

Major Project Report On **Heart Disease Prediction Model**

Under the Supervision of

DR. Rishav Shingh

Assistant prof

National Institute of Technology Delhi

submitted

in partial fulfilment of the requirement for the degree of

Bachelor of Technology

By

Shashi Raj & Anand Ravidas

(181210050) (181210010)



Department of Computer Science and Engineering
National Institute of Technology Delhi

CERTIFICATE

This is to certify that the project report entitled **Heart Disease Prediction Model** submitted by **Shashi Raj & Anand Ravidas** to the National Institute of Technology, Delhi , in partial fulfillment for the award of the degree of **B. Tech in (Computer Science and Engineering)** is a *bona fide* record of project work carried out by him/her under my/our supervision. The contents of this report, in full or in parts, have not been submitted to any other Institution or University for the award of any degree or diploma.

<Signature> <Name of the supervisor>
Department of Computer Science and Engineering

DECLARATION

I declare that this project report titled **Heart Disease Prediction Model** submitted in partial fulfillment of the degree of **B. Tech in (Computer Science and Engineering)** is a record of original work carried out by me under the supervision of **Dr. Rishav Singh**, and has not formed the basis for the award of any other degree or diploma, in this or any other Institution or University. In keeping with the ethical practice in reporting scientific information, due acknowledgements have been made wherever the findings of others have been cited.

<Place>

<Date>

<Signature> <Name of
the Candidate>
<ENrollment No.>

Acknowledgement

*The satisfactory completion of this project will remain incomplete without the heartfelt thanks to everyone whose constant guidance, support and encouragement made this work successful. First of all, I would like to express my wholehearted gratitude to my mentor **DR. RISHA V SINGH** Sir for constantly motivating and guiding us through this challenging project.*

This project may never have come into existence, if it was not for his excellent supervision.

Our next essence of gratitude is to my fellow batchmates who spent hours with me in the lab, helping me get through intricate concepts and supporting me during the entire project.

*Our essence spirit of gratitude to **National Institute of Technology, Delhi** for providing me with such excellent facilities and wonderful hospitality. It will always be one of my most cherish able experiences.*

Last but not the least, We would like to convey our love and gratitude to our parents for supporting us throughout this project in everyway possible.

With Regards

Shashi Raj &

Anand

Ravidas.

ABSTRACT

Cardiovascular disease (CVD) is a main reason for morbidity and mortality in today's world. Identification of CVD is very important and it is very sophisticated task as it needs good efficiency, effectiveness which can be made successful with automation. As a human beings, Not all doctors have same skill and also some may lack knowledge whereas some lacks experience and hence problems may arise in diagnosis of CVD so An automated system would enhance medical diagnosis and it can also reduce costs. In this Project, we have designed a Model which can efficiently identify the probability of CVD diseases and can also identify the factors which have effect on CVD. The performance of the Model is evaluated with the help of confusion matrix and the results shows that the Model is working efficiently and has great potential in predicting the heart disease risk level. Index Terms—Cardiovascular disease, HDPM, XGboost

INDEX

Content	Pages
cerunuaie	3
Declaration	4
Acknowledgement	5
Abstract	6
List of figures	8
List of tables	9
Introduction	10
Literature Survey	10
Methodology	10-13
Result and Analysis	13-14
Conclusion and References	15

LIST OF FIGURES

FIGURE	TITLE	PAGE NUMBER
1.	PCC values of each attribute with class attribute of Cleveland dataset proposed mechanism.	4
2.	Information gain of each attribute of statlog dataset.	4
3.	Information gain of each attribute of Cleveland dataset	4
4.	Plot between attribute thalach and age (red points are outlier points) of statlog dataset.	4
5	Plot between attribute thalach and age (red points are outlier points) of Cleveland dataset.	5
6..	Dataset(statlog) before applying SMOTEENN	5
7.	Dataset(statlog) After applying SMOTEENN	5
		5
8.	Dataset(cleveland) before applying SMOTEENN	5
9.	Dataset(cleveland) After applying SMOTEENN	5
.		

LIST OF TABLES

TABLE	TITLE	PAGE NUMBER
1.	FEATURES DESCRIPTION	4
2.	Performance evaluation for statlog	14
3	Performance evaluation for cleveland	14

Heart Disease Prediction Model

1. INTRODUCTION

Heart disease is a big problem nowadays everywhere around the globe and heart attacks are often common. It is observed that this disease do not occur all of a sudden but a continuous process and is the result of being on a particular lifestyle for a long time and also results after giving some basic and common symptoms being occurring all of a sudden. In case of heart attack, the heart is unable to pump the required amount of blood to different parts of the body and moreover it itself also does not get enough blood supply due to blocked arteries in the heart chambers and thus results in heart failure and deaths. The rate of heart disease is very high in countries like India. India is having a very high rate of deaths due to heart diseases. In this paper a lot of symptoms of heart diseases are included which can be used as the features that could be used to find the accurate diagnosis of the patient. So it is necessary to start timely diagnosis of heart diseases to improve the security of the heart and life. And thus it is important to predict the heart diseases before it is too late. Heart Disease Prediction Model can help Predicting the heart diseases and can be used for early diagnosis of heart and results in saving a million lives.

2. LITERATURE SURVEY

A lot of studies are done on HDPM with Machine learning techniques to improve heart diagnosis prediction . Statlog and Cleveland are two publicly available dataset which are majorly used in model creation and researchers widely uses them for model accuracy. Different types of algorithms are

used by researchers to generate best possible model and they have got accuracy around 96% for statlog dataset and 98% for cleveland dataset. And they have compared there model with six other preexisting model. They found that there model was very efficient and getting very better result in term of accuracy. In the paper they have performed step by step methods to proceed towards training of the model. As data preprocessing is very essential and we know that it affects the accuracy of the model, so we need to do it very sincerely. They have focused on data preprocessing , and data balancing . But they have not used Genetic algorithm for the parameter tuning , so we thought of doing parameter tuning using Genetic algorithm , so that the best parameter will give better accuracy. And also we need to avoid overfitting the model. We compare our model with ten other models.

3. METHODOLOGY

In HPDM , we have applied different ML technique and a step by step procedure to make this model. Using dataset and information provided by the user , model will predict the presence and absence of the heart disease. As we know that dataset have some not useful information and outlyer , so we have to perform data cleaning to remove them. During data preprocessing we have selected attribute on which we have performed operation which has been done on the basis of PCC and information value. We have done data balancing using SMOTEENN and for data training applied Genetic Algorithm and we have trained model using XGboost. Performance of the model have been checked by using Performance Matrix.

3.1 DATA SOURCE

To implement our model we have taken heart disease dataset from ML report of UCI . we have used two dataset :

1. Statlog-No of data point (270),output class(1),attribute(13)(120 subject positively labbled and 150 negitively labeled)
2. Cleveland - No. of datapoint(303),output class (1),Attribute(13),(138 subject positevly labbled and 165 negatvely labbled)

3.2 FEATURES DESCRIPTION

Features	Description
age	age(in years)
sex	gender(M/F etc)
cp	type of chest pain
trestbps	rest blood pressure(bp)
chol	serum cholesterol(chol)
fbs	fasting blood sugar
restecg	resting electrocardiographic(ecg) result
thalach	maximum heart beat rate
exang	exercise induced angina
oldpeak	ST depression due to exercise compared to rest
slope	peak exercise ST segment slope
ca	count of major vessels colored with the help of fluoroscopy
thal	type of defect

3.3. DATA PREPROCESSING

In this process of feature selection , we have made correlation between attribute which have great use like to determine the relationship between the attribute . In HDPM model ,we have used PCC(Person's correlation co effecient) to predict correlation between attribute . the PCC varies from -1 to +1.

- If PCC= -1 (negative relation)
- PCC=+1(positive relationship)
- If PCC = 0(no relationship)

We have calculated PCC class level and all other attribute of dataset. So , if PCC value of any attribute is close to zero then that attribute has no any contribution towards prediction .Hence, we have dropped that attribute from our dataset. By using this method we have dropped (chol,fbs) which have PCC value close to zero. By applying PCC, at the end we have ten future attribute with one class attribute as the dataset.

	age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	presence						
1.	000	0.09	0.097	0.273	0.220	0.123	0	.128	0.	.40	0.	.098	0.194	0.	.160	0.356	0.	.106	0.	.212
0.09		1.000	0.035	-0.06	-0.20	0.042	0	.039	0.	.08	0.	.180	0.097	0.	.051	0.087	0.	.391	0.	.298
0.	.097	0.035	1.000	-0.04	0.090	-0.10	0	.074	0.	.32	0.	.353	0.167	0.	.137	0.226	0.	.263	0.	.417
0.	.273	-0.06	-0.04	1.000	0.173	0.156	0	.116	0.	.04	0.	.083	0.223	0.	.142	0.086	0.	.132	0.	.155
0.	.220	-0.20	0.090	0.173	1.000	0.025	0	.168	-0	.02	0.	.078	0.028	-6	1.01	0.127	0.	.029	0.	.118
0.	.123	0.042	-0.10	0.156	0.025	1.000	0.	.053	0.022	0.00	-0.03	0.	.044	0.124	0.	.049	0.02			
0.	.128	0.039	0.074	0.116	0.168	0.053	1	.000	-0.	.07	0.	.095	0.120	0.	.161	0.114	0.	.007	0.	.182
-0.40	-0.08	-0.32	-0.04	-0.02	0.022	-0.07	1.000	-6	1.38	0.35	-e	1.39	0.27	-0.25	-0.42					
0.	.098	0.180	0.353	0.083	0.078	0.00	0	.095	0.	.38	1.	.000	0.275	0.	.256	0.153	0.	.321	0.	.419
0.	.194	0.097	0.167	0.223	0.028	-0.03	0	.120	0.	.35	0.	.275	1.000	0.	.610	0.255	0.	.324	0.	.418
0.	.160	0.051	0.137	0.142	-0.01	0.044	0	.161	0.	.39	0.	.256	0.610	1.	.000	0.109	0.	.284	0.	.338
0.	.356	0.087	0.226	0.086	0.127	0.124	0.	.114	0.	.27	0.	.153	0.255	0.	.109	1.000	0.	.256	0.	.455
0.	.106	0.391	0.263	0.132	0.029	0.049	0	.007	0.	.25	0.	.321	0.324	0.	.284	0.256	1.	.000	0.	.525
0.	.212	0.298	0.417	0.155	0.118	0.02	0	.182	0.	.42	0.	.419	0.418	0.	.338	0.455	0.	.525	1.	.000

Fig.1. PCC values of each attribute with class attribute of Statlog dataset

3.4. OUTLIER DETECTION AND REMOVAL

As we know outliers are the data point that differs significantly from observation or the desired observation. To remove it from our dataset we used DBSCAN algorithm and we also use Optimal eps and mean point value for DBSCAN algorithm . After removing outliers , we get dataset which have not any incomplete data , noise , outlier etc.

age sex ep trestbps chol fbs restecg thalach exang oldpeak stope ca that presence

1.000	0.100	0.070	0.279	0.214	0.121	0.120	0.400	0.097	0.210	0.170	0.276	0.068	0.230
0.100	1.000	0.050	0.060	0.200	0.145	0.060	0.040	0.142	0.096	0.030	0.118	0.210	0.280
0.070	0.050	1.000	0.048	0.080	0.094	0.044	0.296	0.390	0.150	0.120	0.180	0.160	0.434
0.279	0.060	0.048	1.000	0.123	0.178	0.110	0.050	0.068	0.193	0.120	0.101	0.062	0.140
0.214	0.200	0.080	0.123	1.000	0.113	0.150	0.010	0.067	0.054	0.000	0.071	0.099	0.090
0.121	0.045	0.094	0.178	0.113	1.000	0.080	0.010	0.026	0.006	0.060	0.138	0.030	0.030
0.120	0.060	0.044	0.110	0.150	0.080	1.000	0.044	0.070	0.060	0.093	0.070	0.010	0.137
0.400	0.040	0.296	0.050	0.010	0.010	0.044	1.000	0.380	0.340	0.387	0.210	0.100	0.422
0.097	0.142	0.390	0.068	0.067	0.026	0.070	0.380	1.000	0.288	0.260	0.116	0.207	0.440
0.210	0.096	0.150	0.193	0.054	0.006	0.060	0.340	0.288	1.000	0.580	0.223	0.210	0.430
0.170	0.030	0.120	0.120	0.000	0.060	0.093	0.387	0.260	0.580	1.000	0.080	0.100	0.346
0.270	0.118	0.180	0.101	0.071	0.138	0.070	0.210	0.116	0.223	0.080	1.000	0.152	0.390
0.068	0.210	0.160	0.062	0.099	0.030	0.010	0.100	0.207	0.210	0.100	0.152	1.000	0.340
0.230	0.280	0.434	0.140	0.090	0.030	0.137	0.422	0.440	0.430	0.346	0.390	0.340	1.000

Fig. 2. PCC values of each attribute with class attribute of Cleveland dataset

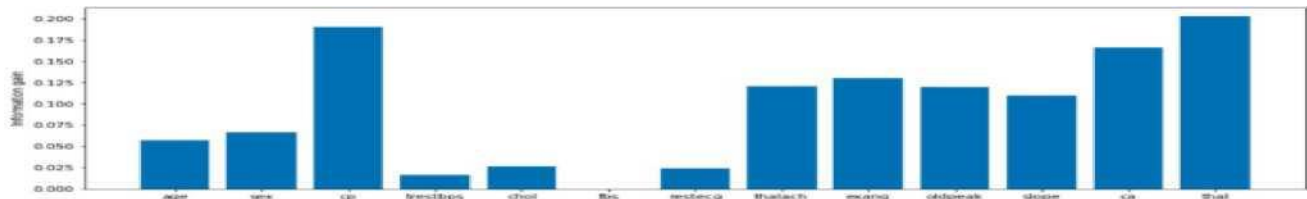


Fig. 3. Information gain of each attribute of statlog dataset

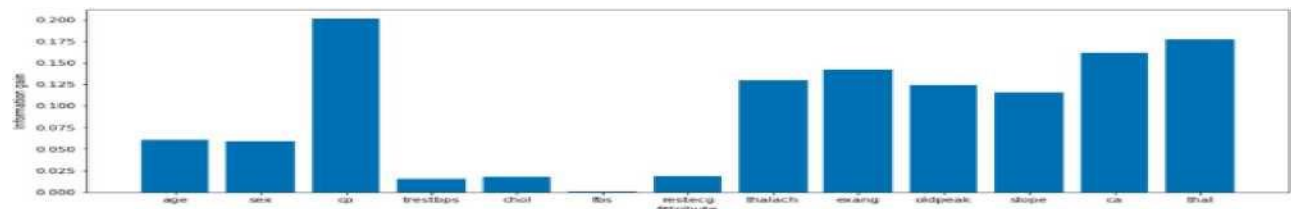


Fig. 4. Information gain of each attribute of Cleveland dataset

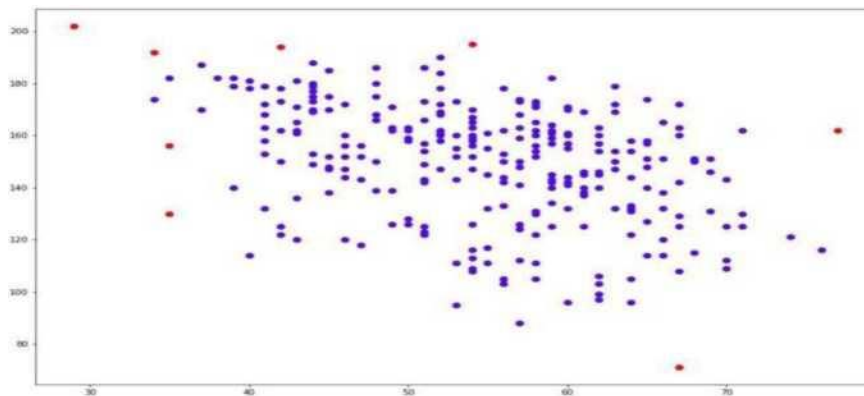


Fig. 5. Plot between attribute thalach and age (red points are outlier points) of statlog dataset.

3.5. DATA BALANCING

To get better results we need balance of dataset .we have some technique like : 1.Over-sampling(increasing data point of minority class)

2. Under-sampling

3. Hybrid

In our case we have used SMMOTEENN that works on the principle of symmetric minority oversampling which edited the nearer neighbor

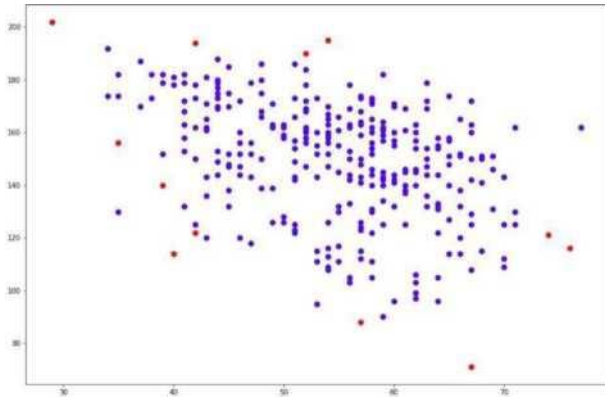


Fig. 6. Plot between attribute thalach and age (red points are outlier points) of Cleveland dataset.

Firstly we do over sampling , but after performing oversampling there may be some overlapping data point , so to remove it again perform under sampling. By applying sampling , we get evenly balanced class distribution which is ready for training.

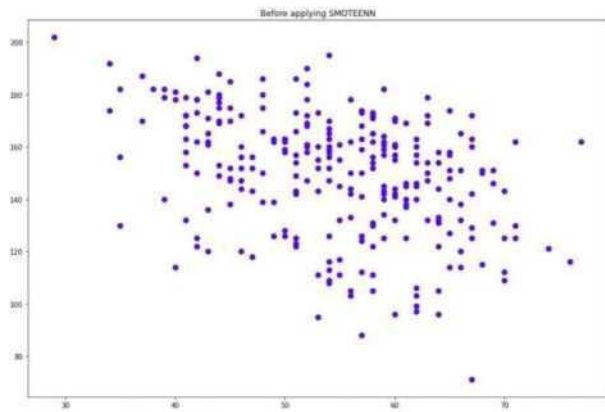


Fig. 7. Dataset(cleveland) before applying SMOTEENN

After applying sampling the dataset is evenly balanced class distributions.

3.6. GENETIC ALGORITHM

Firstly we have performed parameter training before training og model. We use genetic algorithm for parameter training. After applying gnrntic algorithm , we use the output for training our model.

3.7. TRAINING OF MODEL

We used XGboost algorithm for training of our model.It is a gradient boosting decision tree algorithm.It follows sequential model. It is basically designed for giving a high computational speed along with better model efficiency.We

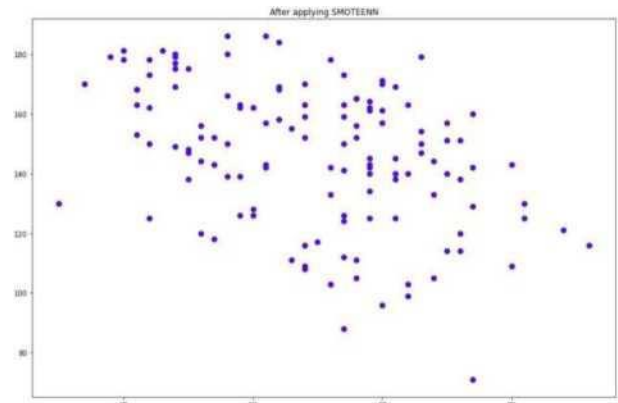


Fig. 8. Dataset(cleveland) After applying SMOTEENN

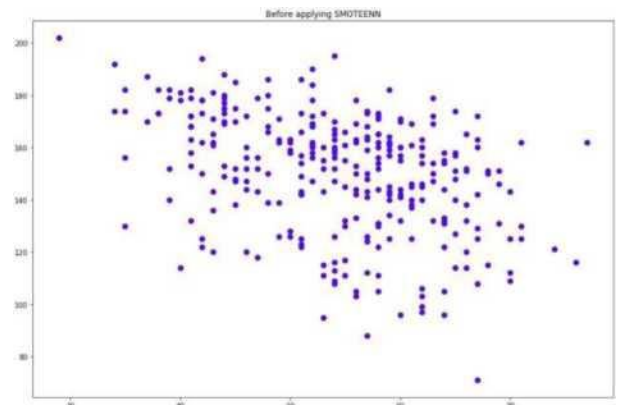


Fig. 9. Dataset(cleveland) before applying SMOTEENN

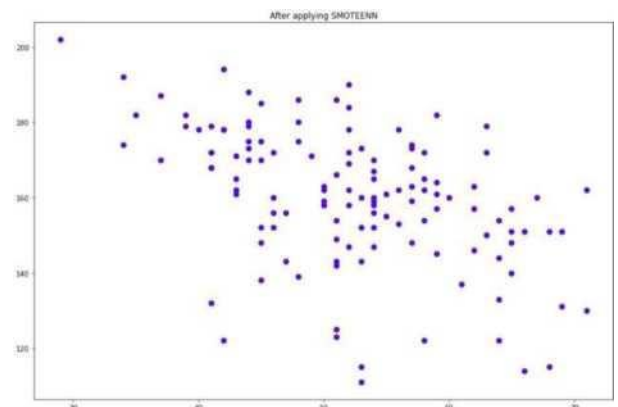


Fig. 10. Dataset(cleveland) After applying SMOTEENN

need to specify many parameter (i.e learing rate, n_estimator, gamma,regalpha etc) at the time of training.The output of genetic algorithm will be used as the parameters for the XGboost while training of the model.

3.8. PERFORMANCE EVALUATION

We used NINE performance metrics to evaluate the performance of the proposed model.accuracy is the number of correct prediction per total number of prediction.F1_score is

```

searchspace = list()
search_space.append(Real(1e-6, 1.0, 'log-uniform', name='learning rate'))
search_space.append(Integer(1, 5000, name='n_estimators'))
search_space.append(Integer(1, 100, name='max_depth'))
search_space.append(Integer(1, 100, name='min_child_weight'))
search_space.append(Real(1e-6, 100.0, 'log-uniform',
name='gamma'))
search_space.append(Real(1e-6, 1.0, 'log-uniform',
name='subsample'))
search_space.append(Real(1e-6, 1.0, 'log-uniform',
name='colsample_bytree'))
searchspace.append(Categorical(['binary:logistic'],
name='objective'))
search
space.append(Categorical(['auc', 'rmse'], name='eval_metric'))
search_space.append(Real(1e-6, 1.0, 'log-uniform', name='reg
alpha'))
search_space.append(Real(1e-6, 1.0, 'log-uniform',
name='reg_lambda'))
<cause named args(search_space) def evaluate_model(**params):
model = XGBClassifier() model.setparams(*params)
cv = RepeatedStratifiedKFold(n_splits=10, n_repeats=3, random_state=1) result =
cross_val_score(model, X, y, cv=cv, n_jobs=-1,
scoring='accuracy') estimate = mean(result) return 1.0 - estimate
# perform optimization
result = gp_minimize(evaluate_model, search_space) print('Best Accuracy: %.3f' %
(1.0 - result.fun)) print('Best Parameters: %s' % (result.x))

```

Fig. 11. Applying GA for parameter tuning.

```

XGBClassifier(base_score=0.5, boosters='gbtree', colsample_bylevel=1,
colsample_bynode=1, colsample_bytree=1, gamma=0.0,
learning_rate=0.0024674117613464074, max_delta_step=0,
max_depth=1, max_leaf_nodes=None, missing=None,
n_estimators=5000, n_jobs=1, nthread=None,
objective='binary:logistic', random_state=0,
reg_alpha=0.015474296722426327, reg_lambda=1, scale_pos_weight=1,
seed=None, silent=None, subsample=1.0, verbosity=1)

```

Fig. 12. The Optimal parameter for XGboost

the harmonic mean of precision and recall of the model. Then metric like precision score, recall score is used. True positive rate is called as sensitivity of the model. True negative rate is called as specificity of the model. Matthews correlation coefficient is used, value 1 signify perfect prediction and value -1 signify inverse prediction and value 0 signify average prediction. Metric like TPR, TNR, FPR, FNR are also used for the performance evaluation.

Metrics are:

- Accuracy
- F1 Score
- Recall Score
- Precision Score
- MCC(Matthews correlation coefficient)
- TPR(true positive rate)
- TNR(true negative rate)
- FPR(false positive rate)
- FNR(false negative rate)

We are getting around 97% accuracy for the statlog dataset and around 98% accuracy for the cleveland dataset.

4. RESULT AND ANALYSIS

The proposed HDPM was applied to both datasets and showed positive results for increasing the prediction accuracy as compared to other models. We selected 10 ML techniques that have been widely used in the research community and have a proven track record for accuracy and efficiency for comparison.

Accuracy: 96.88%

F1 Score: 0.9675612964327055 Recall score: 0.9632418300653594 Precision score: 0.9742640692640692 MCC: 0.9387365353669066 TPR(sensitivity): 0.9657352941176469 TNR(specificity): 0.9742640692640692 FPR: 0.03426470588235294 FNR: 0.025735930735930734

Fig. 13. Result for statlog dataset

Accuracy: 97.02%

F1 Score: 0.9706031617010465 Recall score: 0.9747042424967579 Precision score: 0.9679695531502936 MCC: 0.9410976312565761 TPR(sensitivity): 0.9732734477832031 TNR(specificity): 0.9679695531502936 FPR: 0.02672655221679653 FNR: 0.03203044684970584

Fig. 14. Result for cleveland dataset

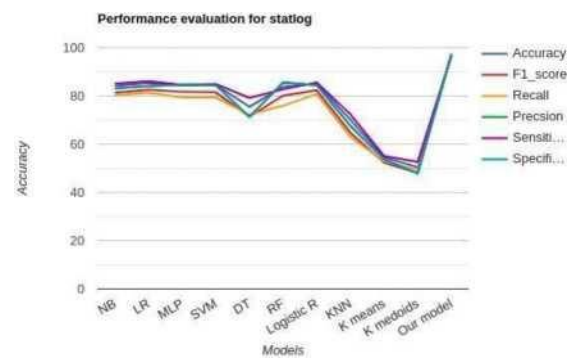


Fig. 15. Line chart for performance evaluation on Statlog dataset

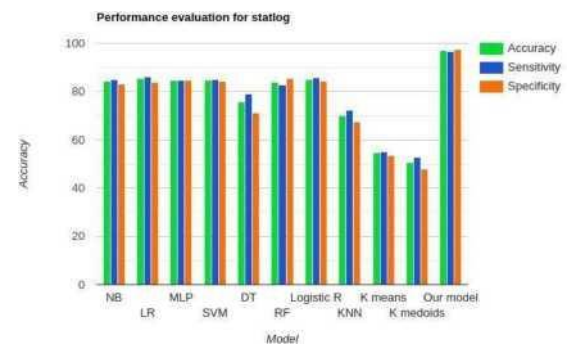


Fig. 16. Bar chart for performance evaluation on Statlog dataset

Model	Accuracy(%)	F1 score(%)	Recall (%)	precision (%)	MCC	TPR(%)	TNR(%)	FPR(%)	FNR(%)
NB	84.20	81.37	80.16	83.14	0.67	85.11	83.14	14.88	16.86
LR	85.30	82.60	81.38	84.07	0.70	86.16	84.07	13.83	15.92
MLP	84.54	81.72	79.41	84.70	0.68	84.57	84.69	15.42	15.30
SVM	84.59	81.53	79.34	84.45	0.68	84.92	84.45	15.08	15.54
DT	75.53	71.40	72.30	71.27	0.50	79.16	71.27	20.83	28.73
RF	83.81	80.13	75.90	85.57	0.67	82.77	85.57	17.22	14.42
Logistic R	85.05	82.35	80.81	84.45	0.69	85.64	84.45	14.35	15.54
KNN	70.16	64.80	63.05	67.60	0.39	72.35	67.60	27.65	32.40
K meanscluster	54.54	52.32	52.90	53.67	0.086	55.12	53.67	44.87	46.32
K medoids	50.41	48.01	49.31	47.86	0.005	52.65	47.86	47.34	52.13
Our Model	96.88	96.75	96.32	97.42	0.94	96.57	97.42	3.42	2.57

Table 1: Performance evaluation for statlog

Model	Accuracy(%)	F1 score(%)	Recall (%)	precision (%)	MCC	TPR(%)	TNR(%)	FPR(%)	FNR(%)
NB	82.09	83.72	84.82	82.89	0.64	81.22	82.89	18.77	17.10
LR	83.44	84.83	85.57	84.27	0.67	82.50	84.27	17.49	15.72
MLP	70.18	66.70	66.03	79.79	0.45	70.74	79.79	29.26	20.20
SVM	81.92	84.09	89.38	79.66	0.64	85.53	79.66	14.46	20.33
DT	76.13	78.19	79.18	77.67	0.51	74.57	77.67	25.42	22.32
RF	82.48	84.78	89.36	80.94	0.65	85.16	80.94	14.83	19.05
Logistic R	82.78	84.79	88.18	81.95	0.65	84.41	81.95	15.58	18.04
KNN	65.39	68.83	70.84	67.52	0.30	63.11	67.52	36.88	32.47
K meanscluster	51.03	50.08	51.01	50.21	0.017	51.44	50.21	48.55	49.78
K medoids	52.03	52.43	52.50	53.33	0.042	50.80	53.33	49.20	46.66
Our Model	97.02	97.06	97.47	96.80	0.94	97.32	96.80	2.67	3.20

Table 2: Performance evaluation for Cleveland

Performance evaluation for statlog

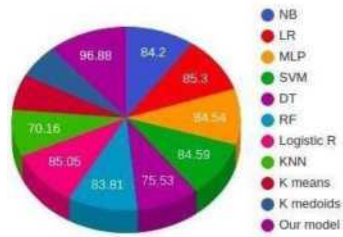


Fig. 17. Pie chart for accuracy on statlog dataset

Performance evaluation for cleveland

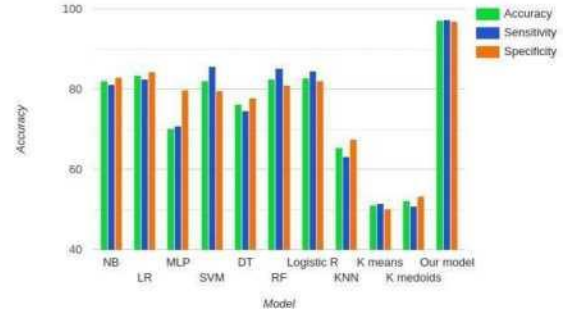


Fig. 19. Bar chart for performance evaluation on Cleveland dataset

Performance evaluation for cleveland

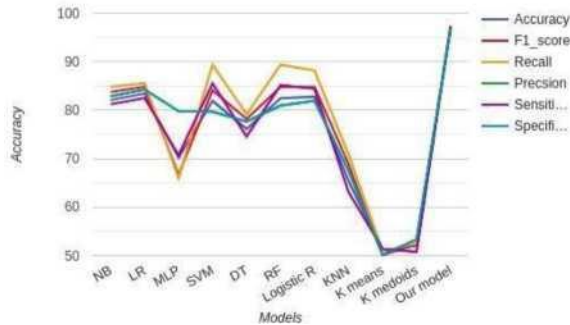


Fig. 18. Line chart for performance evaluation on Cleveland dataset

Performance evaluation for cleveland

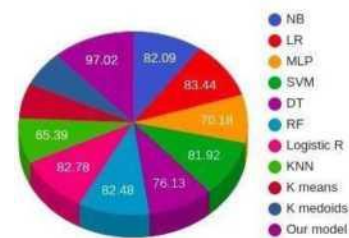


Fig. 20. Pie chart for accuracy on Cleveland dataset

5. CONCLUSION

Heart disease prediction is a very wonderful technique to the mankind. By using this many life's can be saved. In this modern era, with help of technology we can make impossible things as possible. So our heart disease model is a hybrid model combination of different ML technique like DBSCAN, SMOTEENN, GA, XGboost to get a efficient model with high accuracy in prediction. If the people start using it, then the rate of people dying due to heart disease will reduced eventually.

REFERENCES

- [1] World Health Organization. (2017). Cardiovascular Diseases (CVDs).
- [2] E. J. Benjamin et al., "Heart disease and stroke statistics—2019 Statistics Korea. (2018). Causes of Death Statistics in 2018. [Online].
- [4] World Health Organization. (2017). Cardiovascular Diseases (CVDs).
- [5] P. Greenland, J. S. Alpert, G. A. Beller, E. J. Benjamin, M. J. Budoff, Z. A. Fayad, E. Foster, M. A. Hlatky, J. M. Hodgson, F. G. Kushner, M. S. Lauer, L. J. Shaw, S. C. Smith, A. J. Taylor, W. S. Weintraub, and N. K. Wenger, "2010 ACCF/AHA guideline for assessment of cardiovascular risk in asymptomatic adults.
- [6] J. Perk et al., "European guidelines on cardiovascular disease prevention in clinical practice (version 2012)
- [7] G.-M. Park and Y.-H. Kim, "Model for predicting cardiovascular disease: Insights from a Korean cardiovascular risk model," *Pulse*, vol. 3, no. 2, pp. 153-157, 2015, doi: 10.1159/000438683.
- [8] G. J. Njie, K. K. Proia, A. B. Thota, R. K. C. Finnie, D. T. Lackland, and T. E. Kottke, "Clinical decision support systems and prevention," *Amer. J. Preventive Med.*, vol. 49, no. 5, pp. 784-795, Nov. 2015.
- [9] V. Sintchenko, E. Coiera, J. R. Iredell, and G. L. Gilbert, "Comparative impact of guidelines, clinical data, and decision support on prescribing decisions: An interactive Web experiment with simulated cases,"
- [10] K. Kawamoto, C. A. Houlihan, E. A. Balas, and D. F. Lobach, "Improving clinical practice using clinical decision support systems: A systematic review of trials to identify features critical to success,"
- [11] J.R. Quinlan. 1993, C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo-California.
- [12] Machine learning based decision support systems (DSS) for heart disease diagnosis: a review. Online: 25 March 2017 DOI: 10.1007/s10462-01.
- [13] Alexander, Cheryl Wang, Lidong. (2017). Big Data Analytics in Heart Attack Prediction. *Journal of Nursing Care*. 06. 10.4172/2167-1168.1000393
- 14 Data Science Bowl. 2020. Transforming how I measure heart disease. [ONLINE] Available at: <https://datasciencebowl.com/transforming-how-we-diagnose-heart-disease/>. [Accessed 2 February 2020]
- 15 Wynne Hsu, Mong-Li Lee, Bing Liu, Tok Wang Ling, 2000, "Exploration mining in diabetic patients databases: findings and conclusions", *KDD 2000*: pp: 430-436.