.0) <0.001 Yes 12 (0.9%) 6 (2.9%) NEU(10¹⁰/L) 4.1 (3.3–5.3) 4.6 (3.4–6.0) 0.001 PAROXYSMAL <0.001 N.((%) 62.6 (55.9–68.9) 67.7 (61.6–74.4) <0.001 No 1260 (93.2%) 161 (76.7%) ALT (U/L) 19.0 (13.3–29.0) 16.0 (11.0–28.0) 0.005 Yes 92 (6.8%) 49 (23.2%) DBIL (μmol/L) 4.3 (2.7–6.2) 5.3 (3.6–7.6) <0.001 LR <0.001 HDLC (μmol/L) 1.0 (0.9–1.2) 1.1 (0.9–1.2) 0.050 No 1163 (86.0%) 127 (60.5%) LP(a) (mg/L) 120.0 (85.6–228.3) 143.8 (89.7–265.5) 0.036 Yes 189 (14.0%) 83 (39.5%) BUN(mmol/L) 5.4 (4.4–6.6) 6.8 (5.3–9.6) <0.001 MR <0.001 CR (mmol/L) 76.0 (64.0–90.0) 93.3 (74.0–117.4) <0.001 No 195 (14.4%) 20 (9.5%) UA (μmol/L) 337.0 (278.0–410.8) 401.0 (317.3–527.5) <0.001 Little 915 (67.7%) 101 (48.1%) K (mmol/L) 4.0 (3.8–4.3) 4.1 (3.8–4.5) 0.001 Moderate 212 (15.7%) 73 (34.8%) CL (mmol/L) 104.0 (101.4–107.0) 102.5 (100.0–105.9) <0.001 Massive 30 (2.2%) 16 (7.6%) NTPROBNP/ng/L) 384.5 (124.5–1198.0) 2012.0 (501.0–5334.0) <0.001 PVS1AI <0.001 ORS (ms) 96.0 (86.0–106.0) 98.0 (88.0–120.0) <0.001 No 877 (64.9%) 85 (40.5%) OTC (ms) 428.0 (402.0–458.0)

450.0 (419.8–478.0) <0.001 Little 413 (30.5%) 101 (48.1%) LA (mm) 37.0 (34.0–40.0) 40.0 (37.0–45.0) <0.001 Moderate 53 (3.9%) 20 (9.5%) RA1 (mm) 42.0 (39.0–46.0) 45.0 (41.0–49.3) <0.001 Massive 9 (0.7%) 4 (1.9%) LVDD (mm) 51.0 (47.0–57.0) 54.0 (48.0–61.0) <0.001 PVI 0.025 EF (%) 56.0 (45.0–64.0) 46.0 (38.0–57.0) <0.001 No 1327 (98.2%) 200 (95.2%) DRINKING 0.017 Little 21 (1.6%) 9 (4.3%) Never 1026 (75.9%) 175 (83.3%) Moderate 4 (0.3%) 1 (0.5%) Ever 326 (24.1%) 35 (16.7%) Massive 0 0 OCCUPATION <0.001 AS 0.004 Manual worker 498 (36.8%) 45 (21.4%) No 1339 (99.0%) 203 (96.7%) Mental worker 854 (63.2%) 165 (78.6%) Yes 13 (1.0%) 7 (3.3%) PCI 0.001 INFECTION <0.001 No 1053 (77.9%) 185 (88.1%) No 1217 (90.0%) 145 (69.0%) Yes 299 (22.1%) 25 (11.9%) Yes 135 (10.0%) 65 (31.0%) CABG <0.001 EDEMA <0.001 No 1144 (84.6%) 198 (94.3%) No 1298 (95.9%) 188 (89.5%) Yes 208 (15.4%) 12 (5.7%) Yes 56 (4.1%) 22 (10.5%) NYHA classification <0.001 ANTIFREEZING <0.001 II 695 (51.4%) 49 (23.3%) No 27 (2.0%) 14 (6.7%) III~IV 657 (48.6%) 161 (76.7%) Yes 1325 (98.0%) 196 (93.3%) ARRHYTHMIA <0.001 STATINS <0.001 No 1194 (88.3%) 151 (71.9%) No 103 (7.6%) 35 (16.7%) Yes 158 (11.7%) 59 (28.1%) Yes 1249 (92.4%) 175 (83.3%) HISTORYOFO 0.003 NITRATES 0.023 No 1135 ((83.9%) 159 (75.7%) No 691 (51.1%) 125 (59.5%) Yes 217 (16.1%) 51 (24.3%) Yes 666 (48.9%) 85 (40.5%)

Values are median (inter-quartile range) or n (%).

Table 2
Results of the ML models for mortality over 3 years' follow-up in patients with HF caused by CHD [Mean (95%CI)].

| | Indicators LR KNN SVM NB MLP XGBoost ^b P value ^a | | | | | |
|--|--|---|---|---|---|---|
| Sensitivity 0.8215 (0.8086,0.8345) | F1-score 0.3823 (0.3770,0.3876) | (0.5826,0.6125) 0.3481 (0.3404,0.3558) | (0.7216,0.834) 0.3039 (0.2889,0.319) | (0.7877,0.8119) 0.3910 (0.3857,0.3962) | (0.6917,0.732) 0.3835 (0.3759,0.390) | (0.7356,0.760) 0.4476 (0.4407,0.456) |
| Specificity 0.6159 (0.6100, 0.6218) | 0.7156 (0.7104,0.7208) | 0.5298 (0.4976,0.561) | 0.6446 (0.6393,0.6499) | 0.6888 (0.6767,0.709) | 0.7520 (0.7471,0.759) | <0.001 <0.001 <0.001 |
| AUROC 0.7933 (0.7866,0.8001) | 0.6819 (0.6748,0.891) 0.7097 (0.7033,0.712) | 0.7865 (0.7799,0.7930) | 0.7681 (0.7597,0.775) | 0.8207 (0.8143,0.822) | <0.001 | |
| Brier 0.2508 (0.2475, | 0.2540 0.3054 | (0.3009,0.3100) 0.1448 | (0.1422,0.143) 0.2625 | (0.2592,0.2658) 0.2507 | (0.2444,0.250) 0.1960 | (0.1926,0.195) <0.001 |

^a P value is the result of one-way analysis of variance for the five indicators of the six models. ^b After multiple comparisons of Least-significant difference(LSD), XGBoost is significantly different from other models.

available risk scores. These promising results suggest that ML has the potential for clinical implementation to improve risk assessment. Meanwhile, using SHAP values and SHAP plots, we proved that the ML method can illustrate the key features and establish a high-accuracy mortality prediction model in patients with HF caused by CHD. The illustration of cumulative domain-specific feature importance and visualized interpretation of feature importance can allow doctors to

intuitively understand the key features in XGBoost. Generally, several contributions were made in this study. First, in this study, we introduced the XGBoost algorithm, which has attracted widespread attention in recent years for its fast calculation speed, strong generalization ability, and high predictive performance [33–35]. In combination with other advanced ML knowledge, such as missForest-based missing value filling, RFECV-based feature selection,

Fig. 3. Categorization threshold of Prediction score (A) and Kaplan–Meier estimator for population with low and high machine learning risk (B).

Fig. 4. The model's interpretation. (A): The importance ranking of the top 20 variables according to the mean ($|\text{SHAP value}|$); (B): The importance ranking of the top 20 risk factors with stability and interpretation using the optimal model. The higher SHAP value of a feature is given, the higher risk of death the patient would have. The red part in feature value represents higher value.

GridSearchCV-based hyperparameter optimization, and SMOTE + ENN re-sampling techniques were also used. The results demonstrate that using these methods can effectively improve the prediction of 3-year all-cause mortality in patients with HF caused by CHD.

Second, in our analysis, we focused exclusively on patients with HF caused by CHD and generated models to identify clinical characteristic patterns in this particular HF group. Moreover, many pre-existing scores provide risk assessments within 30 days or 1 year of discharge [36,37]. Our goal was to establish a model that could assess the risk of death at 3 years, and a subgroup analysis was performed for different sexes. Furthermore, the study found that models constructed from data collected using the CHF-CRF can accurately predict all-cause mortality in patients with HF caused by CHD during a 3-year follow-up period. If combined with rigorous clinical trial information and bioomics information, it may achieve better prediction results.

Third, it is always a challenge to correctly interpret the prediction model of ML and visually present the predicted results to clinicians. Therefore, we applied SHAP values to XGBoost to achieve the best predictive effect and interpretability. The SHAP value evaluates the importance of the output containing all combinations of features and provides consistent and locally accurate attribute values for each feature in the prediction model. This interpretation is applied to

XGBoost's

black-box tree integration model to help users better understand the decision-making process of the model. Detailed information described in the results and explanations of risk factors provide doctors with more insight, helping them to make more informed decisions, rather than blindly trusting the results of the algorithm. Furthermore, individual explanations can help doctors understand why the model makes specific recommendations for high-risk decisions. In summary, considering key risk factors, the model can intuitively explain to clinicians which specific characteristics of patients with HF caused by CHD predispose them to a higher (or lower) risk of death. Such a prediction breakdown on a subject-by-subject basis has potential in clinical practice by personalizing prevention and potentially driving and reinforcing therapeutic strategies [38].

Fourth, although numerous studies have demonstrated the prognostic capability of clinical factors for adverse consequences of HF, this study further identified the important predictors of all-cause mortality in patients with HF caused by CHD. The importance of variables showed that clinical characteristics, demographic characteristics, and treatment status were important to provide an optimal risk assessment. Consistent with previous literature and clinical experience, age, NT-proBNP concentration, NYHA classification, EF, lung disease, weight, BMI, and other factors [39–43] remain important in the prediction of death in

Fig. 5. Multivariable Cox regression for 3-year all-cause death prediction.

Fig. 6. Variable importance in ML classification for men (A, N = 1023) and women (B, N = 539). Kaplan–Meier curves for subjects with high and low ML risk in mam (C) and woman (D).

5. Conclusion

In this study, the ML-based risk stratification tool was able to accurately assess and stratify the risk of 3-year all-cause mortality in patients with HF caused by CHD. A combination of ML and SHAP could provide an explicit explanation of individualized risk prediction, allowing physicians to intuitively understand the influence of key features in the model, thus helping clinicians better understand the decision-making process for disease severity assessment. With further validation, this paradigm of personalized interpretability could be used to improve risk assessment in the context of other diseases.

Author contributions

All authors made substantial contributions to conception and design, acquisition of data, or analysis and interpretation of data; took part in drafting the article or revising it critically for important intellectual content; agreed to submit to the current journal; gave final approval of the version to be published; and agree to be accountable for all aspects of the work.

Ethical approval

The study complies with the Declaration of Helsinki and has been approved by the Medical Ethics Committee of Shanxi Medical University. All patients were informed about the purpose of the study and provided written informed consent.

Consent for publication

Not applicable.

Availability of data and materials

The datasets during and/or analysed during the current study available from the corresponding author on reasonable request.

Funding

This work was supported by the National Natural Science Foundation of China under Grant [number 82173631 and 81872714]; Shanxi Provincial Key Laboratory of Major Diseases Risk Assessment under Grant [number 201805D111006] and Youth Science and Technology Research Foundation of Shanxi Province under Grant [number 201801D221423].

Declaration of competing interest

No conflict of interest exists in the submission of this manuscript. All authors have contributed to the work, reviewed the final version of the manuscript and approved its submission to Computers in Biology and Medicine.

Acknowledgments

We thank Emily Woodhouse, PhD, from Liwen Bianji (Edanz) (www.liwenbianji.cn/), for editing the English text of a draft of this

Fig. 7. The interpretation of model prediction results with the two samples. (The values of each variable are normalized values.)

patients with HF caused by CHD. Notably, our rank of variable importance broadly corresponded to differences in clinical variables observed in subjects with and without death in our study. For example, patients who died during the follow-up period tended to be older; have a higher NT-proBNP concentration, NYHA classification, RA1, and serum CR concentration; and have a lower EF, weight, and BMI. These were all factors that had high variable importance according to ML. Additionally, risk factors, including occupation, PVSIAI-1, infection, DBIL, RBC count, and RBC volume distribution width, were included in the top 20 important variables in our study, which have been rarely reported in previous literature on HF. These results suggest that these factors may only be effective independent death predictors of HF caused by CHD, indicating that the death predictors of HF are different between subgroups. The value of these factors in predicting the mortality of patients with HF caused by CHD is worthy of clinicians' attention. In particular, occupation in this study was ranked fifth among all patient predictors (3rd among men, and 19th among women), indicating the importance of occupation in predicting death in patients with HF caused by CHD.

4.1. Limitations and development

First, although this was a multi-center study, only patients from two hospitals in the Shanxi Province of China were included in this study, which may have caused a certain bias. Meanwhile, our study lacked external validation by an independent cohort, which could further verify the superiority of our model. We will further expand the research to include patients in different regions and hospitals, and we will use data from different regions for external validation. Second, we focused only on the modeling of commonly used ML methods; we did not compare our results with those obtained using a well-validated risk calculator. Moreover, with the development of artificial intelligence, deep learning has been reported to be used to construct medical models. In the future, we will try to establish a deep learning model to predict the prognosis of HF, and combine more extensive data and information for different levels of research. Third, the information collected in this study was structured data; further research is needed to mine unstructured data and integrate all relevant clinical risk indicators, imaging biomarkers, environmental factors, living habits and other factors to improve predictions.

K. Wang et al.

manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.combiomed.2021.104813>.

References

- [1] L. Yan, H. Zirui, X. Chun, et al., Association of serum total cholesterol and left

8

- ventricular ejection fraction in patients with heart failure caused by coronary heart disease, *Arch. Med. Sci.* 14 (5) (2017) 988–994.
- [2] A. Alba, T. Agoritsas, M. Jankowski, et al., Risk prediction models for mortality in ambulatory patients with heart failure: a systematic review, *Circulation Heart Failure* 6 (5) (2013) 881–889.
- [3] D.K. Arnett, R.S. Blumenthal, M.A. Albert, et al., ACC/AHA guideline on the primary prevention of cardiovascular disease: executive summary, *Circulation* 2019 (2019) e1–171.
- [4] J. Tian, J. Yan, Q. Zhang, et al., Analysis of Re-hospitalizations for patients with heart failure caused by coronary heart disease: data of first event and recurrent event, *Therapeut. Clin. Risk Manag.* 15 (2019) 1333–1341.
- [5] F. Jiang, Y. Jiang, H. Zhi, Y. Dong, et al., Artificial intelligence in healthcare: past, present and future, *Stroke Vasc Neurol* 2 (4) (2017) 230–243.
- [6] A. Rajkomar, J. Dean, I. Kohane, Machine learning in medicine, *N. Engl. J. Med.* 380 (14) (2019) 1347–1358.

- [7] M. Motwani, D. Dey, D.S. Berman, et al., Machine learning for prediction of all-cause mortality in patients with suspected coronary artery disease: a 5-year multicentre prospective registry analysis, *Eur. Heart J.* 38 (2016) 500–507.
- [8] C. Frederic, P.J. Slomka, G. Markus, et al., Machine learning to predict the long term risk of myocardial infarction and cardiac death based on clinical risk, coronary calcium, and epicardial adipose tissue: a prospective study, *Cardiovasc. Res.* 116 (14) (2019) 2216–2225.
- [9] B. Saa, C. Bjm, D. Ag, et al., Machine learning prediction of mortality and hospitalization in heart failure with preserved ejection fraction, *JACC (J. Am. Coll. Cardiol.): Heart Fail.* 8 (1) (2020) 12–21.
- [10] E. Zihni, V.I. Madai, M. Livne, et al., Opening the black box of artificial intelligence for clinical decision support: a study predicting stroke outcome, *PLoS One* (2020) 15.
- [11] M. Athanasiou, K. Sfrintzeri, K. Zarkogianni, et al., An Explainable XGBoost-Based Approach towards Assessing the Risk of Cardiovascular Disease in Patients with Type 2 Diabetes Mellitus[C]/2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE), IEEE, 2020.
- [12] S.M. Lundberg, B. Nair, M.S. Vavilala, et al., Explainable machine-learning predictions for the prevention of hypoxaemia during surgery, *Nature Biomedical Engineering* 2 (10) (2018) 749–760.
- [13] F. Cabitza, R. Rasoini, G.F. Gensini, Unintended consequences of machine learning in medicine, *J. Am. Med. Assoc.* 318 (2017) 517–518.
- [14] S. Lundberg, S.I. Lee, A Unified Approach to Interpreting Model Predictions[C]/NIPS, 2017, pp. 4765–4774.
- [15] Mortazavi BJ, Downing NS, Bucholz EM, et al. Analysis of machine learning techniques for heart failure readmissions. *Circ Cardiovasc Qual Outcomes* 2016;9: 629–640.
- [16] P.Y. Tseng, Y.T. Chen, C.H. Wang, et al., Prediction of the development of acute kidney injury following cardiac surgery by machine learning, *Crit. Care* 24 (1) (2020).
- [17] M'arton Tokodi, W.R. Schwertner, Attila Kovács, et al., Machine learning-based mortality prediction of patients undergoing cardiac resynchronization therapy: the SEMMELWEIS-CRT score, *Eur. Heart J.* 41 (18) (2020).
- [18] S.J. Pocock, C.A. Ariti, J.J.V. McMurray, et al., On behalf of the Meta-Analysis Global Group in Chronic Heart Failure. Predicting survival in heart failure: a risk score based on 39,372 patients from 30 studies, *Eur. Heart J.* 34 (2013) 1404–1413.
- [19] Zile M.R., Koehler J., Sarkar S., et al. Prediction of worsening heart failure events and all-cause mortality using an individualized risk stratification strategy. *ESC Heart Fail.* (7)(2020): 4277–4289.
- [20] E.D. Adler, A.A. Voors, L. Klein, et al., Improving risk prediction in heart failure using machine learning, *Eur. J. Heart Fail.* 22 (1) (2020).
- [21] J.L. Koyner, K.A. Carey, D.P. Edelson, M.M. Churpek, The development of a machine learning inpatient Acute kidney injury prediction model, *Crit. Care Med.* 46 (7) (2018) 1070–1077.
- [22] R.J. Delahanty, D. Kaufman, S.S. Jones, Development and evaluation of an automated machine learning algorithm for in-hospital mortality risk adjustment among critical care patients, *Crit. Care Med.* 46 (6) (2018) e481–e488.
- [23] C.A. Zizza, K.J. Ellison, C.M. Wernette, Total water intakes of community-living middle-old and oldest-old adults, *The journals of gerontology Series A, Biological sciences and medical sciences* 64 (4) (2009) 481–486.
- [24] Heart failure group. C.B., Chinese medical association Chinese guidelines for the diagnosis and treatment of heart failure 2018, *Chin. J. Cardiol.* 46 (10) (2018) 760.
- [25] C.W. Yancy, M. Jessup, B. Bozkurt, et al., ACC/AHA/HFSA focused update of the 2013 ACCF/AHA guideline for the management of heart failure: a report of the American college of cardiology/American heart association task force on clinical practice guidelines and the heart failure society of America, *J. Am. Coll. Cardiol.* 68 (13) (2017) 1476–1488, 2016.
- [26] D.J. Stekhoven, P. Buhlmann, MissForest—non-parametric missing value imputation for mixed-type data, *Bioinformatics* 28 (1) (2012) 112–118.
- [27] S. Okada, M. Ohzeki, S. Taguchi, Efficient partition of integer optimization problems with one-hot encoding, *Scientific Reports*, 2019.
- [28] A. Singh, N. Thakur, A. Sharma, A review of supervised machine learning algorithms, in: 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), 2016, 1310-5. IEEE; 2016.
- [29] T. Chen, C. Guestrin, XGBoost: a scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, New York, 2016, pp. 785–794.
- [30] G.E.A.P.A. Batista, R.C. Prati, M.C. Monard, A study of the behavior of several methods for balancing ML training data, *ACM SIGKDD Explorations Newsletter* 6 (1) (2004) 20.
- [31] K. Wang, J. Tian, C. Zheng, et al., Improving risk identification of adverse outcomes in chronic heart failure using SMOTE+ENN and machine learning, *Risk Manag. Healthc. Pol.* 14 (2021) 2453–2463.
- [32] S.M. Lundberg, G.G. Erion, S.-L. Japa Lee, Consistent Individualized Feature Attribution for Tree Ensembles, 2018.
- [33] Dalakleidi, Kalliopi, Zarkogianni, et al., Comparative assessment of statistical and machine learning techniques towards estimating the risk of developing type 2 diabetes and cardiovascular complications, *Expert Syst.* 34 (6) (2017) 1–8.
- [34] A. Azeez, Adeola Ogunleye, et al., XGBoost model for chronic kidney disease diagnosis, *IEEE ACM Trans. Comput. Biol. Bioinf* 17 (6) (2020) 2131–2140.
- [35] M. Li, X. Fu, D. Li, Diabetes prediction based on XGBoost algorithm, *IOP Conf. Ser. Mater. Sci. Eng.* 768 (7) (2020), 072093 (7pp).
- [36] C. Weber, J. Hung, S. Hickling, et al., 028 unplanned 30-day readmission and risk of one-year mortality following index hospitalisation with heart failure: a western Australia linked population study, *Heart Lung Circ.* 29 (2) (2020) S50.
- [37] A.G. Au, F.A. Mcalister, J.A. Bakal, et al., Predicting the risk of unplanned readmission or death within 30 days of discharge after a heart failure hospitalization, *Am. Heart J.* 164 (3) (2012) 365–372.
- [38] Lundberg S.M., Erion G., Chen H., et al. From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* 2(1)(2020)56-67.
- [39] D.H.D. A, K.T. A, J.Z. B, et al., Global mortality variations in patients with heart failure: results from the International Congestive Heart Failure (INTER-CHF) prospective cohort study, *Lancet Global Health* 5 (7) (2017) e665–e672.
- [40] F.M. Cunha, J. Pereira, A. Ribeiro, et al., Age affects the prognostic impact of diabetes in chronic heart failure, *Acta Diabetol.* 55 (10) (2018) 1–8.
- [41] Z.V. Babayan, R.L. Mcnamara, N. Nagajothi, et al., Predictors of cause-specific hospital readmission in patients with heart failure, *Clin. Cardiol.* 26 (9) (2003) 411–418.
- [42] W. Ouwkerk, A.A. Voors, A.H. Zwiderman, Factors influencing the predictive power of models for predicting mortality and/or heart failure hospitalization in patients with heart failure, *JACC Heart Failure* 2 (5) (2014) 429–436.
- [43] D. Chicco, G. Jurman, Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone, *BMC Med. Inf. Decis. Making* 20 (1) (2020) 16.

Computers in Biology and Medicine 137 (2021) 104813



Comparison of machine learning algorithms for clinical event prediction (risk of coronary heart disease)

Juan-Jose Beunza^{a,b,*}, Enrique Puertas^{a,c}, Ester García-Ovejero^{a,d}, Gema Villalba^{a,e}, Emilia Condes^a, Gergana Koleva^a, Cristian Hurtado^{a,f}, Manuel F. Landecho^{a,g}

^a Machine Learning Health Working Group, Faculty of Biomedical and Health Sciences, Universidad Europea de Madrid, Madrid, Spain ^b Department of Medicine, Faculty of Biomedical and Health Sciences, Universidad Europea de Madrid, Madrid, Spain

^c Department of Computer Science and Technology, School of Architecture, Engineering and Design, Universidad Europea de Madrid, Madrid, Spain ^d Department of Nursing and Psychology, Faculty of Biomedical and Health Sciences, Universidad Europea de Madrid, Madrid, Spain ^e Indra, Madrid, Spain

^f Department of Pharmacy and Biotechnology, Faculty of Biomedical and Health Sciences, Universidad Europea de Madrid, Madrid, Spain ^g Department of Internal Medicine, Clínica Universidad de Navarra, Pamplona, Spain

Conclusions: ML algorithms can reinforce the diagnostic and prognostic capacity of traditional regression techniques. Differences between the applicability of those algorithms and the results obtained with them were a function of the software platforms used in the data analysis.

ARTICLE INFO

Keywords:

Machine learning

Supervised machine learning Support vector machines

Research techniques

Area under curve

Diagnostic techniques and procedures **1. Introduction**

ABSTRACT

Aim: The aim of this study is to compare the utility of several supervised machine learning (ML) algorithms for predicting clinical events in terms of their internal validity and accuracy. The results, which were obtained using two statistical software platforms, were also compared.

Materials and methods: The data used in this research come from the open database of the Framingham Heart Study, which originated in 1948 in Framingham, Massachusetts as a prospective study of risk factors for cardiovascular disease. Through data mining processes, three data models were elaborated and a comparative methodological study between the different ML algorithms – decision tree, random forest, support vector machines, neural networks, and logistic regression – was carried out. The global selection criterion for choosing the right set of hyperparameters and the type of data manipulation was the area under a curve (AUC). The software tools used to analyze the data were R-Studio® and RapidMiner®.

Results: The Framingham study open database contains 4240 observations. The algorithm that yielded the greatest AUC when analyzing the data in R-Studio was neural network applied to a model that excluded all observations in which there was at least one missing value (AUC = 0.71); when analyzing the data in RapidMiner and applying the same model, the best algorithm was support vector machines (AUC = 0.75).

data availability (Big Data).

In the last three years these types of algorithms and techniques have

The algorithms and techniques deployed in machine learning (ML) can be framed within a more general process known as *knowledge discovery in databases* or simply *data mining*. Some of these techniques were described more than 50 years ago [1], however in recent years interest in and about them has surged dramatically, driven in part by major advances in algorithmic programming, increasing processing capacity of modern computers (Graphics Processor Unit for video and graphics cards and Tensor Processing Unit for neural learning), and growth of

begun to be applied to clinical environments, including within the fields of diagnostic radiology [2–4], cardiac electrophysiology [5], diabetes [6], dermatology [7] and psychiatry [8,9]. Given their practicality and accessibility, and the impressive results obtained so far, we expect an explosion in ML applications in healthcare settings in 2019.

Recent publications of the Ministry of Science, Innovation and Universities of the Government of Spain identify the incorporation of artificial intelligence in healthcare as a priority [10]. Official

Abbreviations: ML, machine learning; ACC, accuracy; AUC, area under curve; SE, sensitivity; SP, specificity; PPV, positive predictive value; NPV, negative predictive value; NA, not applicable

* Corresponding author at: Universidad Europea de Madrid, Calle Tajo, s/n, 28670 Villaviciosa de Odón, Madrid, Spain.

E-mail address: juanjo@juanjobeunza.com (J.-J. Beunza).

<https://doi.org/10.1016/j.jbi.2019.103257>

Received 27 May 2019; Received in revised form 21 July 2019; Accepted 22 July 2019

Available online 30 July 2019

1532-0464/ © 2019 Elsevier Inc. All rights reserved.

J.-J. Beunza, et al. *Journal of Biomedical Informatics* 97 (2019) 103257

documents of the European Commission and strategic plans of other European countries point in the same direction [11,12]. The Machine Learning development group Salud-UEM was created in November 2018 at University Europea de Madrid with the aim of applying ML techniques to the study of health-related problems, generating evidence of their potential utility, and eventually incorporating them in the teaching curriculum of students and health professionals. It brings together a heterogeneous mix of health professionals (doctors, nurses, psychologists, pharmacists and biotechnologists), computer specialists specialized in big data, and eHealth consultants. As a first test of the possibilities that ML can offer in the development of predictive models in health, it was decided to test various algorithms to see whether and to what extent their prediction scores approximate or improve upon the results obtained in the original Framingham model [13], one of the most important cardiovascular risk prediction tables from the point of view of clinical practice [14].

Although other researchers have tried to improve upon the Framingham model's predictive value [15], the focus of the present study is purely methodological, given our overarching goal of exploring the applicability of ML models in health. Specifically, the primary objective was to compare in terms of internal validity and accuracy several supervised ML algorithms applied to the prediction of clinical events, using structured data; the secondary objective was to compare the utility, usability and results obtained by means of two software tools, R-studio (script code) and RapidMiner (graphic interface).

2. Materials and methods

A comparative methodological study was carried out between the most commonly used supervised classification algorithms in ML: decision tree, random forest, support vector machines, and neural networks, in addition to traditional logistic regression.

The following outcome variables were selected for comparison: accuracy, sensitivity (recall), specificity, positive predictive value (precision), negative predictive value, and area under the curve (AUC). The global criterion for selecting the right set of hyperparameters and the type of data manipulation was the AUC.

The data used came from the Framingham database available on Kaggle [16], a Google-owned platform for data scientists that published ML-based prediction competitions using publicly available databases.

2.1. Data preparation

A descriptive analysis of the database was performed, attending to the nature of the variables: mean, standard deviation and extreme values were described for quantitative variables; absolute and relative frequencies for qualitative variables. The number of missing values was identified for each variable, since for some of the algorithms to work it was anticipated that it may be necessary to eliminate observations with missing values.

To guide the selection of independent variables to be included in the prediction model for coronary risk at 10 years, an automatic "stepwise" technique was applied, using multiple logistic regression (stepAIC function of the MASS library in R-studio) and establishing as a cut-off point the significance level of $p < 0.10$. Nonetheless, the algorithms identified as "best" used $p < 0.05$ as the significance level.

For the comparison of the algorithms, three different data models were prepared. Model A included the original variables in the Framingham Heart Study database without any modification. Model B included the same variables but excluded all observations in which there was at least one missing value in any of the variables (analysis of complete cases). Finally, in model C missing values in any of the continuous variables were imputed using the average obtained from the rest of the non-missing values present in that variable so as to avoid reducing the number of observations, notwithstanding that some minimal alterations introduced by the imputation.

2

In all three models the data were randomly divided into two subgroups: training set (*train*), which contained 80 percent of the observations, for the training of the algorithm, and test set (*test*), comprising the remaining 20 percent of observations, for the evaluation of the algorithm's capacity for prediction. Different base models were elaborated by normalizing (subtracting the mean and dividing by the standard deviation, which resulted in means of zero in all the

variables) and standardizing (subtracting the minimum value of the variable and dividing by the range, which resulted in minimum values of 0 and maximums of 1 in all the variables) the values in both the *train* and the *test* sets in order to homogenize their range, an important condition for some algorithms, (e.g. the neural network algorithm).

Finally, the number of positive events or results, referred to as *labels* in informatics terminology, was balanced using the ROSE library and the functions *over* (duplicating subjects with positive cardiovascular event), *under* (eliminating subjects with negative cardiovascular event), *both* (combining both techniques) and *rose* (artificially generating completely new subjects based on the distribution of variables of the original database). This resulted in a change from the original prevalence of coronary events of 15% to a prevalence of 50%. The purpose of the balancing, as commented on in the discussion section, is to improve the algorithm's prediction capability. In addition, each algorithm was optimized by way of adjusting its hyperparameters (the internal values of the algorithm that determine its learning function and therefore directly impact on the final result it produces).

The participants were not stratified by sex, as the original authors of the Framingham study had done, so as not to reduce the sample size for training the algorithm, something to which ML algorithms are very sensitive. As a workaround, the variable *sex* was included in the model, because it is strongly associated with the event under study. The accuracy values for both sexes from the original Framingham publication were calculated using the formula [(sensitivity * prevalence) + (specificity * (1 - prevalence))].

2.2. Software used

Two software tools were used in carrying out this research: J.-J. Beunza, et al. *Journal of Biomedical Informatics* 97 (2019) 103257

Table 1
Exploratory descriptive analysis of the Framingham Heart Study open database [16].

| Variable Categories n % Missing | |
|--|---|
| Sex | Women 2420 57.1 0 Men 1820 42.9 |
| Educational level (n = 4135) | 1720 41.6 105 1253 30.3 |
| Some High School | 689 16.7 |
| High School or GED | |
| Some College or Vocational School | place because it is freely accessible and is easy to use for teaching objectives. Secondly, because due to college 473 11.4 |
| Current smoker | No 2145 50.6 0 Yes 2095 49.4 |
| 1600 mHz DDR3 memory). In this way it was possible to avoid having | |
| Antihypertensive treatment (n = 4187) | No 4063 97 53 Yes 124 3 technical complications |
| Prevalent stroke No 4215 99.4 0 Yes 25 0.6 | |
| Hypertension No 2923 68.9 0 Yes 1317 31.1 | |
| Diabetes No 4131 97.4 0 Yes 109 2.6 | |
| advantages. Firstly, it could manage variables with missing values and therefore required less manipulation of the original data. In addition, | |
| Coronary events at 10 years | cemia. Finally, it was a |
| No 4131 84.8 0 Yes 644 15.2 | execution and design, r |
| SD Max Min | processing times (almost |
| Age 49.6 8.6 70 32 0 Daily cigarettes (n = 4211) | 9.01 11.9 70 0 29 Cholesterol (n = 4190) 236.7 44.6 696 107 50 Systolic blood pressure 132.4 22.0 83.5 295 0 Diastolic blood pressure 82.9 11.9 48 142.5 0 Body mass index 25.8 4.1 56.8 15.5 19 Heart rate (n = 4239) 75.9 12.0 143 44 1 Glycaemia (n = 3852) 82.0 24.0 394 40 388 |

x̄: mean; SD: standard deviation; Max: maximum value; Min: minimum value.

The variables with missing data were: educational level, anti hypertensive treatment, daily cigarettes, body mass index, heart rate, and glycaemia (Table 1).

By sex, the 10-year prevalence of coronary heart disease was calculated as 19% (n = 343) in men and 12% (n = 301) in women. Using the formula already described, accuracy values of 0.78 for men and 0.81 for women were obtained.

The automatic “stepwise” selection of variables to include in the models determined the following features to be key predictors for

- (a) R-Studio open source, version 1.1.463 with R open source version 3.5.2 (2018-12-20, “Eggshell Igloo”) for data preparation and algorithm training and evaluation. The libraries and functions used in training the different algorithms were the following: for the decision tree and the boosted decision tree (the most powerful version of the decision tree) algorithms, the *C50* library and the *C5.0* function; for random forests, the *randomForest* library and the *randomForest* function; for the support vector machines, the *kernlab* library and the *ksvm* function; for the neural network, the *neuralnet* library and the *neuralnet* function. The complete code is available on Github [17] and will be published as open source in *markdown* format (html) with explanatory comments with teaching objectives.
- (b) RapidMiner version 9.2.0 to compare it with the results obtained with RStudio. Rapid Miner code is available on Github [18].

3. Results and discussion

The Framingham open database consisted of 4240 observations, of which 57.1% corresponded to women (n = 2420) and 42.9% to men (n = 1820). The mean age was 49.6 years (standard deviation (SD) = 8.6). The percentage of patients who were smokers was 49.4% (n = 2095), 31.1% had hypertension (n = 1317) and 2.6% had diabetes (n = 109). The mean total cholesterol was 236.7 (SD = 44.6), the mean systolic blood pressure was 132.4 mm Hg (SD = 22), and the mean diastolic blood pressure was 82.9 mm Hg (SD = 11.9). Discordant extreme values were seen in blood glucose, whose mean was 82 mg/dl (SD = 24), with a minimum of 40 and a maximum of 394. In total, 15.2% of participants suffered coronary events at 10 years (n = 644).

0.85 and 0.84, respectively. However, direct comparisons between results should be approached with caution, given that the original article used Cox regression (time to event), calculated the probability of being in the fifth risk quintile while using more of the available predictor variables, and differed in the number of observations and variables used compared to the open database.

Working with the Framingham database was selected in the first its relatively small size for ML studies, reasonable processing times (for dies, as a result of which the example, it does not exceed 10 min obtained results were probably not using an Intel Core i7 2.5 GHz optimal, it allows for the processor and 16 GB algorithms to be executed in almost any personal computer with to manage clusters of local computers or to use a cloud, with the added

Reviewing the algorithms one by one, the decision tree had several it offered a prioritization of the variables, something that is of value for clinicians in their decision-making. Interestingly, the variable most frequently used (in 100% of cases) in the design of the trees was gly

coronary risk at 10 years: sex, age, cigarettes per day, prevalent cerebral infarction, prevalent hypertension, total serum cholesterol, arterial systolic pressure, and glucose serum (p < 0.10).

After data preparation for each model, the sizes of the training sets (n-train) and test sets (n-test) were as follows: model A, gross (n-train: 3392; n-test: 848); model B, analysis of complete cases (n-train: 3053; n-test: 764); model C, imputation of the mean (n-train: 3392; n-test: 848). The results in Table 2 were calculated with these numbers.

The algorithm that showed the highest AUC when performing the analysis with R-studio was neural network applied to model B (AUC = 0.71). Among the results obtained with RapidMiner, the algorithm that behaved best was support vector machines applied to model B (AUC = 0.75). Table 2 also summarizes the best outcomes in accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and area under the curve obtained for each prediction algorithm model, considering the three data models and the two com

puter tools R-studio and RapidMiner.

The ML algorithms applied to the Framingham database have obtained comparable, and in some cases superior, results to those described in the original article of the Framingham Coronary Risk Scale [13]. The sensitivity values described for each sex in the original study were 0.49 for men and 0.6 for women, and the specificity values were

models, each comprising between one and three layers and between one and nine nodes per layer; some processing times exceeded 10 min; however, the neural network also offered the best result. Lastly, the traditional logistic regression offered similar results to the previously described algorithms.

In general, all the results of the algorithms improved after normalization/standardization of the data, an almost mandatory requirement in some cases such as support vector machines and neural networks. On the other hand, data balancing was another key factor that led to improved results; however, it is rarely done in traditional statistical analyses.

In Model A, acceptable accuracy values (85%) were almost universally obtained, which could indicate that the algorithm did a good job because it correctly guessed the classification of 85 out of 100 participants. However, what the algorithms were doing in almost all cases was to classify all subjects as non-events. In this way the accuracy was good, but the prediction of “positive events” was almost null (sensitivity close to 0, specificity close to 100). This happens usually when the incidence of the event in studies is not balanced, that is, when it is far from 50%, which is something very frequent in the study of medical or healthcare events. In our case, the incidence was 15%. The way to solve this was to balance the base by increasing the number of observations with positive events, which can be achieved by cloning observations or creating them again by artificial methods, while respecting the distribution of variables; decreasing the number of

3
an algorithm generally considered as having low prediction power. The boosting, which adds a greater number of trees (around 10 in our case) made almost no difference. A very aggressive pruning process in single tree algorithms may incur in overfitting, but this problem is mitigated when using decision tree ensembles like Random Forest or Boosted Trees. When using default hyperparameters of the models, we did not observe symptoms of overfitting or underfitting. Different performance results may be obtained using full or pruned trees.

The random forest algorithm improved the values of the decision tree, although its interpretation was more complex (hence it being called a “black box”) and the processing time was prolonged beyond five minutes in some configurations of the hyperparameters. The support vector machines algorithm improved the AUC a little more, but again extended the processing times and code complexity. The neural network was the algorithm that required the most programming time – around 1250 lines of code – and it ran 73 different architecture

J.-J. Beunza, et al. Journal of Biomedical Informatics 97 (2019) 103257

Table 2
Comparison and evaluation of the different algorithms with R-studio and RapidMiner.

| R-Studio RapidMiner | | | | | | | | | | | | | |
|---------------------|-----|----|----|-----|-----|------|-----|----|----|-----|-----|------|--|
| Algorithms | ACC | SE | SP | PPV | NPV | AUC | ACC | SE | SP | PPV | NPV | AUC | |
| Decision tree | | | | | | | | | | | | | |
| Model A | 84 | 8 | 98 | 48 | 85 | 0.53 | 85 | 3 | 99 | 44 | 85 | 0.53 | Model B 67 39 72 21 86 0.55 54 4 99 83 55 0.5 Model C 84 8 98 44 85 0.53 62 4 99 83 61 0.5 |

Boosted decision tree
Model A 85 6 99 57 85 0.53 63 73 53 61 66 0.67 Model B 81 28 91 37 87 0.60 64 54 70 51 72 0.7 Model C 84 5 99 44 85 0.52 62 53 67 51 69 0.69

Random forests
Model A NA NA NA NA NA NA NA NA NA NA NA NA Model B 79 35 88 39 86 0.63 65 9 97 65 64 0.71 Model C 78 30 87 31 87 0.59 63 14 95 62 63 0.69

Support vector machines
Model A NA NA NA NA NA NA NA NA NA NA NA NA Model B 69 67 69 29 92 0.68 69 42 84 61 71 0.75 Model C 68 69 68 28 92 0.68 68 49 81 62 71 0.71

Neural network
Model A NA NA NA NA NA NA NA NA NA NA NA NA Model B 67 70 66 28 92 0.71 69 36 90 67 70 0.73 Model C 71 64 72 29 92 0.68 68 56 77 61 73 0.72

Logistic regression
Model A 84 5 99 50 85 0.5 63 47 79 69 60 0.68 Model B 68 69 67 29 92 0.68 68 43 83 59 71 0.73 Model C 66 69 66 27 92 0.68 67 46 81 61 70 0.73

ACC: accuracy; SE: sensitivity; SP: specificity; PPV: positive predictive value; NPV: negative predictive value; AUC: area under curve; NA: not applicable (model did not converge).

negative events by eliminating observations; or a combination of both. This type of data manipulation almost certainly introduces significant biases in causality studies. For example, the imputation of missing values with the mean of the variable is something frequently done in prediction work yet is viewed with suspicion in traditional cause-and-effect analysis because of the bias it introduces in case the participants with missing values are intrinsically different from those without missing values.

Model B, which is the one that behaved the best in terms of predictive capacity, is nevertheless the one that follows the most discouraged strategy from the point of view of analytical power. The best approach to handling of missing data remains a problem without a universally agreed-upon solution in epidemiology so, at a minimum, it needs to be included in the analysis plan and communicated in the results [19]. It is a consideration to consider because in risk prediction studies and ML studies in general, the prediction objective allows us to manipulate the data more freely if the objective (prediction in a different database, evaluation) is satisfactory. It implies a change of mentality for health professionals trained in cause-and-effect studies.

Model C handled missing values by imputating mean values. Other classic imputation options are to use the most frequent value for that variable, zero or constant imputation, k nearest neighbors (k-NN) through feature similarity, multivariate imputation by chained equation (MICE), deep learning, stochastic regression imputation, extrapolation and interpolation and hot-deck imputation [20]. Regardless of the chosen method, there is no perfect way to compensate for missing values in a data set and it is up to the knowledge and intuition of the analyst to select the best method for that specific dataset and context. In our case, we considered that imputation by means – which is important in the analysis of relatively small data sets because it “prevents” discarding observations due to missing values – offered comparison value of algorithms without losing participants.

Finally, differences were found among the results produced by the software programs used in this research. R-studio proved to be more flexible and powerful a priori than RapidMiner, although the

4

programming time it required was much higher. However, its ability to keep track of all the changes in a reproducible file (script) was more replicable and amenable to correction, which provided a sense of assurance. RapidMiner, on the other hand, was more intuitive and simpler, which is something very useful in simple or preliminary analyses; however, when the process becomes more complex, it seems easier to make mistakes. The analysis with RapidMiner is more difficult to replicate by an external person and more difficult to correct. It seems therefore an excellent tool for teaching purposes and for selecting algorithms (piloting), but it does not replace code-based programs from our point of view. The differences with respect to the results obtained with R-Studio may be due to the different implementations of the algorithms and the default hyperparameters used by the libraries of both testing environments. Therefore, comparing tools and algorithms often requires a judgment call, given that a good hyperparameter tuning is more an art than a mechanical process and should be based on both the characteristics of the dataset and the knowledge domain.

The present study has several strengths. The principal one is that it was managed by an interprofessional team that included clinical and informatics experts, which has proven to be extremely enriching. Further, a large amount of sensitivity analysis (multiple models of variables management, feature engineering) and model and hyperparameters design was carried out. The fact that the data are public

and open and that the code has been released also allowed us to review and replicate our results, as well as to use the present study as a gateway to the world of health ML in particular and to the programming of prediction models in general. Finally, since the analysis drew upon a relatively small database (4240 observations with 16 variables), it lends itself to being conducted on any personal computer.

J.-J. Beunza, et al. *Journal of Biomedical Informatics* 97 (2019) 103257

Further and somewhat unexpectedly, some variables that are clearly predictive of coronary disease, such as blood levels of LDL-cholesterol, were not included in the dataset; a much more precise model could be obtained by using more information of risk factors of the participants. The possibility of requesting such information from those responsible for the Framingham study was considered but ruled out because of the likelihood that readers would be excluded from access to the additional data. Finally, the fact that stepwise selection was done through logistic regression – and the results applied to all models for the sake of simplicity – could have led to a missed opportunity to obtain “better” results by using other models. It would have been beneficial to do feature selection for each algorithm.

We obtained low levels of AUC values (close for 0.5 for decision tree and 0.6–0.73 for others). We did not find other published results on the same dataset (available from Kaggle) for comparison. However, similar approaches with other larger data sets (378,256 patients from UK family practices) yielded similar results [21]. Failure to obtain better results could be due to the low number of subjects or because the available variables do not have a higher predictive capacity. It could also be due to the fact that we may not have been able to adequately refine the hyperparameters; however, this is unlikely, given that one of the researchers in the group is an expert in big data and has won several national ML awards [22,23]. Nonetheless, it is still possible to refine the algorithms used, as we must not forget that, like medicine, ML is both an art and a science.

The next steps in the development of ML algorithms in healthcare settings is to apply them more widely to different environments and populations. A potential handicap is that as the adoption of advanced data analytics and machine learning in healthcare accelerates and databases grow in volume, the ever-rising number of observations makes it increasingly likely that researchers may need to use a special computer, a cluster, or the cloud to host and process the data; however, the upside to working with more data is the hope to obtain much more accurate results with clinical applicability. In our experience, although today it is difficult to obtain large volumes of health data in some countries, it is possible to apply ML to smaller data volumes (small data) without compromising validity or applicability, especially when the data quality is high. The benefits for the health system can be significant, beginning with the potential improvements of diagnostic, prognostic, and therapeutic tools. In addition, analyzing and learning from smaller datasets now will allow our clinical research teams to become familiar with these techniques and be prepared for when big data containing clinical and healthcare information truly becomes a reality. However, it is very important to highlight that each method has its own merits and may be preferable to others under certain conditions or assumptions. Therefore, it is not possible to compare them in general terms. Results will depend on the data, context, user, processes of imputation, variable selection, parameter tuning, re-balancing, and data partitioning. And although one method may outperform others in some metrics, different scenarios will probably change comparisons again.

Another one of our research group's projects is to prepare a digital manual of ML algorithms with applicability to healthcare, which will be published in late 2019 so that health professionals can start their ML learning journey in a simple and friendly way. Meanwhile, all our code for this study is available on Github [17].

4. Conclusions

The use of ML algorithms when working with small databases (around 4000 participants) is relatively simple. These algorithms can enhance the diagnostic and prognostic capacity of more traditional regression techniques.

R-Studio is a powerful tool for conducting complex ML analytics with high reliability in creating a record of all changes. RapidMiner

In terms of weaknesses, the main drawback of this research is that when using data presented as part of an ML competition, neither its clinical reliability nor the quality of the results can be guaranteed. In addition, as mentioned earlier, the limited number of observations likely diminishes the prediction capacity of the trained models, designed to be working with much larger numbers of observations.

runs and visualizes ML algorithms using a very simple and intuitive graphic interface, although its capacity for manipulating the

parameters can be smaller and less reliable in the case of complex analyses.

Mixed research teams, comprising healthcare professionals and computer scientists or mathematicians, are optimal for the conceptualization and development of ML projects.

Author contributions

JJB designed the article, obtained the database, performed the main analysis and drafted the text; EP performed the analysis with Rapid Miner and co-directed the main analysis with R-Studio; EG performed the descriptive analysis of the Framingham database and adapted the article to the requirements of the journal; ML collaborated in the overall design of the study. All the authors have reviewed and contributed revisions up to the final draft.

Financial support received

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors of this study wish to express their gratitude to the team behind the Framingham Heart Study for providing open access to its data; to Elena Gazapo, Dean of the Faculty of Biomedical Sciences and Health at Universidad Europea de Madrid, for her support for the creation and development of the Machine Learning Health-UEM working group; to Andrew NG for making the machine learning algorithms he has developed accessible and comprehensible; to Brett Lantz and Hadley Wickham for their creativity and empowering us, as well as countless other researchers, through R; and to R-Ladies, Group R of Spain, and Machine Learning Spain, for being sources of constant inspiration for the work done by our group dedicated to teaching and research in machine learning applied to health.

References

- [1] A. Samuel, Some studies in machine learning using the game of checkers, IBM J. Res. Dev. 3 (1959) 210–229, <https://doi.org/10.1147/rd.33.0210>.
- [2] P. Rajpurkar, J. Irvin, K. Zhu, et al., CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning, 2017, < <https://arxiv.org/abs/1711.05225> > (accessed 20 mar 2018).
- [3] M. Grewal, M. Muktabh, P. Kumar, et al., RADNET: Radiologist Level Accuracy using Deep Learning for HEMORRHAGE detection in CT Scans, 2018, < <https://arxiv.org/abs/1710.04934> > (accessed 20 mar 2018).
- [4] Z. Li, C. Wang, M. Han, et al., Thoracic Disease Identification and Localization with Limited Supervision, 2018. < <https://arxiv.org/abs/1711.06373> > (accessed 20 mar 2018).
- [5] P. Rajpurkar, A.Y. Hannun, M. Haghpasahi, et al., Cardiologist-Level Arrhythmia Detection with Convolutional Neural Networks, 2017, < <https://arxiv.org/abs/1707.01836> > (accessed 20 mar 2018).
- [6] D.S.W. Ting, C.Y. Cheung, G. Lim, et al., Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes, JAMA 318 (2017) 2211–2223, <https://doi.org/10.1001/jama.2017.18152>.
- [7] A. Esteva, B. Kuprel, R.A. Novoa, et al., Dermatologist-level classification of skin cancer with deep neural networks, Nature 542 (2017) 115–118, <https://doi.org/10.1038/nature21056>.
- [8] T. Alhanai, M. Ghassemi, J. Glass, Detecting depression with audio/text sequence modeling of interviews, Interspeech 2522 (2018) 1716–1720.
- [9] Y.H. Huang, L.H. Wei, Y.S. Chen, Detection of the prodromal phase of bipolar disorder from psychological and phonological aspects in social media, 2017, < <https://arxiv.org/pdf/1712.09183.pdf> > (accessed 20 mar 2018).
- [10] Secretaría General de Coordinación de Política Científica del Ministerio de Ciencia, Innovación y Universidades y al Grupo de Trabajo en Inteligencia Artificial GTIA. Estrategia Española de I+D+I en inteligencia artificial. Secretaría General Técnica del Ministerio de Ciencia, Innovación y Universidades, 2019, p. 48, < http://www.ciencia.gob.es/stfls/MICINN/Ciencia/Ficheros/Estrategia_Inteligencia_Artificial

las consecuencias de la inteligencia artificial para el mercado único (digital), la producción, el consumo, el empleo y la sociedad» (2017/C 288/01). Diario Oficial de la Unión Europea. Comité Económico y Social Europeo.

- [13] R. D'Agostino, R.S. Vasan, M.J. Pencina, et al., General cardiovascular risk profile for use in primary care the Framingham heart study, Circulation 117 (2008) 743–753, <https://doi.org/10.1161/CIRCULATIONAHA.107.699579>.
- [14] C. Brotons Cuixart, J.J. Alemán Sánchez, J.R. Banegas Banegas, et al., Grupo de Prevención Cardiovascular del PAPPs, Recomendaciones preventivas cardiovasculares

[IDI.pdf](#) > (accessed 13 mar 2019).

- [11] M. Craglia, A. Annoni, P. Benczur, et al., Artificial Intelligence – A; Luxembourg: European Perspective. Joint Research Centre (JRC), the European Commission's Publications Office of the European Union, 2018, 140p, doi: 10.2760/11251.
- [12] Dictamen de iniciativa 526° pleno del cese de 31 de mayo y 1 de junio de 2017. Dictamen del Comité Económico y Social Europeo sobre la «Inteligencia artificial:

lares, Actualización PAPPS 2018, Aten Primaria 50(Suppl 1) (2018) 4–28, doi: 10.1016/S0212-6567(18)30360-3.

- [15] J. Marrugat, I. Subirana, E. Comin, et al., Validity of an adaptation of the Framingham cardiovascular risk function: the VERIFICA study, *J. Epidemiol. Commun. Health Rev.* 61 (2007) 40–47.
- [16] Aman Ajmera, Framingham Heart study dataset [online dataset]. Kaggle Inc; Publicado y actualizado 7 nov 2017. URL < <https://www.kaggle.com/amanajmera1/framingham-heart-study-dataset> > (accessed 8 mar 2019),

- [17] J.J. Beunza, R-studio code for Machine Learning algorithms applied to 10-year coronary risk in the Framingham Heart Study database, GitHub Inc. URL < <https://github.com/Juanjobeunza/Aprendizaje-Automatico-FRAMINGHAM> > (published on March 28, 2019), Updated and accessed on March 28, 2019.
- [18] E. Puertas, Comparison of machine learning algorithms for the prediction of coronary heart disease by using the Framingham data set, GitHub Inc. URL < https://github.com/epuertas/framingham_Rapidminer > (published on March 22, 2019), Updated on March 22, 2019, (accessed on March 22, 2019).
- [19] K.M. Lang, T.D. Little, Principled missing data treatments, *Prev. Sci.* 19 (2018) 284–294.
- [20] R.J. Little, D.B. Rubin, *Statistical Analysis with Missing Data*, John Wiley & Sons, Inc., Hoboken, NJ, 2019.
- [21] S.F. Weng, J. Reps, J. Kai, et al., Can machine learning improve cardiovascular risk prediction using routine clinical data? *PLoS ONE* 12 (2017) 1–14.
- [22] E. Puertas, Enrique Puertas Ganador del Premio FUJITSU OPEN DATA, URL < <http://projectbasedschool.universidadeuropea.es/Enrique+Puertas+Ganador+del+Premio+FUJITSU+OPEN+DA> > (published 9 Jun 2015), (accessed 28 mar 2019).
- [23] Ganadores del II hackathon de tecnologías del lenguaje, URL < http://projectbasedschool.universidadeuropea.es/II_HackathonTecnologiaLenguaje > (accessed 28 mar 2019).

Healthcare Analytics 2 (2022) 100016

Contents lists available at [ScienceDirect](#)

Healthcare Analytics

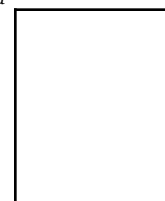
journal homepage: www.elsevier.com/locate/health

An artificial intelligence model for heart disease detection using machine learning algorithms

Victor Chang^{a,*}, Vallabhanent Rupa Bhavani^b, Ariel Qianwen Xu^b, MA Hossain^c

^a Department of Operations and Information Management, Aston Business School, Aston University, Birmingham, UK

^b Cybersecurity, Information Systems and AI Research Group, School of Computing, Engineering and Digital Technologies, Teesside University, Middlesbrough, UK ^c Vice President Office, Cambodia University of Technology and Science, Phnom Penh, Cambodia



ARTICLE INFO

Keywords:

Artificial intelligence
Heart disease detection system Machine learning
Predictive analytics
Random forest classifier algorithm

1. Introduction

1.1. Introduction

ABSTRACT

The paper focuses on the construction of an artificial intelligence-based heart disease detection system using machine learning algorithms. We show how machine learning can help predict whether a person will develop heart disease. In this paper, a python-based application is developed for healthcare research as

it is more reliable and helps track and establish different types of health monitoring applications. We present data processing that entails working with categorical variables and conversion of categorical columns. We describe the main phases of application developments: collecting databases, performing logistic regression, and evaluating the dataset's attributes. A random forest classifier algorithm is developed to identify heart diseases with higher accuracy. Data analysis is needed for this application, which is considered significant according to its approximately 83% accuracy rate over training data. We then discuss the random forest classifier algorithm, including the experiments and the results, which provide better accuracies for research diagnoses. We conclude the paper with objectives, limitations and research contributions.

easier time gaining crucial information for treating and diagnosing patients. Heart disease is mainly an incorrect symptom of coronary artery disease. It is also known as a cardiac disease; therefore, it is not with cardiovascular disease, which is any blood vessel disease.

Python is a programming language with a high level of object oriented abstraction with a spirited, energetic collection of building options and quick development cycles. As per Loku et al. [1] analysis, it is regarded as one of the safest programming languages with numerous applications in the medical field. Furthermore, it is regarded as a well-liked and well-accepted programming language with applications traversing over AI-based software developments and several other web

* Corresponding author.

ing heart diseases, clinicians and institutions can provide better and improvised outcomes for the patients through scalable and dynamic applications. However, the coding packages and libraries used in this project are Pandas, Matplotlib, IPython, Numpy, Python, SciPy, and many others.

1.2. Problem statement

Currently, the health care sector is generating information from several facilities and patients. By applying the best usage of this data, doctors can easily anticipate superior methods for treatment and enhance the complete delivery system of the health care sectors [4]. One of the most important uses is that the python framework can help make sense and encourage computational facilities in extracting valuable insights from the information over the health care sectors. Moreover, Python is considered to be one of the most renowned programming languages all around the globe. 32% of the UK individuals considered this programming language a secured language for developing healthcare

applications. As per the suggestion of Mathur [2], the python framework is used easily for creating a desktop or web-based application. As per the depiction of Guleria and Sood [3], with the application of python programming in the health care sectors, especially for detect

Heart diseases are often used in exchange for cardiovascular diseases. These kinds of diseases mainly refer to the conditions of blocked or narrowed blood vessels, resulting in a stroke, chest pain or angina, and heart attack. Other kinds of heart conditions, such as those affecting the rhythm, valve, or muscle of the heart, are other types of heart diseases. On the other hand, machine learning is crucial for determining whether anyone has suffered from heart disease. In either case, if these are predicted ahead of time, doctors would have a much

E-mail addresses: victorchang.research@gmail.com (V. Chang), rupabhavani22@gmail.com (V.R. Bhavani), qianwen.ariel.xu@gmail.com (A.Q. Xu), alamgir@camtech.edu.kh (MA Hossain).

<https://doi.org/10.1016/j.health.2022.100016>

Received 12 September 2021; Received in revised form 14 November 2021; Accepted 2 January 2022

Available online xxx

2772-4425/© 2022 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

V. Chang, V.R. Bhavani, A.Q. Xu et al. *Healthcare Analytics 2 (2022) 100016*

Fig. 1. Percentage of people preferring Python programming language for healthcare applications.
Source: [5].

applications [5], see Fig. 1. High levels of LDL cholesterol, or “bad” cholesterol, can cause the most common form of heart disease,

coronary artery disease (CAD). It is a plaque that has developed up in the arteries of the patient's heart. CAD has no symptoms in its early stages. Patients can experience symptoms, such as chest pain,

shortness of breath, and fatigue when plaque grows large enough to obstruct blood flow.

Additionally, the health care projects made using the Python language must deal with HIPPA (Health Insurance Portability and Accountability Act) requirements for dealing with healthcare records. In this context, as per Nithya and Ilango [6] depiction, Python supports computer security, as it has built-in tools that provide software-defined security. However, according to McPadden et al. [7], Python is currently used in the health care field for data science and machine learning applications that improve patient outcomes. As per the opinion of Panesar [8], the algorithms of machine learning encourage health care analytics to use Python, as developers can easily establish tracking and health monitoring applications. Thus, in this case, also, python programming is used for detecting heart disease.

1.3. Aims and objectives

1.3.1. Research aim

The research aims to detect heart disease using the python programming language.

1.3.2. Research objectives

The objectives of the study are as follows:

To critically analyze the ways python language is used to detect heart disease.

To critically investigate the previous activities and apply a suitable methodological approach for superscribing the identified problem. To critically apply data interpretation strategies in python language for health problem detection.

To critically assess the artifact or product with the help of cybersecurity approaches using appropriate methods and identifying the limitations and strengths of the work.

1.4. Research questions

The research questions are —

How would Python language help in detecting heart diseases among the patients?

How can the previous activities be critically investigated for applying appropriate methodological approaches towards addressing the identified issues?

How can the strategies for data interpretation be applied and can

V. Chang, V.R. Bhavani, A.Q. Xu et al. *Healthcare Analytics 2* (2022) 100016

language also complies with the HIPAA checklist for assuring the safety of medical information. The major causes of heart disease are diabetes, obesity, unhealthy diet, overweight, excessive alcohol use, and physical inactivity. Therefore, heart disease includes arrhythmia that is considered as atherosclerosis is the hardening of the arteries caused by a heart rhythm abnormality. During a heart attack, some people experience these symptoms. Additionally, pain that spreads to the arm, dizziness or light headedness, throat, snoring, and sweating can occur. Heart attacks, strokes, and coronary heart disease, also known as heart failure and coronary artery disease, are much more common in people over 65 than in younger people.

2.2. Demonstration of a deep understanding of an area of an individual interest associated with specialized computing in the health care sectors

One of the most well-known machine learning algorithms tasks is the classification of data. Machine learning tends to be an essential function in this case for extracting knowledge from business activity datasets and transferring it to larger databases. The majority of the machine learning methods rely on a huge number of features that explain the algorithm's behavior, resulting in the model's complexity, indirectly or directly [10]. Many algorithms such as hybrid methods are used in conjunction with logistic regression, naive Bayes, K-nearest neighbor, and neural networks to integrate the heart disease diagnostic algorithms mentioned earlier. Thus, in this case, the system was trained and implemented over the python platform with the help of the UCI (Unique Client Identifier) machine learning deposited benchmark dataset.

Coronary artery disease, arrhythmias (heart rhythm problems), heart abnormalities (such as congenital heart defects), and a variety of

the findings be interpreted for achieving rational and logical arguments? How will the product or the artifact's insistence assist in evaluating third parties with the assistance of appropriate methods?

1.5. Research hypothesis

Python has wide applications in detecting the heart diseases

Python does not have wide applications in detecting the heart diseases

1.6. Sound justification of evidence

One of the most common diseases is heart disease and the most important reason for death in both developed and developing countries. Davenport and Kalakota's [9] review looked at several research findings, including the use of the Python programming language for detection and prediction mechanisms for cardiovascular disease. The Python programming language is being used in disease detection systems, especially for heart diseases, to improve other healthcare-related systems.

2. Literature Review

2.1. Introduction

The project comprises of detecting the presence of heart diseases using Python. The dataset comprised several factors, such as Cholesterol, sex, age, and others. Several other import libraries, such as matplotlib, Numpy, Pandas, warnings, and many others, were used for the project. Correlation matrix, histogram, support vector classifier, K Neighbors Classifier, Random Forest Classifier, and Decision Tree Classifier were used for assessing the outcomes of the specified dataset using a python programming language. Additionally, Python is also considered an open-source language that encourages developing innovative solutions for the health care sectors and supplies better outcomes for the patients, resulting in enhanced care delivery. However, the

2

other disorders are included in the category of heart diseases. Cardiomyopathy and heart infections are among the conditions that fall under this category. The most common measure of heart risk is chest pain, which is a symptom of cardiovascular disease. After that, it has symptoms of Nausea, Indigestion, Heartburn, or Stomach Pain. The paper will exhibit how a program can be created in Python to analyze whether or not an individual is suffering from cardiovascular disease or not [11]. In this paper, the system uses a dataset comprising fourteen characteristics of the test outcomes, carried on around 100 persons. However, the patient suffering from heart disease symptoms will be diagnosed using binary digits, 1 and 0, where 1 will indicate the true value (The patient has heart disease, in other words.) and 0 will indicate the false value (that is, the patient does not have any kind of heart disease). Additionally, co-relation and trends of the obtained features will also be recognized with the help of several features, such as gender, age, cp (chest pain type), chol (cholesterol level), FBS (fasting blood sugar level), exang (exercise-induced angina), thalach (maximum achieved heart rate), old peak (ST depression persuaded by exercise respective to rest), thal (maximum achieved heart rate), ca (number of major vessels).

In this project, initially, the libraries will be imported. Then, the dataset will be loaded, and it will be stored within a variable for printing the information. Finally, the dataset will be imported and the data will be processed. However, after analyzing the outcomes, it is seen that the K-neighbor classifier algorithm showed an 87% score, whereas the support-vector, decision tree, and random forest classifier displayed 83%, 79%, and 84%, respectively [13]. See Fig. 2.

Contrarily, in this case, a co-relation matrix will be used for evaluating the connections within several types of variables. A positive correlation exists between the predictor and the chest pain variable, indicating that the amount of chest pain is directly proportional to the

probabilities of suffering from heart diseases. In this case, chest pain is considered a statistical feature with four values: value 1, value 2, value 3, and value 4, referring to atypical angina, typical angina, asymptomatic and non-anginal pain, respectively [14]. A negative correlation among these variables would indicate that more amount of blood is required by the heart.

2.3. Development of an approach for addressing the significant research areas or practices over specialized computing areas in health care sectors

However, a major benefit of Python within the health care sector is that it assists in making sense of the information by working with Machine Learning and AI within the healthcare sectors. As per the analysis of Ozgur et al. [15], the development services of Python is a suitable option for a strong and powerful language to encourage computational abilities in obtaining valuable insights from the information of the patients suffering from heart diseases, that will, in turn, help in supporting healthcare based applications. It is convenient in case one has to deliver the diversity of developing something with the help of an internet connection or has autonomously worked without any internet connection. As per the opinion of Srinath [16], the pliability of running over numbers of operating systems is compounded by a large district and a distinct syntax. Moreover, Python proved to be a suitable language for evaluating huge datasets, with the help of machine learning algorithms in receiving significant insights [17]. The language is also favored by data scientists due to the availability of extensive libraries, such as SciPy, Pandas, Numpy, and many others.

2.4. Demonstration of the capability to evaluate, synthesize, and search the information's from the appropriate sources in health care sectors

In this project, the information was gathered from outside databases and a logistic regression was performed during Python. As per the analysis of Jiang et al. [18], several pieces of information are also used for determining the attributes of datasets. For instance, induced angina for the exercise, maximum heart rate, resting blood pressure, resting electrocardiographic measurements, fasting sugar level, thalassemia level, induced depression, number of major vessels,

V. Chang, V.R. Bhavani, A.Q. Xu et al. *Healthcare Analytics 2* (2022) 100016

and many others were used for representing the datasets comprising several values. However, the sex of a person can be evaluated using two values, either 0 and 1, where 0 indicates female and 1 refers to male. Contrarily, the chest pain categories will be evaluated with the help of four values, 0, 1, 2 and 3, indicating asymptomatic condition, atypical angina, non-anginal pain, and typical angina, respectively. However, a confusion matrix is also used for generating false positive and negative outcomes. Moreover, as opined by van den Burg et al. [19], the details for the regression analysis are obtained from adequate CSV files. On the other hand, the classification scores for detecting heart disease can also be obtained. In contrast, help vector classifiers, decision tree classifiers, random forest classifiers, and a variety of other machine learning algorithms are only a few examples. However, in this case, the data wrangling procedures will also be used for determining the relation between the negative and positive binary predictor. As per the depiction of Holdgraf [20], this self-service data wrangling equipment helps deal with more complicated information rapidly and generates accurate outcomes to reach superior decisions.

Additionally, the features are also compared with positive and negative heart patients. From investigating all the information, it has been found that the positive patients experienced increased heart rates and displayed around one-third of the ST depression's amount persuaded by exercise associated with old peak [14]. Thus, developers can effectively use Python to build the required models in predicting heart diseases before they become severe.

2.5. Critical application of the cybersecurity techniques to ensure conformities with networking configurations and management system of information security within the healthcare sectors

Python programming language is mostly chosen to be used within the health care sectors as the cybersecurity professionals can accomplish the project efficiently. As shown in Fig. 3, as per the opinion of Calix et al. [21], the language is also used for decoding and sending

Fig. 2. Conceptual framework.
Source: [12].

packets, network scanning, port scanning, accessing servers, discovering hosts, and analyzing malware. Additionally, it is also useful to conduct a stream of cybersecurity applications, such as malware analysis and scanning. Moreover, as the health care sector comprises a huge number of confidential information of the patients, this language is well suited for developing an application within this sector [22]. It has an executive and well-defined immaculate method. Moreover, the availability of a huge number of libraries also decreases the amount of effort required to conduct specific tasks, such as cyber threat analysis, detection, penetration testing. The language also has a simple syntax that can easily be picked up by the new developers entering this cybersecurity field [23]. Thus, this language is chosen for developing a system for detecting heart diseases.

2.6. Literature gap

The gap in the literature is the availability of minimal information associated with the creation of heart disease detection. Python is a programming language used by the system within the health care systems. As per the depiction of Bau et al. [24], python programming, being a concept-oriented programming language, numerous equipment is involved in developing a prophetic model. However, the selection of the appropriate technology or the appropriate tool can help in the development of a proficient model. Complexity is another issue that is being faced. Additionally, there might also be a lack of the desired

resources.

Python language is the most appropriate language for detecting patients suffering from heart diseases. It is one of the robust languages that foster computational capabilities in gaining valuable insights from the information of the patients suffering from heart diseases. Thus, it is apt for the health care sector. Moreover, it also complies with the HIPAA regulations that ensure medical information safety. However, the project will be using a database from external sources and the libraries will be imported. Loading of the datasets will occur following it, and it will be seized within a variable to reproduce the information. At last, the dataset will be imported with specific import libraries, followed by data processing.

3. Research Methodology

3.1. Overview of methodology

Machine Learning (ML) is important in predicting the existence or absence of heart arrhythmia, locomotor disorders, heart diseases, and other conditions. It was expected well to provide significant insights to physicians, allowing them to adjust their diagnosis and care on a patient-by-patient basis. This project follows the **random forest algorithm** for developing heart disease detection, and this algorithm is used for the methodology to detect heart disease using Python.

Fig. 3. Advantages of python programming language in cybersecurity scanning.
Source: Influenced by [21].

A model regarding ML has played a significant role in creating accuracy and determining results with the aid of training data. This model is considered very significant due to its 83% (approximately) accuracy rate over training data [25]. This accuracy level has been significantly highlighted with the aid of a confusion matrix in terms of using accuracy score calculations. Based on the view of Iwendi et al. [26], through the implementation of the confusion matrix, the accuracy is printed as well as displayed on the screen. This model tends to exhibit an approximate 70% of accuracy rate in terms of evaluating test data. It is effective for receiving higher accuracy levels of scores in relation to existing ones. Based on this context, it can be highlighted that selection of “random forest” is implemented based on a particular dataset as well as a decision tree. A vote for the forecasted outcome can be found in this research. Furthermore, the final forecast is based on the highest number of votes that will be presented as the final result. See Fig. 4.

This algorithm establishes the robust as well as high accuracy of the data. This algorithm also allows the number of participative decision trees within the process. It does not lead to over-fitting issues. The reason behind that is that it considers the average of all predictions and cancels out the prejudices. This random forest algorithm is also useful for regression as well as classification issues [28]. Using Machine Learning algorithms, we can predict possible Heart Diseases in people. There are a variety of Machine Learning algorithms, including the K Neighbors Classifier, Support Vector Classifier, Decision Tree Classifier, and Random Forest Classifier. The Random Forest algorithm is a set of algorithm models demonstrating many decision trees using bootstrap ping, random subsets of tools, and average voting to make predictions. For a classification issue, Random Forest provides the probability of belonging to the class.

3.2. Research method

Here, import all essential libraries used in the project, such as NumPy that works with arrays and the pandas libraries work with the CSV files and data frames. After that, matplotlib creates charts using pyplot. This library is to define parameters using rcParams as well as color them with cm.rainbow. After that, split the dataset into training as well as testing data.

Machine learning is the science of programming a system that can learn from different types of data, according to Larsen et al. [29].

Python is the most commonly used programming language for this

type of project. It will also replace lots of languages in the industry and have a huge amount of collection of libraries such as Numpy, scipy, pandas, scikit-learn, matplotlib and many more. When doing exploratory data analysis, Pandas dataframe.info() function is useful for providing a concise overview of the data frame. After that, import the dataset and use read_csv() that reads the dataset as well as will save it to the dataset variable. Pandas describe() is also used to display certain simple statistical explanations like percentile, mean, standard deviation, and many others. A set of numeric values if not part of a data frame. This method accepts a series of strings and returns a variety of results. After that, use a correlation matrix that understands the data. pyplot was used to display the xticks and yticks in the correlation matrix and the addition of names for the correlation matrix colorbar () displays the matrix's colorbar.

3.3. Use of algorithm with justification

This project is based on the random forest algorithm because this algorithm is considered as flexible as well as easy to use in machine learning. Even without hyper-parameter tuning, this algorithm achieves excellent results in the majority of cases. It is also one of the most widely used algorithms due to its versatility and simplicity. Random forests are considered a supervised learning algorithm, according to Amini et al. [30]. This algorithm is used to categorize trustworthy loan applicants, detect fraudulent behavior, and predict diseases.

The random forest algorithm belongs to the machine learning group, is a useful learning method under supervision. The ensemble learning theorem is its foundation, which states that it is a tool for solving a complex problem by integrating several classifiers and improving the model's accuracy. The Random Forest algorithm is a classification algorithm that uses a random forest to classify data which combines the results of several decision trees into a single result. The aim is to apply to different subsets of a dataset to increase the dataset's predictive accuracy. The random forest is formed in two steps: The first is to mix and match to make the random forest, you will need $\frac{1}{2}$ decision trees in total, and the second step is to make predictions for each of the trees you made in the first step. The following steps and diagram can be used to illustrate the working process:

Fig. 4. Algorithm of random forest.

Source: Influenced by [27].

- Step 1: Pick K data points at random from the training collection.
- Step 2: Build decision trees for the data points you've chosen (Subsets).
- Step 3: Choose the number $\diamond\diamond$ for the number of decision trees you want to build.
- Step 4: Repetition of Steps 1 and 2.
- Step 5: Find the predictions of each decision tree for new data points, and allocate the new data points to the group with the most votes.

The steps for implementation are listed below.

The first phase is data pre-processing. The training set is then equipped with the Random forest algorithm, and the test result is predicted. The training set should be fitted to the Random forest algorithm. To make it fit, we will use the RandomForestClassifier class from the sklearn.ensemble library. To visualize the training set results, we will plot a graph for the Random forest classifier. Last but not least, assess the accuracy of the results to build the uncertainty matrix and then visualize the test and training set outcomes.

After that, data is preparing for training and scaling. After data scaling training has been generated for random forest algorithm.

```
fromsklearn.ensemble import RandomForestRegressor regressor =
RandomForestRegressor(n_estimators = 20, random_state = 0)
regressor.fit(X_train, y_train)
y_pred = regressor.predict(X_test)
```

Muñoz et al. [31] also mentioned that the regression approach is used in the 'sklearn.ensemble' library's RandomForestRegressor class. Additionally, the RandomForestClassifier class of the 'sklearn.ensemble' library was classified. The 'n_estimators' is a parameter in the RandomForestClassifier class. The number of trees in our random forest is determined by this parameter and the search information for all of the RandomForestClassifier parameters.

After that, classification faced challenges. The metrics used to eval⁶

3.6. Use of the required software with justification

This project required the dataset and jupyter notebook editor that used Python's package manager, pip instead of anaconda. Jupyter notebook is a free, open-source, interactive web component, also known as

uate an algorithm are accuracy, confusion matrix, precision recall, and F1 values.

```
fromsklearn.metrics import classification_report, confusion_matrix,
accuracy_score
print(confusion_matrix(y_test,y_pred))
print(classification_report(y_test,y_pred))
print(accuracy_score(y_test, y_pred))
```

The accuracy is achieved by our random forest classifier. To enhance the accuracy using parameters of the RandomForestClassifier class as well as to see improvement on the results.

3.4. Data analysis

Data analysis works with categorical variables and it will break particular categorical columns 1s and 0s are placed in dummy columns. Furthermore, column Gender has values of 1 for males and 0 for females, which can be divided into two columns, one with the value 1 for true and the other with the value 0 for false. The idea of decision trees is taken to the following level with this random forest classifier and building a forest of trees, each of which is made up of a random selection of features from the total features. Here is a list of how many trees will be used to estimate the class.

3.5. Data output

There are many classifiers; therefore, random forest classifiers give better accuracy to this project. For prediction, this project uses a variety of medical parameters such as age, sex, blood pressure, cholesterol, and obesity. Aside from that, the EHDPs predicts a patient's heart disease. Those patients develop heart disease because of their condition. This project work allows for substantial information, health causes, relationships linked to heart disease, and trends to be identified.

a computational notebook, as explained by Reich et al. [32]. Software code, numerical performance, explanatory text, and multimedia tools can all be combined into a single document by the researchers. This app allows everyone to view and share documents containing images, live code, calculations, visualizations, narrative text, data cleaning and transformation, numerical simulation, mathematical modeling, data

V. Chang, V.R. Bhavani, A.Q. Xu et al. Healthcare Analytics 2 (2022) 100016

visualization, machine learning, and other features. Normally, Python programming allows building code for multi-purposes. Apart from this, the anaconda is considered most likely preferred for Data science projects. It will also provide pre-built libraries to help projects like machine learning get up and running quickly. Jupyter is a data science platform designed for Python-based data science applications. Besides, it would lower the barriers for data scientists, as Jupyter has simplified documentation, data visualization, and caching a lot easier. Apart from this, Mendez et al. [33] have described that Jupyter helps programmers write code and display the results in real-time without waiting for other parts of the program to finish. If it is a code that is training an ML model or a code that is downloading gigabytes of data from a remote server, Jupyter caches the output of every cell that is running. Jupyter Notebook is a platform-agnostic and language-agnostic programming environment. Jupyter can be interpreted in a variety of languages. Apart from this, Yin et al. [34] have narrated that Jupyter supports visualizations. It also includes rendering some of the datasets such as graphics and charts that are generated from codes with the help of modules such as Matplotlib, Plotly and many more. Sample documentation code is available in Jupyter notebook. The more Jupyter is used, the easier it is for developers to describe their codes line-by-line with suggestions attached all along the way. The better the Jupyter is used, the more interactivity and explanations can be added once the fully functional code.

4. Data findings and analysis

4.1. Introduction

Python is widely regarded as the most effective and useful programming language. It contains a number of libraries that are used in this machine learning project. A subset of the Artificial Intelligence model is this approach to Machine Learning. Python libraries are used to make predictions with SKLEARN, which is a machine learning prediction tool.

4.2. Critical analysis regarding the description of heart disease detection

The healthcare industry generates massive quantities of data, also known as big data. The most common cause of death is heart disease, worldwide and a significant public health issue. In medical science, the detection of heart disease in its early stages has become one of the most serious problems. According to Ramalingam et al. [35], the RR interval, QT interval, and QRS interval are some of the characteristics that are studied in heart disease detection. This method determines whether or not the patient is well. This method determines whether the patient is normal or not.

On the other hand, Subhadra and Vikas [36] have opined that heart disease is considered a fatal human disease that repeatedly enhances the world in both undeveloped and developed countries and increases with consequently and last of all it causes death. This system helps to classify a complex as well as large medical dataset along with that to detect heart disease. Along with that, it detects some steps that use a map-reduce algorithm to both detect the disease and reduce the size of the dataset.

This output represents a histogram of the dataset of heart disease detection. Here, use the code of `df.hist(figsize=(10,10))`. This histogram is used to interpret discrete data visually and to summarize it. By showing the number of data points that fall within a given range of values, necessitates focusing on the most significant points, or facts, of numerical data. See Fig. 5. Exploratory Data Analysis (EDA) finds relevant data, detects the mistake, checks assumptions, and determines

the correlation between these explanatory variables. This EDA allows data analysis that excludes statistical models as well as inferences. Apart from this, Ambale-Venkatesh et al. [37] have narrated that the risk factors are because the random forest algorithm is used to consider and forecast heart disease, and the study is done using publicly accessible heart disease data. There are 304 records in this dataset, each with 14 attributes, including age, gender, and more. In order to predict heart disease, a random forest algorithm is used for data visualization as well as data analytics. Along with that, Dogan et al. [38] have described that this research paper discusses classification performances, pre processing methods, and evaluation metrics. Furthermore, the result of the visualized data shows that the prediction is considered correct. The precision of this method was 83 percent. The ANN-based three-stage method for predicting heart disease. Heart disease prediction using deep neural networks as well as his proposed model performed well. Apart from this, it also produced good outcomes. See Fig. 6.

4.3. Data interpretation of the selected dataset

Data interpretation is considered a process that reviews data over some predefined processes. It helps to assign some data. It involves the outcomes of data analysis and makes inferences on the relations. Mahdavi et al. [39] described that data interpretation also helps with data analysis, data collection, and data presentation in row form. A machine learning approach's interpretation is described as the process of attempting to comprehend a machine learning model's predictions. Python libraries are helping to build the data interpretation through a machine learning approach. To understand why the classifier chooses a specific class, interpret the models using the predictions as well as the parameters. The data analysis of machine learning models facilitates the method of attempting to comprehend a machine learning model's predictions. This involvement of heart disease detection has two significant phases: it has monitored the evaluation metric and tried various ideas of algorithms selection that enhance and develop the more robust approach. It is also essential to interpret the models using the parameters as well as predictions to understand the classification chose the exact class. On the other hand, Tauzin et al. [40] have declared that data interpretation is a part of data analysis in the modern past, cost-effective technologies, and it acts as an essential part of the health sector involves emergency situations as well as outbreaks of disease. Data interpretation by the term "machine learning" refers to a form of data processing. The construction of analytical models is automated using this data analysis. Data analysis is an artificial intelligence specialization focused and based on the assumption that computers can learn and recognize patterns in data. It also makes decisions with little to no human input. See Fig. 7.

4.4. Data interpretation strategies using python language in detecting heart problem

Stats models are considered a python model that helps the users explore data perform statistical tests and estimate statistical models. It has a go-to language for data analysts. According to Peters et al. [41], it depends on important data analysis libraries, data pre-processing, and, last of all, exploratory data analysis. Data analysis libraries contain libraries and packages, and those are open-source and widely used to crunch data.

Fundamental Scientific Computing

Numpy is a numeric library using Python that can perform linear algebra, Fourier transforms, and a random number. Along with that,

SciPy is a scientific library of Python, and this library involves a high-level science module.

The criteria of machine learning approach interpretation processes as well as explore the techniques for interpretation dependent on the scope. This study is focused on an in-depth understanding of current model interpretation approaches and their drawbacks and challenges. It will also go through the age-old trade-off between model accuracy and model interpretability. Finally, consider some of the most popular model interpretation strategies.

4.5. Data manipulation and data visualization

Pandas are used in the form of data frames. The panda's library also provides data from various file formats, such as CSV, Excel, plain text, JSON, SQL and many more. Peters et al. [41] explained that data manipulation is considered the process of data transformation, formatting, and structuring. Apart from this, the Matplotlib library is used for plotting as well as visualizing data. This library allows plot graphs, Heatmaps, line plots,

Fig. 6. Heart disease detection using a deep learning approach.
Source: [38].

Fig. 7. Data interpretation.
Source: [40].

histograms as well as a lot more. It is embedded in GUI toolkits. See Fig. 8.

This bar graph of the counterplot represents the target variable of the dataset. It means the sex of the patient who faced much heart disease. 0 represent female and 1 represents male of the patients. Here, used of the code is `sns.set_style('whitegrid')`

```
sns.countplot(x='target',data=df,palette='flare')
```

Machine Learning

V. Chang, V.R. Bhavani, A.Q. Xu et al. *Healthcare Analytics 2 (2022) 100016* 4.9. Assessment of artifact regarding cybersecurity approach

As Carleo et al. [42] explained, Scikit Learn of machine learning is considered a free machine learning library that helps to build on NumPy, SciPy, and Matplotlib. This library contains efficient components for statistical model development. It also runs different classification, regression, as well as clustering algorithms. It is also integrated well with pandas through working on data frames.

The cybersecurity consultants are deploying a revolution as users can transition from managing the perimeter that is extracting as well as analyzing any residue left by cyber thieves on every endpoint device such as laptop, desktop and many more. As narrated by Jia et al.

Fig. 8. Count Plot the sex.

Importing Libraries and Loading the Data

This approach imports libraries as NumPy and pandas and then loads the dataset. The CSV is the most generally used format for that machine learning data is considered as presented. This CSV file is used for automatically assigning names. Otherwise, it is labeled if the file does not have a header, each column of the dataset manually names the attributes.

4.6. Exploratory data analysis

The EDA data analysis is used to get a better understanding of data and look for the data. For statisticians, it is similar to a kind of storytelling. It allows for the discovery of trends and observations within data using visual methods. Aside from that, EDA is frequently used as the first step in the data modeling process. It will explore the dataset as well as perform the exploratory data analysis. It follows handling missing value, outlier treatment, encoding categorical variables and normalizing, and finally, scaling and removing duplicates.

4.7. Evaluation of data

Evaluation of data introduces Python, and it is a mathematical data processing language that uses Pandas Data Frame objects to store data. Importing, washing, and converting data in preparation for review is the work involved in data analysis.

4.8. Data cleaning

Data cleaning is considered as preparing data for analysis by cleaning the raw data that prepare the data visualizing data and predicting data. It is the method of removing false, corrupted, improperly formatted, and redundant data from a dataset that would otherwise be incorrect, corrupted, and incomplete.

Correlation Matrix Plot

According to Harper et al. [43], a correlation Matrix Plot is a covariance matrix that is a metric called the correlation that defines the strength of the linear association. The Correlation matrix sums up the power and direction of a linear relationship between two variables, and it allows values between -1 as well as $+1$. The feature of the correlation matrix displays the correlation between the coefficients. A particular random variable is considered correlated with each of its other values. This represents an excellent way to check correlations among features by visualizing the correlation matrix as a heat map. See Fig. 9.

The relationship between age, sex, cp, trestbps, chol, FBS, restecg, thalach, exang, oldpeak, slope, and ca is expressed as a graph. The linear relationship between two continuous variables is defined using the dataset's correlation.

vices include investigative activities, and when assessments are¹⁰
V. Chang, V.R. Bhavani, A.Q. Xu et al. *Healthcare Analytics 2* (2022) 100016

[44], they focus on finding artifacts that convey every user as well as applications. These applications are ever interacted with by the system and find these artifacts deep in the OS system files, memory, file systems, as well as more systems. Artifacts can reveal evidence even when the perpetrators proclaim innocence. It can also show the cyber criminal's intent by showing their Internet searches and websites visited.

On the other hand, Galinec et al. [45] have narrated that concern about cybersecurity services and artifacts provides important clues about unauthorized access by unauthorized entities. Cybersecurity ser drawn, the artifacts help corroborate the findings. Mainly, the root cause of a cyber-attack is considered as never discovered, nor are the threat actors ever found.

4.10. Different methods for cybersecurity

Computers, mobile devices, networks, servers, and data can be protected by cybersecurity systems, which are similar to security gatekeepers to ensure that the information on the computers and the data stored on them are not vulnerable to external threats. According to Fan et al. [46], cybersecurity has various security methods, such as cloud security based on the cloud. Because of its increased privacy, its storage has been a common choice over the last decade. Data is stored in the cloud, which is considered more reliable because it is protected by a software program that tracks activity and can warn users if anything unusual occurs with their cloud accounts. Following that, Network Protection protects an internal network from external threats by enhancing network security. Firewalls supervised internet access, anti-spyware, and antivirus programs are all part of network security. Apart from this, Jin et al. [47] have opined that application Security protects the data. This information is kept on the applications that are used to operate the business. Applications are more accessible across various networks that are particularly vulnerable to cyber-attacks, and the safety of applications with cybersecurity antivirus programs, firewalls, and encryption services is critical. It is also essential for the control access to carefully manage the physical access to that premises and the computer network. Control access restricts access to outside threats as well as unauthorized users. Apart from this, Anees and Hussain [48] have explained that it has limited access to the data. Otherwise, it also has limited access to the services or application controls. Control access is limited to sending as well as receiving certain kinds of email attachments. Hence, firewalls are effectively gatekeepers between the system as well as the internet. It has one of the most effective defenses against cyber threats, including malware and viruses. The firewall of the device properly checks them regularly and it ensures that it is up to date in terms of applications and firmware. Otherwise, it will not be completely successful. Apart from this, use for security software such as antivirus, anti-malware, and anti-spyware that help detect the malicious code after detection will be removed from it. Along with that, Monitor for intrusion that detects unusual network activity as well as monitor the system. If a monitoring system detects a possible security breach, it may issue a warning based on its discovered features, such as an email alert. Apart from that, raising knowledge has to assist in the security of the company and ensure that understanding of the user's role and any relevant policies and procedures that provide them with regular cybersecurity awareness and training. Besides this, many important security enhancements are included in updates to help defend against various threats. This is referred to as bugs or glitches, and it ensures that apps and systems are up to date to avoid being targeted by criminals. See Fig. 10.

Fig. 9. Correlation of each feature in the dataset using the heatmap.

Fig. 10. The architecture of the cybersecurity knowledge base.
Source: [47].

4.11. Critical evaluation regarding the identified methods of cybersecurity

The traffic passing through cybersecurity is monitored by a firewall. This network allows information to flow in the form of packets. Apart from that, the firewall examines each of these packets separately for

V. Chang, V.R. Bhavani, A.Q. Xu et al. Healthcare Analytics 2 (2022) 100016

course of action is to prevent those hosts from gaining access to the device. If a user believes that they need protection from this type of unauthorized access, an access policy may be implemented. One of a user's most important characteristics is privacy. Hackers are constantly

any potentially dangerous attacks. If the firewall comes across these threats by chance, it will automatically block them. Firewalls also come with an access policy. Apart from this, Anees and Hussain [48] have narrated that it can support development for certain hosts as well as services. Following that, attackers exploit certain hosts, and the best

11

on the lookout for data privacy to gather knowledge about the user. As a result of the use of a firewall, various services are provided where the domain name service and the finger, which are used by a website, are blocked. On the other hand, Jin et al. [47] have described that the hackers have no chance of getting a privacy description. Furthermore,

firewalls will prevent the site system’s DNS information from being accessed. According to, the names and the IP address are not available to the attackers. Firewalls, on the other hand, require an expenditure that varies depending on the type of cost. Hardware firewalls are believed to be more costly than software firewalls. Hardware firewalls often necessitate installation as well as maintenance, which can be expensive.

Access control allows an effective path of the unwanted control entry of the logical as well as physical assets. On the other hand, access control is more damaging; it is caused when the key is compromised, which is the disadvantage.

Client security gives them peace of mind by making them aware of the security system. In addition, cybersecurity is becoming increasingly important, with the advantage of potentially increasing revenue and marketability. On the other hand, Security flaws in software are commonly caused by bugs. Many types of bugs can be found in software. Minor issues, such as incorrect print output rendering, are among them. Otherwise, it is an error message that has not been formatted properly. Some bugs are security flaws that could allow unauthorized access to sensitive information. As a result of these vulnerabilities, attackers will take advantage of security flaws. Apart from that, Fan et al. [46] have mentioned that the protection and safety of their confidential personal details is one of the users’ main concerns, which makes them reluctant to share their information and, as a result, make online transactions.

On the other hand, authentication allows the process of ascertaining. User and session authentication in software that has been improperly configured is extremely vulnerable. Aside from that, it has security solutions to ensure that customers’ information is secure in the users’ system. The best business practices can increase the number of buyers, increase revenue, and create a positive consumer reputation. In the software development industry, security misconfiguration is a

determines accuracy flawlessly after importing CONFUSION MATRIX. The random forest algorithm creates and merges many decision trees based on the given scenario to generate a more accurate and stable prediction. The Random forest classifier is a bagging classifier with common issue.

$$f(x) = \frac{1}{n} \sum_{i=1}^n f(x_i)$$

4.12. Role of python in detecting heart disease

This project predicts the heart disease of the patient by extracting the patient’s medical history. It leads to fatal heart disease from a dataset. It contains the patient’s medical history, such as sugar level, blood pressure, chest pain, and age and gender. Based on the given scenario, Python provides an accuracy of the heart disease detection and achieves 83% accuracy using the random forest classifier. This classifier shows a collection of decision trees drawn randomly from the training set described by Mehmood et al. [49]. It combines the votes of different decision trees are used to determine the final class of the test object. It works by building a large number of decision trees during training and then generating the class that represents the mode of the classes; otherwise, the average of heart disease detection prediction mean is calculated. Therefore, there is not much point in enhancing this number of estimators to enhance the accuracy further [Refers to appendix].

4.13. Conclusion

The other library used in this prediction is SEABORN, which links all of the attributes together. Last but not least, the confusion matrix

5.3. Decision tree classifier

V. Chang, V.R. Bhavani, A.Q. Xu et al. *Healthcare Analytics 2* (2022) 100016

hyperparameters, much like a decision tree.

5. Discussion

5.1. Introduction

In this section, discuss the method that is developed using Python. This language allows lots of packages and libraries that are dependent on “machine learning algorithms”. These “machine learning algorithms” as well as their outcomes are compared with the proposed model.

5.2. K-Neighbors classifier

The k-Neighbors classifier recognizes a vector based on the plurality vote of its neighbors. The classification route of this classifier is very important. According to Tjahjadi and Ramli [50], this method is used based on the neighbor’s majority vote. Along with that, the k-neighbor is given a weight of age. It is much more distant than the others. This distance between the vector and the neighbor is represented by d, and the weight assigned is constituted by 1/d. Apart from this; For instance based learning, this classification technique is used. According to, the classification of heart disease detection is when the estimation also occurs.

On the other hand, Pedrozo et al. [51] have opined that the K Nearest Neighbor is considered a machine learning algorithm that can solve issues of classification and regression. The K-Nearest Classifier is used to determine the data’s accuracy. As a result, it can determine which model should be the most suitable among the following models for usage in the future. In Eq. (1), the Euclidean distance is symbolized as

$$d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (1)$$

The decision tree classifier creates a tree from data item observations. This data item is made up of branches that connect the branches, which are made up of target values that make up the leaves. Assegie and Nair [52] narrated that the decision tree is used to classify heart disease detection after that. The leaves show the class labels and branches, which show the features that are the leading features of the labels for the classes. It is used to identify data objects using visual components. The decision tree model’s accuracy value is calculated by inputting the XTrain parameters as well as the YTrain, which is the fit shape The Decision tree model’s score is then discovered, bypassing the XTest and YTest parameters to the system score() function, which searches for the Decision tree model’s score. Apart from this, Herbold [53] has stated that the decision tree algorithm is considered as a flowchart such as a tree structure. It has an internal mode that displays the outcome and the branches that constitute a decision rule. Along with that, every leaf of the node includes the outcome. Apart from this, the top node node is known as the root node of the decision tree algorithm. See Fig. 11.

Fig. 11. Decision tree classifier.
Source: [53].

5.4. Support vector machines

A Support Vector Machine (SVM) algorithm generates a model that classifies new examples into one of several categories. Aside from that, it is divided into two groups. The SVM model can be referred to as a “non-probabilistic classifier”, since it can be thought of as a

right (j) = child node from right split on node j

sub () is used because subscript is not available in medium. After that, the value of each function on a decision tree is determined as follows:

\sum

node splits on feature

=

\sum

space-based characterization of instances. A hyper-plane is considered a

all nodes (3)

building that divides the various examples into various categories. SVM supervises the machine learning algorithm used for both classifications and regression problems, as Zahariev et al. [54] define. The Support Vector simplifies the coordinates of separate observations. After that, the SVM classifier is considered as a frontier that best segregates the two classes. The performance is called the SVM model’s level of accuracy that inputs the parameters of the XTrain as well as the YTrain that is the method of fit, according to Wang and Liang [55]. After that, the model’s score is input into the system of the score using the parameters XTest and YTest. The score of the SVM model can be found using this tool.

normfi sub i = the feature’s normalized importance i

fi sub i = the significance of i

By dividing by the number of all function importance values, these can be normalized to a value between 0 and 1.

=

all features (4)

At the Random Forest stage, the final feature importance is the average of all the trees. The total number of trees is divided by the amount of the feature’s significance value on each tree. \sum

=

5.5. Random forest classifier

all trees

(5)

Random forest classification is considered an ensemble learning approach used for solving machine learning challenges such as classification and regression. This random forest classification of heart disease detection algorithm works by constructing multiple decision trees. Apart from this, this classifier uses a technique called “bootstrap aggregation”.

The likelihood of a reduction in node impurity is weighted, and it can be improved by hitting the node to determine feature significance. With only two child nodes, Scikit-learn calculates a node’s importance using Gini Importance. Between Eqs. (2) and (6), we show key formulas for random forest classifiers.

$$= \text{left} - \text{right} \quad (2)$$

ni sub (j) = node j significance

W sub (j) = weighted number of samples hitting node j

C sub (j) = node j impurity value j

left (j) = child node from left split on node j

RFfi sub (i) = the importance of feature i calculated from all trees in the Random Forest model

normfi sub (ij) = the normalized feature importance for i in tree j

T = total number of trees

As Huljanah et al. [56] stated, the classification has produced the class that the majority of the decision trees predicted in the forest. On the other hand, For regression outputs, the class with mean predictions of the individual trees. Based on the given scenario, trees and entropy criteria are used for developing the classifier. The entropy criterion is represented by

$$\text{Entropy} = -\sum \log_2 \quad (6)$$

A “Random forest classifier”, according to Mehrang et al. [57], shows a collection of decision trees built from randomly selected sub sets of the training set. The sepsis data is then aggregated from various decision trees, determining the final class of the test object. The Random Forest Prediction depicts the model’s pictorial representation as

well as its precision for the dataset's data collection. Using the Random Forest, the accuracy of the heart disease detection system is 0.821875. Furthermore, by inputting the XTrain and YTrain parameters required for the method of fit, the accuracy value of the "Random Forest Score model" is obtained. Then, to determine the score of this model, use the parameters XTest and YTest in the method score() to determine the Random Forest Score model's score.

5.6. Logistic regression

A widely used classification methodology is the logistic regression model. The variables of the logistic regression show the class. This class is a variable of categorical dependency. Ge et al. [58] explained that the logistic regression includes the logistic equation, with the model as the dependent variable. Along with that, according to Saif et al. [59], a supervised learning classification algorithm called logistic regression is used to estimate the likelihood of a target variable. This algorithm of the target variable is dependent on the categorical method that is used. Aside from that, the dependent variable has a dichotomous nature, which means it can be classified into two groups.

5.7. Summary

These models are compared depending on precision, specificity, f measure, accuracy as well as sensitivity. The VLRANK achieved the highest accuracy when it is compared with algorithms such as naive Bayes, random forest classification, logistic regression and decision trees. With the observation of this approach, Random forest classification gives the parameters for the algorithms are based on the best accuracy score of the data reading.

6. Conclusion

6.1. Conclusion

Based on the given scenario, the first section discusses heart disease prediction using Python. Python is object-oriented as well as it is also a high-level programming language that has quick development cycles and spirited, energetic building options. This language helps better to be able to predict the heart disease pathway accurately. The heart care industry is generating the data from several facilities as well as patients over applying the best use of the strategy of this data. Apart from this, doctors are easily demonstrating this superior model for treatments and it will be improving the complete delivery system of the healthcare sector. This prediction model of heart disease is especially used in heart diseases, clinicians, and institutions that can provide better along with that improvised outcomes of the patients over scalable and dynamic applications and conclude the problem of this model. Apart from this, chapter two discussed detecting the presence of heart diseases using Python. This application depends on the heart disease dataset that involves data of the patients, which are age, sex, chol, treetops, and many more.

On the other hand, individually import libraries such as matplotlib, Numpy, Pandas, warnings, and many more are used in this application. This python language is one of the robust languages that foster computational capabilities in enhancing valuable insights from the information of the patients suffering from heart diseases. However, it also complies with the HIPAA regulations that ensure medical information safety. Aside from that, chapter three delves into the technique for detecting heart disease using machine learning, including several algorithms. Machine learning is important for predicting the existence of a threat. This project follows the random forest algorithm for developing heart disease detection. This algorithm is used for the methodology

to detect heart disease using Python. This application is considered ¹⁴
V. Chang, V.R. Bhavani, A.Q. Xu et al. *Healthcare Analytics 2* (2022) 100016

significant according to its 83% (approximately) accuracy rate over training data.

Along with that, a model regarding ML has played a significant role in creating accuracy and determining results with the aid of training data. According to the scenario, it can be highlighted that the selection of random forest classification is developed based on the exact dataset as well as the decision tree. In this section, the emphasis for discussion is on data analysis. It works with categorical variables along with that; it will break particular categorical columns into dummy columns with 1s and 0s. Apart from this, this part also mentioned that the data output of this application is used for many medical parameters such as age, gender, blood pressure, cholesterol, and obesity for prediction and requirement of the software used in the development application. Besides this, the Machine learning application using Python is a subset of the Artificial Intelligence model and the python libraries are the prerequisites for making predictions that SKLEARN is normally used in machine learning prediction. Except for the Decision Tree, the best values provided by the ML model are provided by Random Forest. This is the simplest method for predicting heart disease to produce precise results.

6.2. Recommendation

Recommendation 1: The aim is to introduce a dataset auditing technological setup for removing issues within the structure, as shown in Table 1.

Recommendation 2: The aim is to require proper training regard ing this machine learning approach for users, see Table 2.

6.3. Linking to objectives

Objective 1:

The first goal defines critical analysis of the python language used to detect heart disease.

In the literature review, Section 2.3 has highlighted the development approach of addressing the significant research in the health care sector. The major beneficial factor of Python within the health care sector is that it assists in making sense of the information by working with Machine Learning and AI within the healthcare sectors. Thus it is related to the first objective.

Section 4.2 has described the analysis according to the details of heart disease detection. Therefore, this objective is successfully met in the research.

Objective 2:

The second objective defines investigating the previous activities critically and applying a suitable methodological approach for super scribing the identified problem.

Section 2.4 has highlighted the capability to evaluate, synthesize, and search the information from the appropriate sources in health care sectors in the literature review. Thus, it is related to the second objective. Section 4.3 has described the data interpretation of the selected dataset. Therefore, the objective is successfully met.

Objective 3:

The third objective defines critically applied data interpretation strategies in python language for health issue detection. In the literature review, Section 2.2 highlighted a deep learning understanding of individual interest associated with specialized computing in healthcare. Thus it is related to the third objective. Section 4.4 has described different data interpretation strategies based on python language.

Objective 4:

The fourth objective defines critically assessing the artifact or product with the help of cybersecurity approaches using appropriate methods and identifying the limitations and strengths of the work.

Table 1

Summary of recommendation 1.

S-specific To introduce a data auditing technological setup

M-measurable It can be measured through the success rate of doctors' treatment in detecting heart disease.

A-attainable A daily examination of the selected dataset will be required to detect heart disease without any error.
R-realistic It will be effective for removing barriers in the dataset and reducing the chances of error within it.
T-time plan 5 months

Table 2

Summary of recommendation 2.

S-specific To introduce a machine learning training

M-measurable It can be measured over success rate according to the patients

A-attainable Arrange the weekly training for the users

R-realistic It will be effective for removing the boundary in training as well as decreases the error in the training

T-time plan 6 months

In the literature review, Section 2.5 has highlighted the Critical application of the cybersecurity techniques that conform to networking configurations within the healthcare industries and the information security management system. Section 4.5 has described various methods of cybersecurity. Thus, it is related to the fourth objective.

6.4. Limitations

Traditional invasive-based approaches and angiography are also considered well-known appropriate techniques for diagnosing cardiac conditions in this study. It is a drawback in heart disease prediction. Apart from this, intelligent learning is dependent on computational techniques that are upright and efficient in predicting the occurrence of heart disease. This prediction method is presented here for the purposes of heart disease prediction and diagnosis. This classification technique depends on different pruning as well as data cleaning techniques. This technique is prepared as well as developed a dataset that is suitable for data mining and selects the proper technique, which provides much better accuracy of this application.

6.5. Research contributions

In this research, we develop a healthcare application to help detect heart diseases among patients and those with symptoms. Based on the random forest algorithm, our work provides better accuracies. It has a very low cost for development. Additionally, research outputs can be used as valuable data for further analyses. To achieve rational and logical arguments, we can then develop better diagnoses to detect heart diseases, including findings and interpretations. The artifact developed from this research will assist in evaluating third parties with the assistance of this appropriate method. Our research is aimed at offering both theoretical and practical contributions to healthcare.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work is partly supported by VC Research (VCR 0000154).

Appendix. Codes

V. Chang, V.R. Bhavani, A.Q. Xu et al. *Healthcare Analytics 2* (2022) 100016

- [2] P. Mathur, Overview of machine learning in healthcare, in: *Machine Learning Applications using Python*, A Press, Berkeley, CA, 2019, pp. 1–11. [3] P. Guleria, M. Sood, Intelligent learning analytics in healthcare sector using machine learning, in: *Machine Learning with Health Care Perspective*, Springer, Cham, 2020, pp. 39–55. [4] V.V. Kumar, *Healthcare Analytics Made Simple: Techniques in Healthcare Computing using Machine Learning and Python*, Packt Publishing Ltd., 2018, Available at: <https://books.google.com/books?hl=en&lr=&id=nwZnDwAAQBAJ&>

```
import numpy as np
import pandas as pd
df = pd.read_csv('dataset.csv')
df.describe()
import matplotlib.pyplot as plt
from matplotlib import rcParams
from matplotlib.cm import rainbow
%matplotlib inline
import warnings
warnings.filterwarnings('ignore')
import seaborn as sns
# get correlations of each features in dataset
corrmat = df.corr()
uniform_data = corrmat.index
plt.figure(figsize=(20,20))
# plot heat map
g = sns.heatmap(df[uniform_data].corr(),annot=True,cmap="rocket")
df.hist(figsize=(20,20))
sns.set_style('whitegrid')
sns.countplot(x='target',data=df,palette='flare')
data=pd.get_dummies(df, columns=['sex', 'cp', 'fbs', 'restecg', 'exang', 'slope', 'ca', 'thal'])
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
standardScaler = StandardScaler()
columns_to_scale = ['age', 'trestbps', 'chol', 'thalach', 'oldpeak']
data[columns_to_scale] =
standardScaler.fit_transform(data[columns_to_scale]) data.head(10)
y = data['target']
X = data.drop(['target'], axis = 1)
from sklearn.neighbors import KNeighborsClassifier
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import cross_val_score
knn_scores = []
for k in range(1,11):
knn_classifier = KNeighborsClassifier(n_neighbors = k)
score = cross_val_score(knn_classifier,X,y,cv=10)
knn_scores.append(score.mean())
plt.plot([k for k in range(1, 11)], knn_scores, color = 'red')
for i in range(1,11):
plt.text(i, knn_scores[i-1], (i, knn_scores[i-1]))
plt.xticks([i for i in range(1, 11)])
plt.xlabel('Number of Neighbors (K)')
plt.ylabel('Scores')
plt.title('K Neighbors Classifier scores for different K values')
from sklearn.ensemble import RandomForestClassifier
randomforest_classifier = RandomForestClassifier(n_estimators=10)
score = cross_val_score(randomforest_classifier,X,y,cv=10)
score.mean()
```

References

- [1] L. Loku, B. Fetaji, A. Krstev, M. Fetaji, Z. Zdravev, Using python programming for assessing and solving health management issues, *South East Eur. J. Sustain. Dev.* 4 (1) (2020).
[5] BelltSoft, Python in healthcare, BelltSoft (2017) Available at: <https://beltssoft.com/custom-application-development-services/healthcare-software-development/python-healthcare>, [Accessed on 5th March, 2021].
[6] B. Nithya, V. Ilango, Predictive analytics in health care using machine learning tools and techniques, in: *2017 International Conference on Intelligent Computing and Control Systems (ICICCS)*, IEEE, 2017, pp. 492–499.

- [7] J. McPadden, T.J. Durant, D.R. Bunch, A. Coppi, N. Price, K. Rodgerson, C.J. Torre Jr., W. Byron, A.L. Hsiao, H.M. Krumholz, W.L. Schulz, Health care and precision medicine research: analysis of a scalable data science platform, *J. Med. Internet Res.* 21 (4) (2019) e13043–e13045.
- [8] A. Panesar, Machine Learning and AI for Healthcare (1–73), Apress, Coven try, UK, 2019, Available at https://iedu.us/wp-content/uploads/edd/2020/01/Arjun_Panesar_Machine_Learning_and_AI_for_Health-iedu.us_.pdf, [Accessed on 5th March, 2021].
- [9] T. Davenport, R. Kalakota, The potential for artificial intelligence in healthcare, *Future Healthcare J.* 6 (2) (2019) 94.
- [10] Analytics Vidhya, Commonly used machine learning algorithms (with python and R codes), 2021, Analytics Vidhya. Available at: <https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/>, [Accessed on 5th March, 2021].
- [11] H. Mayfield, C. Smith, M. Gallagher, M. Hockings, Use of freely available datasets and machine learning methods in predicting deforestation, *Environ. Model. Softw.* 87 (2017) 17–28.
- [12] P. Barot, Why use python in healthcare applications? BoTree Technologies. (2020) Available at: <https://www.botreetechnologies.com/blog/python-in-healthcare-application/>, [Accessed on 5th March, 2021].
- [13] K. Bhanot, Predicting presence of heart diseases using machine learning, Towards Data Sci. (2019) Available at: <https://towardsdatascience.com/predicting-presence-of-heart-diseases-using-machine-learning-36f00f3edb2c>, [Accessed on 5th March, 2021].
- [14] J. Zaidi, Project: Predicting heart disease with classification machine learning algorithms, Towards Data Sci. (2020) Available at: <https://towardsdatascience.com/project-predicting-heart-disease-with-classification-machine-learning-algorithms-fd69e6fcd9d6>, [Accessed on 5th March, 2021].
- [15] C. Ozgur, T. Colliou, G. Rogers, Z. Hughes, Matlab vs. Python vs. R, *J. Data Sci.* 15 (3) (2017) 355–371.
- [16] K.R. Srinath, Python—the fastest growing programming language, *Int. Res. J. Eng. Technol.* 4 (12) (2017) 354–357.
- [17] B. Copeland, Advancing opportunities in health care with python-based machine learning, Chief HealthCare Executive (2019) Available at: <https://www.chiefhealthcareexecutive.com/view/advancing-opportunities-in-healthcare-with-pythonbased-machine-learning>, [Accessed on 5th March, 2021].
- [18] W. Jiang, M. Zhuang, C. Xie, J. Wu, Sensing attribute weights: A novel basic belief assignment method, *Sensors* 17 (4) (2017) 721.
- [19] G.J. van den Burg, A. Nazábal, C. Sutton, Wrangling messy CSV files by detecting row and type patterns, *Data Min. Knowl. Discov.* 33 (6) (2019) 1799–1820. [20] C. Holdgraf, Case study 7: Feature extraction and data wrangling for predictive models of the brain in python, in: *The Practice of Reproducible Research*, University of California Press, 2017, pp. 139–148.
- [21] R.A. Calix, S.B. Singh, T. Chen, D. Zhang, M. Tu, Cyber security tool kit (Cyber SecTK): A python library for machine learning and cyber security, *Information* 11 (2) (2020) 100.
- [22] H. Cui, F. Li, Andes: A python-based cyber–physical power system simulation tool, in: 2018 North American Power Symposium (NAPS), IEEE, 2018, pp. 1–6. [23] Python.org, What Is Python? Executive Summary, Python.org, 2020, <https://www.python.org/doc/essays/blurb/#:text=Python%20is%20an%20interpreted%2C%20object,programming%20language%20with%20dynamic%20semantics.&text=Python's%20simple%2C%20easy%20to%20learn,program%20modularity%20and%20code%20reuse>, [Accessed on 5th March, 2021].
- [24] D. Bau, J. Gray, C. Kelleher, J. Sheldon, F. Turbak, Learnable programming: blocks and beyond, *Commun. ACM* 60 (6) (2017) 72–80.
- [25] Medium, Heart disease detection using machine learning in python, Medium (2021) Available at: <https://randerson112358.medium.com/heart-disease-detection-using-machine-learning-python-a701f39396cb>, [Accessed on 5th March, 2021].
- [26] C. Iwendi, A.K. Bashir, A. Peshkar, R. Sujatha, J.M. Chatterjee, S. Pasupuleti, R. Mishra, S. Pillai, O. Jo, COVID-19 patient health prediction using boosted random forest algorithm, *Front. Public Health* 8 (2020) 357.
- [27] A. Navlani, Understanding random forests classifier in python, DataCamp (2018) Available at: <https://www.datacamp.com/community/tutorials/random-forests-classifier-python>, [Accessed on 5th March, 2021].
- [28] Z. Noshad, N. Javaid, T. Saba, Z. Wadud, M.Q. Saleem, M.E. Alzahrani, O.E. Sheta, Fault detection in wireless sensor networks through the random forest classifier, *Sensors* 19 (7) (2019) 1–21.
- [29] A.H. Larsen, J.J. Mortensen, J. Blomqvist, I.E. Castelli, R. Christensen, M. Dulak, J. Friis, M.N. Groves, B. Hammer, C. Hargus, E.D. Hermes, The atomic simulation environment—a python library for working with atoms, *J. Phys.: Condens. Matter* 34 (2022) 100016.
- [30] S. Amini, S. Homayouni, A. Safari, A.A. Darvishsefat, Object-based classification of hyperspectral data using random forest algorithm, *Geo-Spat. Inf. Sci.* 21 (2) (2018) 127–138.
- [31] P. Muñoz, J. Orellana-Alvear, P. Willems, R. Céleri, Flash-flood forecasting in an andean mountain catchment—Development of a step-wise methodology based on the random forest algorithm, *Water* 10 (11) (2018) 1–18.
- [32] M. Reich, T. Tabor, T. Liefeld, H. Thorvaldsdóttir, B. Hill, P. Tamayo, J.P. Mesirov, The GenePattern notebook environment, *Cell Syst.* 5 (2) (2017) 149–151.
- [33] K.M. Mendez, L. Pritchard, S.N. Reinke, D.I. Broadhurst, Toward collaborative open data science in metabolomics using Jupyter Notebooks and cloud computing, *Metabolomics* 15 (10) (2019) 1–16.
- [34] D. Yin, Y. Liu, H. Hu, J. Terstriep, X. Hong, A. Padmanabhan, S. Wang, Cybergis jupyter for reproducible and scalable geospatial analytics, *Concurr. Comput.: Pract. Exper.* 31 (11) (2019) 1–14.
- [35] V.V. Ramalingam, A. Dandapath, M.K. Raja, Heart disease prediction using machine learning techniques: a survey, *Int. J. Eng. Technol.* 7 (2.8) (2018) 684–687.
- [36] K. Subhadra, B. Vikas, Neural network based intelligent system for predicting heart disease, *Int. J. Innov. Technol. Explor. Eng.* 8 (5) (2019) 484–487. [37] B. Ambale-Venkatesh, X. Yang, C.O. Wu, K. Liu, W.G. Hundley, R. McClelland, A.S. Gomes, A.R. Folsom, S. Shea, E. Guallar, D.A. Bluemke, Cardiovascular event prediction by machine learning: the multi-ethnic study of atherosclerosis, *Circ. Res.* 121 (9) (2017) 1092–1101.
- [38] M.V. Dogan, I.M. Grumbach, J.J. Michaelson, R.A. Philibert, Integrated genetic and epigenetic prediction of coronary heart disease in the framingham heart study, *PLoS One* 13 (1) (2018) 1–18.
- [39] M.S. Mahdavinnejad, M. Rezvan, M. Berekatain, P. Adibi, P. Barnaghi, A.P. Sheth, Machine learning for internet of things data analysis: A survey, *Digit. Commun. Netw.* 4 (3) (2018) 161–175.
- [40] G. Tausin, U. Lupo, L. Tunstall, J.B. Pérez, M. Caorsi, A.M. Medina-Mardones, A. Dassatti, K. Hess, Giotto-tda: A topological data analysis toolkit for machine learning and data exploration, *J. Mach. Learn. Res.* 22 (39) (2021) 1–6.
- [41] B. Peters, E. Haber, J. Granek, Neural networks for geophysicists and their application to seismic data interpretation, *The Leading Edge* 38 (7) (2019) 534–540.
- [42] G. Carleo, I. Cirac, K. Cranmer, L. Daudet, M. Schuld, N. Tishby, L. Vogt-Maranto, L. Zdeborová, Machine learning and the physical sciences, *Rev. Modern Phys.* 91 (4) (2019) 1–47.
- [43] R. Harper, S.T. Flammia, J.J. Wallman, Efficient learning of quantum noise, *Nat. Phys.* 16 (12) (2020) 1184–1188.
- [44] Y. Jia, Y. Qi, H. Shang, R. Jiang, A. Li, A practical approach to constructing a knowledge graph for cybersecurity, *Engineering* 4 (1) (2018) 53–60. [45] D. Galinec, D. Možnik, B. Guberina, Cybersecurity and cyber defence: a national level strategic approach, *Automatika: Časopis za automatiku, Mjerenje, Elektroniku, Računarstvo i Komunikacije* (2017).
- [46] Y. Fan, J. Li, D. Zhang, J. Pi, J. Song, G. Zhao, Supporting sustainable maintenance of substations under cyber-threats: An evaluation method of cybersecurity risk for power CPS, *Sustainability* 11 (4) (2019) 1–30.
- [47] G. Jin, M. Tu, T.H. Kim, J. Heffron, J. White, Evaluation of game-based learning in cybersecurity education for high school students, *J. Educ. Learn.* 12 (1) (2018) 150–158.
- [48] A. Anees, I. Hussain, A novel method to identify initial values of chaotic maps in cybersecurity, *Symmetry* 11 (2) (2019) 140.
- [49] F. Mehmood, H.U. Rashidkayani, F. Hussain, Chronic diseases modelling—python environment, *FUUAJST J. Biol.* 10 (1) (2020) 31–38.
- [50] H. Tjahjadj, K. Ramli, Noninvasive blood pressure classification based on photoplethysmography using K-nearest neighbors algorithm: A feasibility study, *Information* 11 (2) (2020) 1–18.
- [51] D. Pedrozo, F. Barajas, A. Estupiñán, K.L. Cristiano, D.A. Triana, Data analysis for a set of university student lists using the k-Nearest Neighbors machine learning method, *J. Phys. Conf. Ser.* 1514 (1) (2020) 1–8.
- [52] T.A. Assegie, P.S. Nair, Handwritten digits recognition with decision tree classification: a machine learning approach, *Int. J. Electr. Comput. Eng.* 9 (5) (2019) 1–4.
- [53] S. Herbold, Autorank: A python package for automated ranking of classifiers, *J. Open Source Softw.* 5 (48) (2020) 1–4.
- [54] A. Zahariev, M. Zveryakov, S. Prodanov, G. Zaharieva, P. Angelov, S. Zarkova, M. Petrova, Debt management evaluation through support vector machines: on the example of Italy and Greece, *Entrepreneurship Sustain. Issues* 7 (3) (2020) 1–12.
- [55] C. Wang, C. Liang, Msipred: a python package for tumor microsatellite instability classification from tumor mutation annotation data using a support vector machine, *Sci. Rep.* 8 (1) (2018) 1–10.
- [56] M. Huljanah, Z. Rustam, S. Utama, T. Siswantining, Feature selection using random forest classifier for predicting prostate cancer, *IOP Conf. Ser.: Mater. Sci. Eng.* 546 (5) (2019) 1–9.
- [57] S. Mehrang, J. Pietilä, I. Korhonen, An activity recognition framework deploying the random forest classifier and a single optical heart rate monitoring and triaxial accelerometer wrist-band, *Sensors* 18 (2) (2018) 613.
- [58] J. Ge, X. Li, H. Jiang, H. Liu, T. Zhang, M. Wang, T. Zhao, Picasso: A sparse learning library for high dimensional data analysis in R and python, *J. Mach. Learn. Res.* 20 (44) (2019) 1–5.
- [59] M.A. Saif, A.N. Medvedev, M.A. Medvedev, T. Atanasova, Classification of online toxic comments using the logistic regression and neural networks models, in: *AIP Conference Proceedings*, Vol. 2048 (1) 2018, pp. 1–6.

10 III March 2022 <https://doi.org/10.22214/ijraset.2022.40768>

International Journal for Research in Applied Science & Engineering Technology (IJRASET) *ISSN: 2321-9653;*
IC Value: 45.98; SJ Impact Factor: 7.538 Volume 10 Issue III Mar 2022- Available at www.ijraset.com

Prediction of Heart Disease Using Machine Learning Algorithms

Abstract: *In living organisms, the heart plays an important function. Diagnosis and prediction of heart diseases necessitates greater precision, perfection, and accuracy because even a minor error will result in fatigue or death. There are multiple death cases related to the heart, and the number is growing rapidly day by day. The scope of this study is restricted to discovering associations in CHD data using three supervised learning techniques: Logistic Regression, K-Nearest Neighbour, and Random Forest, in order to improve the prediction rate. As a result, this paper conducts a comparative analysis of the results of various machine learning algorithms. The trial results verify that Logistic Regression algorithm has achieved the highest accuracy of 89% compared to other ML algorithms implemented.*

Keywords: *Machine Learning, Logistic Regression, K-Nearest Neighbour, Random Forest, Python, Heart Disease, Prediction model, Healthcare*

I. INTRODUCTION

Heart disease has risen to become one of the leading causes of death all over the world. According to the World Health Organization, cardiac illnesses claim the lives of 17.7 million people each year, accounting for 31% of all fatalities worldwide. Heart disease has become the top cause of death in India as well. As a result, it is essential to be able to forecast heart-related disorders in a reliable and precise manner. Data on various health-related concerns is compiled by medical institutions all over the world. These data can be used to gain significant information utilizing a variety of machine learning techniques. However, the amount of data collected is enormous, and it is frequently noisy.

II. PROBLEM-STATEMENT

We analyzing the various machine learning algorithms and finding the best to predict the presence or absence of heart disease. The target we will be exploring is binary classification which is 0 to show the absence of heart disease and 1 to show the presence of heart disease.

III. PROPOSED METHOD

We are going to use various machine learning algorithms to predict the target. We will be using a number of different features about a person to predict whether they have heart disease or not. The dependent variable is whether or not a patient has heart disease, while the independent variables are the patient's many medical characteristics. The various machine learning algorithms used for our model will be Logistic Regression, K-Nearest Neighbours, and Random Forest. We will compare the scores of all these models by splitting our data into training and testing in an approximate 80:20 ratio. We will also tune the hyper parameters for all these models to yield the best results. And finally conclude the best prediction model for our heart disease dataset.

Flow Diagram

Fig 1-Flow diagram

IV. LITERATURE SURVEY

- 1) [Archana singh, Rakesh Kumar (2020) heart disease prediction using machine learning algorithms in this particular publication various machine learning algorithms such as linear regression, decision tree, support vector machine and k- nearest neighbour is used. When they performed the analysis of algorithms on the basis of the dataset whose attributes are shown in a research paper and on the basis of the confusion matrix, they found KNN is the best one. [7.]SVM has an accuracy of 87%, decision tree of 79%, and k-nearest neighbour has an accuracy of 74%. For the future scope more machine learning approach will be used for the best analysis of heart diseases and for earlier prediction of diseases so that the rate of a number of deaths can be

reduced if people are informed of the illness.

- 2) Jaymin Patel, Prof. Tejalupadhyay, Dr. Samir Patel (2016) used machine learning and data mining techniques to predict cardiac disease. In this research paper, they have analyzed the experimental results, it is concluded that j48 tree technique turned out to be the best classifier for heart disease prediction because it contains more accuracy and the least total time to build. Weka is an open-source software tool that consists of an accumulation of machine learning algorithms for data mining undertakings. It contains apparatuses for information pre-processing, regression, visualization, classification, association rules and clustering. [8.] The best algorithm is j48 based on UCI data has the highest accuracy i.e. 56.76% and the total time to build the model is 0.04 seconds while LMT algorithm has the lowest accuracy i.e. 55.77% and the total time to build a model is 0.39 seconds. There is an only marginal success is achieved in the creation of predictive model for heart disease patients and hence there is a need for combinational and more complex models to increase the accuracy of predicting the early onset of heart disease.
- 3) Rajesh n, T manesha, Shaik hafeez, Hari krishna (2018) prediction of heart disease using machine learning algorithms. The Naive Bayes treats every variable as independent which helps it to predict even if variables don't have proper relation. We used decision trees and a combination of algorithms for the prediction of heart disease based on the above attributes. When the dataset is small Naive Bayes algorithm gives accurate results and when the dataset is large decision trees give the accurate results. Naïve Bayes will not give accurate results every time we need to consider the results of different algorithms and by all its results if a prediction is made it will be accurate.
- 4) V.v. Ramalingam*, Ayantan Dandapath, M Karthik raja (2018). They published a paper named; 'A survey on the use of machine learning techniques to forecast heart disease'. Algorithms and techniques used are -. Naive Bayes, support vector machine, random forest, ensemble model, decision tree, and k – nearest neighbour. Models based on Naïve Bayes classifier were computationally very fast and have also performed well. In the vast majority of cases, SVM performed admirably. Because they employ many algorithms to overcome the problem of overfitting, random forest and ensemble models have fared exceptionally well. A lot of research may be done on the best algorithm ensemble to employ for a specific sort of data.
- 5) Marjia sultana; afrin haider; Mohammad shorif uddin (22-24 Sept. 2016). They published a paper [6]. In this paper, two data sets (collected and UCI standard) are used separately for each data mining technique. This paper performed an experiment using diverse data mining techniques to find out a more accurate technique for heart disease prediction. [6.] Their findings show that for heart disease prediction the performances of Bayes net and SVM classifiers are the optimum among the investigated five classifiers: Bayes net, sma, kstar, mlp and j48. The prediction of heart disease requires a huge size of data that is too complex and massive to process and analyze by conventional techniques. Various experts employ a variety of data mining approaches to solve various problem.

V. METHODOLOGY IMPLEMENTATION

A. Preprocessing

We have collected data from various reliable sources from the internet. After analysing various factors, we have reached to a conclusion that 13 independent variables will determine 1 target variable. To do this we will have to split the target variable from the rest. If we can reach 96% accuracy at predicting whether or not a patient has heart disease during the proof of concept, we'll pursue this project.

Fig 2-Correlation Matrix

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |
International Journal for Research in Applied Science & Engineering Technology (IJRASET) ISSN: 2321-9653;
IC Value: 45.98; SJ Impact Factor: 7.538 Volume 10 Issue III Mar 2022- Available at www.ijraset.com

VI. TRAINING AND TEST SPLIT

The train and split procedure is used to divide the data the dataset into two halves.

1) Train split

2) Test split

The model designed will first train on the train split where it tries to learn the patterns in the data. Then based on the patterns it has learnt it will tested on the test split. In this entire process choosing the test split size is also very important. A rule thumb is to use 80% of your data to train on and the other 20% to test on.

VII. MACHINE LEARNING MODELS

Machine learning models are majorly classified as supervised and unsupervised. If the model is supervised, it is divided into two categories: regression and classification. We will focus on the following machine learning models:

1) *Logistic Regression*: It is a basic classification algorithm which predicts the probability of a target variable.

Fig 3-Logistic Regression

2) *K-nearest Neighbours*: It's a machine learning algorithm that's supervised. The idea behind nearest neighbour methods is to find a predetermined number of training samples that are closest in distance to the new point and use them to predict the mark. It makes no assumptions about the data and is typically used for classification tasks where little to no prior knowledge of the data distribution is available. Finding the k closest data points in the training set to the data point for which a target value is unavailable and assigning the average value of the identified data points to it is the aim of this algorithm.

Fig 4-KNN

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |
International Journal for Research in Applied Science & Engineering Technology (IJRASET) ISSN: 2321-9653;
IC Value: 45.98; SJ Impact Factor: 7.538 Volume 10 Issue III Mar 2022- Available at www.ijraset.com

3) *Random Forest*: Random forest is a supervised machine learning algorithm that can be used to solve problems in both classification and regression. It builds decision trees out of data samples, then gets predictions from each of them before voting on the best solution.

Fig 5-Decision Tree

VIII. RESULTS OBTAINED BY MACHINE LEARNING MODELS

- 1) 'Logistic regression': 0.8852459016393442,
- 2) 'knn': 0.6885245901639344,
- 3) 'random forest': 0.8360655737704918

IX. HYPER-PARAMETER TUNING AND CROSS VALIDATION

A hyperparameter is a parameter whose value is set before the model is allowed to train on the train split. Tuning the hyper parameters helps to increase the efficiency of a model. Not all the hyperparameters are to be considered any context. Choosing the right hyperparameters is also an important task.

The best accuracy obtained for KNN is when the number of nearest neighbours is 11 with an accuracy score of 0.7540983606557377

Fig 6-Tuning of Hyperparameter for KNN

The best parameter found for logistic regression is {'solver': 'liblinear', 'c': 0.23357214690901212} with a accuracy score of 0.8852459016393442

The best parameter found for random forest is {'n_estimators': 210, 'min_samples_split': 4, 'min_samples_leaf': 19, 'max_depth': 3} with a accuracy score of 0.8688524590163934

X. COMPARE WITH YOUR EXISTING MODEL

| Sno. | Algorithm | Accuracy | Paper |
|-------------|-------------|------------|-------|
| Found By | 1. Logistic | Regression | |
| Us | 89% | -- | |
| Accuracy Of | Base | | |
| Research | | | |

2. Random Forest 87% --
3. Decision Tree -- 79%
4. KNN 75% 74%
5. SVM --87%

In our base research (Paper 1) we found that the machine learning algorithms used were KNN, SVM, Decision Tree and the highest accuracy achieved was 87%. Also there was a lack of tuning of hyperparameters. In our re- search paper we worked on ensemble learning algorithms like Random Forest , Logestic Regression, KNN. And after tuning the hyperparameters we found that the highest accuracy is achieved through Logistic Regression with a accuracy rate of 89%

XI. RESULTS

After tuning the hyper parameters for KNN, Logistic Regression, Random forest and selecting the best ones we found the following results for accuracy:

KNN: 0.6885245901639344

Logistic Regression: 0.8852459016393442

Random Forest: 0.8360655737704918

Fig 7-Accuracy Comparison

Among these we can see that random forest with a certain set of hyperparameters Logistic Regression performs the best. Now we will find the other metrics for the logistic regression model.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |
International Journal for Research in Applied Science & Engineering Technology (IJRASET) ISSN: 2321-9653;
IC Value: 45.98; SJ Impact Factor: 7.538 Volume 10 Issue III Mar 2022- Available at www.ijraset.com

A. ROC Curve

The metric compares the true positive rate with the false positive rate.

The True Positive Rate (TPR) is defined as follows:

$$TPR = \frac{TP}{TP + FN}$$

The False Positive Rate (FPR) is defined :

$$FPR = \frac{FP}{FP + TN}$$

It also provides us with AUC scores which denotes the area underneath the ROC curve

Fig 8-ROC Curve

B. Confusion Matrix

A confusion matrix is a table that is used to describe the output of a classification model/classifier by comparing the true values of the training and test datasets. It is divided into four parts, each of which is defined as follows:

- 1) True positives (TP): These are cases in which we expected yes (they have the disease) and they do.
- 2) Real negatives (TN): We predicted they wouldn't have the disorder, and they don't.
- 3) False positives (FP): We expected that they will have the disease, but they don't. (This is often referred to as a "Type I error.")
- 4) False negatives (FN): We expected that they will not have the disorder, but they do. (This is often referred to as a "Type II error.")

Fig 9-Confusion Matrix

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |
International Journal for Research in Applied Science & Engineering Technology (IJRASET) ISSN: 2321-9653;
IC Value: 45.98; SJ Impact Factor: 7.538 Volume 10 Issue III Mar 2022- Available at www.ijraset.com

C. Classification Report

The Classification report is used to find the quality of predictions from a classification algorithm. It helps us to find how many predictions are correct and how many are wrong. More specifically, it gives us an understanding of True negatives and False Negatives, True Positives and False Positives, and uses them to predict the metrics of a classification. The main metrics found by the Classification report are accuracy, precision, recall, and f1-score.

The model's accuracy is expressed in decimal form. Precision refers to a classifier's ability to avoid labelling a negative occurrence as positive. Recall - This metric indicates the percentage of true positives that were successfully classified. The F1 score is a weighted harmonic mean of precision and recalls, with 1.0 being the highest and 0.0 being the poorest. $F1 \text{ Score} = 2 * (\text{Recall} * \text{Precision}) / (\text{Recall} + \text{Precision})$ Support - The number of samples used to calculate each metric. Support - The number of samples used to calculate each metric.

Fig 10-Classification Report

D. Cross Validation Score

The statistical method of cross-validation is majorly used for measuring the skill of machine learning models. The k-fold cross validation is used to test how a machine learning model performs with different sets of data. As our data set consists of 303 entries using 5-folds of cross-validation along with the Logistic Regression model and with the best hyperparameters yielded the following results:

Fig 11-Cross Validation Metrics

E. Feature Importance

It refers the techniques that assign a score to the input attributes/features with respect to the fact that which feature has the highest contribution in predicting the results for a given machine learning model. For finding it we will use the `coef_` attribute. The `coef_` attribute is the coefficient of the features in the decision function. We can note that negative `coef_` attribute denotes the presence of negative correlation.

Fig 12-Feature Importance

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |
International Journal for Research in Applied Science & Engineering Technology (IJRASET) ISSN: 2321-9653;
IC Value: 45.98; SJ Impact Factor: 7.538 Volume 10 Issue III Mar 2022- Available at www.ijraset.com

XII. CONCLUSION

With the rising number of deaths due to heart disease, it is becoming increasingly important to build a system that can effectively and accurately forecast heart disease. The motivation for the study was to find the most efficient ML algorithm for detection of heart diseases. This study compares the accuracy score of KNN, Logistic Regression and Random Forest for predicting heart disease using UCI machine learning repository dataset. The result of this study indicates that the Logistic regression algorithm is the most efficient algorithm with accuracy score of 89% for prediction of heart disease. Accuracy of the algorithms in machine learning depends upon the dataset that used for training and testing purpose.

XIII. FUTURE SCOPE

In the future, the work could be improved by creating a web application premised on the logistic regression algorithm and by using a larger dataset than the one used in this study, which would help to provide better outcomes and aid health professionals in predicting heart disease efficiently and effectively.

REFERENCES

- [1] Singh, A., & Kumar, R. (2020, February). Heart disease prediction using machine learning algorithms. In 2020 international conference on electrical and electronics engineering (ICE3) (pp. 452-457). IEEE.
- [2] Patel, J., TejalUpadhyay, D., & Patel, S. (2015). Heart disease prediction using machine learning and data mining technique. *Heart Disease*, 7(1), 129-137. [3] Rajesh, N., T. M., Hafëez, S., & Krishna, H. (2018). Prediction of Heart Disease Using Machine Learning Algorithms. *International Journal of Engineering & Technology*, 7(2.32), 363-366. doi:http://dx.doi.org/10.14419/ijet.v7i2.32.15714
- [4] Ramalingam, V. V., Dandapath, A., & Raja, M. K. (2018). Heart disease prediction using machine learning techniques: a survey. *International Journal of Engineering & Technology*, 7(2.8), 684-687
- [5] Kaur, A., & Arora, J. (2018). HEART DISEASE PREDICTION USING DATA MINING TECHNIQUES: A SURVEY. *International Journal of Advanced Research in Computer Science*, 9(2).
- [6] "Sultana, M., Haider, A., & Uddin, M. (2016). Analysis of data mining techniques for heart disease prediction. 2016 3rd International Conference on Electrical Engineering and Information Communication Technology (ICEEICT), 1-5.
- [7] Deekshatulu, B. L., & Chandra, P. (2013). Classification of heart disease using k-nearest neighbor and genetic algorithm. *Procedia technology*, 10, 85-94. [8] Learning, M. (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.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

**HBRP
PUBLICATION**

Advancement in Image Processing and
Pattern Recognition
Volume 3 Issue 2

Heart Attack Prediction and Analysis System Using Decision Tree Algorithm

Mayuri Asabe^{1*}, Shweta Shilwant², Nagnath Dolare³, Sulakshana Chorghade⁴, K. R. Pathak⁵

^{1,2,3,4}Student, Computer Science, Rajgad Dnyanpeeth Technical Campus Dhangawadi, Pune, India

⁵Assistant Professor, Department of Computer Science, Rajgad Dnyanpeeth Technical Campus Dhangawadi, Pune, India

***Corresponding Author**

ABSTRACT

Heart Attack Prediction using Machine Learning Technique in Big data analytics has started to play an important role in the healthcare practices and research. heart attack prediction will be found primarily on real-time processing, distributed and real-time classification and distribution, storage so; databases can be easily modified by the doctors. If you know all the attributes related to our health we can check easily how much chance to the Heart attack risk, using the system applications. It was recently used to train classification models. After that using extract the features that is condition to be find to be classified by Decision Tree (DT). Compared to existing; algorithms provides better performance. After classification, performance criteria including accuracy, precision, F-measure is to be calculated. If you are concern about the heart attack risks, you might be referred to a heart specialist. Some attributes are Heart Attack risk factors including which is the High blood pressure, high cholesterol and diabetes, increases your risk even more. Hence we are also checking your symptoms of heart attack and take about prevention.

Keywords:- Decision Tree, Machine Learning, QA System, Heart Attack prediction.

INTRODUCTION

The heart is a muscle and its role is to pump blood throughout the body. This makes the body a major staple. Heart disease is one of the biggest health risks for association today. According to the World Health Organization (WHO), stroke and heart attack are the most common cause of global death (85%). Therefore, the availability of data and data mining techniques, especially machine learning and early detection of Heart Attack, can help patients to anticipate a potential disease response. In the healthcare field, it is becoming more and more common nowadays to source large amounts of data (big data), streaming machines, advanced healthcare services, high throughput

instruments, sensor networks, Internet of Things, mobile application applications, data archiving and processing, from many areas.

PROBLEM STATEMENT

Design and Implement the Heart Attack Prediction and Detection System using machine learning techniques.

MOTIVATION

Heart Attack is one of the huge health risks for human's healthy life. big data growth in medical and healthcare association today, early solution and accurate analysis of medical data benefits through patient care and community services. If the quality of medical data

HBRP Publication Page 1-10 2020. All Rights Reserved **Page 1**

**HBRP
PUBLICATION**

some data are incomplete, the accuracy of the analysis decreases.

RELATED WORK

Previous research studies have examined the use of machine learning techniques to predict and classify heart disease. However, these studies focus on the specific effects of specific machine learning techniques.

This work analyses the predictive system

Advancement in Image Processing and Pattern Recognition Volume 3 Issue 2

for heart disease. In this work, medical terms like sex, blood pressure, and cholesterol are used to describe the possibility of heart disease in patients with 12 points. So far, 13 attributes are used for forecasting. Two more points have been added to this research work - obesity and smoking. Data mining classification algorithms, decision trees, navy bias and neural networks are analysed in the cardiovascular database [1].

Medical diagnostic systems play an important role in medical practice and are

used by medical professionals for diagnosis and treatment. In this work, the medical diagnostic system is defined to indicate the risk of cardiovascular disease. The system is built by combining the relative advantages of genetic mechanisms and neural networks. Multi-layered feed forward neural networks typically adapt to complex classification problems. The weight of the nerve space is determined using the genetic technique because it finds a good set of good weights at low repetitions [2].

The condition of the heart is explained in

detail by a thorough examination of the features of the Electrocardiogram report. It is valuable to automatically remove the features of the time plane to detect essential heart disease. This function introduces a multi-resolution wavelet transform based system to detect 'P', 'Q', 'R', 'S', 'T' peaks complexes from the original ECG signal.[3].

SYSTEM ARCHITECTURE

Our proposed system will have including some parameters to completed the structure of the prediction or analysis process.

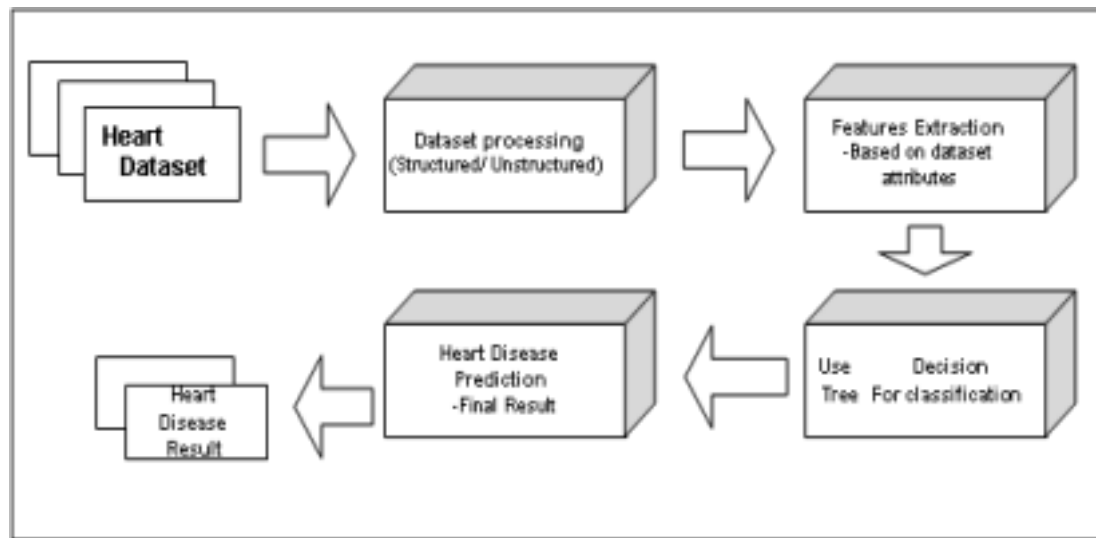


Fig.1:-System Architecture

The focus of this paper is primarily on real-time processing, distributed and real time classification and distribution storage

so databases can be easily modified by other databases. It was recently used to train classification models. After that using

extract the features that is condition to be find to be classified by Decision Tree (DT). Compared to existing; algorithms provides better performance. After classification, performance criteria including accuracy, precision, F-measure is to be calculated.

Our proposed system will have three modules:

List of Modules:

- Admin

- Doctor

- Patient

Admin Module

- Admin is having predefined username and password.
- Admin can log in to the system and can patient login is active or disactive.
- Admin will upload heart attack patient dataset in database for training.
- Admin will have other general rights as to view number of user's data, their details etc.

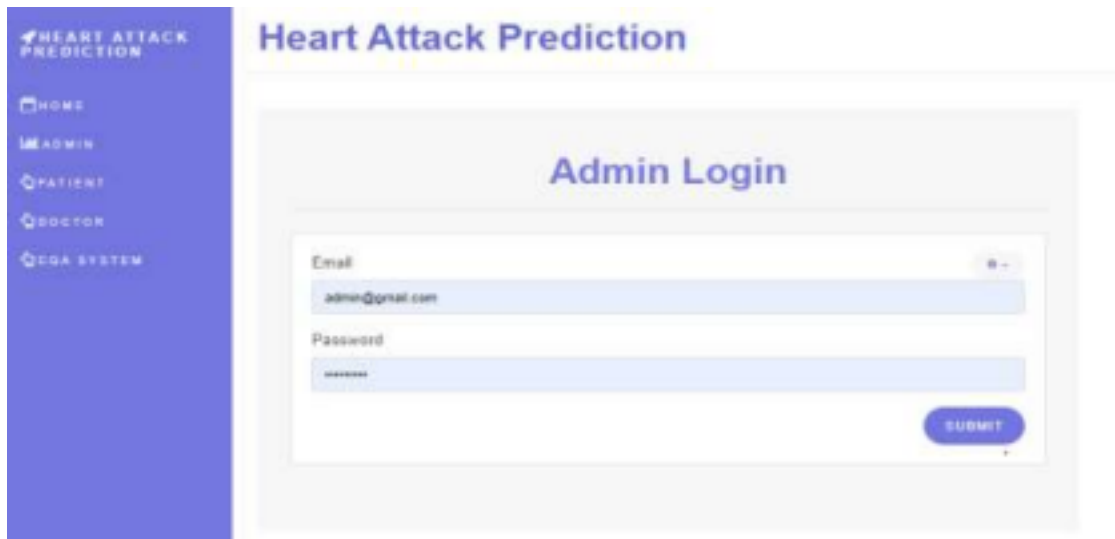


Fig.2:-Admin Login Page



Fig.3:-Admin Home Page

Doctor Module

- Doctor is the expert in medical field.
- Doctor will resolve the queries of the Patients.

Patient Module

- Anyone can register and can become a part of the system as a User or Patient.
- Patient have to fill up one form after



registration with the details like i.e. Cholesterol, Blood Pressure, Age,

Advancement in Image Processing and Pattern Recognition Volume 3 Issue 2
family history, body mass index etc.,
this Parameter and some other
parameters.

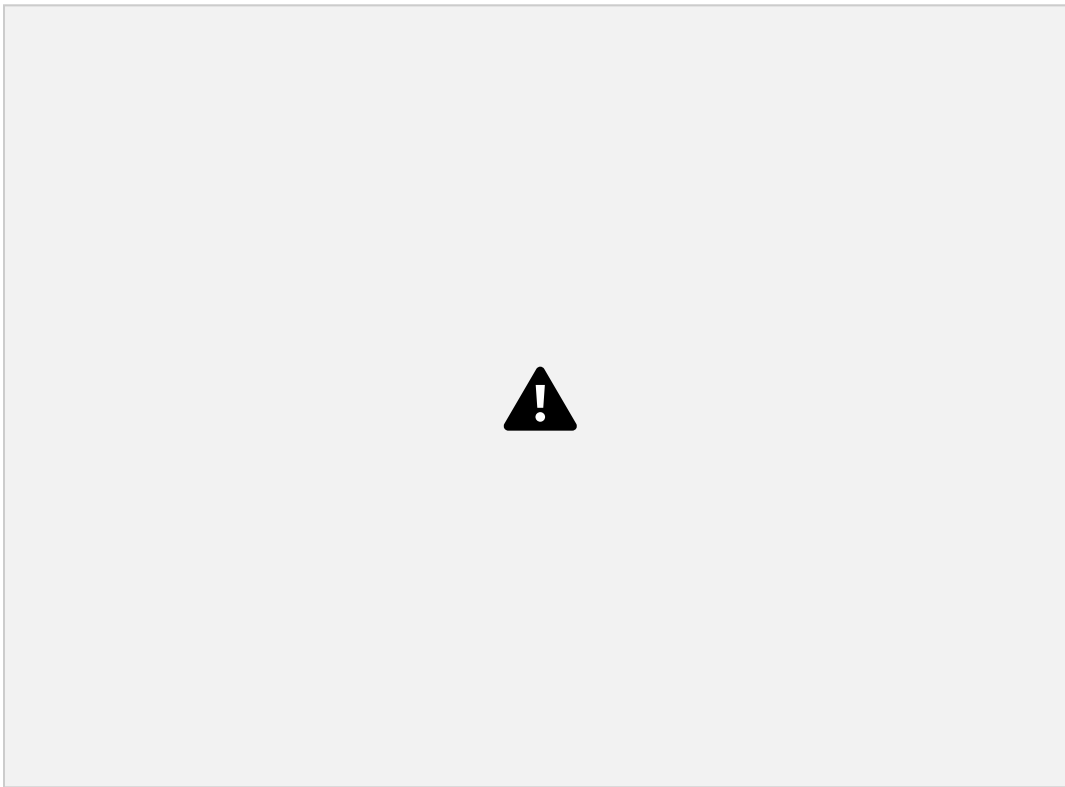


Fig.4:-Patient Login Page

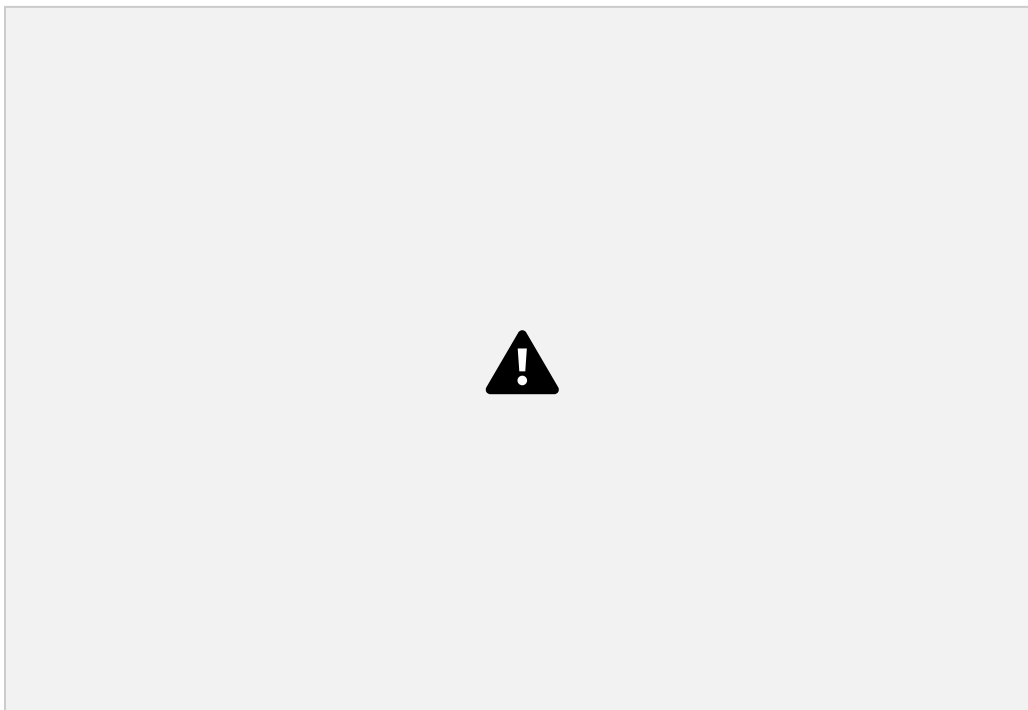


Fig.5:-Patient Home Page



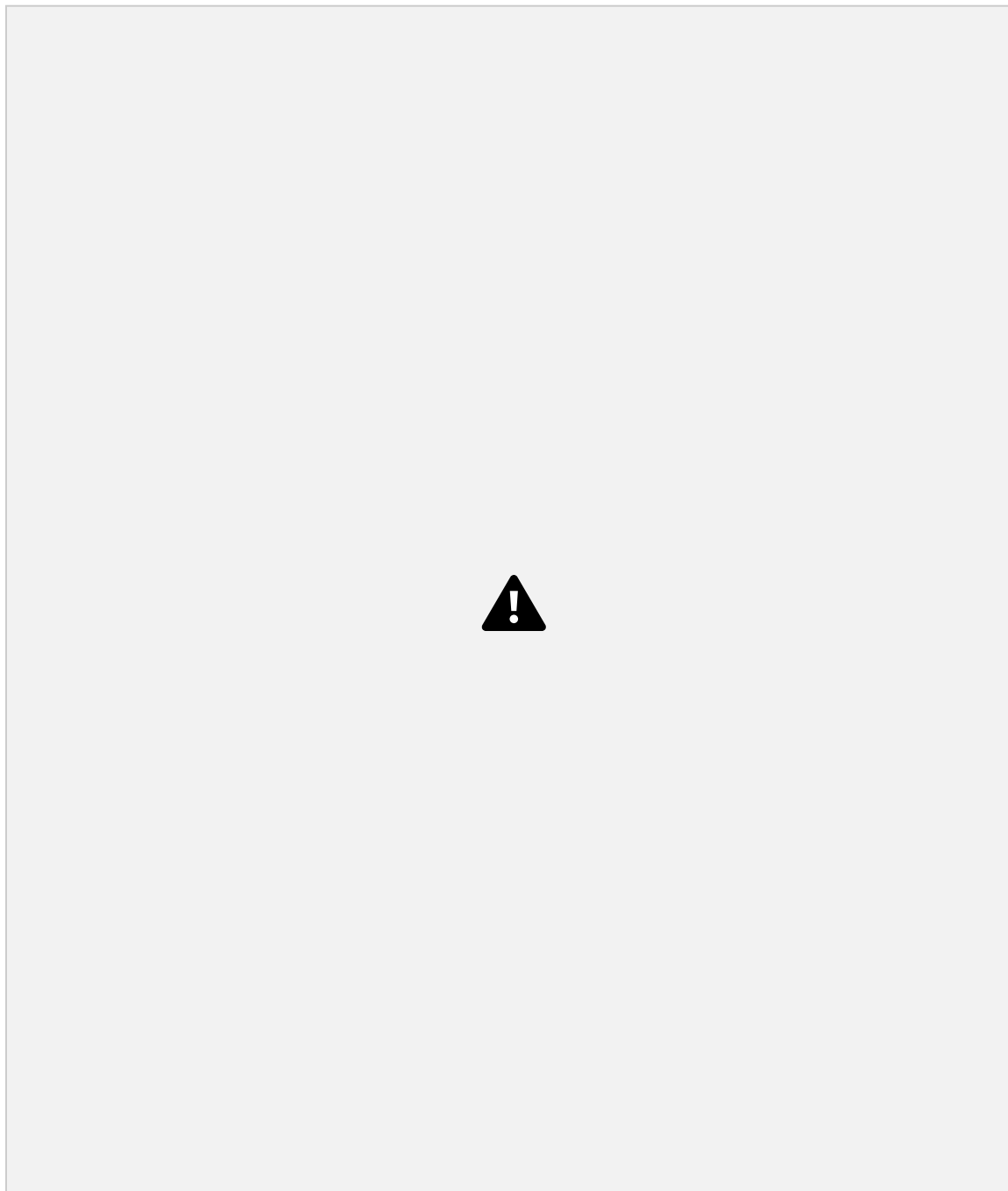


Fig.6:-Patient Attribute Page

PROCEDURE

This section covers our featured selection and classification system. The main structure of the proposed system is shown in Figure. 1 Our system includes speed based on feature selection, Also Feature extraction procedure.

DATA SET AND ATTRIBUTES Each

element attributes that's name

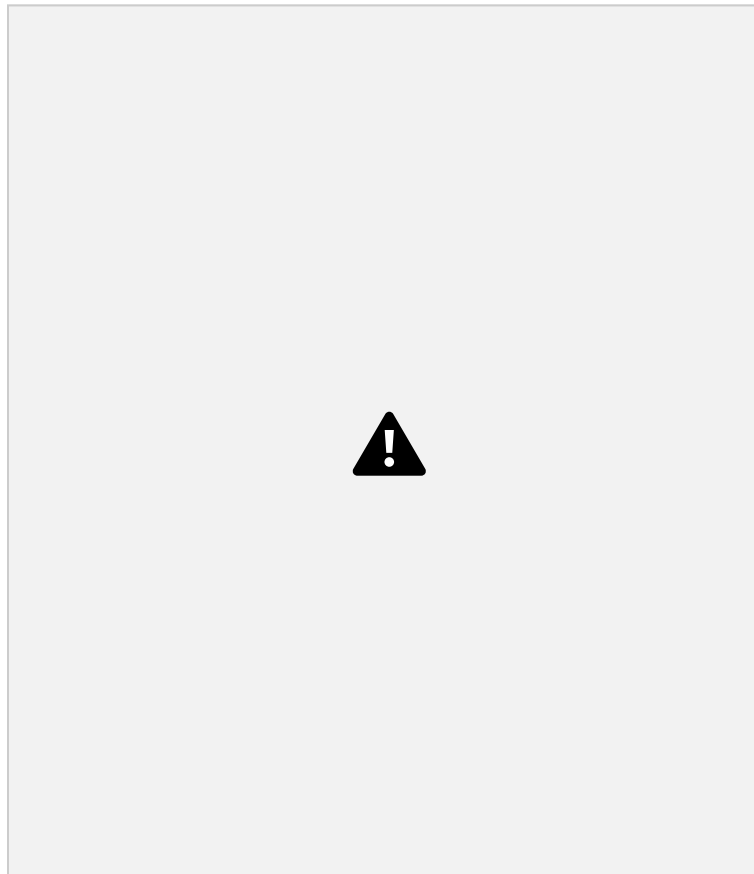
beginning with the data-is a data attribute. Data attribute can be storage of the information contain of any element suppose the continuously changing the values like scores in the any type of games.

Data is collected from the UCI Machine Learning. The data set is named Heart Attack Attributes.

When we talk about the machine learning technique and data mining, data knowledge, it's very necessary to finding the data attributes and data objects also. Data's knowledge is necessary because the

helping for the finding its relationships each other in the data. Data objects are the most important part in the database. Data objects are in short, it is an attribute group. When the data objects are listed in the database they are known as the data tuples.

Table 1:-Heart attack attributes



CLASSIFICATION FUNCTIONS From the point of view of machine learning technique, a heart attack prediction can be viewed as a classification or clustering problem. In this section, the theoretical reference to all the methods used in this research is given. For the purpose of comparative analysis, 5 machine learning technique algorithms are also used. K-Nearest Neighbour (KNN), Random Forest (RF), Support Vector Machine (SVM), Naïve Bayes and Artificial Neural Network (ANN) Decision Tree (DT) are different machine learning (ML) algorithms. The reason for choosing this algorithm is based on their popularity.

Feature Selection (Attributes)

Feature selection used to predict the output value; feature importance helps to evaluate significance of every feature of multi dimensional dataset. Decision trees offer a different approach based on the impurity reduction determined by every single feature.

Decision Tree (Dt)

Decision tree has leaf node and decision nodes. Every branches of the last node is a leaf node, which represents the class label. And the second type node which is decision node is used to take the class label after some trials.it has a leaf node or

sub tree. The calculation of the value of absolute information does not depend on decision making. Usually when calculating using expected values you calculate the expected value of each choice and choose the one that has the highest expected value.

J. R. Quinlan C. The basic algorithm for making decision trees called, which takes top-down, greedy exploration in place of non-backtracking potential branches. C4.5 enters entropy and captures information for decision trees.

will be added proportionally to get the A decision tree is made from the root node to the top-down and divided into subdivisions in the data that have the same values (homogeneous). The expected amount of information (in bits) needed to assign a class to randomly drawn object in s under the optimal, shortest length code.

A. Entropy (E) used for frequency of one attribute:

B. Entropy (E) used for frequency of two attributes:

Entropy (E)

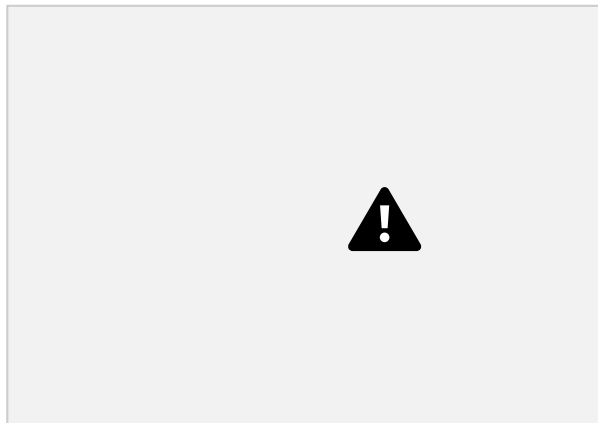


Fig.7:-A Simple Decision Tree.

Information Gain

Dividing a dataset by an attribute gives information based on the reduction in entropy. Making a decision tree means finding the most informative properties (i.e. the most homogeneous branch).

1: Calculating the entropy (E) of the target attribute.

2: The dataset will be split into different attributes.

Entropy is calculated for each attribute. It total entropy for the division.

Resulting entropy (E) is subtracted from the entropy (E) before division.

$Gain(T, X) = Entropy(T) - Entropy(T, X)$ Decision Tree Working Steps:

Input:

Step 1: Upload training dataset.

Step 2: symptoms set is the set of input attributes

Step 3: disease is the set of output attributes

Step 4: sample is a set of training data.

Function Iterative Di-Chotomiser returns a decision tree

1. Create root node for the tree.
2. If
(All inputs are positive, return leaf node positive)
If Else

(If all inputs are negative, return leaf node negative)

Else (Some inputs are positive and some inputs are negative, check condition (Positive>negative||Positive<negative), Then return result)



Decision Tree Flowchart

3. Calculate the entropy of current state $H(S)$.
4. For each attribute, calculate the entropy with respect to the attribute 'X'
Denoted by $H(S, X)$
5. Select the attribute which has maximum value of $IG(S, X)$.
6. Remove the attribute that offers highest value from the set of attributes.
7. Repeat until we run out of all attributes

or the decision tree has all leaf nodes.

Advantages

- Predict heart disease for Structured Data using machine learning algorithms i.e., Decision Tree (DT).
- Find reliable answer using this system • To achieve better accuracy.

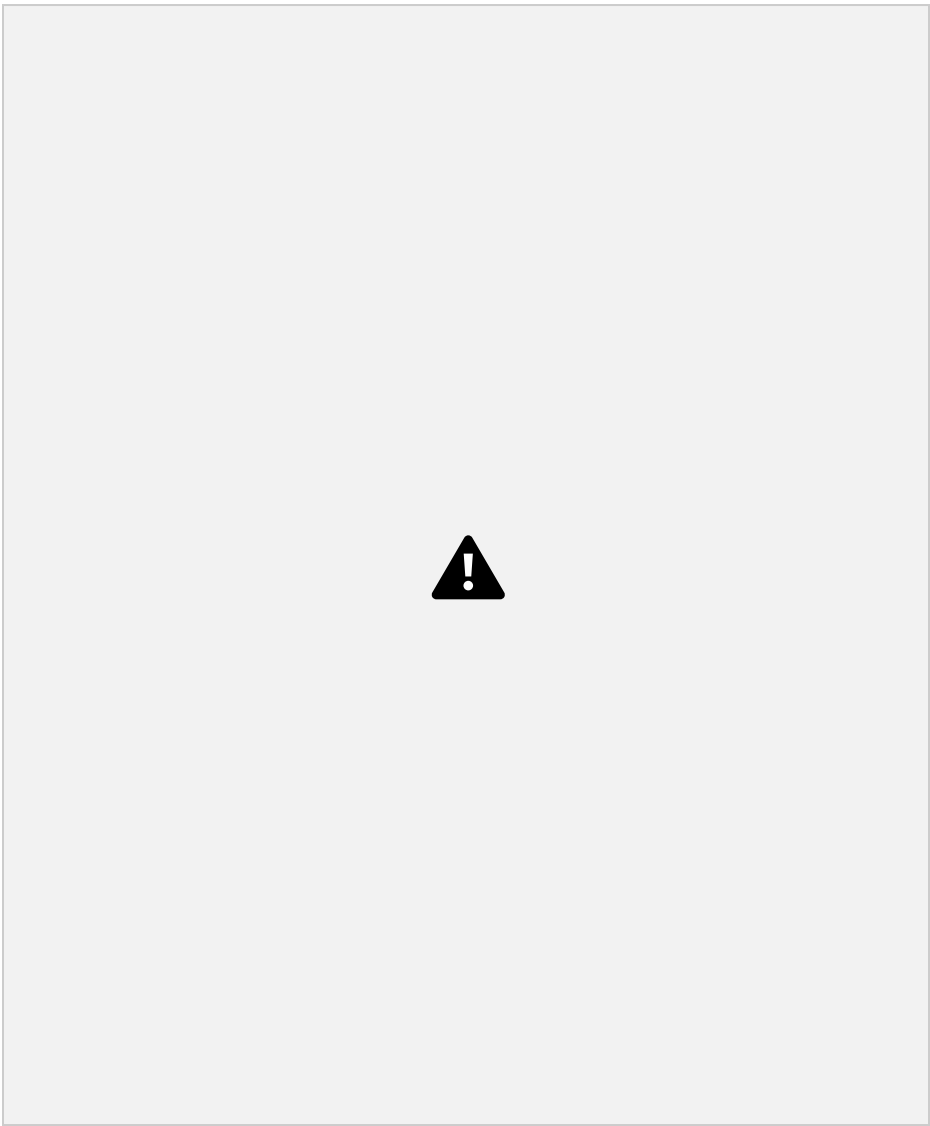
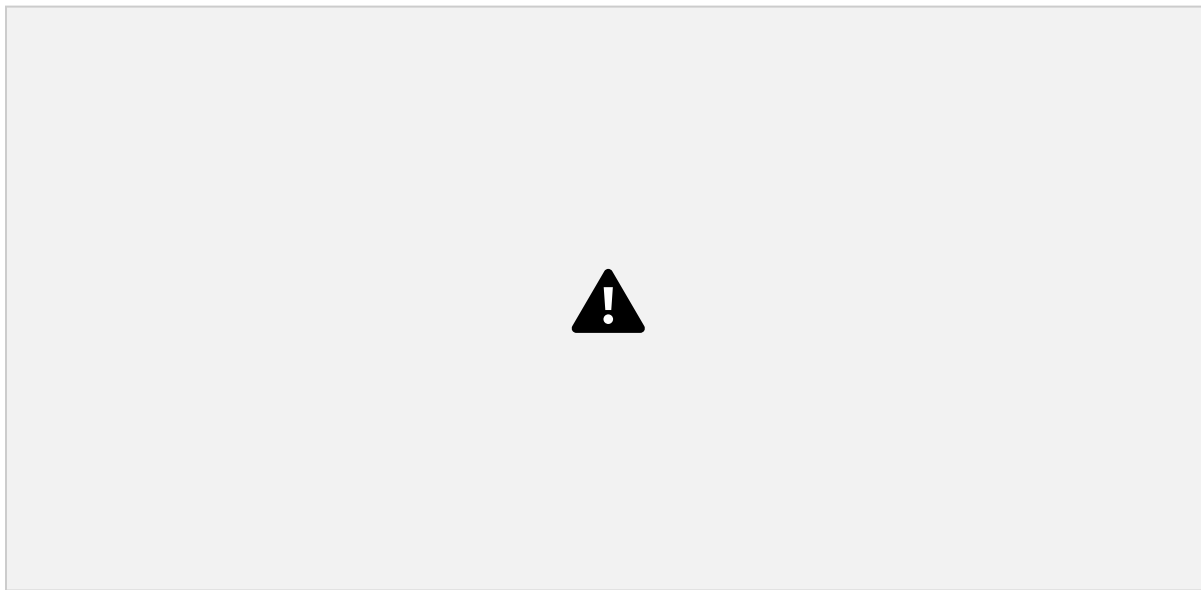


Fig.8:-Decision Tree Flowchart

Table 2:-Accuracy Graph for Heart Dataset

| | Existing System | Proposed System |
|-----------|-----------------|-----------------|
| Precision | 0.825 | 0.9 |
| Recall | 0.825 | 0.9 |
| F-Measure | 0.825 | 0.9 |





Graph 1: Decision Tree Performance

Flow Diagram

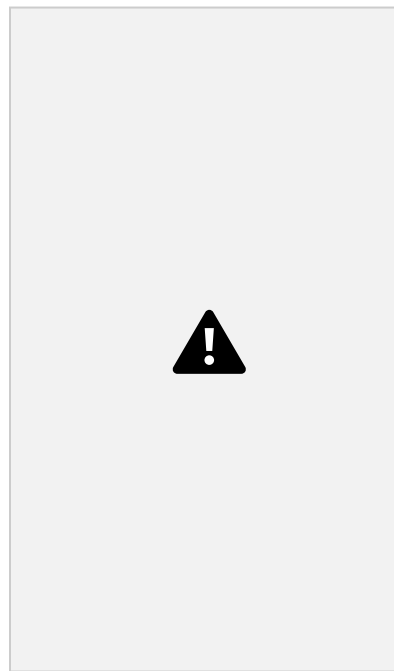


Fig.9:-Flow Diagram

CONCLUSION

The experiment is organized with the dataset of Heart Disease by machine learning algorithms. Heart Disease dataset is taken and analysed to predict the asperity of the disease. A Decision tree approach is used to predict the disease.

The data in the dataset is pre-processed to make it suitable for classification. The Decision tree approach to generate efficient classification rules is proposed. To perform classification task of medical data, the network is trained using Convolutions technique. This paper



real-time healthcare analytics system using traditional analytical tools is extremely complex, while exploiting open source big data technologies can do it in a simpler and more effective way.

REFERENCES

1. Khourdifi, Y., & Bahaj, M. (2019). Heart Disease Prediction and Classification Using Machine Learning Algorithms Optimized by Particle Swarm Optimization and Ant Colony Optimization. *International Journal of Intelligent Engineering & Systems*, 12(1).
2. Ed-Daoudy, A., & Maalmi, K. (2019, April). Real-time machine learning for early detection of heart disease using big data approach. In *2019 International Conference on Wireless Technologies, Embedded and Intelligent Systems (WITS)* (1-5). IEEE.
3. Deekshatulu, B. L., & Chandra, P. (2013). Classification of heart disease using k-nearest neighbor and genetic algorithm. *Procedia technology*, 10, 85-94. 3.
4. Kamkar, I., Akbarzadeh-T, M. R., & Yaghoobi, M. (2010, October). Intelligent water drops a new optimization algorithm for solving the vehicle routing problem. In *2010 IEEE International Conference on Systems, Man and Cybernetics* (4142-4146). IEEE.
5. Wilson, B., & Das, J. P. (2013). A survey of non-local means based filters for image denoising. *International Journal of Engineering Research & Technology*, 2(10), 3768-3771.
6. Chaitrali S Dangare. *Improved Study of Heart Disease Prediction System Using Data Mining Classification Techniques*. International Journal of Computer Applications.2012.47(10).
7. Amma, N. B. (2012, February). Cardiovascular disease prediction system using genetic algorithm and neural network. In *2012 International Conference on Computing, Communication and Applications* (1-5). IEEE.
8. Mukhopadhyay, S., Biswas, S., Roy, A. B., & Dey, N. (2012). Wavelet based QRS complex detection of ECG signal. *arXiv preprint arXiv:1209.1563*.



**A Mini-Project Report
On
“Heart Disease Prediction”**

[COMP 484]

Machine Learning

Submitted by

Nirusha Manandhar (31)

Sagun Lal Shrestha (53)

Ruchi Tandukar (57)

Submitted to

Dr. Bal Krishna Bal

Associate Professor

Department of Computer Science and

Engineering Submission Date: 11th March

2020

Table of Contents

| | |
|--------------------------|----------------|
| ABSTRACT | |
| i LIST OF FIGURES: | |
| ii LIST OF TABLES: | |
| iii LIST | OF |
| | ABBREVIATIONS: |
| iv | CHAPTER 1: |

| | |
|--|----|
| INTRODUCTION | 1 |
| 1.1 Problem Definition | 1 |
| 1.2 Motivation | 1 |
| 1.3 Objectives | 2 |
| CHAPTER 2: RELATED WORKS..... | 3 |
| CHAPTER 3: DATASETS | 4 |
| CHAPTER 4: METHODS AND ALGORITHMS USED..... | 5 |
| 4.1 Logistic Regression | 5 |
| 4.2 Backward Elimination Method: | 5 |
| 4.3 Recursive Feature Elimination using Cross-Validation (RFECV) | 6 |
| CHAPTER 5: EXPERIMENTS | 7 |
| 5.1 Data Preparation..... | 7 |
| 5.2 Exploratory Analysis: | 8 |
| 5.3 Feature Selection | 9 |
| 5.4 Training and testing | 10 |
| CHAPTER 6: EVALUATION METRICS | 11 |
| 6.1 Confusion Matrix | 11 |
| 6.2 Accuracy | 11 |
| 6.3 Recall | 12 |
| 6.4 Precision | 12 |
| CHAPTER 6: DISCUSSION ON RESULTS | 13 |
| CHAPTER 7: CONTRIBUTIONS | 14 |
| CHAPTER 9: CODE | 15 |
| 9.1 Libraries used: | 15 |
| CHAPTER 10: CONCLUSION | 16 |

ABSTRACT

This report represents the mini-project assigned to seventh semester students for the partial fulfillment of COMP 484, Machine Learning, given by the department of computer science and

engineering, KU. Cardiovascular diseases are the most common cause of death worldwide over the last few decades in the developed as well as underdeveloped and developing countries. Early detection of cardiac diseases and continuous supervision of clinicians can reduce the mortality rate. However, it is not possible to monitor patients every day in all cases accurately and consultation of a patient for 24 hours by a doctor is not available since it requires more sapience, time and expertise. In this project, we have developed and researched about models for heart disease prediction through the various heart attributes of patient and detect impending heart disease using Machine learning techniques like backward elimination algorithm, logistic regression and REFCV on the dataset available publicly in Kaggle Website, further evaluating the results using confusion matrix and cross validation. The early prognosis of cardiovascular diseases can aid in making decisions on lifestyle changes in high risk patients and in turn reduce the complications, which can be a great milestone in the field of medicine.

Keywords: Machine Learning, Logistic regression, Cross-Validation, Backward Elimination, REFCV, Cardiovascular Diseases.

LIST OF FIGURES:

| | |
|---|---|
| Figure 1: Original Dataset Snapshot | |
| 4 Figure 2: Bar Graph of the Target Classes After Dropping | 7 |
| Figure 3: Bar Graph of the Target Classes Before Dropping | 7 |
| Figure 4: Dataset after Scaling and Imputing | 7 |
| Figure 5: Correlation Matrix Visualization..... | 8 |
| Figure 6: Result from Feature Selection using Backward Elimination Method | 9 |
| Figure 7: Dataset | |

| | | |
|--|----|---------------|
| After Dropping Columns after Feature Selection | 9 | Figure 8: Top |
| 10 important features supported by RFECV | 10 | |

LIST OF TABLES:

| | |
|---|----|
| Table 1: Confusion Matrix Obtained after training the data (feature selection by backward elimination) | |
| 11 Table 2: Confusion Matrix Obtained after training the data (feature selection by RFECV method) | |
| 11 Table 3: Comparison between the feature selection models after training and testing through LogisticRegression model | |
| 13 Table 4: Work Division | |
| 14 Table 5: Major modules and classes used from Sklearn | 15 |

LIST OF ABBREVIATIONS:

1. IDE: Integrated Development Environment
2. REFCV: Recursive Feature Elimination using Cross-Validation
3. CV: Cross Validation
4. RFE: Recursive Feature Elimination

CHAPTER 1: INTRODUCTION

According to the World Health Organization, every year 12 million deaths occur worldwide due to Heart Disease. The load of cardiovascular disease is rapidly increasing all over the world from the past few years. Many researches have been conducted in attempt to pinpoint the most influential factors of heart disease as well as accurately predict the overall risk. Heart Disease is even highlighted as a silent killer which leads to the death of the person without obvious symptoms. The early diagnosis of heart disease plays a vital role in making decisions on lifestyle changes in high-risk patients and in turn reduce the complications. This project aims to predict future Heart Disease by analyzing data of patients which classifies whether they have heart

disease or not using machine-learning algorithms.

1.1 Problem Definition

The major challenge in heart disease is its detection. There are instruments available which can predict heart disease but either they are expensive or are not efficient to calculate chance of heart disease in human. Early detection of cardiac diseases can decrease the mortality rate and overall complications. However, it is not possible to monitor patients every day in all cases accurately and consultation of a patient for 24 hours by a doctor is not available since it requires more sapience, time and expertise. Since we have a good amount of data in today's world, we can use various machine learning algorithms to analyze the data for hidden patterns. The hidden patterns can be used for health diagnosis in medicinal data.

1.2 Motivation

Machine learning techniques have been around us and has been compared and used for analysis for many kinds of data science applications. The major motivation behind this research-based project was to explore the feature selection methods, data preparation and processing behind the training models in the machine learning. With first hand models and libraries, the challenge we face today is data where beside their abundance, and our cooked models, the accuracy we see during training, testing and actual validation has a higher variance. Hence this project is carried out with the motivation to explore behind the models, and further implement Logistic Regression

1

model to train the obtained data. Furthermore, as the whole machine learning is motivated to develop an appropriate computer-based system and decision support that can aid to early detection of heart disease, in this project we have developed a model which classifies if patient will have heart disease in ten years or not based on various features (i.e. potential risk factors that can cause heart disease) using logistic regression. Hence, the early prognosis of cardiovascular diseases can aid in making decisions on lifestyle changes in high risk patients and in turn reduce the complications, which can be a great milestone in the field of medicine.

1.3 Objectives

The main objective of developing this project are:

1. To develop machine learning model to predict future possibility of heart disease by implementing Logistic Regression.
2. To determine significant risk factors based on medical dataset which may lead to heart disease.

3. To analyze feature selection methods and understand their working principle.

CHAPTER 2: RELATED WORKS

With growing development in the field of medical science alongside machine learning various experiments and researches has been carried out in these recent years releasing the relevant significant papers. The paper [1] propose heart disease prediction using KStar, J48, SMO, and Bayes Net and Multilayer perceptron using WEKA software. Based on performance from different factor SMO (89% of accuracy) and Bayes Net (87% of accuracy) achieve optimum performance than KStar, Multilayer perceptron and J48 techniques using k-fold cross validation. The accuracy performance achieved by those algorithms are still not satisfactory. So that if the performance of accuracy is improved more to give better decision to diagnosis disease.

[2]In a research conducted using Cleveland dataset for heart diseases which contains 303 instances and used 10-fold Cross Validation, considering 13 attributes, implementing 4 different algorithms, they concluded Gaussian Naïve Bayes and Random Forest gave the maximum accuracy of 91.2 percent.

[3]Using the similar dataset of Framingham, Massachusetts, the experiments were carried out using 4 models and were trained and tested with maximum accuracy K Neighbors Classifier: 87%, Support Vector Classifier: 83%, Decision Tree Classifier: 79% and Random Forest Classifier: 84%.

CHAPTER 3: DATASETS

The dataset is publicly available on the Kaggle Website at [4] which is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. It provides patient information which includes over 4000 records and 14 attributes. The attributes include: age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting, sugar blood, resting electrocardiographic results, maximum heart rate, exercise induced angina, ST depression induced by exercise, slope of the peak exercise, number of major vessels, and target ranging from 0 to 2, where 0 is absence of heart disease. The data set is in csv (Comma Separated Value) format which is further prepared to data frame as supported by pandas library in python.

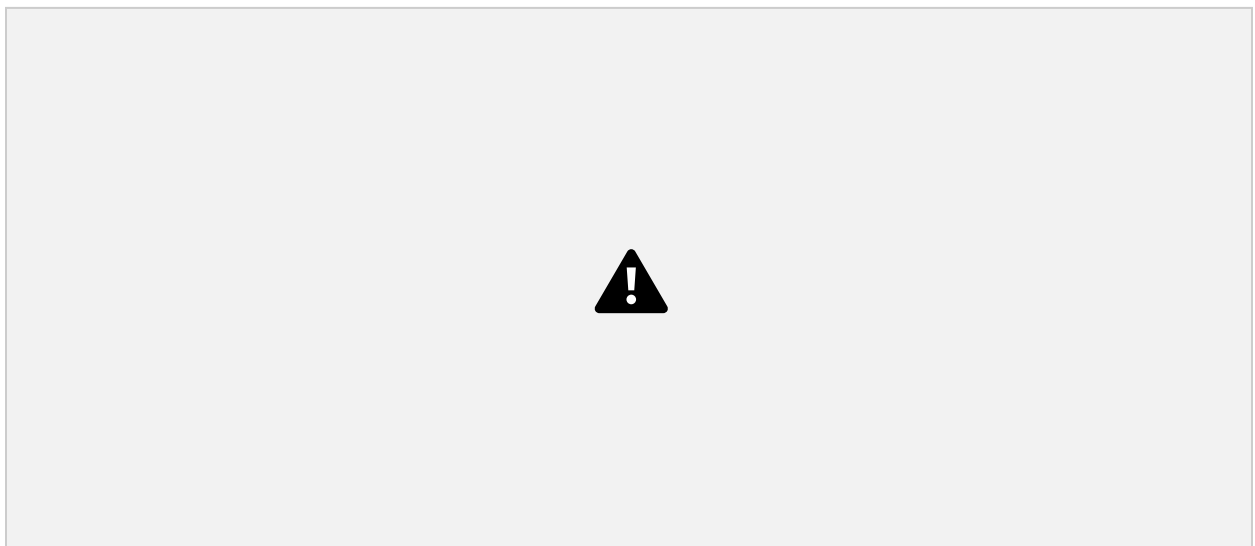


Figure 1: Original Dataset Snapshot

The education data is irrelevant to the heart disease of an individual, so it is dropped. Further with this dataset pre-processing and experiments are then carried out.

CHAPTER 4: METHODS AND ALGORITHMS USED

The main purpose of designing this system is to predict the ten-year risk of future heart disease. We have used Logistic regression as a machine-learning algorithm to train our system and various feature selection algorithms like Backward elimination and Recursive feature elimination. These algorithms are discussed below in detail.

4.1 Logistic Regression

Logistic Regression is a supervised classification algorithm. It is a predictive analysis algorithm based on the concept of probability. It measures the relationship between the dependent variable (TenyearCHD) and the one or more independent variables (risk factors) by estimating probabilities using underlying logistic function (sigmoid function). Sigmoid function is used as a cost function to limit the hypothesis of logistic regression between 0 and 1 (squashing) i.e. $0 \leq h_{\theta}(x) \leq 1$.

In logistic regression cost function is defined as:

$$J(h_{\theta}(x), y) = \begin{cases} -\log(h_{\theta}(x)) & \text{if } y = 1 \\ -\log(1 - h_{\theta}(x)) & \text{if } y = 0 \end{cases}$$

Logistic Regression relies highly on the proper presentation of data. So, to make the model more powerful, important features from the available data set are selected using Backward elimination and recursive elimination techniques.

4.2 Backward Elimination Method:

While building a machine learning model only the features which have a significant influence on the target variable should be selected. In the backward elimination method for feature selection, the first step is selecting a significance level or P-value. For our model, we have chosen a 5% significance level or P-value of 0.05. The feature with high P-value is identified, and if its P-value is greater than the significance level it is removed from the dataset. The model is fit again with a new dataset, and the process is repeated till all remaining features in dataset is less than the

significance level. In this model, factors male, age, cigsPerDay, prevalentStroke, diabetes, and sysBP were chosen as significant ones after using the backward elimination algorithm.

5

4.3 Recursive Feature Elimination using Cross-Validation (RFECV) RFECV is greedy optimization algorithm which aims to find the best performing feature subset. Recursive Feature Elimination (RFE) fits a model repeatedly and removes the weakest feature until specified number of features is reached. The optimal number of features is used with RFE to score different feature subsets and select the best scoring collection of features which is RFECV. The main issue of this algorithm is that it can be expensive to run. So, it is better to reduce the number of features beforehand. Since correlated features provide the same information, such features can be eliminated prior to RFECV. To address this, correlation matrix is plotted and the correlated features are removed.

The arguments for instance of RFECV are:

- a. estimator - model instance (RandomForestClassifier)
- b. step - number of features removed on each iteration (1)
- c. cv – Cross-Validation (StratifiedKFold)
- d. scoring – scoring metric (accuracy)

Once RFECV is run and execution is finished, the features that are least important can be extracted and dropped from the dataset. Top 10 features ranked by the RFECV technique in our model listed below from least importance to highest importance.

1. prevalentStroke
2. diabetes
3. BPMeds
4. currentSmoker
5. prevalentHyp
6. male
7. cigsPerDay
8. heartrate
9. glucose
10. diaBP

6

CHAPTER 5: EXPERIMENTS

5.1 Data Preparation

Since the dataset consists of 4240 observations with 388 missing data and 644 observations to be risked for heart disease, two different experiments were performed for data preparation. First, we checked by dropping the missing data, leaving with only 3751 data and only 572 observations risked for heart disease.



Figure 3: Bar Graph of the Target Classes Before Dropping Figure 2: Bar Graph of the Target Classes before Dropping Figure 3: Bar Graph of the Target Classes after Dropping Figure 2: Bar Graph of the Target Classes After Dropping

This leads to reduced number of the observations providing irrelevant training to our model. So, we progressed with imputation of data with the mean value of the observations and scaling them using SimpleImputer and StandardScaler modules of Sklearn.

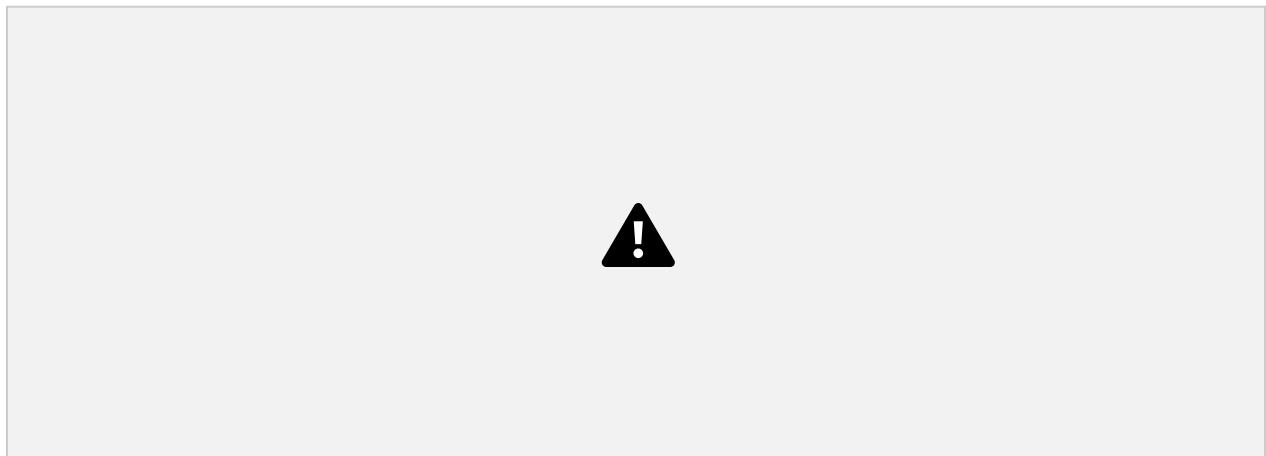


Figure 4: Dataset after Scaling and Imputing

5.2 Exploratory Analysis:

Correlation Matrix visualization Before Feature Selection shows



Figure 5: Correlation Matrix Visualization

It shows that there is no single feature that has a very high correlation with our target value. Also, some of the features have a negative correlation with the target value and some have positive. The data was also visualized through plots and bar graphs.

5.3 Feature Selection

Feature Selection using Backward Elimination (P-value) algorithm:

Further the data was passed through the backward elimination function to select the most relevant features which gave following result:

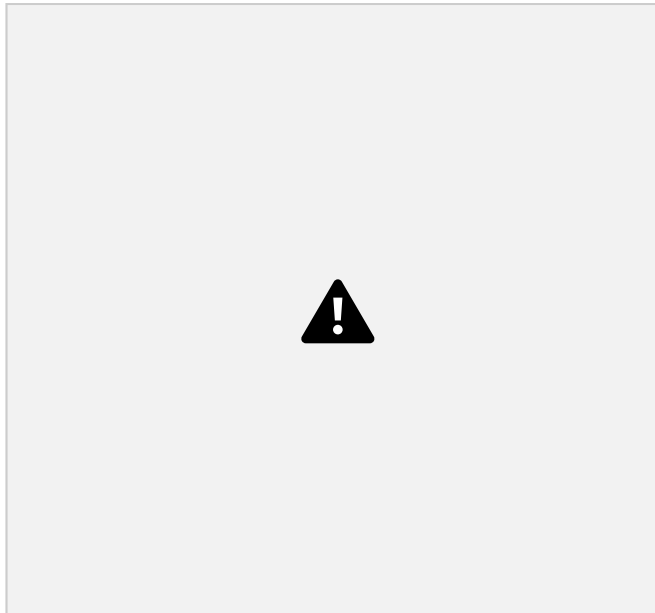


Figure 6: Result from Feature Selection using Backward Elimination Method

According to the result above the columns were dropped.

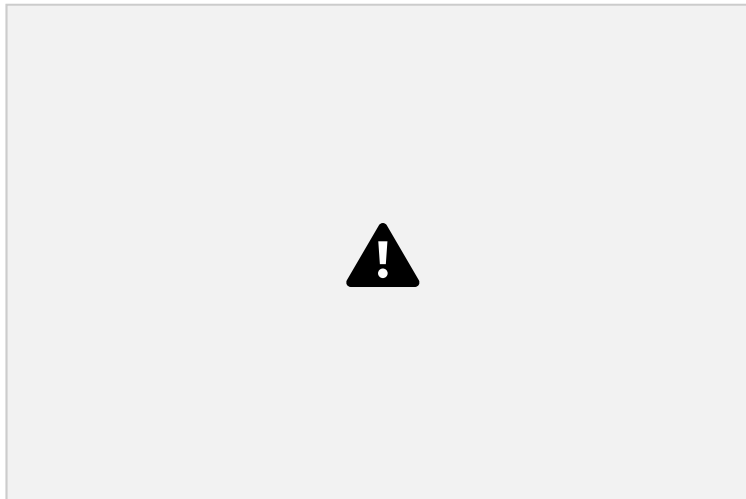


Figure 7: Dataset After Dropping Columns after Feature Selection

Feature Selection using Recursive Feature Elimination and Cross-Validated selection method:

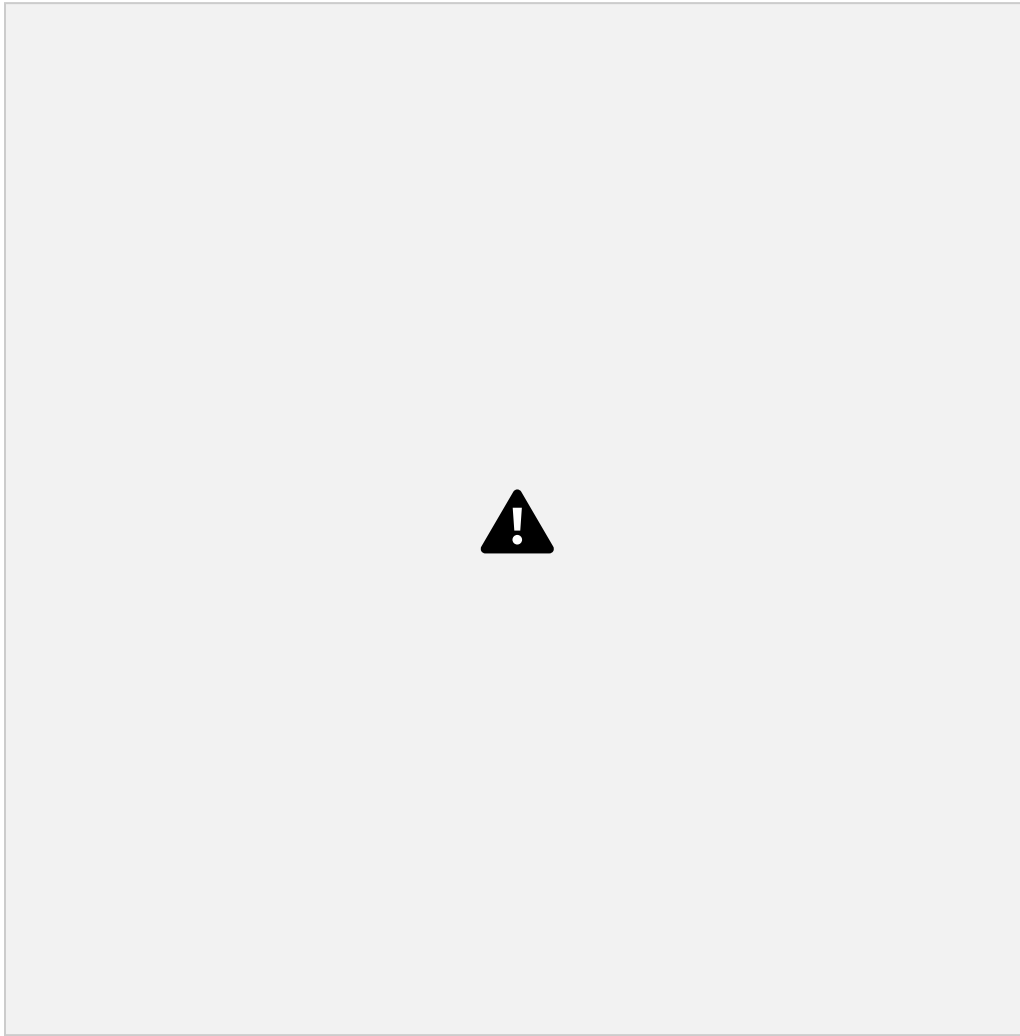


Figure 8: Top 10 important features supported by RFECV

5.4 Training and testing

Finally, this resulting data split into 80% train and 20% test data, which was further passed to the LogisticRegression model to fit, predict and score the model.

CHAPTER 6: EVALUATION METRICS

For the evaluation of our output from our training the data, the accuracy was analyzed “Confusion matrix”.

6.1 Confusion Matrix

A confusion matrix, also known as an error matrix, is a table that is often used to describe the performance of a classification model (or “classifier”) on a set of test data for which the true values are known. It allows the visualization of the performance of an algorithm. It allows easy

identification of confusion between classes e.g. one class is commonly mislabeled as the other. The key to the confusion matrix is the number of correct and incorrect predictions are summarized with count values and broken down by each class not just the number of errors made.

| | |
|---------|-------|
| TP=3569 | FP=27 |
| FN=599 | TN=45 |

Table 1: Confusion Matrix Obtained after training the data (feature selection by backward elimination)

| | |
|---------|-------|
| TP=3582 | FP=14 |
| FN=600 | TN=44 |

Table 2: Confusion Matrix Obtained after training the data (feature selection by RFECV method)

6.2 Accuracy

The accuracy is calculated as:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}}$$

Where,

- True Positive (TP) = Observation is positive, and is predicted to be positive. •
- False Negative (FN) = Observation is positive, but is predicted negative. • True
- Negative (TN) = Observation is negative, and is predicted to be negative. • False
- Positive (FP) = Observation is negative, but is predicted positive

The obtained accuracy during training the data after feature selection using backward elimination was 86 % and during testing was 83%.

The obtained accuracy during training the data after feature selection using RFECV method was 86 % and during testing was 85 %.

6.3 Recall

Recall can be defined as the ratio of the total number of correctly classified positive examples divide to the total number of positive examples. High Recall indicates the class is correctly recognized (a small number of FN). Recall is calculated as:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

The obtained recall during training the data after feature selection using backward elimination was and during testing was 0.99.

The obtained recall during training the data after feature selection using REFCV method was 1.00 and during testing was 0.99.

6.4 Precision

To get the value of precision we divide the total number of correctly classified positive examples by the total number of predicted positive examples. High Precision indicates an example labelled as positive is indeed positive (a small number of FP). Precision is calculated as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

The obtained precision during training the data after feature selection using backward elimination was 0.86 and during testing was 0.84.

The obtained precision during training the data after feature selection using REFCV method and during testing was 0.86.

CHAPTER 6: DISCUSSION ON RESULTS

When performing various methods of feature selection, testing it was found that backward elimination gave us the best results among others. The various methods tried were Backward Elimination with and without KFold, Recursive Feature Elimination with Cross Validation. The accuracy that was seen in them ranged around 85% with 85.5% being maximum. Though both methods gave similar accuracy but it was seen that in Backward Elimination we found that the number of misclassifications of True Negative was more and it was observed that the accuracy had more variance compared to RFEV. The precision of Backward Elimination and RFEV are 84% and 86% respectively. And the recalls are 0.99 and 1 respectively. The precision and recall also shows that the number of misclassifications is less in REFCV than in Backward Elimination.

| Evaluation Metrics | Backward Elimination | REFCV |
|--------------------|----------------------|-------|
|--------------------|----------------------|-------|

| | | |
|-----------|------|------|
| Accuracy | 83% | 85% |
| Recall | 0.99 | 0.99 |
| Precision | 0.84 | 0.86 |

Table 3: Comparison between the feature selection models after training and testing through LogisticRegression model

CHAPTER 7: CONTRIBUTIONS

| Members | Nirusha Manandhar | Sagun Lal Shrestha | Ruchi Tandukar |
|-----------------------------|-------------------|--------------------|----------------|
| Task | | | |
| Data Imputation and Scaling | | | |
| Data Cleaning | | | |
| Exploratory Analysis | | | |
| Feature Selection | | | |
| Building Model | | | |

| | | | |
|-----------------------------------|--|--|--|
| Result analysis and Accuracy Test | | | |
| Documentation | | | |

Table 4: Work Division

CHAPTER 9: CODE

The coding portion were carried out to prepare the data, visualize it, pre-process it, building the model and then evaluating it. The code has been written in Python programming language using Jupyter Notebook as IDE. The experiments and all the models building are done based on python libraries. The code is available in the Git repository given in following link:

https://github.com/ruchi-032/heart_disease_prediction_LogisticRegression

9.1 Libraries used:

1. NumPy
2. SciPy
3. Matplotlib (pyplot, rcparams, matshow)
4. Statsmodels
5. Pandas
6. Tkinter
7. Sklearn

| Modules used: | Imported class from respective modules: |
|--------------------------|---|
| a. Sklearn.impute | SimpleImputer |
| b. Sklearn.preprocessing | StandardScaler |

| | |
|------------------------------|-----------------------------------|
| c. Sklearn.pipeline | Pipeline |
| d. Sklearn.feature_selection | |
| e. Sklearn.ensemble | RandomForestClassifier |
| f. Sklearn.model_selection | Train_test_split, StratifiedKFold |
| g. Sklearn.linear_model | LogisticRegression, |
| h. Sklearn.utils | Shuffle |
| i. Sklearn.metrics | Accuracy_score, confusion_matrix |

Table 5: Major modules and classes used from Sklearn

CHAPTER 10: CONCLUSION

The early prognosis of cardiovascular diseases can aid in making decisions on lifestyle changes in high risk patients and in turn reduce the complications, which can be a great milestone in the field of medicine. This project resolved the feature selection i.e. backward elimination and RFECV behind the models and successfully predict the heart disease, with 85% accuracy. The model used was Logistic Regression. Further for its enhancement, we can train on models and predict the types of cardiovascular diseases providing recommendations to the users, and also use more enhanced models.

REFERENCES

- [1] A. H. M. S. U. Marjia Sultana, "Analysis of Data Mining Techniques for Heart Disease Prediction," 2018.
- [2] M. I. K. ., A. I. ., S. Musfiq Ali, "Heart Disease Prediction Using Machine Learning Algorithms".
- [3] K. Bhanot, "towarddatascience.com," 13 Feb 2019. [Online]. Available: <https://towardsdatascience.com/predicting-presence-of-heart-diseases-using-machine-learning-36f00f3edb2c>. [Accessed 2 March 2020].
- [4] [Online]. Available: <https://www.kaggle.com/ronitf/heart-disease-uci#heart.csv>.. [Accessed 05 December 2019].

- [5] M. A. K. S. H. K. M. a. V. P. M Marimuthu, "A Review on Heart Disease Prediction using Machine Learning and Data Analytics Approach".



Heart Disease Prediction using Machine Learning Algorithms

Devara Sandhya¹, Dr. Kamalraj R²

¹PG STUDENT, Department of Computer Application, JAIN(Deemed-To-Be) University Bangalore, Karnataka, India

²Assistant Professor, Department of CS and IT, JAIN(Deemed-To-Be) University, Karnataka, India

Abstract -In This Project Cardiovascular diseases are the most common cause of death worldwide over the last few decades in the developed as well as underdeveloped and developing countries. Early detection of cardiac diseases and continuous supervision of clinicians can reduce the mortality rate. However, it is not possible to monitor patients every day in all cases accurately and consultation of a patient for 24 hours by a doctor is not available since it requires more sapience, time and expertise. In this project, we have developed and researched about models for heart disease prediction through the various heart attributes of patient and detect impending heart disease using Machine learning techniques like backward elimination algorithm, logistic regression and REFCV on the dataset available publicly in Kaggle Website, further evaluating the results using confusion matrix and cross validation. The early prognosis of cardiovascular diseases can aid in making decisions on lifestyle changes in high risk patients and in turn reduce the complications, which can be a great milestone in the field of medicine.

Validation, Backward Elimination, REFCV, Cardiovascular Diseases.

1 Introduction

According to the World Health Organization, every year 12 million deaths occur worldwide due to Heart Disease. The load of cardiovascular disease is rapidly increasing all over the world from the past few years. Many researches have been conducted in attempt to pinpoint the most influential factors of heart disease as well as accurately predict the overall risk. Heart Disease is even highlighted as a silent killer which leads to the death of the person without obvious symptoms. The early diagnosis of heart disease plays a vital role in making decisions on lifestyle changes in high-risk patients and in turn reduce the complications. This project aims to predict future Heart Disease by analyzing data of patients which classifies whether they have heart disease or not using machine-learning algorithms.

Keywords: Machine Learning, Logistic regression, Cross

2 Problem Statement

The major challenge in heart disease is its detection. There are instruments available which can predict heart disease but either they are expensive or are not efficient to calculate chance of heart disease in human. Early detection of cardiac diseases can decrease the mortality rate and overall complications. However, it is not possible to monitor patients every day in all cases accurately and consultation of a patient for 24 hours by a doctor is not available since it requires more sapience, time and expertise. Since we have a good amount of data in today's world, we can use various machine learning algorithms to analyze the data for hidden patterns. The hidden patterns can be used for health diagnosis in medicinal data.

3 Proposed Solution

section depicts the overview of the proposed system and illustrates all of the components, techniques and tools are used for developing the entire system. To develop an intelligent and user friendly heart disease prediction system, an efficient software tool is needed in order to train huge datasets and compare multiple machine learning algorithms. After choosing the robust

16 top features. After that applied ANN and Logistic algorithm individually and compute the accuracy. Finally, we used proposed Ensemble Voting method and compute best method for diagnosis of heart disease.

Data Pre-processing:

This file contains all the pre-processing functions needed to process all input documents and texts. First we read the train, test and validation data files then performed some preprocessing like tokenizing, stemming etc. There are some exploratory data analysis is performed like response variable distribution and data quality checks like null or missing values etc.

Feature:

Extraction In this file we have performed feature extraction and selection methods from sci-kit learn python libraries. For feature selection, we have used methods like simple bag-of-words and n-grams and then term frequency like tf-idf weighting. We have also used word2vec and POS tagging to extract the features, though POS tagging and word2vec has not been used at this point in the project.

Classification:

Here we have built all the classifiers for the breast cancer

algorithm with best accuracy and performance measures, it will be implemented on the development of the smart phone-based application for detecting and predicting heart disease risk level. Hardware components like Arduino/Raspberry Pi, different biomedical sensors, display monitor, buzzer etc. are needed to build the continuous patient monitoring system.

The below figure shows the process flow diagram or proposed work. First we collected the Cleveland Heart Disease Database from UCI website then pre processed the dataset and select 16 important features.

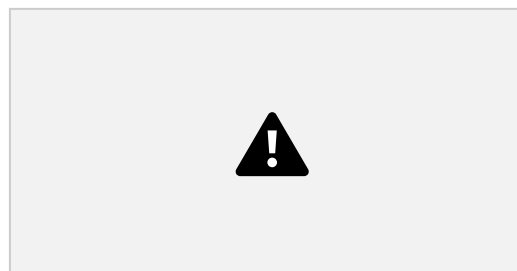


Fig : System Architecture

For feature selection we used Recursive feature Elimination Algorithm using Chi2 method and get

diseases detection. The extracted features are fed into different classifiers. We have used Naive bayes, Logistic Regression, Linear SVM, Stochastic gradient decent and Random forest classifiers from sklearn. Each of the extracted features was used in all of the classifiers. Once fitting the model, we compared the f1 score and checked the confusion matrix. After fitting all the classifiers, 2 best performing models were selected as candidate models for heart diseases classification. We have performed parameter tuning by implementing GridSearchCV methods on these candidate models and chosen best performing parameters for these classifier. Finally selected model was used for heart disease detection with the probability of truth. In Addition to this, we have also extracted the top 50 features from our term-frequency tfidf Vectorizer to see what words are most and important in each of the classes. We have also used Precision-Recall and learning curves to see how training and testset performs when we increase the amount of data in our classifiers.

Prediction:

Our finally selected and best performing classifier was algorithm which was then saved on disk with name final_model.sav. Once you close this repository, this model will be copied to user's machine and will be used by prediction.py file to classify the Heart diseases. It takes a news article as input from user then model is used for final classification output that is shown to user along with probability of

truth.

4 Literature Study

Machine Learning techniques are used to analyze and predict the medical data information resources. Diagnosis of heart disease is a significant and tedious task in medicine. The term Heart disease encompasses the various diseases that affect the heart. The exposure of heart disease from various factors or symptom is an issue which is not complimentary from false presumptions often accompanied by unpredictable effects. The data classification is based on Supervised Machine Learning algorithm which results in better accuracy. Here we are using the Random Forest as the training algorithm to train the heart disease dataset and to predict the heart disease. The results showed that the medicinal prescription and designed prediction system is capable of prophesying the heart attack successfully.

Machine Learning techniques are used to indicate the early mortality by analyzing the heart disease patients and their clinical records .have brought about the two Machine Learning techniques, k nearest neighbor model and existing multi line arregression to predict the stroke severity index of the patients. Their study show that k-nearest

neighbor performed better than Multi Linear Regression model. Have suggested various Machine Learning techniques such as support vector machine, penalized logistic regression to predict the heart stroke. Their results show that Support vector machine produced the best performance in prediction when compared to other models. Boshra Brahmi et al, developed different Machine Learning techniques to evaluate the prediction and diagnosis of heart disease. The main objective is to evaluate the different classification techniques, Decision Tree, KNN and Naïve Bayes. After this, evaluating some performance in measures of accuracy, precision, sensitivity, specificity are evaluated.

5 Future Scope

As illustrated before the system can be used as a clinical assistant for any clinicians. The disease prediction through the risk factors can be hosted online and hence any internet users can access the system through a web browser and understand the risk of heart disease. The proposed model can be implemented for any real time application .Using the proposed model other type of heart disease also can be determined. Different heart diseases asrheumatic heart disease, hypertensive heart disease, ischemic heart disease,

© 2022 IRJET | Impact Factor value: 7.529 | ISO 9001:2008 Certified Journal | Page 1545

International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056 Volume: 09 Issue: 03 | Mar 2022 www.irjet.net p-ISSN: 2395-0072

cardiovascular disease and inflammatory heart disease can be identified. Other health care systems can be formulated using this proposed model in order to identify the diseases in the early stage. The proposed model requires an efficient processor with good memory configuration to implement it in real time. The proposed model has wide area of application like grid computing, cloud computing, robotic modeling, etc. To increase the performance of our classifier in future, we will work on ensembling two algorithms called Random Forest and Adaboost. By ensembling these two algorithms we will achieve high performance.

6 Conclusion

This project, we introduce about the heart disease prediction system with different classifier techniques for the prediction of heart disease. The techniques are Random Forest and Logistic Regression: we have analyzed that the Random Forest has better accuracy as compared to Logistic Regression. Our purpose is to improve the performance of the Random Forest by removing unnecessary and irrelevant attributes from the dataset and only picking those that are most informative for the classification task.

References

[1] P.K. Anooj, --Clinical decision support system:

Risk level prediction of heart disease using weighted fuzzy rules; Journal of King Saud University – Computer and Information Sciences (2012) 24, 27 – 40. Computer Science & Information Technology (CS & IT) 59

[2] Nidhi Bhatla, Kiran Jyoti "An Analysis of Heart Disease Prediction using Different Data Mining Techniques". International Journal of Engineering Research & Technology

[3] Jyoti Soni Ujma Ansari Dipesh Sharma, Sunita Soni. "Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction".

[4] Chaitrali S. Dangare Sulabha S. Apte, Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques" International Journal of Computer Applications (0975 – 888)

[5] Dane Bertram, Amy Volda, Saul Greenberg, Robert Walker, "Communication, Collaboration, and Bugs: The Social Nature of Issue Tracking in Small, Collocated Teams".

[6] M. Anbarasi, E. Anupriya, N.Ch.S.N.Iyengar, —Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm; International Journal of Engineering Science and Technology, Vol. 2(10), 2010.

[7] Ankita Dewan, Meghna Sharma, "Prediction of Heart Disease Using a Hybrid Technique in Data Mining Classification", 2nd International Conference on Computing for Sustainable Global Development IEEE 2015 pp 704-706. [2].

[8] R. Alizadehsani, J. Habibi, B. Bahadorian, H. Mashayekhi, A. Ghandeharioun, R. Boghrati, et al., "Diagnosis of coronary arteries stenosis using data mining," J Med Signals Sens, vol. 2, pp. 153-9, Jul 2012.

[9] M Akhil Jabbar, BL Deekshatulu, Priti Chandra, "Heart disease classification using nearest neighbor classifier with feature subset selection", Anale. Seria Informatica, 11, 2013

[10] Shadab Adam Pattekari and Asma Parveen, "PREDICTION SYSTEM FOR HEART DISEASE USING NAIVE BAYES", International Journal of Advanced Computer and Mathematical Sciences ISSN 2230-

9624, Vol 3, Issue 3, 2012, pp 290-294.

[11] C. Kalaiselvi, PhD, "Diagnosis of Heart Disease Using K -Nearest Neighbor Algorithm of Data Mining", IEEE, 2016

[12] Keerthana T. K., "Heart Disease Prediction System using Data Mining Method", International Journal of Engineering Trends and Technology", May 2017.

[13] Data Mining Concepts and Techniques, Jiawei Han and Micheline Kamber, ELSEVIER. Animesh Hazra, Arkomita Mukherjee, Amit Gupta, Prediction Using Machine Learning and Data Mining July 2017, pp. 2137-2159.

HEART DISEASE PREDICTION USING MACHINE LEARNING

¹Rishabh Magar, ²Rohan Memane, ³Suraj Raut

¹Prof. V. S. Rupnar

¹Computer Department,

¹MMCOE, Pune, India.

Abstract : Heart disease is one of the major causes of death throughout the world. It cannot be easily predicted by the medical practitioners as it is a difficult task which demands expertise and higher knowledge for prediction. An automated system in medical diagnosis would enhance medical efficiency also reduce costs. will design a system that can efficiently discover the rules predict the risk

and
We

to

level of patients based on the given parameters about their health. The goal is to extract hidden patterns by applying data mining techniques, which are noteworthy to heart diseases and to predict the presence of heart disease in patients where the presence is valued on a scale. The prediction of heart disease requires a huge size of data which is too complex and massive to process and analyze by conventional techniques. Our objective is to find out the suitable machine learning technique that is computationally efficient as well as accurate for the prediction of heart disease. Data mining combines Statistical analysis machine learning and database technology to extract hidden patterns and relationships from large databases. The implementation of work is done on Cleveland heart diseases data set from the University of California Irvine (UCI) machine learning repository to test on different data mining techniques.

***IndexTerms* - Machine learning (ML), support vector machines (SVM), supervised learning.**

I. INTRODUCTION

A. Basics and backgrounds

Heart disease is considered as one of the major causes of death throughout the world. It cannot be easily predicted by the medical practitioners as it is a difficult task which demands expertise and higher knowledge for prediction. An automated system in medical diagnosis would enhance medical efficiency and also reduce costs. We will design a system that can efficiently discover the rules to predict the risk level of patients based on the given parameters about their health. The goal is to extract hidden patterns by applying data mining techniques, which are noteworthy heart diseases and to predict the presence of heart disease in patients where the presence is valued on a scale. The prediction of heart disease requires a huge size of data which is too complex and massive to process and analyze by conventional techniques. Our objective is to find out the suitable machine learning technique that is computationally efficient as well as accurate for the prediction of heart disease. Data mining combines statistical analysis, machine learning and database technology to extract hidden patterns and relationships from large databases. The implementation of work is done on Cleveland heart diseases dataset from the University of California Irvine (UCI) machine learning repository to test on different data mining techniques.

B. Literature Survey

Senthilkumar Mohan have suggested mine to discover hidden information for effective decision making. Discovery of hidden patterns and relationships often goes unexploited. Advanced ML techniques can help remedy this situation. This research concludes with various models that can be used for prediction. Anjan N. Repaka stated the performance of prediction for two classification models, which is analyzed and compared to previous work. Experimental results show the improved accuracy percentage of risk prediction of our proposed method compared to other works.

Aditi Gavhane addresses the issue of prediction of heart disease according to input attributes on the basis of various data mining techniques and represented them with their accuracy in tabular format. It proposes to develop an application which can predict the vulnerability of a heart disease given basic symptoms like age, sex, pulse rate etc. The machine learning algorithm neural networks has proven to be the most accurate and reliable algorithm and hence used in the proposed system.

Santhana Krishnan predicts the arising possibilities of Heart Disease. The outcomes of this system the chances of occurring heart disease in terms of percentage. The datasets used are classified in terms of medical parameters. This system evaluates those parameters using data mining classification technique. The datasets are processed in python programming using four main Machine Learning Algorithm Namely Decision Tree, Logistic Regression, Support