Heart Disease Prediction using Ensemble Model

Conference Paper · October 2021

CITATIONS
3

READS
1,600

2 authors, including:

Bikal Adhikari
Pulchowk Campus
4 PUBLICATIONS **8** CITATIONS

SEE PROFILE

Heart Disease Prediction using Ensemble Model

[*]Bikal Adhikari[1,2], [*]Subarna Shakya[3]
[1]Post Graduate Department of Information System Engineering, HIST College, Nepal
[2]Dept. of Computer & Electronics and Communication Engineering, Kathford Int'l College of Engineering & Management, Nepal
[1*]bkl.adh@gmail.com
[2*]bikaladhikari@kathford.edu.np
[3]Department of Electronics & Computer Engineering, Pulchowk Campus, Nepal
[3*]drss@ioe.edu.np

* Corresponding Author

Abstract. Heart is one of the vital organs of the human body. The mortality rate due to cardiovascular or heart disease is in the rising trend. From the review of existing literatures, timely diagnosis of the potential heart disease is the key to prevent the risk of early death. But due to the lack of proper health infrastructure in different parts of the country (specially in the rural area), timely diagnosis or the early diagnosis of the heart disease has been a problem in Nepal. In this thesis, ensemble learning is used to predict the potential heart risk in patient. First different supervised learning models such as Logistic Regression, SVM, Decision Tree, KNN and Gaussian Naive Bayes were built out of UCI heart disease dataset. These models gave accuracy of 82.46 %, 87.34 %, 97.67 %, 89.94 % and 78.57 % respectively. An ensemble model was built by combining aforementioned five supervised ML models. The voting based ensemble model gave an accuracy of 96.10 % while averaging based ensemble model gave an accuracy of 96.43 %. Upon the comparison, the ensemble model gave overall better accuracy. A module capable of predicting heart disease for a patient was implemented.

Keywords: *Heart disease, machine learning, ensemble learning, prediction*

1 Introduction

Heart is one of the vital organs of the human body. Functions of the heart includes pumping oxygenated blood to different part of the body, pumping hormones and other vital substances, receiving deoxygenated blood, carrying metabolic waste products from the body and pumping it to the lungs for oxygenation and maintaining the blood pressure. Healthy heart is the mandatory requirement for living a healthy lifestyle.

Heart disease is the major cause of morbidity and mortality globally: it accounts for more deaths annually than any other cause. According to the WHO, an estimated 17.9 million people died from heart disease in 2016, representing 31% of all global deaths [**1**]. Over three quarters of these deaths took place in low- and middle-income countries. Of all heart diseases, coronary heart disease (aka heart attack) is by far the most common and the most fatal. In the United States, for example, it is estimated that someone has a heart attack every 40 seconds and about 805,000 Americans have a heart attack every year [**2**]. In case of Nepal, the cardiovascular disease mortality rate increased from 124.1 per 100,000 populations in 1990 AD to 164.7 per 100,000 populations in 2017. In 2017, the CVD mortality rate in male population was estimated to be 230.7 and that among female population was estimated to be 104.3 [**3**] [**4**].

This shows that the mortality rate due to cardiovascular or heart disease is in the rising trend. From the review of existing literatures, timely diagnosis of the potential heart disease is the key to prevent the risk of premature death. But due to the lack of proper health infrastructure in different parts of the country (specially in the rural area), timely diagnosis or the early diagnosis of the heart disease has been a problem in Nepal.

In this context, we can make a hypothesis that a Machine Learning model trained with the dataset of the heart patients can be used to make the prediction of the potential heart risk in the patients from the area where there is lack of proper health infrastructure.

In this research, five different supervised machine learning models (Logistic Regression, Support Vector, Decision Tree, K Nearest Neighbor, and Gaussian Naive Bayes) were built using UCI heart disease dataset. The models are then used to build the ensemble model using voting and averaging method. The ensemble model thus built can used to predict the chance of heart disease for a patient.

2 Review of Relevant Works

In a research by [**5**], the authors have used the ensemble model for the prediction of the Parkinson's disease. Different performance measures of machine learning such as accuracy, log loss, F1 score etc have been used to evaluate the models. The authors found that the ensemble model based on stacking and voting gave the better performance.

In a research by [**6**], the authors used the concept of machine learning and deep learning for the prediction of heart disease. Their model was based on UCI heart disease dataset with 14 attributes. The authors reported that they achieved 94.2 % accuracy using the approach of deep learning.

In a research by [**7**], the authors implemented Naïve Bayes, Decision Tree, KNN and Random Forest algorithm based on the dataset provided by UCI heart disease. The authors reported that the highest accuracy was achieved using KNN algorithm.

In a research by [**8**], the authors implemented several supervised machine algorithms such as Decision Tree, KNN, SVM, and Random Forest. They reported the accuracy of 79 %, 87 %, 83 % and 84 % respectively. Similarly, the authors reported AUC value of 71.6 %, 88.5 %, 90.4 and 90.8 % respectively.

In a research by [**9**], the authors found that different classification algorithms gave high accuracy for the prediction of heart disease. The authors also mentioned that for the prediction of heart disease, we can use limited number of attributes instead of considering all the attributes in the dataset.

2 Supervised Machine Learning Models

Machine Learning is a branch of Artificial Intelligence which helps to create different models based on the dataset and helps to solve different problems such as classification problem, clustering problem, prediction problem etc. Machine learning helps to analyze the data without writing codes explicitly. In general, machine learning models can be classified as supervised machine learning models and unsupervised machine learning models. Supervised machine learning models are the models which use labeled data [**10**]. In this research, five different supervised machine learning models namely Logistic Regression, Support Vector Machine, Decision Tree, KNN and Gaussian Naïve Bayes have been implemented using Python based on UCI heart disease dataset. These models are then combined to implement the ensemble model [**11**] [**12**] [**13**].

2 Ensemble Model

In ensemble model, diverse predictors/classifiers are used for classification. Different types of ensemble model are voting, averaging, weighted averaging, stacking, blending, bagging, boosting etc. In this research, the voting and averaging ensemble models have been used. Following figure shows the concept of ensemble model using voting approach:

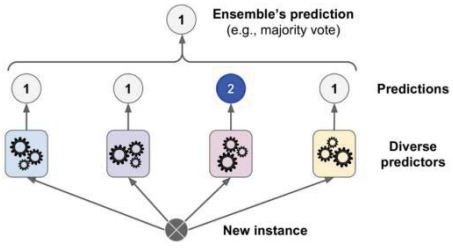

Fig. 1. Ensemble method of prediction using voting technique [**14**]

3 Methodology

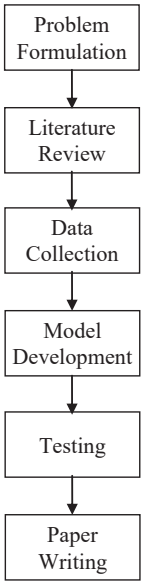

Fig. 2. Flow diagram of methodology adopted

Figure 2 shows the flow diagram of the methodology that was adopted for carrying out this research work. Problem formulation was the first task done. The problem formulation identified the research problems. To justify the problem formulation and to validate the research purpose, detailed study of available literatures on the topic was carried out. In literature review, different articles, books, dissertations etc were studied.

Following literature review, data sets were collected and an ensemble learning model was developed on the basis of training data set. The machine learning modeling process is shown in figure 3 [**15**] [**16**].

After the implementation, the model was tested with the test data set in order to find the accuracy of the model for the given data set.

In this research, first existing classification models such as Logistic Regression, Support Vector Machine, KNN, Decision Tree etc are used. Then ensemble learning models are build upon these models to improvise the accuracy of the model. After the model is ready we can predict the result for individual patient. This is illustrated in following figure 4.

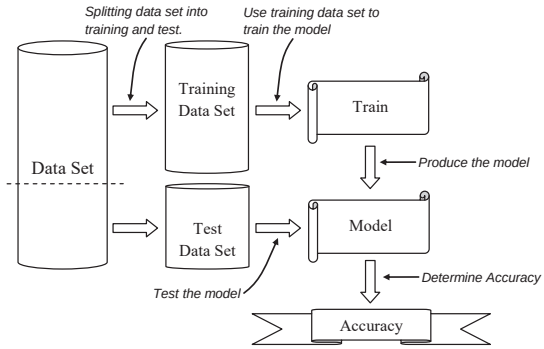

Fig. 3. Machine Learning Process

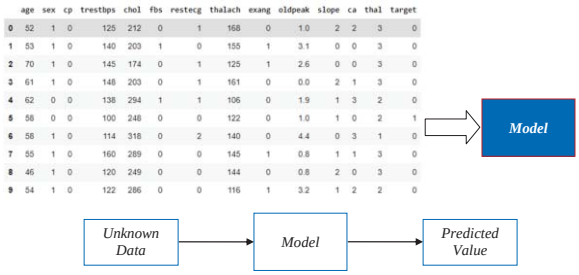

Fig. 4. Use of ML model to predict for unknown data

4 Results, Analysis and Comparison

Following sections presents the results obtained from the application of different classification model to the UCI heart disease dataset:

4.1 Result from Logistic Regression

Following is the result obtained from the Logistic Regression:

	precision	recall	f1-score	support
0	0.85	0.77	0.80	145
1	0.81	0.88	0.84	163
accuracy			0.82	308
macro avg	0.83	0.82	0.82	308
weighted avg	0.83	0.82	0.82	308

Fig. 5. Results Obtained from Logistic Regression

Figure 5 shows the result obtained from the application of logistic regression classification model. For target 0 precision is 0.85, recall rate is 0.77, f1-score is 0.80 and support is 145. Similarly, for target 1, precision is 0.81, recall rate is 0.88, f1 score is 0.84, and support is 163. The accuracy of the model was determined to be 82.46 %.

4.2 Result from Support Vector Machine

Following is the result obtained from the Support Vector Machine (SVM):

	precision	recall	f1-score	support
0	0.86	0.87	0.87	145
1	0.88	0.88	0.88	163
accuracy			0.87	308
macro avg	0.87	0.87	0.87	308
weighted avg	0.87	0.87	0.87	308

Fig. 6. Results Obtained from SVM

Figure 6 shows the result obtained from the application of SVM classification model. For target 0 precision is 0.86, recall rate is 0.87, f1-score is 0.87 and support is 145. Similarly, for target 1, precision is 0.87, recall rate is 0.88, f1 score is 0.88, and support is 163. The accuracy of the model was determined to be 87.34 %.

4.3 Result from Decision Tree

Following is the result obtained from the Decision Tree:

	precision	recall	f1-score	support
0	0.96	0.98	0.97	145
1	0.98	0.96	0.97	163
accuracy			0.97	308
macro avg	0.97	0.97	0.97	308
weighted avg	0.97	0.97	0.97	308

Fig. 7. Results Obtained from Decision Tree

Figure 7 shows the result obtained from the application of Decision Tree based classification model. For target 0 precision is 0.96, recall rate is 0.98, f1-score is 0.97 and support is 145. Similarly, for target 1, precision is 0.98, recall rate is 0.96, f1 score is 0.97, and support is 163. The accuracy of the model was determined to be 97.07 %.

4.4 Result from KNN

Following is the result obtained from the KNN clasffier:

	precision	recall	f1-score	support
0	0.85	0.96	0.90	145
1	0.96	0.85	0.90	163
accuracy			0.90	308
macro avg	0.90	0.90	0.90	308
weighted avg	0.91	0.90	0.90	308

Fig. 8. Results Obtained from KNN classifier

Figure 8 shows the result obtained from the application of KNN classifier. For target 0 precision is 0.85, recall rate is 0.96, f1-score is 0.90and support is 145. Similarly, for target 1, precision is 0.96, recall rate is 0.85, f1 score is 0.90, and support is 163. The accuracy of the model was determined to be 89.94 %.

4.5 Result from Gaussian Naive Bayes

Following is the result obtained from the Naive Bayes classifier:

	precision	recall	f1-score	support
0	0.78	0.76	0.77	145
1	0.79	0.81	0.80	163
accuracy			0.79	308
macro avg	0.79	0.78	0.78	308
weighted avg	0.79	0.79	0.79	308

Fig. 9. Results Obtained from Naive Bayes Classifier

Figure 9 shows the result obtained from the application of Gaussian Naive Bayes classification model. For target 0 precision is 0.78, recall rate is 0.76, f1-score is 0.77 and support is 145. Similarly, for target 1, precision is 0.79, recall rate is 0.81, f1 score is 0.80, and support is 163. The accuracy of the model was determined to be 78.57 %.

4.6 Result from Ensemble Model

Following is the result obtained from the voting based ensemble model:

	precision	recall	f1-score	support
0	0.94	0.98	0.96	145
1	0.98	0.94	0.96	163
accuracy			0.96	308
macro avg	0.96	0.96	0.96	308
weighted avg	0.96	0.96	0.96	308

Fig. 10. Results Obtained from Ensemble Model(Voting based)

Figure 10 shows the result obtained from the application of voting classifier based ensemble model. For target 0 precision is 0.94, recall rate is 0.98, f1-score is 0.96and support is 145. Similarly, for target 1, precision is 0.98, recall rate is 0.94, f1 score is 0.96, and support is 163. The accuracy of the model was determined to be 96.10 %.

Similarly, the result obtained from the average based ensemble model is:

	precision	recall	f1-score	support
0	0.97	0.96	0.96	145
1	0.96	0.97	0.97	163
accuracy			0.96	308
macro avg	0.96	0.96	0.96	308
weighted avg	0.96	0.96	0.96	308

Fig. 11. Results Obtained from Ensemble Model (Average)

Figure 11 shows the result obtained from the application of average classifier based ensemble model. For target 0 precision is 0.97, recall rate is 0.96, f1-score is 0.96and support is 145. Similarly, for target 1, precision is 0.96, recall rate is 0.97, f1 score is 0.97, and support is 163. The accuracy of the model was determined to be 96.43 %.

4.7 Comparison of different models

The comparison between different models used for prediction is given in the figure below:

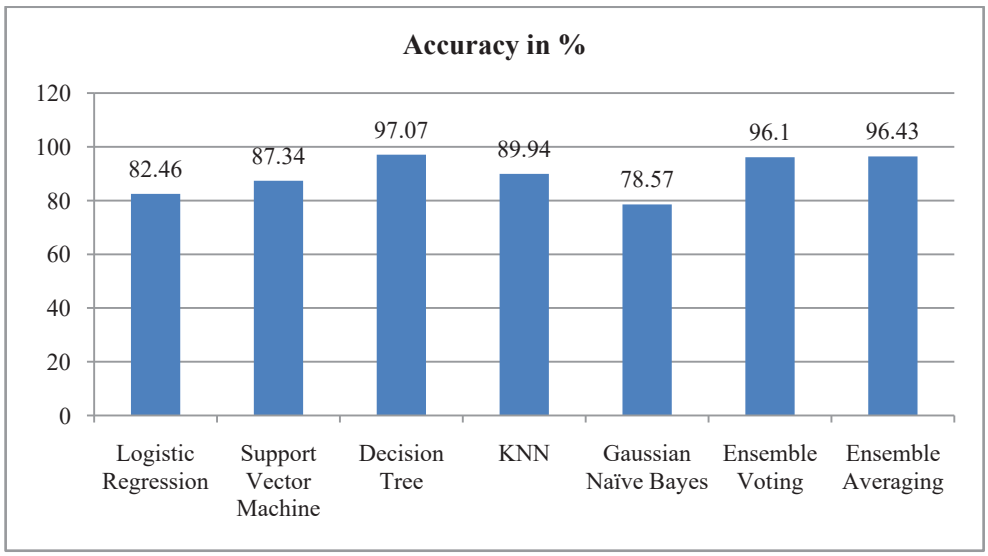

Fig. 12. Comparison between different models used for prediction

Figure 12 shows the comparison between the different prediction models used for the prediction of heart disease. The Decision Tree gave the highest accuracy for the given dataset. However, in totality, the ensemble model gave overall better accuracy. The ensemble models give the better performance in prediction.

4.8 Making prediction for single patient

In this research, a module capable of predicting chances of heart disease using a feature set for single patient has been implemented. Following figures shows the module making prediction for a patient who is fine or not likely to have heart disease:

```
Please Provide the input variable respectively :
52
1
0
125
212
0
1
168
0
1
2
2
3
```

Fig. 13. Taking input the feature set for single patient who is fine

```
The patient is fine!
```

Fig. 14. Output showing that patient is fine

Similarly, following figures shows the module making prediction for a patient who is likely to have heart disease:

```
Please Provide the input variable respectively :
108
0
0
60
148
1
1
100
0
1
1
0
5
```

Fig. 15. Taking input the feature set for single patient likely to have heart disease

```
The patient is likely to have Heart Disease!
```

Fig. 16. Output showing that patient is likely to have heart disease

5 Conclusions

The machine learning model based on existing classifiers such as logistic regression, support vector machine, decision tree, KNN, Gaussian Naive Bayes were built using UCI heart disease dataset. Ensemble model using voting classifier and average classifier were built by using the aforementioned models. The accuracy of the models was determined. Logistic regression yielded an accuracy of 82.46 %, SVM yielded an accuracy of 87.34 %, Decision Tree yielded an accuracy of 97.07 %, KNN yielded an accuracy of 89.94 %, Gaussian Naive Bayes yielded an accuracy of 78.57 %. The ensemble learning model based on voting classifier and probability was developed. Voting classifier based ensemble model yielded an accuracy of 96.10 % and average classifier based ensemble model yielded an accuracy of 96.43 %. Upon the comparison of the models, the ensemble models gave overall better accuracy. A module capable of predicting heart disease using the feature set for single patient was also implemented.

6 Recommendations

Since the ensemble model can help in making the prediction of the potential heart disease, this modeling approach can also be used to predict the chances or likelihood of other diseases as well. These types of models are use useful for the early detection of disease and can be used in the rural sector where there is lack of adequate number of qualified doctors and medical equipments.

References

1. WHO.:Country Cooperation Strategy at a Glance," World Health Organization.(2017)
2. Murphy,S.L.:Mortality in the United States, 2017," Centers for Disease Control and Prevention.(2018)
3. Bhattarai, S.:Cardiovascular Disease Trend in Nepal, IJC Heart and Vasculature.(2017)
4. Anil,O.M.: Prevalence of Cardiovascular Risk Factors in Apparently Healthy Urban Adult Population of Kathmandu," *Journal of Nepal Health Research Council*, vol. 16, no. 41, pp. 438-445.(2019)
5. Sipai, S., Mali, D., Shakya,S., Mali, R.:Parkinson's Disease Data Analysis and Prediction using Ensemble Machine Learning Techniques, *Mobile Computing*

and Sustainable Informatics. Lecture Notes on Data Engineering and Communications Technologies, vol. 68.(2021)

6. Shabaz, M., Dhiman,G., Pande,S., Singh,P., Bharti,R., Khamparia,A.:Prediction of Heart Disease using a Combination of Machine Learning and Deep Learning, *Computational Intelligence and Neuroscience*, vol. 2021.(2021)
7. Bharti,S.K., Shah,D., Patel, S.:Heart Disease Prediction using Machine Learning Techniques, *SN Computer Science*.(2020)
8. Choubey, D.K.:Heart Disease Prediction Using Machine Learning and Data Mining, *International Journal of Recent Technology and Engineering*, vol. 9, pp. 212-219.(2020)
9. Almustafa,K.M.:Prediction of Heart Disease and Classifier's Sensitivity Analysis, *BMC Bioinformatics*.(2020)
10. Theobald,O.: *Machine Learning for Absolute Beginners: A Plain English Introduction*.: Scatterplot Press.(2017)
11. Borges,D.M.: (2021, July) Kdnuggets. [Online]. https://www.kdnuggets.com/2019/01/ensemble-learning-5-main-approaches.html
12. Santarcangelo,J.: Machine Learning With Python: A Practical Introduction, edX course.(2020)
13. Chowdary,D.H.:Towards AI. [Online].(2020) https://towardsai.net/p/programming/decision-trees-explained-with-a-practical-example-fe47872d3b53
14. OGIQ.:OpengenusIQ. [Online]. https://iq.opengenus.org/gaussian-naive-bayes/.(2021)
15. UCI.: UCI Machine Learning Repository. [Online]. https://archive.ics.uci.edu/ml/datasets/heart+disease.(2021)
16. Sayed,S.: An Introduction to Data Science. [Online]. https://www.saedsayad.com/decision_tree.htm.(2021)