Indian Institute of Information Technology, Allahabad

Department of Information Technology

Prediction of heart disease using Ensemble Techniques and Multi Layer Perceptron Neural Network

Under the Guidance of

Prof. O. P. Vyas

By

Hemant Kumar Lader (MIT2019007)

ABSTRACT

- According to the World Health Organization around 17.9 million people die each year due to the cardiovascular heart disease.

- Early prediction of heart disease can prevent many patient deaths and and efficient treatment could be provided at early stages.

- Developing a medical diagnosis system based on machine learning for prediction of heart disease gives more accurate diagnosis results than traditional way.

ABSTRACT...(continued)

- The predictive modeling solution for prediction of heart disease is extremely challenging.

- Some attributes such as age, blood pressure, blood sugar etc. from the UCI cleveland dataset are fed into algorithms which are used to predict the risk of heart attack.

- Many previous work has been done on different data mining techniques, so in this project I will be trying to optimize the result by comparing and combining them using ensemble method and multi layer perceptron

LITERATURE SURVEY

- In [1], the age range was grouped by the K-Nearest Neighbor algorithm into class. The Risk level of each class was identified with the help of ID3 algorithm. The accuracy of prediction was measured by considering different attributes. This gave highest of 80.6% accuracy.

- In [2], authors investigated the heart disease prediction using KStar, J48, SMO, Bayes Net and Multilayer Perceptron through Weka software. Using k-Fold Cross-Validation resampling procedure. The performance of these data mining techniques is measured by combining the results of predictive accuracy, ROC curve

LITERATURE SURVEY...(continued)

- In [3], authors uses algorithms like logistic regression, support vector machine, k-nearest neighbor, Gaussian naïve Bayes, decision tree classifier and random forest classifier. And the prediction accuracy for logistic regression is found to be the highest among all with 88.29% accuracy. They uses 13 attributes of UCI cleveland dataset.

- The prediction method for heart disease using Neural Network has been proposed by Chaitrali S.Dangare.etl [4]. It has mainly three layers, i.e. the input layer, hidden layer and the output layer. The input is given to the input layer and the result is obtained in the output layer.

Problem Statement

To develop a system for predicting heart disease from risk factors using machine learning techniques Ensemble Methods and Multi Layer Perceptron Neural Network. To predict heart disease efficiently and accurately.

Dataset Description

The dataset I am using for this project is “Heart Disease Data Set” from UCI Machine learning repository [5].

- Inputs attributes:
  - Age, Sex, Chest Pain, Resting blood pressure, Serum cholesterol, Fasting blood sugar, Resting electrocardiographic results, Maximum heart rate achieved, Exercise induced angina, ST depression, Slope of the peak exercise ST segment, Number of major vessels colored by fluoroscopy and thal.

- Output Attribute
  - Target : Normal or Heart Problem

Dataset Description...(continued)

1. Age : age in years
2. Sex : (1 = male; 0 = female)
3. Cp : chest pain type
4. Trestbps : resting blood pressure (in mmHg)
5. Chol : serum cholesterol in mg/dl
6. fbs(fasting blood sugar > 120 mg/dl) : (1 = true; 0 = false)
7. Restecg : resting electrocardiographic results
8. Thalach : maximum heart rate achieved
9. Exang : exercise induced angina (1 = yes; 0 = no)
10. Oldpeak : ST depression induced by exercise relative to rest
11. Slope : the slope of the peak exercise ST segment
12. Ca : number of major vessels (0-3) colored by fluoroscopy
13. Thal : 3 = normal; 6 = fixed defect; 7 = reversable defect
14. Target : 1 = Heart Problem or 0 = Normal

Dataset Description...(continued)

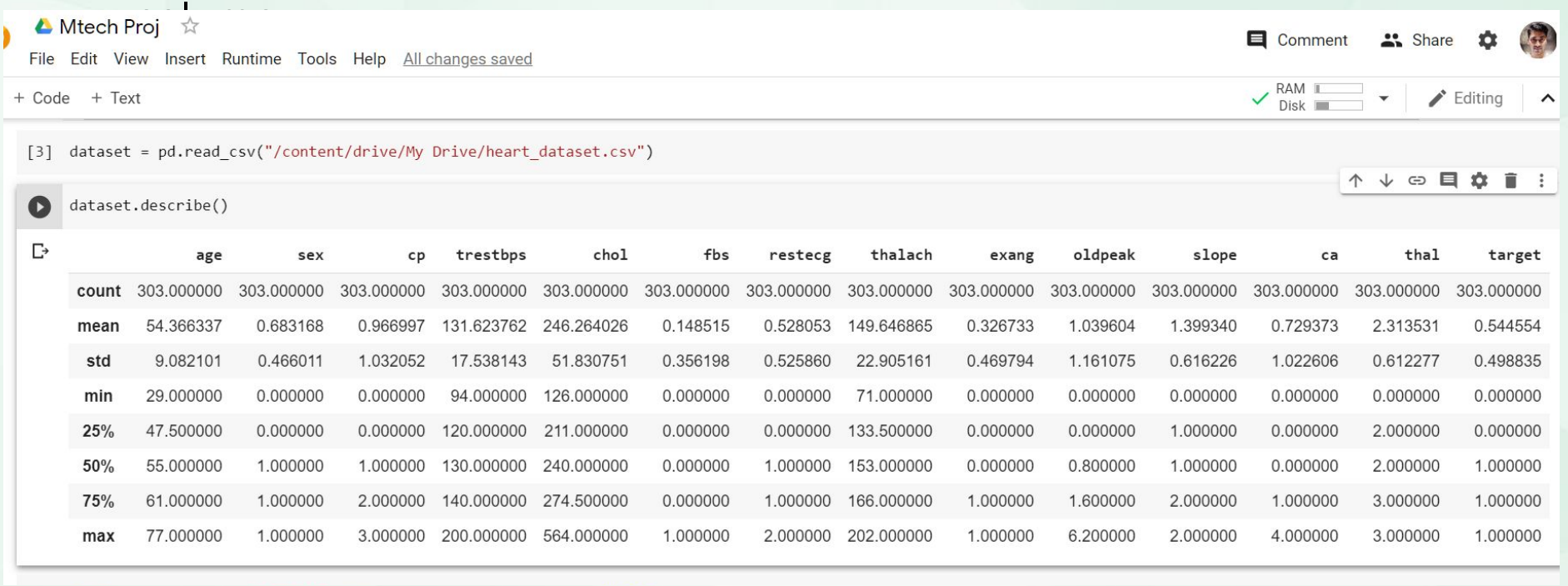
- The heart disease dataset is made up of 75 raw features from which 13 features were published. These features are very vital in the diagnosis of heart diseases.

- The features include fasting blood sugar test which must indicate
  < 120 mg / dl for a patient with absent test result and test result of
  > 120 mg / dl for a patient that has heart disease.

- Also, a patient that has serum cholesterol greater than 180 mg/dl is also considered as heart disease present.

Dataset Description...(continued)

The description of used dataset, 13 feature columns and 1 target



	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
count	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000	303.000000
mean	54.366337	0.683168	0.966997	131.623762	246.264026	0.148515	0.528053	149.646865	0.326733	1.039604	1.399340	0.729373	2.313531	0.544554
std	9.082101	0.466011	1.032052	17.538143	51.830751	0.356198	0.525860	22.905161	0.469794	1.161075	0.616226	1.022606	0.612277	0.498835
min	29.000000	0.000000	0.000000	94.000000	126.000000	0.000000	0.000000	71.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	47.500000	0.000000	0.000000	120.000000	211.000000	0.000000	0.000000	133.500000	0.000000	0.000000	1.000000	0.000000	2.000000	0.000000
50%	55.000000	1.000000	1.000000	130.000000	240.000000	0.000000	1.000000	153.000000	0.000000	0.800000	1.000000	0.000000	2.000000	1.000000
75%	61.000000	1.000000	2.000000	140.000000	274.500000	0.000000	1.000000	166.000000	1.000000	1.600000	2.000000	1.000000	3.000000	1.000000
max	77.000000	1.000000	3.000000	200.000000	564.000000	1.000000	2.000000	202.000000	1.000000	6.200000	2.000000	4.000000	3.000000	1.000000

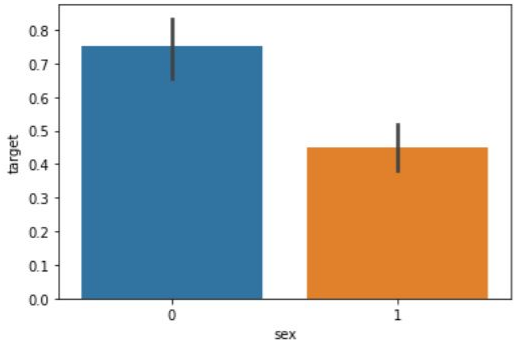
METHODOLOGY

- First I analyse which attributes of dataset are strongly correlated. And how it affect the final prediction.

```
[ ]  tar = dataset["target"]
     # sb.countplot(tar)

[ ]  sb.barplot(dataset.sex,tar)

     <matplotlib.axes._subplots.AxesSubplot at 0x7f23f968ecf8>
```



- By analysing the dataset it is found that females are more likely to have heart problems than males

Fig. 1. Correlation of gender with Heart disease.

METHODOLOGY...(continued)

- Likewise exang attribute is strongly correlated with the target attribute. And this is how other attributes are correlated.

```
[ ]  sb.barplot(dataset.exang,tar)
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f23f9741ac8>
```

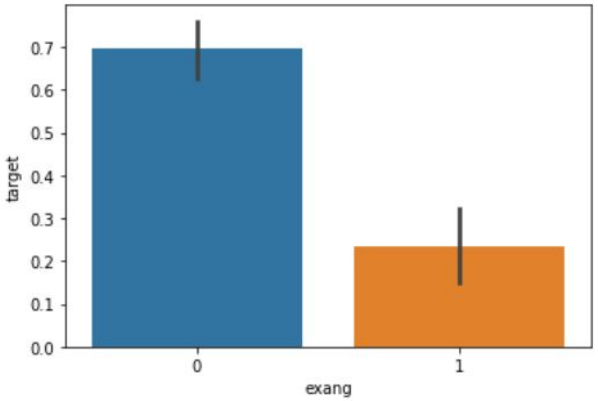


Fig. 2. Correlation of exang with Heart disease.

exang	0.436757
cp	0.433798
oldpeak	0.430696
thalach	0.421741
ca	0.391724
slope	0.345877
thal	0.344029
sex	0.280937
age	0.225439
trestbps	0.144931
restecg	0.137230
chol	0.085239
fbs	0.028046

METHODOLOGY...(continued)

I splitted 70% of dataset as training data and 30% as test data.

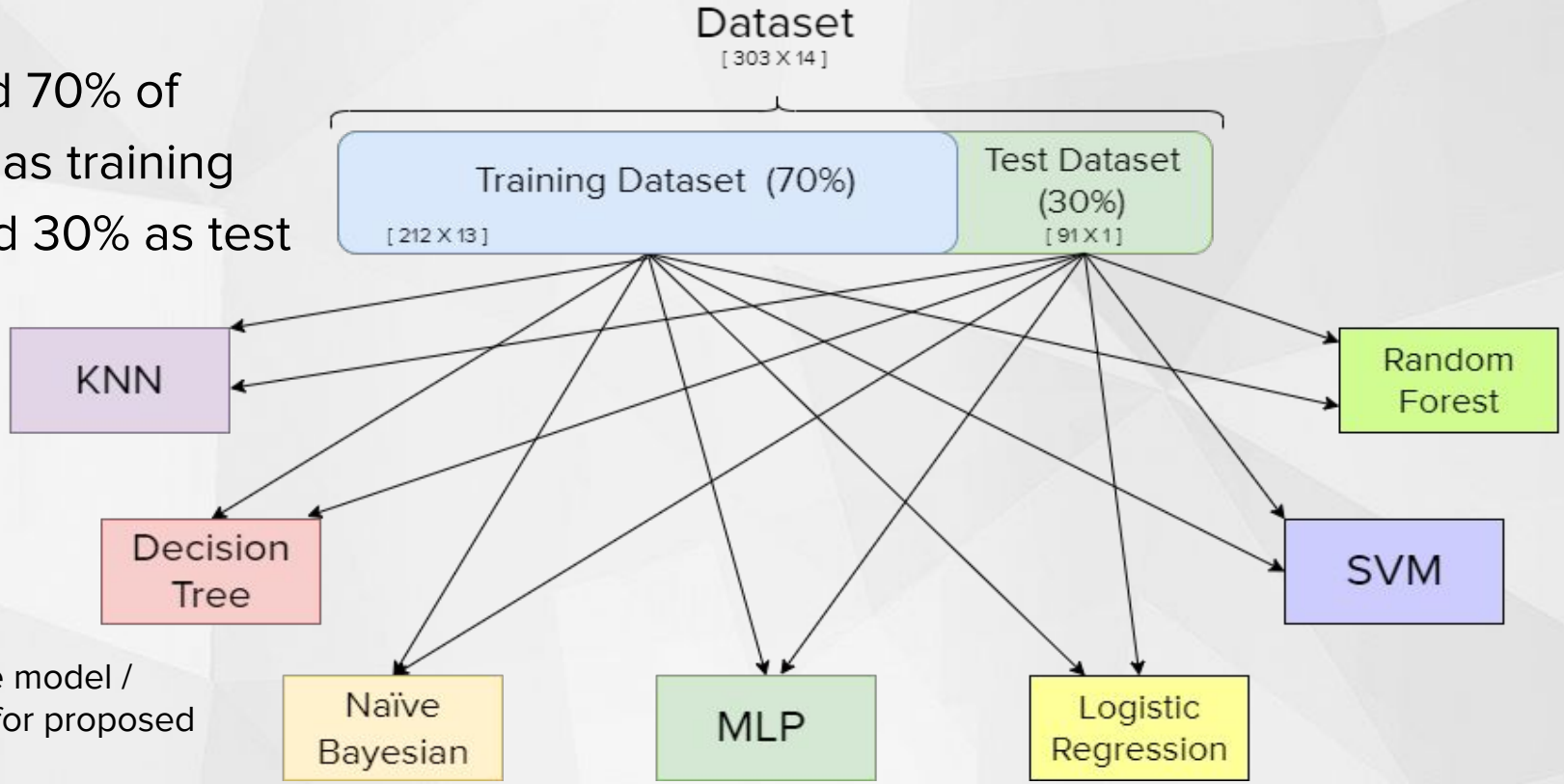


Fig. 3. Base model / Classifiers for proposed approach

METHODOLOGY...(continued)

I calculated individual accuracy percentage of 7 classifiers. To see the performance on prediction.

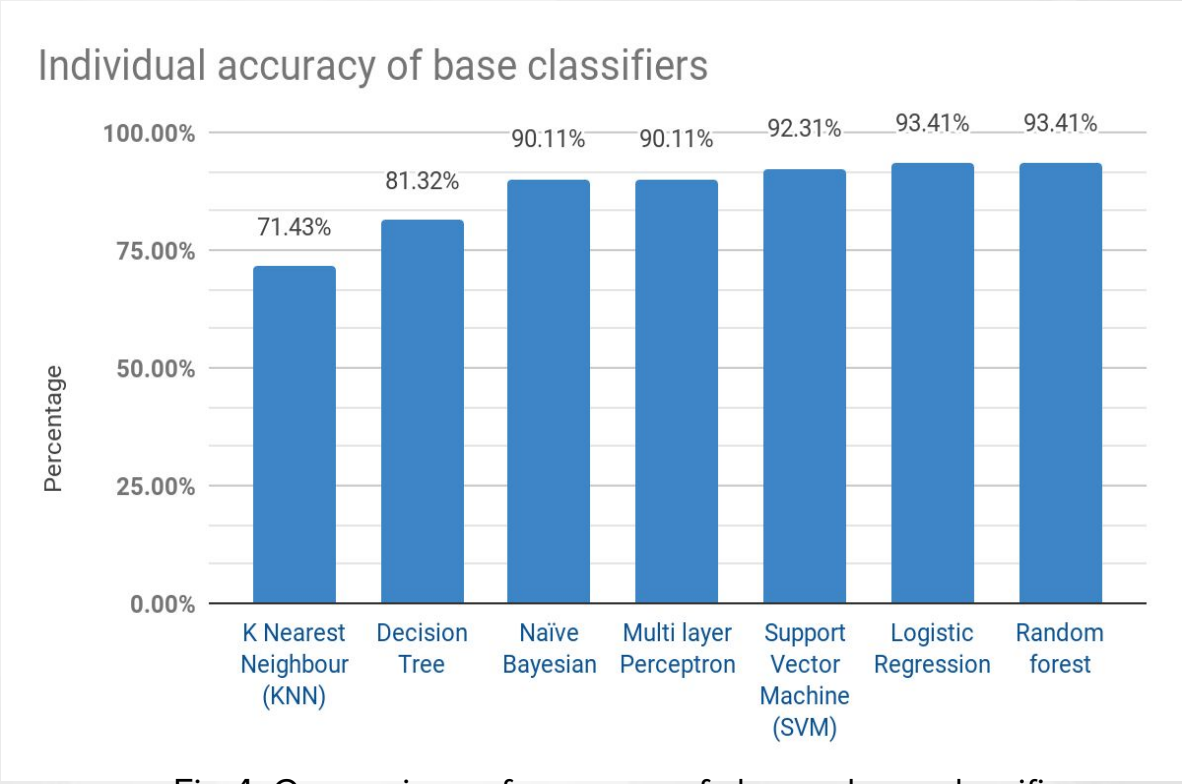


Fig 4. Comparison of accuracy of chosen base classifier

METHODOLOGY...(continued)

- To optimize the prediction accuracy I chose these techniques of classification as base models based on reviewed papers:
  - Logistic Regression
  - Naïve Bayesian
  - Support Vector Machine (SVM)
  - K Nearest Neighbour (KNN)
  - Decision Tree
  - Multi layer Perceptron
  - Random forest

METHODOLOGY...(continued)

- In paper [8], S. Kamley proposed hybrid approach to predict heart disease using ensemble classification techniques like bagging and boosting.

- In this project I used Stacking, Voting Classifier and Adaboost techniques of Ensemble methods and compared the overall performance of the system with individual models.

METHODOLOGY...(continued)

Proposed ensemble approach using Stacking method

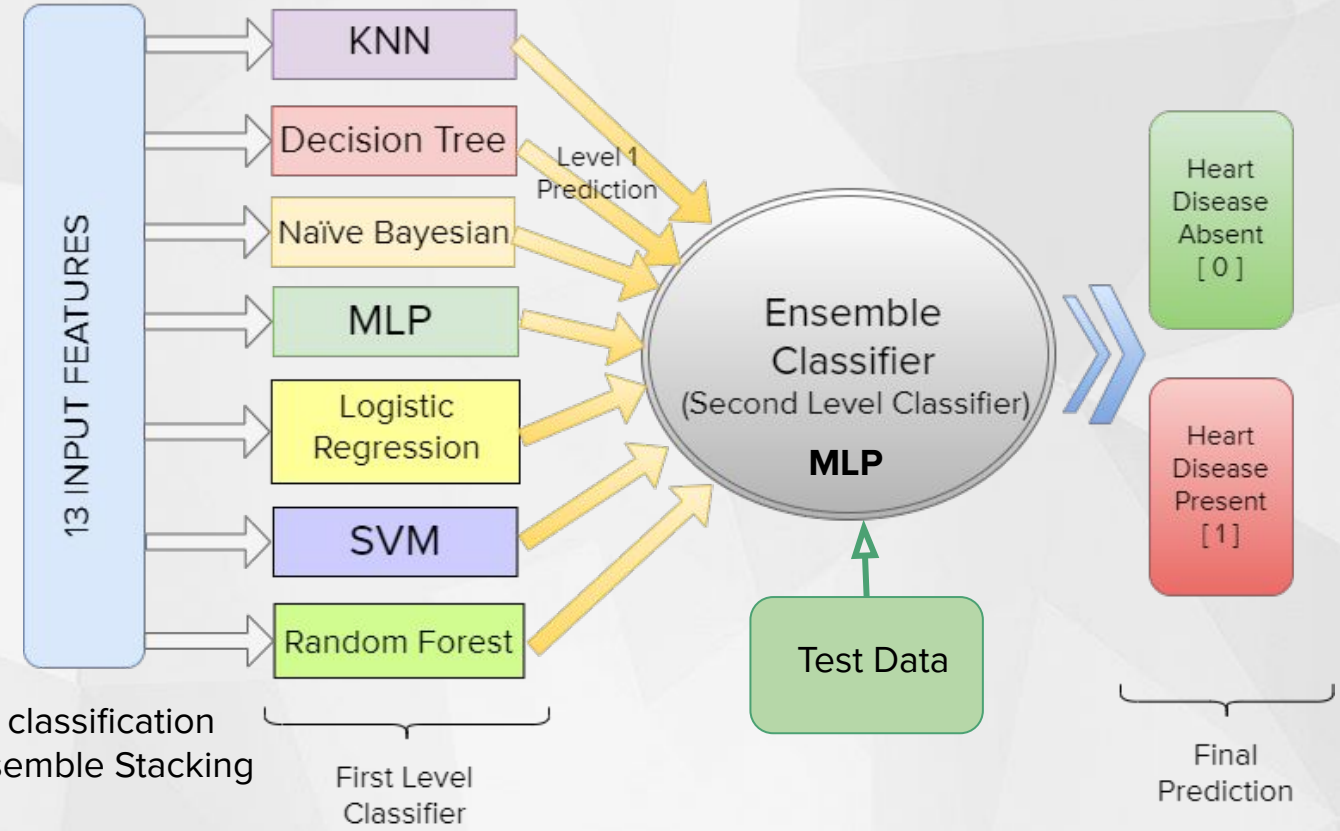


Fig. 5. Proposed classification system using ensemble Stacking technique

METHODOLOGY...(continued)

Stacking Ensemble Technique:

Uses a meta-learning algorithm to learn how to best combine the predictions from two or more base machine learning algorithms [9]

I used Multi-Layer Perceptron as Meta Classifier.

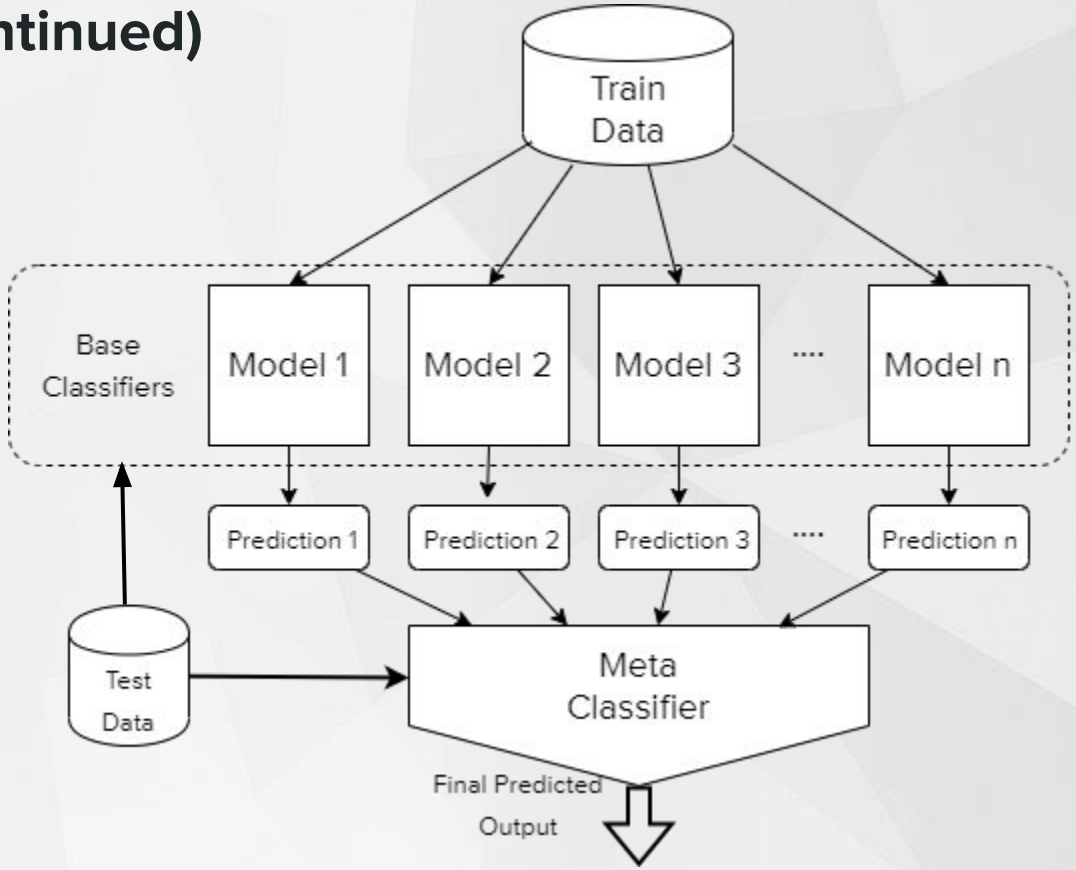


Fig 6. Standard stacking ensemble learning

METHODOLOGY...(continued)

- As in [6] authors used data mining and Artificial Neural Network (ANN) techniques. Multilayer perceptron neural network along with back propagation algorithm is used to develop the system. Because MLPNN model proves the better results. Their experimental result shows that using neural networks the system predicts Heart disease with very high accuracy.

- So to fill the accuracy gap from other techniques I used Multilayer perceptron neural network as meta classifier for proposed approach.

METHODOLOGY...(continued)

Multilayer Perceptron Neural Network (MLPNN):

- Artificial neurons are used in multiple layers
- Has multiple layers input layer, output layer and hidden layers
- Trained by a back propagation algorithm
- Input layer will contain 13 neurons
- Output layer contains 2 neurons for “Disease Presence”/“Disease Absence”.
- Two hidden layer with 7 and 5 neurons for base classifier. Fig 7.a.
- One hidden layer with 5 neurons for meta-classifier. Fig. 7.b.

METHODOLOGY...(continued)

Multilayer Perceptron Neural Network (MLPNN):

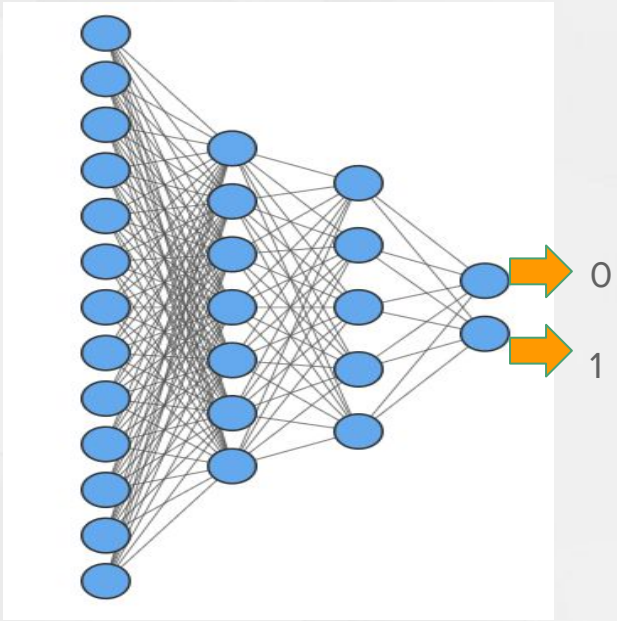


Fig 7. a. MLP for base classifier

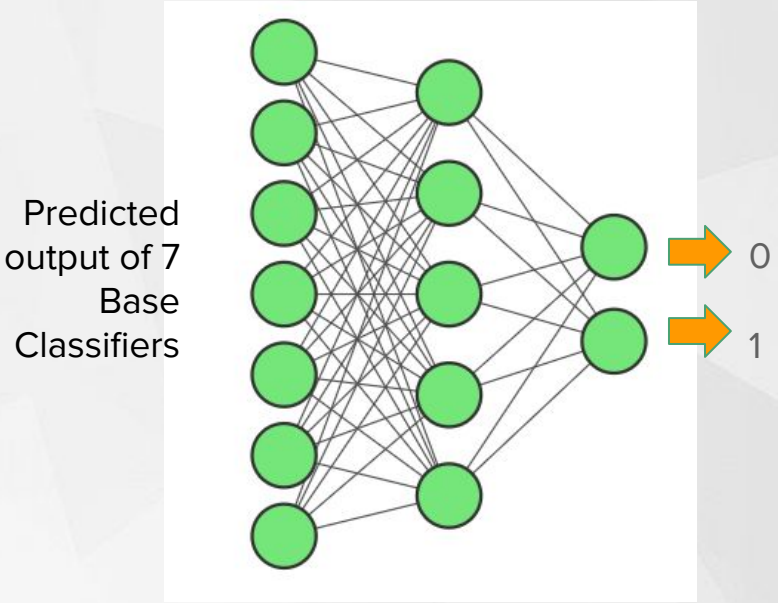


Fig 7. b. MLP for meta classifier

METHODOLOGY...(continued)

- I implemented these approaches in Python programming language in Google Colab.
- In addition to stacking Ensemble technique. I used Adaboost Classifier and Voting Classifier to check if it performs better than stacking approach for this system.
- I trained 7 individual classifiers and created a stack classifier using multi layer perceptron as meta-classifier (final classifier).
- I tried different combinations of random states for mlp and split ratios to get the optimal performance by this system.

METHODOLOGY...(continued)

- This is the comparison of different adaboost classifiers, voting classifier and Stacking classifier.

- Stacking with MLP as final classifier gives higher accuracy among others which is 94.51 %.

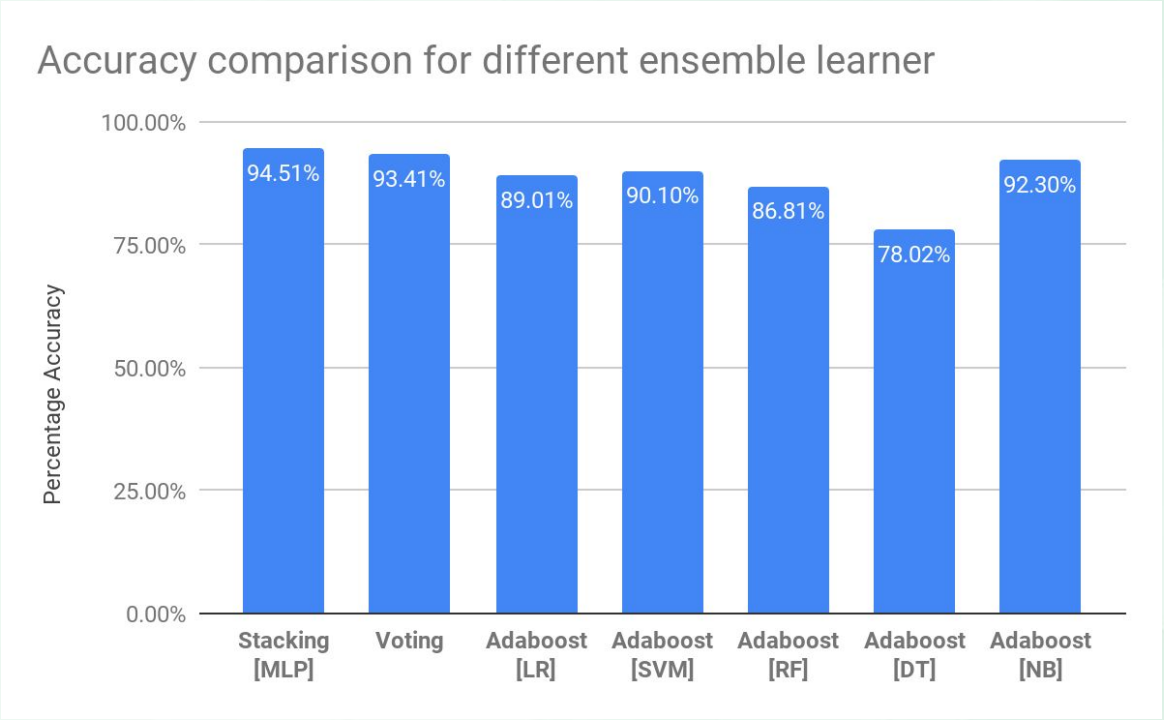


Fig 8. Comparison of different ensemble approach with proposed approach Stacking with MLP as Final classifier (first bar)

RESULT

The proposed approach for heart disease prediction gives higher accuracy than their individual base classifiers (From Fig. 8. and Fig. 4) with the accuracy of 94.51%.

This is the confusion matrix of the predictive model.

True Positive (TP) : Disease actually present, predicted present.
False Negative (FN) : Disease actually present, predicted absent.
True Negative (TN) : Disease actually absent, predicted absent.
False Positive (FP) : Disease actually absent, predicted present.

Table. 1. Confusion Matrix

Heart Disease	Predicted Absent	Predicted Present
Actual Absent	40 TP	3 FN
Actual Present	2 FP	46 TN

RESULT...(continued)

For disease absent

Recall : 0.93

Precision: 0.95

For disease present

Recall : 0.96

Precision: 0.94

```
print("Confusion Matrix:\n", confusion_matrix(Y_test, mlpPred))
print(classification_report(Y_test,mlpPred))
```

```
Confusion Matrix:
 [[40  3]
 [ 2 46]]
              precision    recall  f1-score   support

           0       0.95      0.93      0.94        43
           1       0.94      0.96      0.95        48

    accuracy                           0.95        91
   macro avg       0.95      0.94      0.94        91
weighted avg       0.95      0.95      0.95        91
```

Conclusion

From the result it is clear that using ensemble methods gives better performance than that of their individual learning model. Ensemble methods are meta-algorithms that combine several machine learning techniques into one predictive model in order to decrease variance, bias, or improve predictions (stacking).

Conclusion...(continued)

- The trained model of this system can be used to combined with web-portal or mobile application, It can also be integrated with IoT applications. To alert the doctors prior to the high risk.

- Accuracy for this prediction may increase beyond 94.51% by using extreme gradient boosting. It converts weak model to strong models by adjusting the weights iteratively to reduce bias and increase accuracy.

REFERENCES

1. Thomas, J., and R Theresa Princy. “Human Heart Disease Prediction System Using Data Mining Techniques.” *2016 International Conference on Circuit, Power and Computing Technologies (ICCPCT)*, 2016, doi:10.1109/iccpct.2016.7530265.
2. Sultana, Marjia, et al. “Analysis of Data Mining Techniques for Heart Disease Prediction.” *2016 3rd International Conference on Electrical Engineering and Information Communication Technology (ICEEICT)*, 2016, doi:10.1109/ceeict.2016.7873142.
3. Swain, Debabrata, et al. “An Efficient Heart Disease Prediction System Using Machine Learning.” *Machine Learning and Information Processing Advances in Intelligent Systems and Computing*, 2020, pp. 39–50., doi:10.1007/978-981-15-1884-3_4.
4. Dangare, Chaitrali & Apte, Sulabha. (2012). A Data Mining Approach for Prediction of Heart Disease Using Neural Networks. 3.
5. Amin, Syed Umar, et al. “Genetic Neural Network Based Data Mining in Prediction of Heart Disease Using Risk Factors.” *2013 Ieee Conference On Information And Communication Technologies*, 2013, doi:10.1109/cict.2013.6558288.
6. D. Dua and C. Graff, “*Heart Disease Data Set.*” archive.ics.uci.edu/ml/datasets/Heart Disease.

REFERENCES...(continued)

6. Deshmukh, Rachana, et al. “Heart Disease Prediction Using Artificial Neural Network.” *Ijarcce*, vol. 8, no. 1, 2019, pp. 85–89., doi:10.17148/ijarcce.2019.8119.

7. Alim, Muhammad Affan, et al. “Robust Heart Disease Prediction: A Novel Approach Based on Significant Feature and Ensemble Learning Model.” *2020 3rd International Conference on Computing, Mathematics and Engineering Technologies (ICoMET)*, 2020, doi:10.1109/icomet48670.2020.9074135.

8. Kamley, Sachin. “Performance of Hybrid Ensemble Classification Techniques for Prevalence of Heart Disease Prediction.” *VOLUME-8 ISSUE-10, AUGUST 2019, REGULAR ISSUE International Journal of Innovative Technology and Exploring Engineering*, vol. 8, no. 10, Oct. 2019, pp. 1875–1882., doi:10.35940/ijitee.j9233.0881019.

9. Sikora, Riyaz, and Ola Al-Laymoun. “A Modified Stacking Ensemble Machine Learning Algorithm Using Genetic Algorithms.” *Artificial Intelligence*, pp. 395–405., doi:10.4018/978-1-5225-1759-7.ch016.

Thank you