वैAIद्य

Tanish Pratap Singh,2nd year student, B. tech, Bennett University, Bhavya Gumber 2nd year student, B. tech, Bennett University, Aadi Jain 2nd year student, B.tech , Bennett University

Abstract-In recent years, nearly one death has occurred every minute because of heart-related diseases. With advancement in lifestyle, there is high amount of fast food eaten in today generation leading to high cholesterol level in the body. With help of the Data science and machine learning we can have a huge impact on healthcare industry. It is quite crucial to predict these diseases beforehand, so the patient can go through diagnosis for the same.

Using machine learning we would automate the process of heart disease prediction using the medical history and several other factor. We would use the existing database on such topic to train our model and make it effective in nature using algorithms using supervised learning such as Random Forest, Decision Tree, and unsupervised learning Neural network etc., to predict the likelihood of heart disease and classify patients' risk level. As a result, the final machine learning model will be decided after observing the performance of various specified classification algorithms.

This system would help in enhancing the medical care and making the medical care cost effective and could be instrument help for user.

I. INTRODUCTION

Heart disease is a major cause of death globally, it is the top most cause of death in the whole world. According to WHO, more than 17.5 million people lost their lives from heart diseases in 2016, accounting for around 1/3rd of all fatalities across the world.[1]

There are several projects related to this topic online, about similar interest. Several of the effective predictor has been made using various supervised and unsupervised learning such as KNN, random forest, Neural and etc. Each of which has its own strength and weaknesses.

Nearly 805000 Americans have a heart attack every year according to the data of 2019.[2].

Among all the heart diseases, coronary heart disease which is deadly among all the heart disorders.

We know that heart attacks can be avoided easily and some modest lifestyle changes and quitting smoking, eating healthier and exercising regularly can help us in avoiding these fatal heart diseases. So, it is high time to examine and prevent the death rate by rightly guessing the disease in its beginning stages itself.

In the modern world, with the use of machine learning and AI in medicine field, the opportunities to prosper are truly great. Artificial intelligence and machine learning in medicine has proven to be a boon to mankind. With the use of AI, examination and diagnosis of diseases has become a less challenging task. AI and machine learning can be used throughout complicated treatments and operations.

In our Machine learning model, the major attributes on which we are going to work upon age ,gender of person, chest pain type, maximum heart rate, cholesterol and many other features are used.

Our Target Variable will produce a binary output, binary:1 means “There is a risk”, binary:0 “There is no risk”

II. RELATED WORK

Basically, we plan to create a project that could help a person to predict whether he has risk of getting heart related disease or not, using basic simple input parameter which would help the person to get expert help and advice on the same which would be instrumental for the person wellbeing.

The risk of coronary heart disease is identified using McPherson, which uses unsupervised learning algorithm, Neural networks using it for the prediction for the model to be accurate and efficient in nature.

Other similar work in this field is Diagnosis and prediction of heart disease and blood pressure with aid of neural network was introduced by R. Subramaniam, which include hundreds of hidden

layers, which is one of the layman ways of predicting precise and accurate result.

We look to explore and the current option and the option that could be added at the time using the knowledge of the existing and add some changes to enhance the model further.

The main contribution of our work to the problem would be to build a effective and accurate prediction and also make it easily accessible to the people which is majorly lacking in the current available option as its not commonly available to the common people, this is the gap we want to bridge the most and would also be helpful to hospital for the same reasons.

III. BLOCK DIAGRAM

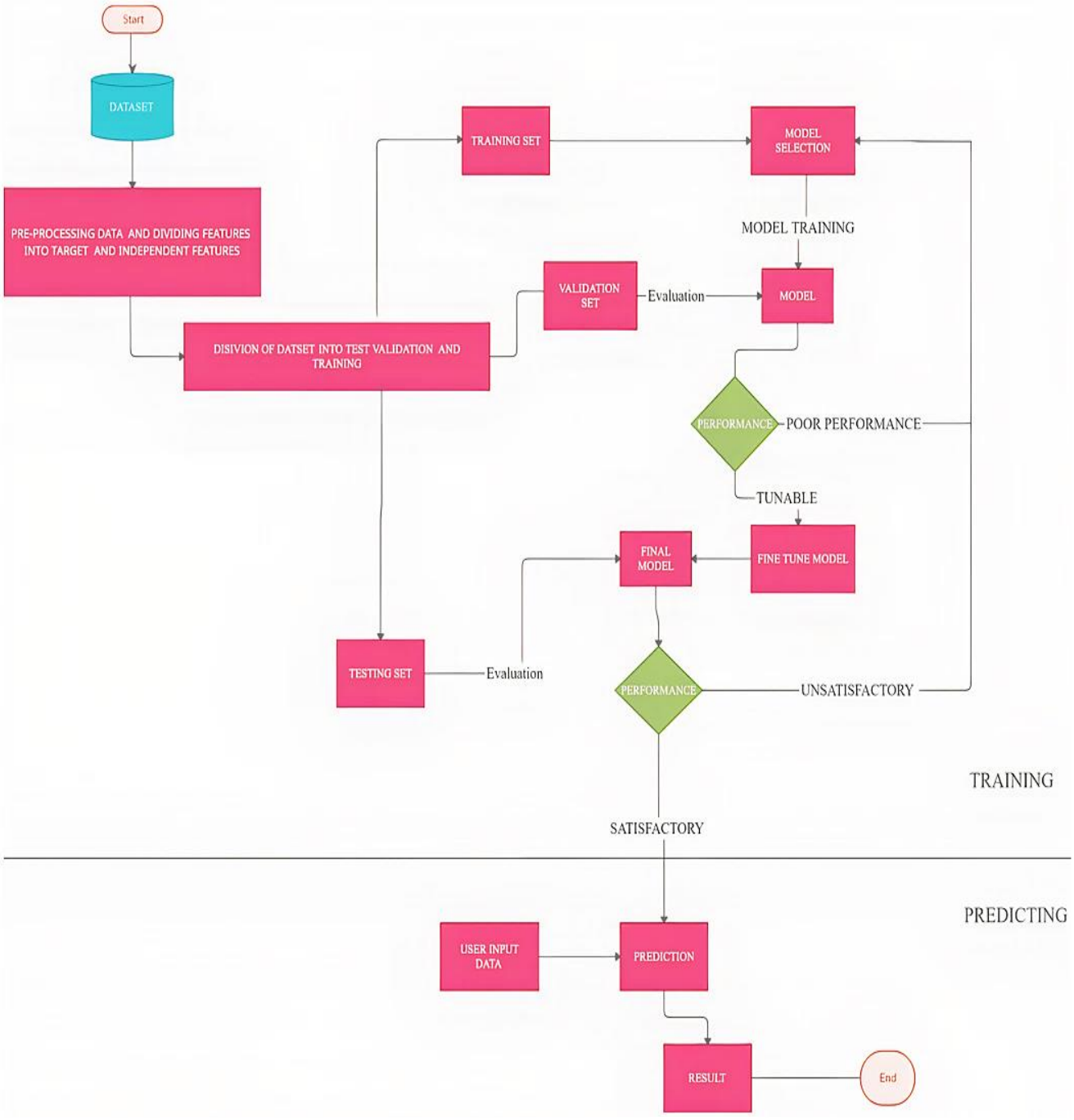

Fig 1-Block diagram of heart disease prediction

The whole process is divided into two parts:

1)Training of model

2)Prediction using the model

Training of model:

- At the start the dataset used would be pre-processed for removal of unwanted data and filling up of empty data. Also, data would be normalized for accurate results.
- Dataset would be divided into independent variable and target variable.
- Later, the dataset would be divided into training, testing and validation.
- Using one of the classifications approach the model would be trained for training dataset and be evaluated by validation dataset.
- Depending on the performance of the model we would either fine tune it or change the model.
- Evaluation using the test data if satisfied would be used in the project ,if not the new model would be selected.
- Process will be repeated till we are satisfied by the performance of the model.

Predicting using model:

- The user would be made input on certain parameters required by the model.
- Using the input, the model will evaluate and predict the output and present it to the user.

IV. PROPOSED METHODOLOGY

The working of the project starts with the collection of data. We used almost all features of the dataset. The required data is then pre-processed into required format. Next, we perform Exploratory Data Analysis on the existing data. We would also try to use feature selection if viable for our model and solution to decrease complexity and overfitting of the model.

We use several machine learning algorithms and test their accuracy. System's accuracy is obtained by testing the system with testing data. Various classification model like logistic regression, XG boost, Support Vector Machine, Random Forest and Stacking classifier are used.

All these models would be train and used to predict and their performance would be calculated ,top few would be checked for over-fitting via N-fold cross validation. The top classifiers would be hyper tuned further to improve their accuracy and performance.
The best model after all these steps would be chosen as the final model which would be used for predicting data. To keep in overfitting in check we will see the performance metrics on training dataset .

V. EXPERIMENTAL SETUP

```
RangeIndex: 1025 entries, 0 to 1024
Data columns (total 14 columns):
 #   Column    Non-Null Count  Dtype
---  ------    --------------  -----
 0   age       1025 non-null   int64
 1   sex       1025 non-null   int64
 2   cp        1025 non-null   int64
 3   trestbps  1025 non-null   int64
 4   chol      1025 non-null   int64
 5   fbs       1025 non-null   int64
 6   restecg   1025 non-null   int64
 7   thalach   1025 non-null   int64
 8   exang     1025 non-null   int64
 9   oldpeak   1025 non-null   float64
 10  slope     1025 non-null   int64
 11  ca        1025 non-null   int64
 12  thal      1025 non-null   int64
 13  target    1025 non-null   int64
dtypes: float64(1), int64(13)
```

Fig 2-Features of dataset

Dataset contain 14 features,**target** being the dependent varaible having value 0 or 1 , 0 represent the 'no risk of heart disease' whereas 1 represent 'risk of heart disease' and other being the independent variable.

The independent features are as follow:

S. No	Observation	Description	Values
1.	Age	Age in Years	Continuous
2.	Sex	Sex of Subject	Male/Female
3.	CP	Chest Pain	Four Types
4.	Trestbps	Resting Blood Pressure	Continuous
5.	Chol	Serum Cholesterol	Continuous
6.	FBS	Fasting Blood Sugar	< ,or > 120 mg/dl
7.	Restecg	Resting Electrocardiograph	Five Values
8.	Thalach	Maximum Heart Rate Achieved	Continuous
9.	Exang	Exercise Induced Angina	Yes/No
10.	Oldpeak	ST Depression when Workout compared to the Amount of Rest Taken	Continuous
11.	Slope	Slope of Peak Exercise ST segment	up/ Flat /Down
12.	Ca	Gives the number of Major Vessels Coloured by Fluoroscopy	0-3
13.	Thal	Defect Type	Reversible/Fixed/Normal

Fig 3-Features description of independent features of dataset

After exploring the data set we need to get to preprocessing firstly we separate the independent and dependent feature into different dataframe for prediction.

As in out case there are no categorical data present hence we wont be required to do encoding for the same.As value in the column have high range so we would standardize these value from range of 0 to 1 to make prediction more accurate and feature not getting overshadowed by the other feature with high range.

We would also see that the train data set have close number of instances of both case otherwise due to this imbalanced dataset would cause model to be ineffective in predicting the lesser trained value.Some insight about data from the EDA:

Fig 4-Divison of dependent variable of dataset

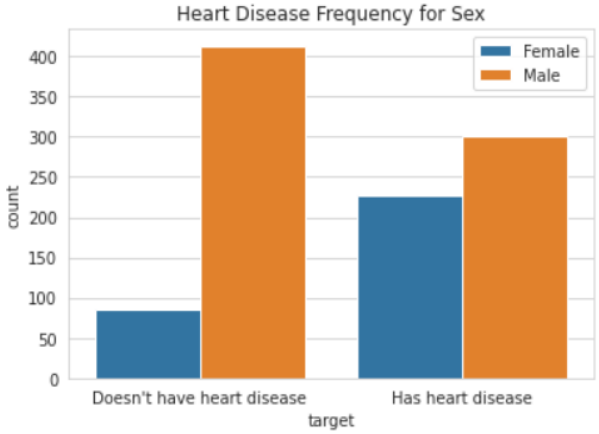

Fig 5- Heart disease on the basis of Sex.

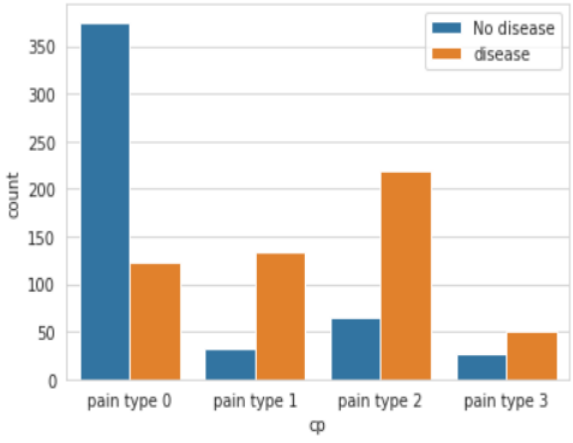

Fig 6-Heart disease on basis of chest pain type

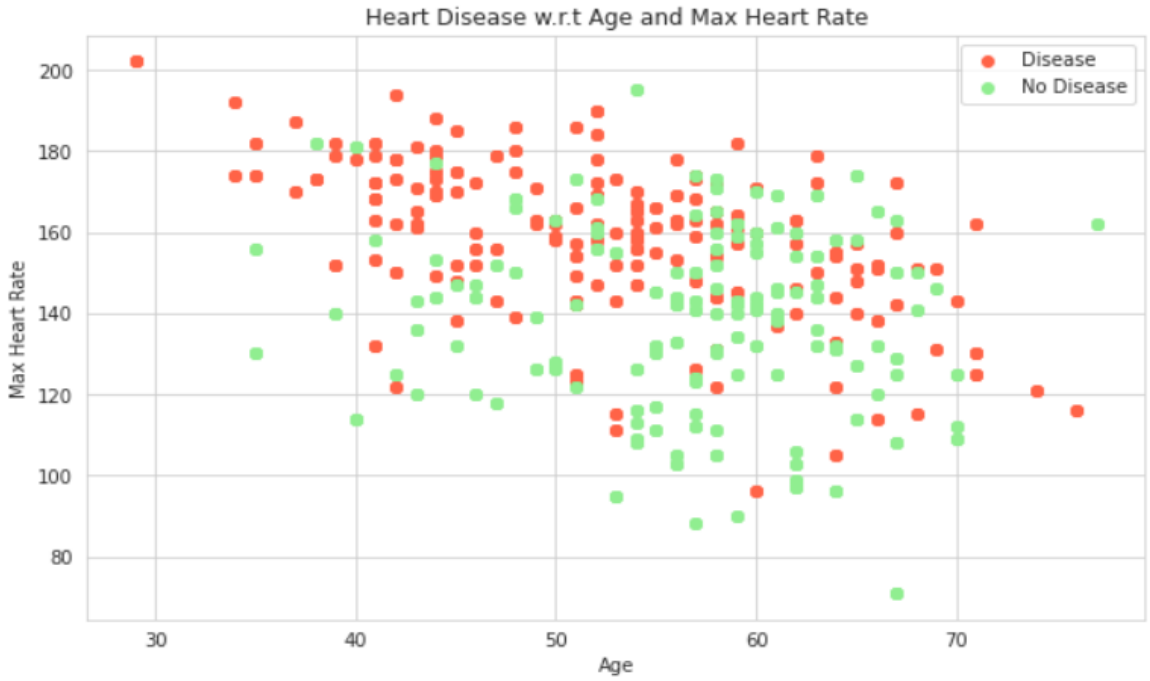

Fig-7 Correlation of heart disease with respect to age and max heart rate

VI. RESULT

As in the first phase several of the choosing a algorithm we put several classification model to test such as Logistic regression, Naïve bayes, Random Forest, Extreme Gradient Boosting, KNN, Decision Tree, SVM and the stacking classifier. In general, the ensembling algorithm like Extreme Gradient Boosting, Random Forest, and stacking classifier performed much better than the rest of the algorithms.

	Model	Accuracy
6	Support Vector Machine	69.080780
4	K-Nearest Neighbour	71.030641
1	Naive Bayes	82.451253
0	Logistic Regression	85.515320
5	Decision Tree	89.693593
2	Random Forest	90.529248
7	StackingClassifier	90.807799
3	Extreme Gradient Boost	91.922006

Fig 8-Accuracy of all models

In the next phase, we pick the top 4 algorithm and check overfitting via K-fold cross validation and pick the best for hyper tunning of parameter.

	Model	Accuracy
0	Decision Tree	89.169568
1	Random Forest	91.802862
2	StackingClassifier	92.289619
3	Extreme Gradient Boost	94.047435

Fig 9-Accuracy of top 4 model in k cross validation.

As we are using boosting method each feature has a importance/weight and using the function we would find the least important feature and remove it to check the performance change which in this case is feature ‘fbs’.

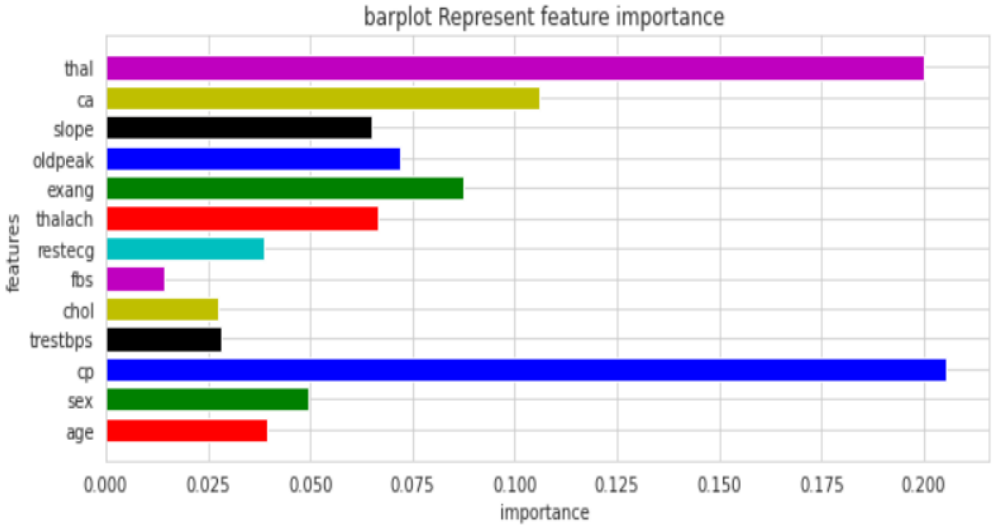

Fig 10-feature importance via XG boost

After various hyper parameter tunning of algorithms, we were able to reach accuracy of 94.15% from 91.92% in holdout technique and 94.73% from 94.04% in k-fold cross validation for XG boost.

After the feature 'fbs' has been removed, we see an overall improvement in our model performance as accuracy, recall and f1 score increases for new independent set of features.

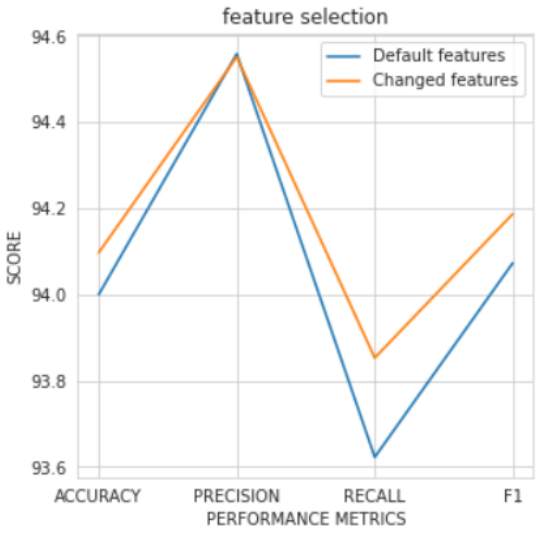

Fig 11-performace comparison of default and changed features(k-fold)

```
For default XG BOOST model:
ACCURACY: 91.92200557103064
PRECISION: 92.26519337016575
RECALL: 91.75824175824175
F1: 92.01101928374656
```

```
For hyper tuned XG BOOST model
Testing dataset:
ACCURACY: 94.15041782729804
PRECISION: 94.97206703910615
RECALL: 93.4065934065934
F1: 94.18282548476455
```

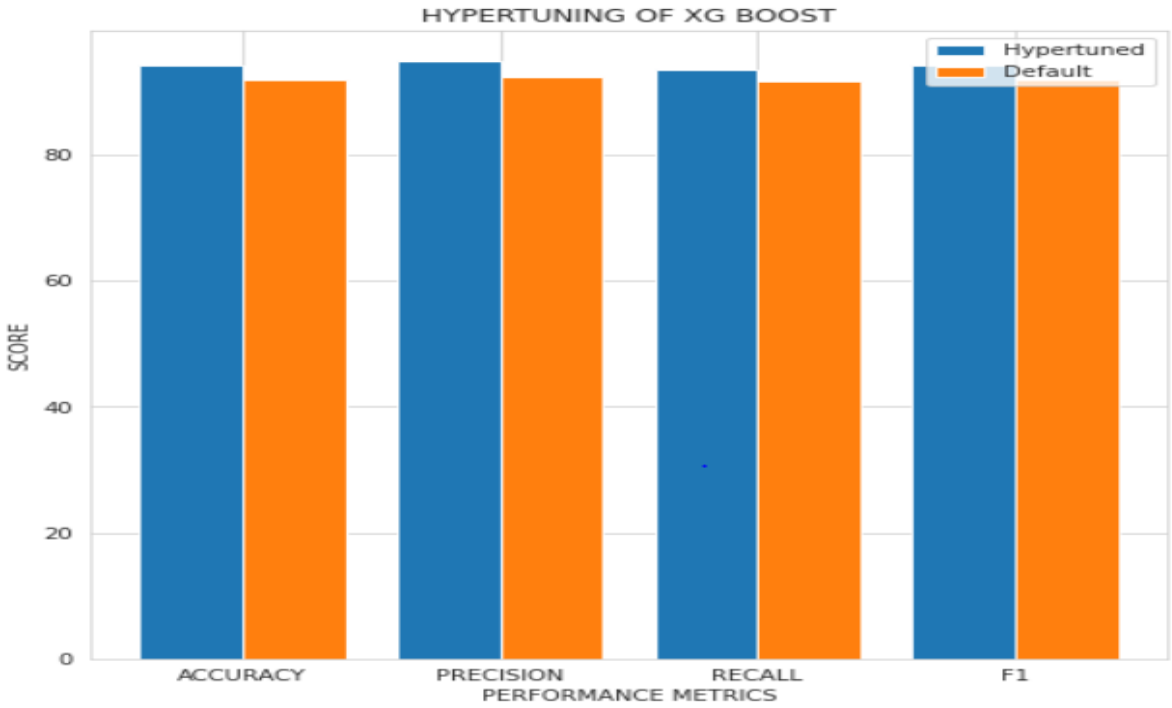

Fig 12-Performance of final model (default vs hyper tuned)

For checking overfitting, we would run the model for training dataset, and we would see that there is not much difference in the accuracy hence we can say that our model is not overfitted.

```
For hyper tuned XG BOOST model
Testing dataset:
ACCURACY: 94.15041782729804
PRECISION: 94.97206703910615
RECALL: 93.4065934065934
F1: 94.18282548476455
```

```
For hyper tuned XG BOOST model
Training dataset:
ACCURACY: 98.49849849849849
PRECISION: 99.11764705882354
RECALL: 97.96511627906976
F1: 98.53801169590642
```

Fig 13-Checking overfitting of model

VII. CONCLUSION

The early detection of cardiovascular diseases can aid in making decisions on lifestyle changes in high-risk patients and in turn reduce the complications. We used various classification models and compared their accuracies to choose most accurate model. During evaluation we found that ensembling model performed better than other.

After the evaluation we found out that the XG boost was the most suited model for the task showing highest performance between all the models. Further, for model enhancement, we did feature selection based upon feature importance in XG boost. To further improve the model, we hyper tuned the model to make it accuracy reach up to 94%. We also did performance evaluation for both testing and training dataset to check overfitting in the model and making sure it was a generalized model.

VIII. FUTURE SCOPE

Our future work, we aim at building a one stop AI doctor, we would create a website which will not only heart disease prediction tools but also performs

accurate prediction for tumor, diabetes, and many other diseases.

We would also extend our project and provide our user to get information about the disease they are having based on symptoms and providing them the doctor available in their locality for getting the treatment in turn saving their time and effort and providing one stop shop reachable in some clicks.

IX. REFERENCES

[1]- WHO annual report [Online]. Available: https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1

[2]- Fryar CD, Chen T-C, Li X. NCHS data brief, no. 103. Hyattsville, MD: National Centre for Health Statistics; 2012.Available: Prevalence of uncontrolled risk factors for cardiovascular disease: United States, 1999–2010 pdf icon[PDF-494K].