International Journal of Application on Sciences, Technology and Engineering
(IJASTE)
Volume 1, Issue 1, 2023.ISSN:2987-2499

Analyzing the Most Relevant Predictors for Adult Coronary Heart Disease using Machine Learning

Muhammad Naufaldi[1, a)], Sunny Jovita[1, b)] and Nunung Nurul Qomariyah[1, c)]

[1]*Department of Computer Science, Faculty of Computing and Media, Bina Nusantara University, Jakarta, Indonesia, 11480.*

[a)]*Corresponding author: muhammad.naufaldi@binus.ac.id*
[b)]*sunny.jovita@binus.ac.id*
[c)]*nunung.qomariyah@binus.ac.id*

Submitted: November-December 2022, Revised: January 2023, Accepted: February 23, 2023

Abstract. Coronary heart disease has been the number one illness to cause death in the world for decades. The healthcare indus-tries generates vast amount of clinical data, driven by medical records of patients, regulatory requirements, and results of medicalexaminations. In order to obtain the most relevant features for coronary heart disease, this study has conducted an experimental evaluation on data-driven diagnosis of coronary heart disease using classification algorithms. A statistical test (Chi-square) is usedto find the most valuable features and risk factors associated with coronary heart disease. The purposed of this univariate feature extraction algorithm is to determine the difference between the observed resuslts with expected results. Furthermore, CHD is predicted using several classification machine learning algorithms including Logistic Regression, Complement Naïve Bayes. andSupport Vector Machine (SVM). This study also evaluates ensemble machine learning algorithms, such as Random Forest and Extreme Gradient Boosting (XGBoost), Gradient Boost, to find the best performance of the classifications algorithms and select essential features from the dataset. Holdout and cross-validations methods are used to separated the dataset into two sets, called thetraining set and the testing set. The performance of proposed algorithm are assessed in terms of certain factors such as specificityand sensitivity. From this study, it is shown that Gradient boost model exhibits the best performance with 0.839 sensitivity.

INTRODUCTION

The most common heart disease, Coronary Heart Disease (CHD) is widely known as one of the major causes of mor-tality. As stated by the World Health Organization (WHO), 17,9 million people died from CHD in 2019, representing32% of all deaths globally [1]. Some heart related diseases are caused due to a number of contributing factors, includ-ing diabetes, high blood pressure, cholesterol, smoking, chronic diseases, and many more. With the advancement of technology, the treatment of coronary heart disease has recently been mentioned in numerous studies that have gottentremendous attention within the healthcare industry. This can be proven by researchers from New York University'sSchool of Global Public Health and Tandon School of Engineering, screened more than 1,600 articles and focused on 48 peer-reviewed studies published in journals between 1995 and 2020 [2]. They found that applying machine learning models improved the ability to predict cardiovascular diseases. Hence, to examine and identify some of the early signs of heart disease, various diagnoses and many data analytics tools have been adopted. However, manuallyestimating the likelihood of substantial coronary heart disease is troublesome to rely upon the risk factors. Therefore, machine learning and various data mining techniques are applied in regards to solve such complicated issues. More-over, advanced machine learning techniques such as tuning, oversampling, feature selection, and ensemble learning will assist us to identify the most relevant features and patterns of coronary heart disease.

Generally, machine learning plays a significant role in the medical industry, particularly in forecasting coronary heart disease. Machine learning can be an effective tool both to predict heart failure symptoms and to detect the most important clinical features that may lead to heart failure [3]. As a matter of fact, machine learning helps in minimizing the diagnostic time and demonstrating accuracy and effectiveness.

In this study, we will present a comparative analysis of which machine learning algorithms can perform better in predicting the initial stage of CHD. Several machine learning techniques are utilized to identify how well can these algorithms classify questions related to coronary heart disease from the dataset. Furthermore, cross validation is employed to separate the test and training datasets and to evaluate the models' performance.

The purpose of this study is to find the most important features which can better predict the heart disease in adult and

https://doi.org/10.24912/ijaste.v1.i1.180-188

to review the performance of classification methods based on sensitivity and specificity in the context of heart disease. Therefore, it can be used as a reference in choosing a method for heart disease prediction of future research. This study is divided into several sections: Section 2 is a journal article on recent research in this field. The methodology and proposed architecture are discussed in Section 3. Section 4 presents experimental results as well as a comparisonof classification techniques. Finally, the study is concluded with suggestions on how to improve the results for future work.

RELATED STUDIES

A number of research studies report the use of machine learning algorithms for predicting heart disease [4, 5, 6,7, 8, 9]. Different data mining techniques and performance methods have been implemented to provide different perspectives on prediction of heart disease. Similar study has been performed by Khan et al. [4], where Khan [4] studied three datasets to diagnose heart disease. Numerous machine learning classifiers are utilized, including SVM,Logistic Regression, Naïve Bayes, and others. Furthermore, the accuracy, recall, and F1-score were calculated to validate the algorithms. When applied to the first dataset, the results show that the machine learning combination hasthe highest accuracy of 88.89%.

Thai-Nghe et al. [5] presented cost-sensitive learning to deal with imbalanced data. They [5] combined and compared several sampling techniques with cost-sensitive analysis using SVM, and also used cost-sensitive analysis by optimizing the cost ratio. Their experimental results show that using cost sensitive method can reduce misclassification costs and improve classifier performance.

Begum et al. [6] dealt with XGBoost, SVM, Random Forest, Logistic Regression, and some regularly algorithms to predict heart disease. The experimental findings indicated that Random Forest machine learning algorithm achieves the most accurate yet reliable algorithm and hence utilized in the proposed system.

Zeng et al. [7] proposed a powerful processing method using hybrid technique, SMOTE with Tomek links technique and then applied it to the imbalanced medical dataset to evaluate the effectiveness of this method. The findings showed that the SMOTE method combined with Tomek links technique is far superior in contrast to using only SMOTE.

Zhu et al. [8] utilized sensitivity, specificity, and accuracy analysis in the context of disease diagnosis. The author said that sensitivity and specificity are widely applied to describe diagnostic tests in the medical fields. They are usedto qualify how good and reliable a test is [8].

From the related studies, we can conclude that the relevant subset features gained from the classifier training truly improves the accuracy of the algorithms. This study however, will try to tackle this problem with a different data. The data that will be used is a result of a survey which mostly contains yes or no questions. Using these answers, thisstudy will find the most important features for adult coronary heart disease and create models accordingly.

METHODOLOGY

The proposed study focuses on discovering relevant features by removing unnecessary and redundant attributes fromthe dataset. The architecture of the proposed framework is depicted in Figure 1. Data gathering, data pre-processing,extraction of features, data splitting, model training with classifiers, and model evaluation are the key components ofthe framework.The steps of the proposed approach are explained in detail in the figure.

Dataset

The dataset was taken from the Behavioral Risk Factor Surveillance System (BRFSS), which was the result of a surveyconducted by the Centers for Disease Control and Prevention (CDC) [10]. Only 1 year was taken from the dataset which was 2019, the reason for this is to avoid any duplicates since there is no identifying data that can separate eachperson. The dataset consisted of 342 columns, 27,724 entries, and 1 target column. A lot of these columns are not relevant to our research therefore we eliminate most of them and only 25 columns that are valuable to CHD. The target column is divided into two categories: 1 implies heart disease and 0 implies non-heart disease. Table I contains information about the features. Another note about this dataset is that it is very imbalanced as shown in Figure 2. These can hinder the process of training model later but there will be solutions that will be explained on the next

https://doi.org/10.24912/ijaste.v1.i1.180-188

sections.

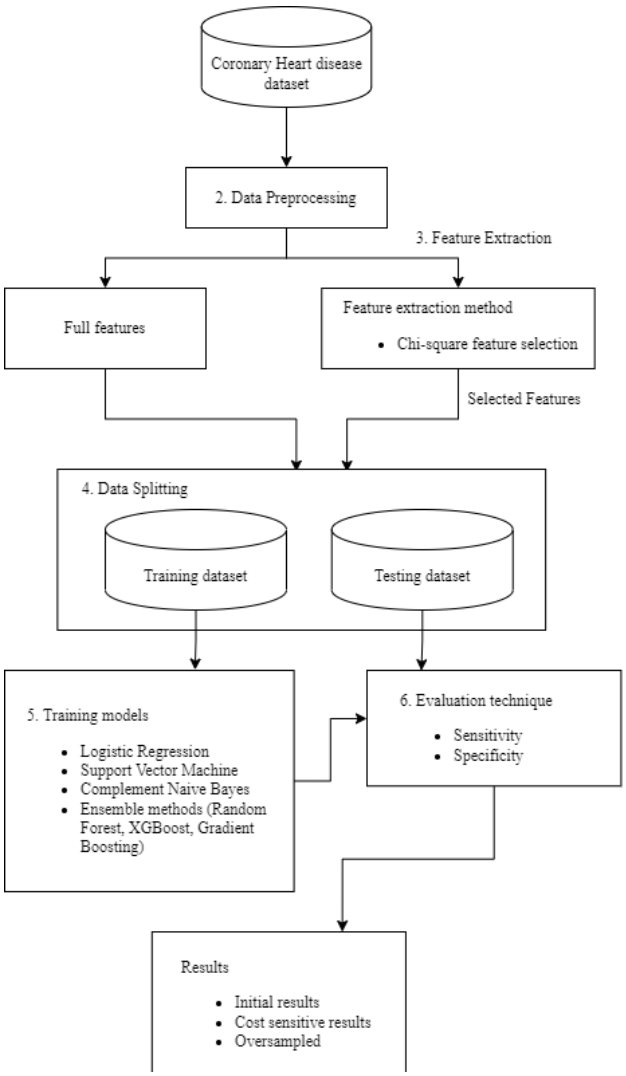

FIGURE 1. The proposed system for prediction CHD

Data Preprocessing

First and foremost, the data needed to be cleaned. These are all categorical data but there are some columns that needto be cleaned. For example, most of the data has the category 7 or 9 which usually indicates that the respondent did not answer or did not know the answer. These categories have no meaning therefore, not needed in the experiment. Finally, to obtain higher accuracy from the imbalance dataset, we applied a hybrid technique using SMOTE and Tomek links [7] to clean up overlapping data for each class distributed in the dataset.

Feature Extraction

In order to assess the effect of feature extraction, the experiments were conducted with and without feature extraction. Feature extraction is a significant stage since unessential features (outliers) frequently affect the classification efficiency of machine learning classifiers [9]. The Chi-square feature selection algorithm method is utilized in this study to select essential features from the dataset. It identifies the most valuable features and risk factors associated with CHD in the dataset.

https://doi.org/10.24912/ijaste.v1.i1.180-188

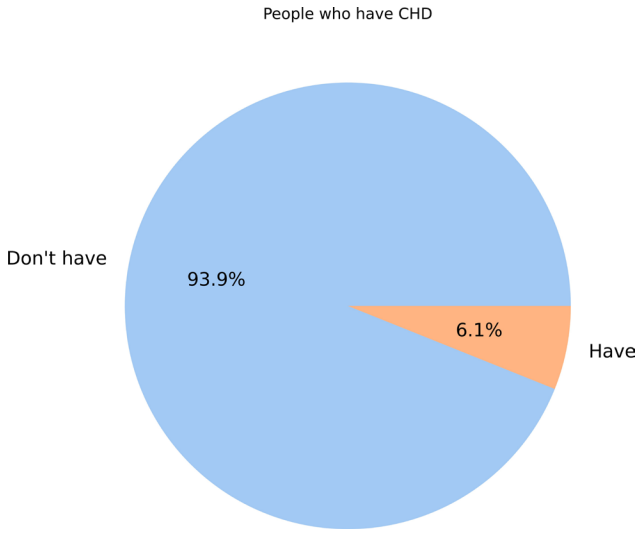

FIGURE 2. Distribution of the target feature

TABLE 1. Dataset features

Feature Name	Description	Value
GenHlth	General health	1-5 Where 1 is excellent and 5 is poor
HlthCare	Have health care	0:Yes, 1:No
MedCost	Have not avoided seeing a doctor due to the cost	0:Yes, 1:No
Stroke	Have stroke	0:No, 1:Yes
OthCncr	Have other cancer besides skin cancer	0:No, 1:Yes
ChronPulmo	Have chronic pulmonary disease	0:No, 1:Yes
DepressDis	Have depressive disorder	0:No, 1:Yes
ChronKidn	Have chronic kidney disease	0:No, 1:Yes
Diabetes	Have diabetes	0:No, 1:Yes
DiffWalk	Have difficulty walking	0:No, 1:Yes
Smoke100	Have smoked 100 tobacco	0:No, 1:Yes
HighBP	Have high blood pressure	0:No, 1:Yes
HighChol	Have high cholesterol	0:No, 1:Yes
CHD	Have coronary Heart Disease (Target)	0:No, 1:Yes
Arthritis	Have athritis	0:No, 1:Yes
Sex	Gender	0:Female, 1:Male
Age	Age	14-level age category starting from 18-24 until 85+
SmokerStatus	Smoker Status	0:No, 1:Yes
HeavyDrink	Heavy Drinker	0:No, 1:Yes
ActivityAerobi	Met aerobic exercise recommendation	0:Yes, 1:No
Fruit	Eating fruit everyday	0:Yes, 1:No
Vegetable	Eating vegetables everyday	0:Yes, 1:No
BMI-tr	BMI	0:Underweight or Normal, 1:Overweight or Obese
Phys	Met muscle strengthening recommendation	0:Yes, 1:No
MI	Have myocardial infarction	0:No, 1:Yes
Skin	Have skin cancer	0:No, 1:Yes

Machine Learning Model

We utilized six machine learning algorithms in this analysis: Logistic Regression (LR), Complement Naïve Bayes

https://doi.org/10.24912/ijaste.v1.i1.180-188

(CNB), Support Vector Machine (SVM), and three types of ensemble techniques: Random Forest and boosting algorithms (XGBoost and Gradient Boost).

Logistic Regression

Logistic regression is a useful analytical method for classification problems, where you are trying to determine ifa new sample fits best into a category. This algorithm works very similar to linear regression, but with a binomial response variable [11]. Logistic regression does not require a linear relationship between input and output variables. This will be one of the best algorithms in predicting whether a patient has a heart disease or not.

Support Vector Machine

Due to its ability to handle multidimensional data, the Support Vector Machine (SVM) is classified as a supervised machine learning technique that divides data into multiple classes using hyperplanes. The basic idea of SVM can be seen in Figure 3. Figure 3 shows the two parallel hyperplanes (Class A and B) are separated by H. H is the hyperplane that maximizes the distance between the Class A and B [12]. The distance of two parallel hyperplanes is called class interval. An assumption is made that the larger the margin or distance between these parallel hyperplanes the better the classifier generalization will be [12].

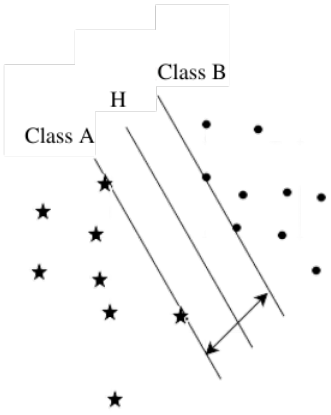

FIGURE 3. Support Vector Machine Hyperplane

To find where the result tends to negative or positive (Class A or B), we can use the Hyperplane's equation and define decision rule as:

$$w.x+b=0 \quad (1)$$

$$w.x+b>0 \quad (2)$$

$$w.x+b<0 \quad (3)$$

where:

w = a vector normal to Hyperplane
b = an offset

If the value of w.x+b<0 then we can say it is a negative point otherwise it is a positive point. For this, we will takefew assumptions that the equation of A is w.x+b=1 and for B it is w.x+b=-1.

https://doi.org/10.24912/ijaste.v1.i1.180-188

Complement Naive Bayes

Complement Naïve Bayes (CNB) is an adaptation of the Multinomial Naïve Bayes (MNB) algorithm. However, MNB does not perform well with imbalanced datasets. Whereas, CNB is particularly suited to work with imbalanced datasets because it is formed to complete disadvantages of NB algorithms such as skewed data [13]. We assume CNB's forecasts will be more effective as it uses more training data per class, which will reduce bias in weight estimates.

Ensemble Learning

To achieve highest classification accuracy, we utilized Random Forest, Gradient Boosting, and XGBoost ensemble methods to reduce drawbacks in prediction due to the imbalanced data.

Random Forest (RF)

Random Forest is a powerful supervised classification tool. This ensemble classifier is applied to create forest of treesthat are not correlated with weak features to strong features and give more accurate results.

Gradient Boosting (GB)

The characteristics of gradient boosting classifier is to optimize a loss function. It relies on the intuition that the nestmodel, when combined with the previous model, will minimize overall prediction errors [14]. Hyperparameter tuning(GridSearchCV) has been performed on the Gradient boosting algorithm to improve the model performance and to reduce overfitting.

XGBoost (XGB)

Similar to the Gradient boosting classifier, XGBoost (Extreme Gradient Boosting) is an ensemble method, which attempts to accurately predict the target feature (CHD) by combining the weaker feature estimates.

The hyperparameters for each models are determined using GridSearchCV with 10-folds for the search parameter. However, CNB does not have any parameters therefore will be skipped during this process. The hyperparameters that GridSearchCV produced are as follows:

- LR
  C = 10, penalty = '12'
- SVM
  C = 0.1, penalty = 'l2'
- RF
  criterion = 'gini', max_depth = 5, max_features = auto, n_estimators = 100
- XGB
  learning_rate = 0.01, n_estimators = 500, gamma = 5, max_delta_step = 0, max_depth = 5, min_child_weight = 1
- GB
  learning_rate= 0.5, max_depth = 2, n_estimators = 1000

Evaluation Technique

The effectiveness and accuracy of the machine learning methods can be evaluated using performance indicators. Theresults of this experiments were evaluated with sensitivity and specificity. The reason for these metrics is that the sensitivity and specificity of a are useful for cross sectional studies [15]. Sensitivity is an approach that identifies people with the coronary heart disease (CHD) (true positive rate) and specificity is an approach that identifies patients

without coronary heart disease (true negative rate) [16]. The experiments will include both a dataset that went through feature selection and the one that does not. Next, for the evaluation, we adopted cost-sensitive learning to deal with prediction errors when training those models. The reason why we implement cost-sensitive learning is because it is

https://doi.org/10.24912/ijaste.v1.i1.180-188

closely related to the problem of imbalanced data classification and skewed class distribution. Although for some models, there are no cost-sensitive options in the parameters. Those models are CNB, XGB, and GB. Therefore, for this process, the mentioned models are skipped. The heart disease dataset is divided into a 70% training set and a 30% testing set. The training set is utilized to train the models, whereas the testing set is utilized to assess the models. A 10-fold cross-validation technique is also used to validate the classifier's training phase.

RESULTS AND DISCUSSION

Before training the models, features should be checked first to see if there are some features that are more important than the others. This process will be done using a Chi-squared test. The results of this test is showed in Figure 4. After seeing this result, it was decided that the top 11 features will be used to train the models.

Different kinds of models and solutions are used in this study. The first one is using normal data with models that are tuned with GridSearchCV. The results for this are shown in Table 2. Most of these results indicate that the models are mostly predicted negative since they have low sensitivity and high specificity.

TABLE 2. Model performance

	LR	SVM	CNB	RF	XGB	GB
Sensitivity	0.292	0.271	0.662	0.283	0.307	0.305
Specificity	0.987	0.988	0.841	0.983	0.986	0.987

After seeing the results of the first models, cost-sensitive learning are used to train some of the models instead to see whether it will improve. The results are shown in Table 3. There is a significant improvement on sensitivity, meaning that the model did predict more positive cases than before. This improvement is great because the models is finally going the right way with the way it is predicting cases. Instead of predicting most cases as negative like before, it is now predicting a significant portion of the test set as positive cases.

TABLE 3. Cost-sensitive model performance

	LR	SVM	RF
Sensitivity	0.776	0.768	0.615
Specificity	0.821	0.827	0.856

Another way to improve the model is to oversample the data itself. Since the data is imbalanced, oversampling is more relevant than usual. The results from the oversampled data in Table 4 shows that there is a little improvement on sensitivity although at the expense of the specificity. It seems like at this point, it is the matter of balancing the sensitivity and specificity.

TABLE 4. Model performance with oversampled data

	LR	SVM	CNB	RF	XGB	GB
Sensitivity	0.821	0.823	0.734	0.805	0.837	0.839
Specificity	0.783	0.78	0.769	0.778	0.766	0.764

https://doi.org/10.24912/ijaste.v1.i1.180-188

After a lot of improvements that were done, the highest performing models are the one that was trained with oversampled data. With the highest sensitivity of 0.839 achieved by Gradient Boost model while not having much difference in specificity with the other models, it can be said that Gradient Boost performed the best during this study.

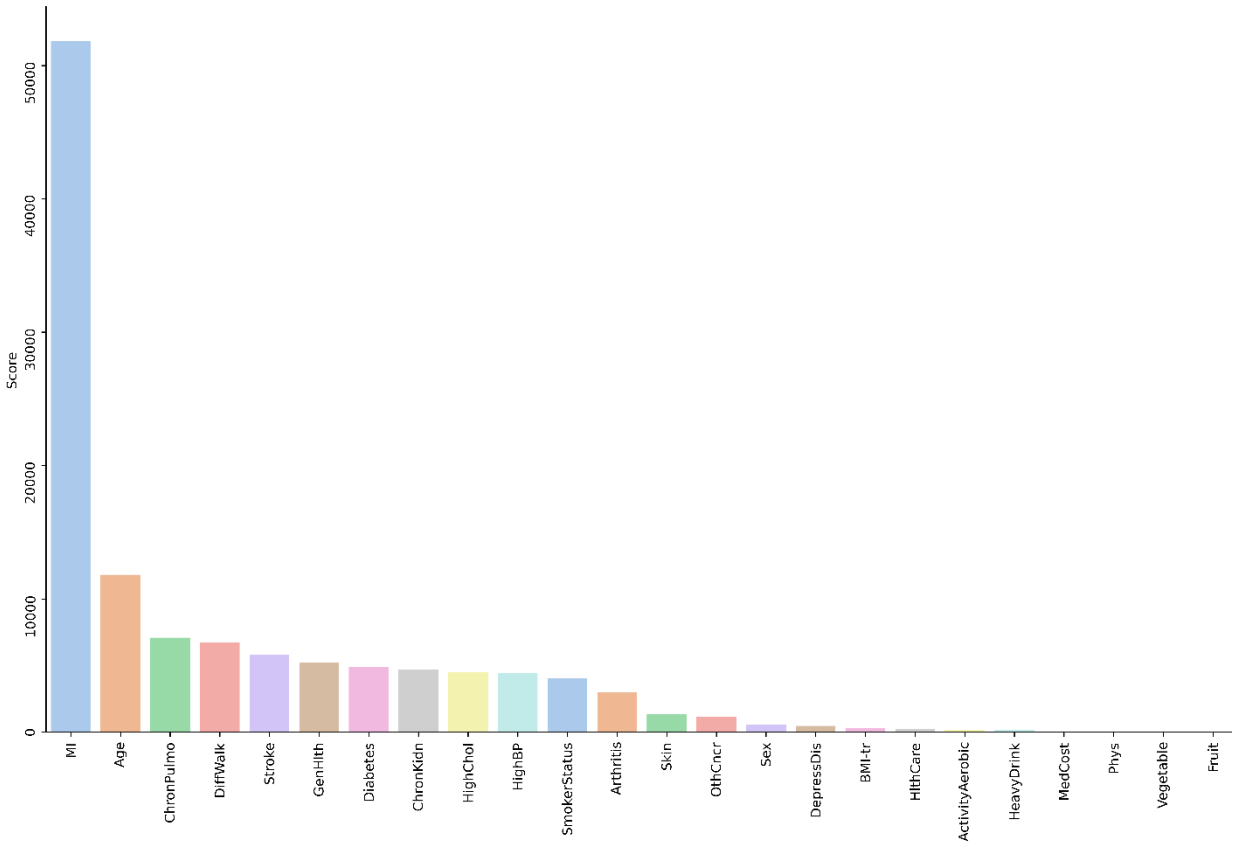

FIGURE 4. Results of the Chi-Squared

CONCLUSION AND RECOMMENDATIONS

The primary goal of this study is to use machine learning algorithms to identify the most essential features of adult coronary heart disease. This analysis was performed using dataset from the Behavioral Risk Factor Surveillance Sys-tem, which was the result of a survey conducted by CDC. However, since the dataset was imbalanced, the experimentswere performed with feature selection. Chi-square was employed as a feature selection algorithm to find attributes that relevant to the target (CHD). The analysis was conducted on six algorithms: Logistic Regression, Complement Naïve Bayes, Support Vector Machine, and three types of ensemble techniques: Random Forest, boosting algorithms(XGBoost and Gradient Boost).

For the evaluation, different kinds of solutions were used in this study, the first one, using normal data with models that are tuned with GridSearchCV. The results of these models are mostly predicted to be negative since they have low sensitivity and high specificity. After seeing the results of the first solution, cost sensitive learning was used to train some models rather than to see if it would improve. As a result, the prediction accuracy has significantly improved the models in terms of sensitivity, meaning that the models successfully predict more positive cases than before. Finally, we used oversampling as another way to improve the models. Since the data is imbalanced, oversampling is very suitable for evaluating models. The results from the oversampled data show that there is a slight increase in sensitivity.

https://doi.org/10.24912/ijaste.v1.i1.180-188

After a lot of enhancements that were done, finally we found Random Forest performed better results with 0.839 sensitivity compared to all other applied machine learning algorithms in this study.

Henceforward, different feature extraction techniques can be performed (other than Chi-square) to select the most relevant subset features to develop models. Furthermore, the experimental findings also suggest that applying Neural Network and real-time medical datasets gathered could be performed for model development. This could enhance the performance with improved accuracy for coronary heart disease prediction.

REFERENCES

1. World Health Organization, "Cardiovascular diseases (cvds)," 2021. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/ cardiovascular-diseases-(cvds)
2. Y. Zhao, E. Wood, N. Mirin, R. Vedanthan, S. Cook, and R. Chunara, "Machine learning for integrating social determinants in cardiovascular disease prediction models: A systematic review," September 2020.
3. D. Chicco and G. Jurman, "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone," *BMC Medical Informatics and Decision Making*, vol. 20, February 2020. [Online]. Available: https://doi.org/10.1186/s12911-020-1023-5
4. M. Khan and M. R. Mondal, "Data-driven diagnosis of heart disease," *International Journal of Computer Applications*, vol. 176, pp. 46–54, July 2020. [Online]. Available: 10.5120/ijca2020920549
5. N. Thai-Nghe, Z. Gantner, and L. Schmidt-Thieme, "Cost-sensitive learning methods for imbalanced data," June 2010, pp. 1–8.
6. R. T. Suriya Begum, Farooq Ahmed Siddique, "A study for predicting heart disease using machine learning," *Turkish Journal of Computer and Mathematics Education*, vol. 12, pp. 4584–4592, April 2021.
7. M. Zeng, B. Zou, F. Wei, X. Liu, and L. Wang, "Effective prediction of three common diseases by combining smote with tomek links technique for imbalanced medical data," pp. 225–228, May 2016.
8. W. Zhu, N. Zeng, and N. Wang, "Sensitivity, specificity, accuracy, associated confidence interval and roc analysis with practical sas ® imple- mentations," *NorthEast SAS users group, health care and life sciences*, January 2010.
9. X.-Y. Gao, A. Ali, H. Shaban, and E. Anwar, "Improving the accuracy for analyzing heart diseases prediction based on the ensemble method," *Complexity*, vol. 2021, pp. 1–10, February 2021.
10. Https://www.cdc.gov/brfss/annual_data/annual_2019.html.
11. S. Sperandei, "Understanding logistic regression analysis," *Biochemia medica*, vol. 24, pp. 12–8, February 2014.
12. D. Srivastava and L. Bhambhu, "Data classification using support vector machine," *Journal of Theoretical and Applied Information Technology*, vol. 12, pp. 1–7, February 2010.
13. B. Seref and G. E. Bostanci, "Performance comparison of naïve bayes and complement naïve bayes algorithms," pp. 131–138, April 2019.
14. E. P. R. Manpreet Kaur, Er. Shailja, "An optimized approach for prediction of heart diseases using gradient boosting classifier," *International Journal of Application or Innovation in Engineering Management*, vol. 9, pp. 130–136, August 2020.
15. A. Hanga, M. Alalyani, I. Hussain, Musa, and M. Almutheibi, "Brief review on sensitivity, specificity and predictivities," *IOSR Journal of Dental and Medical Sciences*, vol. 14, June 2015.
16. S. Arunachalam, "Cardiovascular disease prediction model using machine learning algorithms," *International Journal for Research in Applied Science and Engineering Technology*, vol. 8, pp. 1006–1019, July 2020.

https://doi.org/10.24912/ijaste.v1.i1.180-188