IJRAR.ORG E-ISSN: 2348-1269, P-ISSN: 2349-5138

INTERNATIONAL JOURNAL OF RESEARCH AND ANALYTICAL REVIEWS (IJRAR) | IJRAR.ORG

An International Open Access, Peer-reviewed, Refereed Journal

Design And Deploy Heart Disease Prediction By Using XGBOOST

[1]Vishal, [2]Karan, [3] Gaurav Siwal

[1] Data Science Executive, [2]Data Science Executive, [3]Data Scientist

Orangus Pvt. Ltd, Delhi, India

ABSTRACT : Heart disease is one of the most significant causes of mortality in the world today. Prediction of cardiovascular disease is a critical challenge in the area of clinical data analysis. Machine learning has been shown to be effective in assisting in making decisions and predictions from the large quantity of data produced by the healthcare industry. Heart disease, one of the major causes of mortality worldwide, can be mitigated by early heart disease diagnosis[1]. Exploratory Data Analysis (EDA) detects mistakes, finds appropriate data, checks assumptions and determines the correlation among the explanatory variables. We analyzed the data properly and found those measurements, by not doing which we can avoid heart disease. A hybrid Synthetic Minority Over-sampling Technique-Edited Nearest Neighbor (SMOTE) to balance the training data distribution and XGBoost to predict heart disease. Make different models so that good accuracy can be found. (logistic regression (LR), ridge regression (RR),grid search (GS), naive bayes (NB), decision tree (DT), random forest (RF), gradient boosting(GB), k-nearest neighbor (KNN), artificial neural network (ANN) and xgboost (XGB)). achieving accuracy of 92.27% datasets, respectively.

Data Source :

S.No.	Observation	Descriptions	values
1.	Heart Disease	Heart Disease	Yes/No
2.	BMI	Body Mass Index	Continuous
3.	Smoking	Inhaling and exhaling	Yes/No
4.	Alcohol Drinking	Ethanol or ethyl	Yes/No
5.	Stroke	Block blood supply	Yes/No
6.	Physical Health	Physical activity level	Continuous
7.	Mental Health	Emotional, psychological	Continuous
8.	Diff Walking	Abnormal, uncontrollable	Yes/No
9.	Sex	Characteristics women men	Male/Female

10.	Age Category	Age in Years	Continuous
11.	Race	Physical or social qualities	Six types
12.	Diabetic	Blood Sugar	Yes/No
13.	Physical Activity	Body Movement	Yes/No
14.	Gen Health	Condition of their body	Five types
15.	Sleep Time	Sleep Cycle	Continuous
16.	Asthma	Long-term disease lungs	Yes/No
17.	Kidney Disease	Kidney Failure	Yes/No
18.	Skin Cancer	Abnormal growth of skin	Yes/No

The detailed steps, including dataset and modules description and the performances metrics are presented.

Methodology :

This paper shows the analysis of various machine learning algorithms, the algorithms that are used in this paper are K nearest neighbors (KNN), Logistic Regression, Random Forest Classifiers and xgboost which can be helpful for practitioners or medical analysts for accurately diagnose Heart Disease[2].

The methodology is a process which includes steps that transform given data into recognized data patterns for the knowledge of the users.

After pre-processing of data, classifier is used to classify the pre-processed data the classifier used in the proposed model are KNN, Logistic Regression, Random Forest Classifier and xgboost[3].

Finally, the proposed model is undertaken, where we evaluated our model on the basis of accuracy and performance using various performance metrics[8]. Here in this model, an effective Heart Disease Prediction System has been developed using different classifiers.

The predictions model's performance is evaluated using different parameters, such as accuracy, precision, recall, and F-measure. The model that gives the highest prediction accuracy, precision, recall and F-measure is selected[4]. The accuracy metric assesses the precision or correctness of a machine learning or classifier model's predictions[12]. Mathematically, it given by equation

$$\text{Accuracy} = \frac{\text{true positive (TP)} + \text{true negative(TN)}}{\text{TP} + \text{TN} + \text{false negative (FN)} + \text{false positive(FP)}}$$

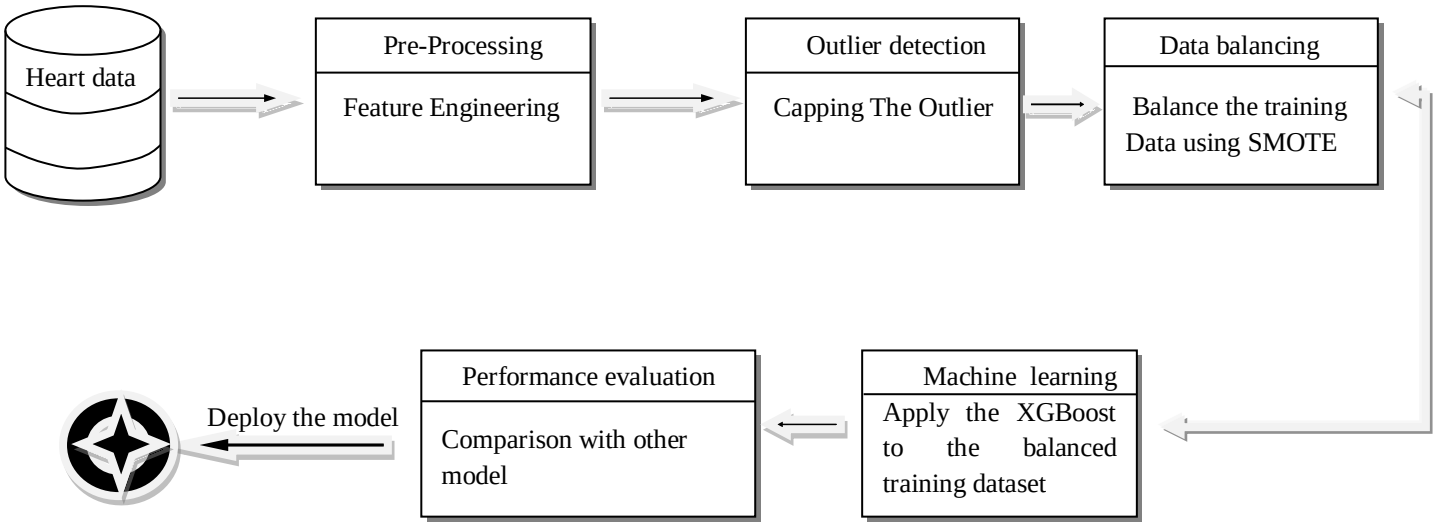
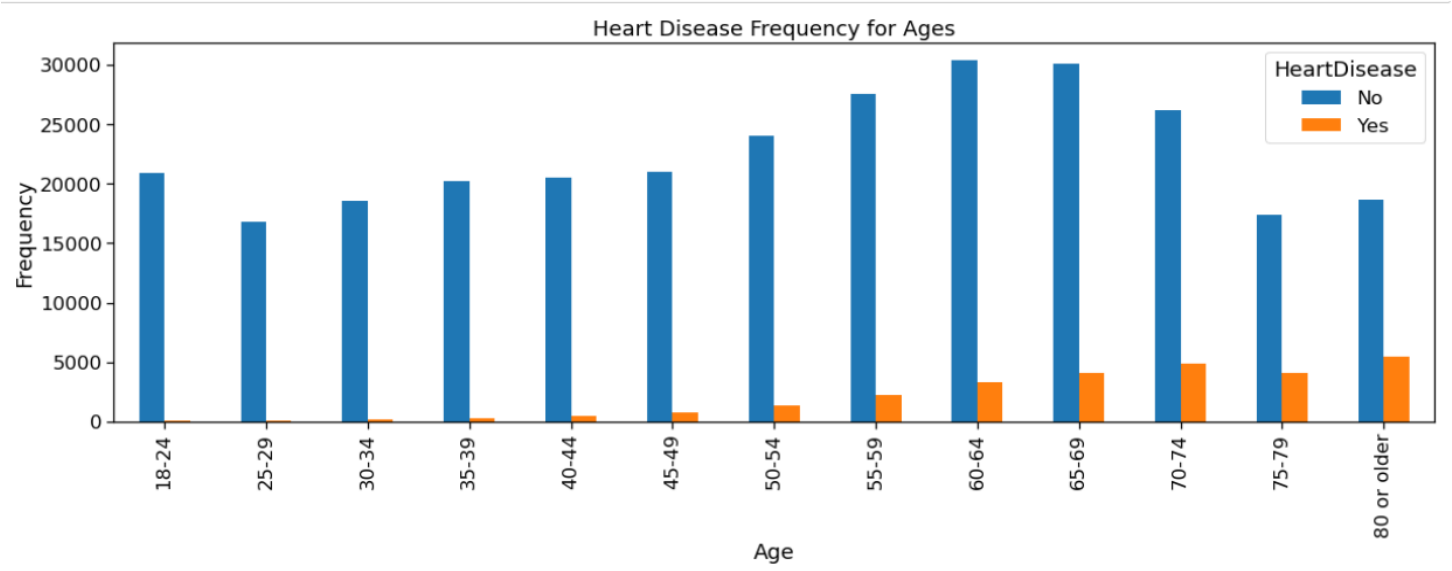


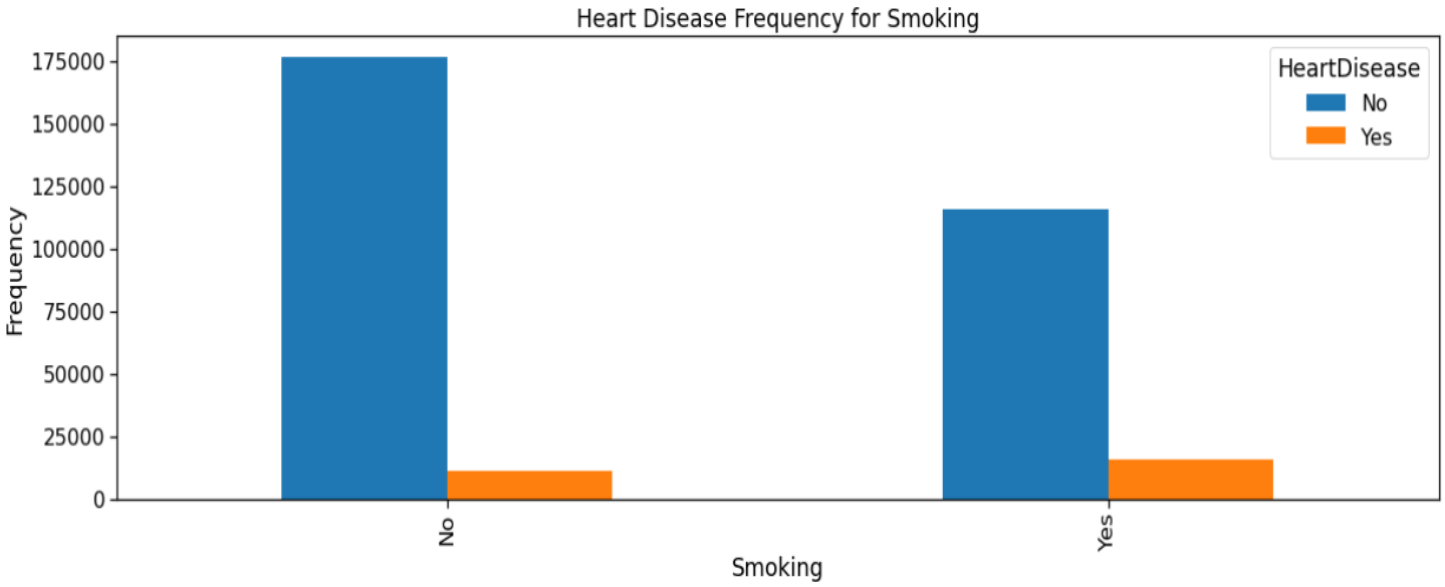
Fig.1 The proposed Heart Disease prediction Model

Manual Exploration:

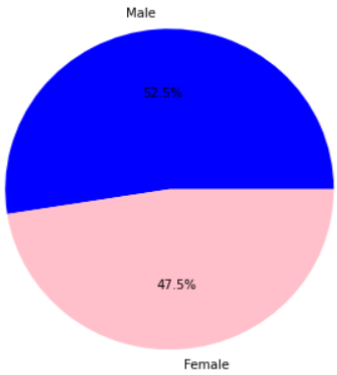
Data exploration or Manual Exploration is the initial step in data analysis, where users explore large data set in an unstructured way to uncover initial patterns, characteristics, and points of interest[5]. This process isn't meant to reveal every bit of information a data set holds, but rather to help create a broad picture of important trends and major points to study in greater detai[9]l.



A bar chart is used when you want to show a distribution of data points or perform a comparison of metric values across different subgroups of your data. From a bar chart, we can see which groups are highest or most common, and how other groups compare against the others. The bar graph shows comparison between Heart disease and Age.



This graph shows smoking effect on heart disease is a major cause of heart and blood vessel disease. Smoking is a major risk factor for heart disease.



A pie chart (or a circle chart) is a circular statistical graphic, which is divided into slices to illustrate numerical proportion. In a pie chart, the arc length of each slice (and consequently its central angle and area) is proportional to the quantity it represents to describe percentage Male and Female on the given dataset.

Model:

From these results we can see that although most of the researcher are using different algorithms such as Logistic Regression, Decision tree, Neural network, KNN, Random forest, XGBOOST[10].

After we balanced the training datasets using smoth[6]. We used the extreme gradient boosting (XGBOOST) algorithm to detect the presence or absence of heart disease. XGBOOST is a type of supervised machine learning used for classification and regression modelling . XGBOOST is an enhanced algorithm based on the implementation of gradient boosting[7]. We implemented XGBOOST using the XGBOOST python library. The outlier data from heart disease training datasets and Smoth is used to balance the training dataset[11]. Finally, XGBOOST is used to learn from the training dataset.

Performance evaluations

Model	accuracy	Recall	precision	F1 Score
LR	78%	80%	78%	79%
RR	79%	80%	78%	79%
GS	79%	76%	80%	79%
NB	70%	90%	64%	75%
DT	87%	89%	86%	88%
RF	90%	92%	88%	90%
GB	91%	89%	93%	91%
KNN	85%	91%	81%	86%
ANN	79%	82%	78%	80%
XGBOOST	92%	92%	93%	92%

Results & Discussions:

A model has been developed using ML classification modeling techniques. The algorithms that we used are more accurate, saves a lot of money it is cost efficient and faster than the algorithms that the previous researchers used. Moreover, the maximum accuracy obtained by XGBOOST are equal to 92% which is greater.

References :

[1] An overview of heart disease abstract. Available: https://ieeexplore.ieee.org/document/9302468

[2] Wolgast G, Ehrenborg C, Israelsson A, Helander J, Johansson E & Manefjord H (2016). *Wireless*

[3] T. Chen and C. Guestrin, ‘‘XGBoost: A scalable tree boosting sys- tem,’’ in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, San Francisco, CA, USA, Aug. 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[4] Powers, David M. W. (2011). "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation". *Journal of Machine Learning Technologies*. **2** (1): 37–63. S2CID 55767944.

[5] FOSTER Open Science, Overview of Data Exploration Techniques: Stratos Idreos, Olga Papaemmonouil, Surajit Chaudhuri.

[6] Simonoff, Jeffrey S. (1998) *Smoothing Methods in Statistics*, 2nd edition. Springer ISBN 978-0387947167[*page needed*]

[7] "Tree Boosting With XGBoost – Why Does XGBoost Win "Every" Machine Learning Competition?". *Synced*. 2017-10-22. Retrieved 2020-01-04

[8] Stehman, Stephen V. (1997). "Selecting and interpreting measures of thematic classification accuracy". *Remote Sensing of Environment*. **62** (1): 77–89. Bibcode:1997RSEnv..62...77S. doi:10.1016/S0034-4257(97)00083-7

[9] Mitchell, Tom (1997). *Machine Learning*. New York: McGraw Hill. ISBN 0-07-042807-7. OCLC 36417892

[10] Hu, J.; Niu, H.; Carrasco, J.; Lennox, B.; Arvin, F., "Voronoi-Based Multi-Robot Autonomous Exploration in Unknown Environments via Deep Reinforcement Learning" IEEE Transactions on Vehicular Technology, 2020.

[11] Ruan, Da; Chen, Guoqing; Kerre, Etienne (2005). Wets, G. (ed.). *Intelligent Data Mining: Techniques and Applications*. Studies in Computational Intelligence Vol. 5. Springer. p. 318. ISBN 978-3-540-26256-5.

[12] Mitchell, Tom (1997). *Machine Learning*. New York: McGraw Hill. ISBN 0-07-042807-7. OCLC 36417892