Proceedings of the 3rd International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/6/20230959

Heart disease prediction based on machine learning algorithms

Mingyuan Xu

National University of Singapore, Singapore, 119077, Singapore

e0950171@u.nus.edu

Abstract. Heart disease is a medical research field in which the outcome can benefit lots of people. Because there are several factors that might raise the risk of heart disease, it is useful to build a prediction model to assist people in assessing their health. This paper makes use of a Kaggle dataset that was derived from CDC (Centers for Disease Control and Prevention). First, 8 components are analyzed using diagrams, and then the dataset is used to train classifiers in machine learning models. This paper conducts a comparative study between different algorithms, including Decision Tree, Logistic Regression, SVM (Support Vector Machine), and Random Forest. Besides, the factors taken into consideration while evaluating performance include accuracy, precision, recall, and f1-score. As a result, the maximum accuracy is reached by SVM with a linear kernel, and logistic regression achieves the highest precision. In addition, the highest recall and f1-score are obtained from the model SVM with an RBF kernel.

Keywords: Heart Disease Prediction, Decision Tree, Logistic Regression, SVM (Support Vector Machine), Random Forest

1. Introduction

According to Rajdhan et al. [1], because the human heart regulates blood flow throughout the body, it plays a significant role in the human body. However, one of the main causes of the majority of fatalities is heart disease, and it is likely to contribute to some complications, such as myocardial infarction. Therefore, it is necessary to understand the characteristics of heart disease and take proactive measures to prevent the disease. Moreover, making an early effort to diagnose heart disease is important since it can lower the likelihood of disastrous outcomes.

This paper concentrates on the analysis of different feature variables of heart disease and applies machine learning algorithms to construct classification models in order to help people predict whether they have the tendency to suffer from this disease. On the one side, this may encourage people to practice healthier lifestyles in an effort to ward off the disease. On the other side, people are able to identify their health status early and obtain suitable medical aid.

In detail, this paper is divided into three sections. The literature review section introduces some relative work. In the second section, material and methods, the dataset of this project is demonstrated and the principles of the applied machine learning algorithms are illustrated. In addition, the parameters used in comparing different algorithms are displayed. The final section focuses on the outcomes of this project, including the exploratory data analysis and machine learning results.

DOI: 10.54254/2755-2721/6/20230959

2. Literature review

In order to identify models for the prediction of heart disease, Rajdhan et al. [1] utilized 4 different machine learning algorithms: decision tree, logistic regression, random forest, and naïve bayes. By computing metrics such as precision, recall, f1-score, and accuracy, these models are assessed. Additionally, random forest had the greatest f1-score (0.909), the highest precision (0.937), and the best accuracy (90.16%). The random forest classifier, according to the study, is the best effective algorithm for detecting heart problems overall.

In their attempt to predict heart diseases, Jagtap et al. [2] used SVM in addition to naïve bayes and logistic regression. After pre-processing and cleaning the dataset, they separated the data according to the ratio 7.5:2.5, and the accuracy of the SVM algorithm was shown to be the best at 64.4%. As a consequence, they developed a website to provide locals a heart health report and help them acquire a predictive analysis of heart problems.

Nagaraj et al. [3] chose to use naïve bayes classification and SVM, but Mean Absolute Error, Sum of Squared Error, and Root Mean Squared Error are applied to compare the effectiveness of various methods. In summary, SVM with radial kernel provides classification accuracy that is superior to naïve bayes.

3. Material and methods

3.1. Data introduction

The dataset used in this project is from CDC, and the original dataset is a part of Behavioral Risk Factor Surveillance System (BRFSS). Every year, this system conducts telephone surveys to gather data on health-related issues. The dataset that was actually used in this project is a cleaned version of the original data, with just 18 instead of the 300 columns in the original data. The number of factors contained in this dataset is 17, and the additional column is the heart disease indicator. A value of 0 denotes the absence of cardiac disease in the patient, whereas a value of 1 denotes the presence of this condition. Additionally, the number of observations is 319795. However, compared to other classes, the number of heart disease patients is out of proportion. Undersampling will be used to provide results that are more reliable. Each of the 17 features in the dataset is explained in detail in Table 1, and the data of these 17 factors will be fed into different models to complete the prediction of heart disease.

3.2. Machine learning classifiers

3.2.1. Decision trees. The first method applied in this project is the decision tree. Using decision rules, the aim of this algorithm is to extract data from a large number of accessible datasets, and the data can be adequately stored and categorised [4]. Additionally, decision tree has the form of a flowchart in which the dataset is signified by the inner node and the outer branches represent the outcome [1].

3.2.2. Logistic regression. The second model used in this research, a part of the regression analysis, is logistic regression. There are three forms of regression analysis, and linear regression is one of them [4]. In sociology, neural networks are associated with logistic regression [4]. Rajdhan et al. [1] regard logistic regression as a technique for restricting the outcome of a linear equation to the range of 0 to 1 instead of a way to fit a straight line or a hyperplane. Furthermore, according to Chang et al. [4], this approach, which is a controlled learning algorithm, yields only binary results (0 or 1, yes or no). The probability will be rounded to 0 if it is less than 0.5, and to 1 if it is more than 0.5.

3.2.3. Support vector machine (SVM). According to Chang et al. [4], SVM is a machine learning technique for both linear and non-linear data and may be used in a variety of domains, including bioinformatics, image pattern identification, and target recognition. The input data is mapped using SVM to a higher dimension where linear separation is practical [5]. The principle of this algorithm is finding an optimal linear hyperplane, which is also called a decision boundary to divide samples into

DOI: 10.54254/2755-2721/6/20230959

two categories by using support vectors [5]. Its advantage is the efficiency in controlling prediction errors [6], whereas the disadvantage is taking a long time in training data [7]. In addition, according to Chen et al. [5], SVM models may use kernels to handle nonlinear data, and a feature space consisting of observations of original samples can be constructed. Then the classification can be processed in this feature space. Some common choices of the kernel are the linear kernel, polynomial kernel, Gaussian radial basis function (RBF), and Sigmoid kernel [5].

Table 1. Introduction of attributes from the dataset.

Attribute Description	Distinct Values
BMI–Body Mass Index	Multiple values between 12.02 and 94.85
Smoking–whether patients have smoked 100 cigarettes in their entire life	0:No, 1:Yes
Alcohol Drinking–whether patients are heavy drinkers	0:No, 1:Yes
Stroke–whether patients have stroke	0:No, 1:Yes
Physical Health–number of days during the past 30 days that physical health is not good	Multiple values between 0 and 30
Mental Health–number of days during the past 30 days that mental health is not good	Multiple values between 0 and 30
Diff Walking–whether patients have serious difficulty in walking or climbing stairs	0:No, 1:Yes
Sex–whether patients are male or female	0:Female, 1:Male
Age Category–3 levels of age	0:18-34, 1:35-59, 2:60 or older
Race–race of the patients	0:white, 1:black, 2:Asian, 3:American/Indian/Alaskan Native, 4:others, 5:Hispanic
Diabetic–whether patients have diabetes	0:No, 1:Yes
Physical Activity–whether patients do physical activity or exercise during the past 30 days other than their regular job	0:No, 1:Yes
Gen Health–level of general health	0:Excellent, 1:Very Good, 2:Good, 3:Fair, 4:Poor
Sleep Time–hours of sleep in a 24-hour period on average	Multiple values between 1 and 24
Asthma–whether patients have asthma	0:No, 1:Yes
Kidney Disease–whether patients have kidney diseases (not including kidney stones, bladder infection or incontinence)	0:No, 1:Yes
Skin Cancer–whether patients have skin cancer	0:No, 1:Yes

DOI: 10.54254/2755-2721/6/20230959

3.2.4. Random forest. Random Forest is the last algorithms applied in this project to build the prediction model. According to Hamrani et al. [8], random forest consists of decision trees, which are based on Bootstrap and Bagging techniques. A decision tree strategy may be used to classify a collection of data, whereas the random forest approach uses several decision trees to produce a forest that performs better than a single decision tree. [9]. Ramalingam [10] considered random forest as an ensemble of decision trees. The following is the general equation representing the random forest algorithm, where x denotes the input variable of the vector, B is the number of decision trees, and $f_i(x)$ stands for a single decision tree [6].

$$\hat{f}(x) = \frac{1}{B}\sum_{i=1}^{B} f_i(x) \tag{1}$$

3.3. Performance measures

In this paper, several parameters are applied to evaluate the performance of different machine learning algorithms, including accuracy, hamming loss, precision, recall, f1-score, and the detailed information. First, four abbreviations will be introduced, TP, TN, FP, and FN, and they are the entries of the confusion matrix. TP (true positive) is for the number of samples for which the model properly recognized the positive category, and TN (true negative) stands for the number of samples for which the model correctly predicted the negative category. Similarly, the number of samples for which the model mistakenly predicted the positive category is denoted by FP (false negative), while the number of samples for which the model incorrectly forecasted the negative category is denoted by FN (false negative). Figure 1 is a schematic of the confusion matrix.

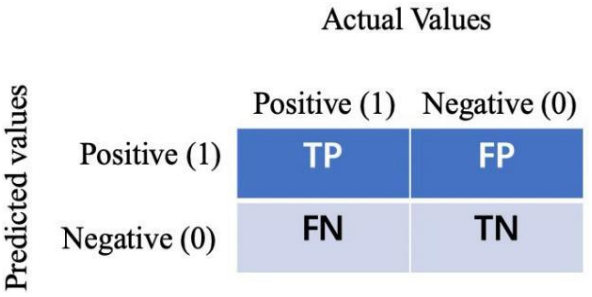

Figure 1. The confusion matrix.

Following the introduction of these four concepts, the performance metrics listed above may be established using the confusion matrix. First, accuracy denotes the proportion of the samples that the model predicts correctly, and hamming loss represents the proportion of samples that the model predicts incorrectly. According to Chang et al. [4], hamming loss is defined as the hamming distance between 'actual' and 'predictions' in multi-class classification, and the lower the loss, the better the model performance.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{2}$$

$$\text{Hamming loss} = 1 - \text{Accuracy} = \frac{\text{FP} + \text{FN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{3}$$

Second, precision indicates the proportion of the samples which are predicted to be positive correctly among the samples predicted to be positive. It is significant to improve the precision, since it can assist in cutting down the amount of money wasted on healthcare when detecting diseases [4].

DOI: 10.54254/2755-2721/6/20230959

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4}$$

Third, recall is calculated in order to reflect the proportion of positive samples that were predicted correctly, which is similar to precision. However, the denominator of recall is the total number of positives, including TP and FN.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5}$$

Finally, F-measure:

$$\text{f1} - \text{score} = \frac{2\text{TP}}{2\text{TP} + \text{FN} + \text{FP}} = \frac{2\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6}$$

In summary, precision assesses how well the model classifies negative samples, and recall evaluates the model in the performance of identifying positive samples. F-measure denotes the harmonic mean between them, precision and recall [5].

4. Results

4.1. Exploratory data analysis

There are 17 factors listed in Table 1, and 8 factors are selected for detailed analysis in this section. General health is measured based on 5 levels, namely, poor, fair, good, very good, and excellent, and they are ranked from 0 to 4. In Figure 2, class 0 represents no heart disease, while class 1 represents patients suffering from heart disease. From Figure 2, it can be determined that for class 0, the majority of patients are mainly dispersed in the first three groups, which signifies good health conditions. In contrast, the density of class 1 is higher in the final three categories, suggesting that people with cardiac disease are not in good condition. Therefore, patients who are not suffering from heart disease are supposed to improve their general health since the risk of developing heart disease may be higher for people in poor health conditions.

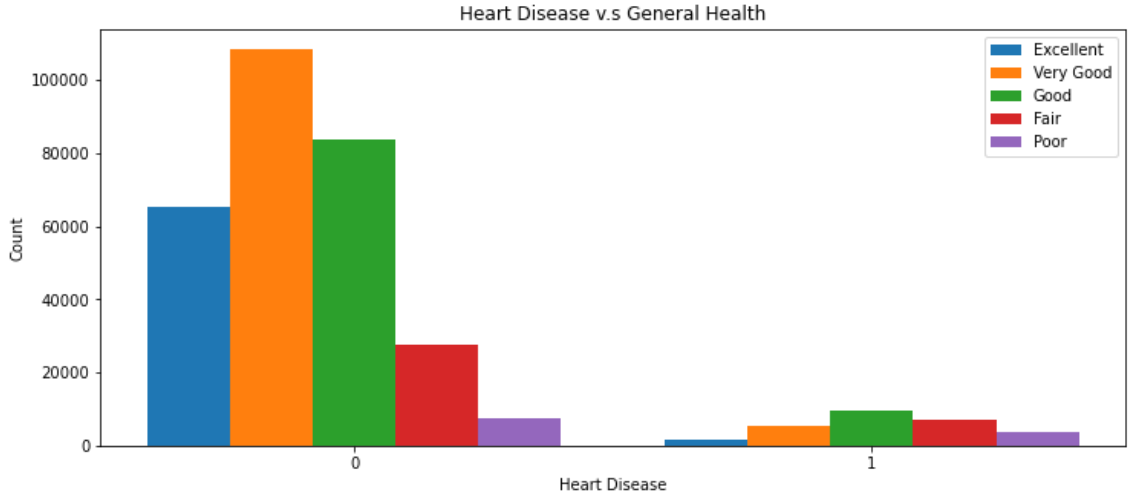

Figure 2. General health comparison between people with and without heart disease.

Figure 3 illustrates how patients with heart disease experience significantly more physical health issues than people without heart disease.

DOI: 10.54254/2755-2721/6/20230959

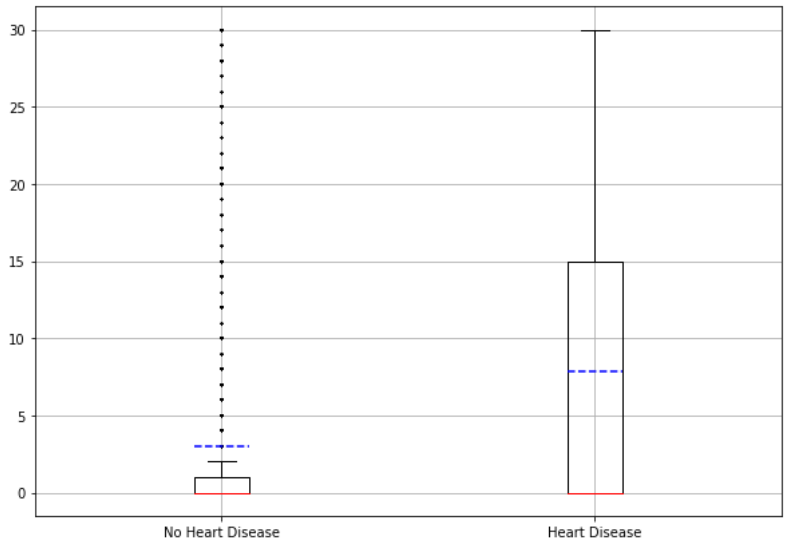

Figure 3. Physical health comparison between people with and without heart disease.

As for the age, the original data divides age into 14 levels, and the number of categories is reduced to 3 in this project. Some of the categories are merged into one level, and the 3 levels are 18-34, 35-59, and 60 or older. In Figure 4, it can be concluded that, for people who suffer from heart disease, the proportion of people who is 60 or older is much larger than that of the healthy group. Therefore, older people are supposed to pay greater attention to their health since they are at an increased risk of developing heart disease.

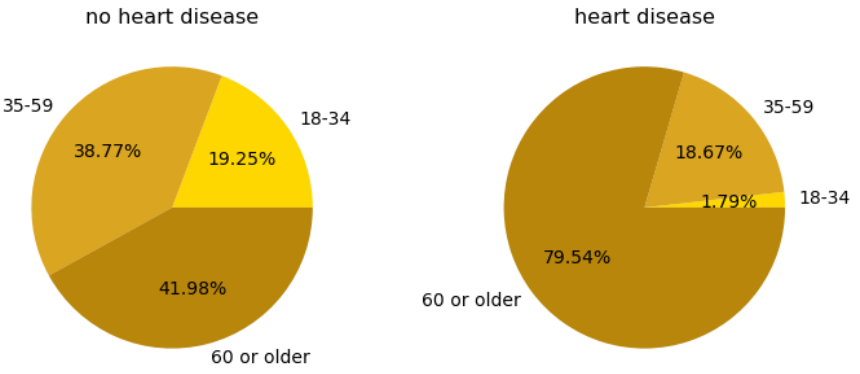

Figure 4. Age comparison between people with and without heart disease.

From Figure 5, it can be concluded that most of the patients do not smoke, and a majority of them do not have a stroke, diabetes, kidney disease, or difficulty in walking. However, as seen in Figure 6, people with heart disease have higher percentages of smoking, stroke, walking problems, diabetes, and renal disease than patients without heart disease. Even though some members of the latter group may have similar problems, the proportion is substantially lower than that of patients who have already been diagnosed with heart disease. Therefore, people who have these characteristics need to take good care of their health, and have timely tests for heart disease.

DOI: 10.54254/2755-2721/6/20230959

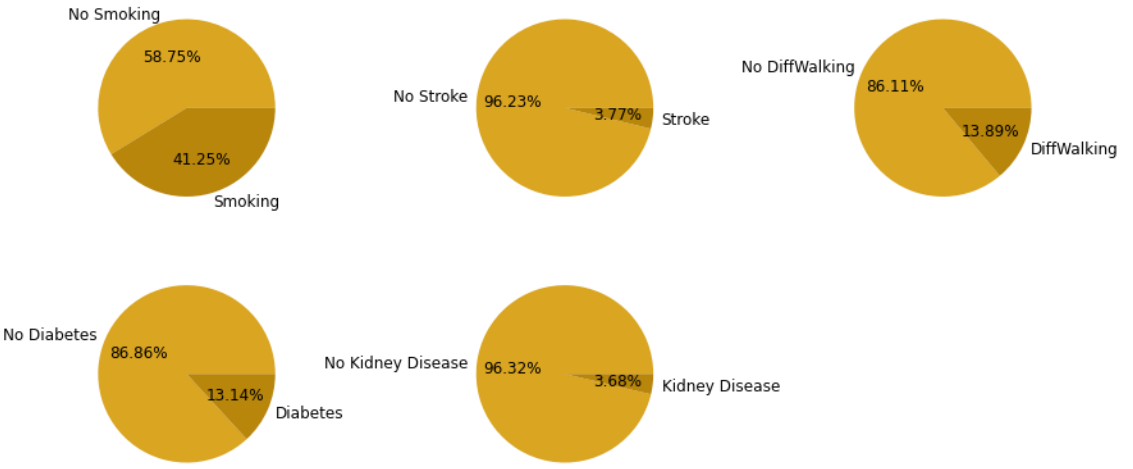

Figure 5. Total percentage on the five factors.

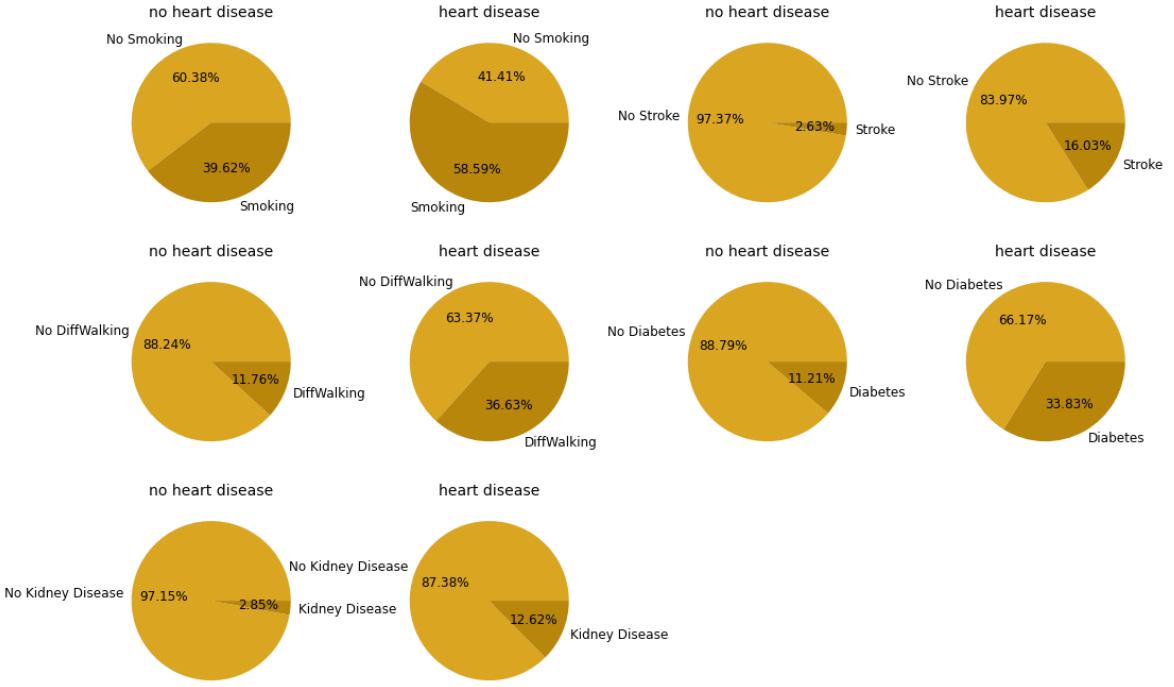

Figure 6. Separate-group comparisons between people with and without heart disease.

4.2. Machine learning results

As mentioned in the data introduction, this dataset is unbalanced. The percentage of patients who do not suffer from heart disease is 91.44%. If this dataset is fed into machine learning models directly, the result will be inaccurate. Therefore, undersampling is used to balance the dataset, and it can lead to equal densities between two target groups.

First, 80% of the training dataset and 20% of the test dataset are separated from the input dataset. The training dataset is used to train models, while the test dataset is used to assess their performance. However, here is the unbalanced dataset being split, so no matter the training dataset or the test dataset, it is still not balanced. The new training set now has 43,562 observations and the new test dataset now contains 11,184 elements after undersampling was applied to these two sets of data. After feeding the processed dataset into different machine learning models, the values of parameters used to evaluate the

DOI: 10.54254/2755-2721/6/20230959

performances of classification can be obtained. The models applied include Decision Tree, Logistic Regression, SVM with three different kernels (linear kernel, polynomial kernel, and RBF), and Random Forest. The performance metrics are displayed in Table 2.

Table 2. Performance metrics.

Algorithm	Accuracy	Precision	Recall	f1-score	hamming loss
Decision Tree	0.65424	0.65690	0.64574	0.65128	0.34576
Logistic Regression	0.74437	**0.74037**	0.75268	0.74648	0.25563
SVM(Linear)	**0.74571**	0.73600	0.76627	0.75083	**0.25429**
SVM(Polynomial)	0.74338	0.73506	0.76109	0.74785	0.25662
SVM(RBF)	0.74365	0.71895	**0.80007**	**0.75734**	0.25635
Random Forest	0.71772	0.70639	0.74517	0.72526	0.28228

According to Table 2, the SVM with a linear kernel has the maximum accuracy, and the greatest value in terms of precision is held by the logistic regression model. In addition, when values of recall and f1-score are contrasted, the SVM (RBF) performs better overall. Furthermore, as seen in Figure 7, SVM (RBF) is more effective at predicting patients without heart disease than it is at predicting patients with this condition since two values of precision 0.77 are higher than 0.72. Contrarily, because two values of recall, 0.69 is less than 0.80, this model is more accurate when diagnosing patients with heart disease from positive samples than when deciding on patients without heart disease.

```
Classification report:

              precision    recall  f1-score   support

           0       0.77      0.69      0.73      5592
           1       0.72      0.80      0.76      5592

    accuracy                           0.74     11184
   macro avg       0.75      0.74      0.74     11184
weighted avg       0.75      0.74      0.74     11184
```

Figure 7. SVM with RBF kernel.

5. Conclusion

In this paper, the dataset applied is derived from CDC, and it has been cleaned into 17 factors from the original dataset. This study aims to develop a model based on these 17 characteristics that might be used to forecast heart disease for residents. 4 models are applied, and they are decision tree, logistic regression, SVM, and random forest, in which the SVM is applied by using 3 different kernel functions, linear, polynomial, and RBF. In terms of comparison, logistic regression has the best precision, 0.74037 while SVM with a linear kernel has the highest accuracy, 0.74571. Additionally, for both the recall and f1-score, SVM with the RBF kernel holds the highest values, they are 0.80007 and 0.75734 separately.

At the same time, useful information can be obtained from the data directly. This paper focuses on 8 factors, age, general health, physical health, smoke, stroke, difficulty in walking, diabetes, and kidney disease. Healthy people are supposed to care about their general health, since patients in poor health may be more possible to develop heart disease. The same condition is true for old people, and they need to pay more attention to their health conditions. Finally, for the rest of the 5 factors, the differences in the proportion between the healthy group and patients with heart disease are obvious.

However, there are still certain restrictions in this paper. The magnitude of the data is one of the obvious limitations. The size of the unprocessed data is around 300,000, but the size of the dataset after undersampling is reduced to 43,562. The issue of an imbalanced dataset is addressed by undersampling, and in the future, more efficient methods can be conducted to avoid the problem of the small size of data,

DOI: 10.54254/2755-2721/6/20230959

such as SMOTE. The accuracy of these models is below 80%, which is another flaw of this research. To increase the accuracy value in the future, additional pertinent datasets can be employed.

References

[1] Rajdhan, A., Agarwal, A., Sai, M., Ravi, D., Ghuli, P. Heart disease prediction using machine learning. International Journal of Research and Technology 9(04), 659-662 (2020).

[2] Jagtap, A., Malewadkar P, Baswat, O., Rambade H. Heart disease prediction using machine learning. International Journal of Research in Engineering. Science and Management 2(2), 352-355 (2019).

[3] Nagaraj, M. L., Chethan, C., Basavaraj, S. P. Prediction of heart disease using machine learning. International Journal of Recent Technology and Engineering 8(2), 474-477 (2019).

[4] Chang, V., Ganatra, M. A., Hall, K., Golightly, L., Xu, Q. W. A. An assessment of machine learning models and algorithms for early prediction and diagnosis of diabetes using health indicators. Healthcare Analytics 2, 100118 (2022).

[5] Chen, T. J. Simone A Ludwig, et al. A performance analysis of dimensionality reduction algorithms in machine learning models for cancer prediction. Healthcare Analytics 100125 (2022).

[6] Javanmard, M. E., Ghaderi, S. F., Hoseinzadeh, M. Data mining with 12 machine learning algorithms for predict costs and carbon dioxide emission in integrated energy-water optimization model in buildings. Energy Conversion and Management 238, 114153 (2021).

[7] Han J. W., Pei, J. Tong, H. H. Data mining: concepts and techniques. Morgan kaufmann (2022).

[8] Hamrani, A. Akbarzadeh, A., Madramootoo, C. A. Machine learning for predicting greenhouse gas emissions from agricultural soils. Science of The Total Environment 741, 140338 (2020).

[9] Wang, Z. Y., Wang, Y. R., Zeng, R., Srinivasan R. S., Ahrentzen, S. Random forest based hourly building energy prediction. Energy and Buildings 171, 11-25 (2018).

[10] Ramalingam, V. V., Dandapath, A., Raja, M. K. Heart disease prediction using machine learning techniques: a survey. International Journal of Engineering & Technology 7(2.8), 684-687 (2018).