Visualizing and Predicting Heart Diseases with an Interactive Dash Board

Project Report

Submitted By:-

ADITYA SAI YOGEESWAR PYDA 20BLC1129 VIT CHENNAI

SRI HARSHAVARDHAN PALLA 20BCE1308 VIT CHENNAI

YELALA NITHIN REDDY 20BCI7305 VIT-AP

PRANAV NUDURUPATI 20BEC0242 VIT VELLORE

VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

1 INTRODUCTION

1.1 Overview

Heart disease is a major public health concern that affects millions of people worldwide. Early detection and accurate prediction of heart disease risk are critical for prevention and timely intervention, particularly in high-risk populations. Traditional risk prediction models rely on clinical risk factors such as age, gender, blood pressure, cholesterol levels, and smoking status. They often focus on symptom relief rather than addressing the underlying causes of the condition. While they can provide significant relief, they are not always effective in treating the root cause of the disease. Additionally, these methods may have side effects and can be costly and time-consuming. However, these models have limitations in accurately predicting heart disease risk, and new approaches are required to improve risk assessment.
In recent years, machine learning techniques have shown promise in improving heart disease risk prediction, leveraging large and diverse datasets to identify novel risk factors and improve accuracy. This paper emphasises the significance of heart disease prediction, the datasets used in these studies, evaluation metrics, and the potential of machine learning in improving risk assessment, early detection, and prevention of heart disease.

1.2 Purpose

Cardiovascular diseases (CVDs) are a group of conditions that affect the heart and blood vessels. They are the leading cause of death worldwide, killing an estimated 17.9 million people each year, accounting for 31% of all deaths worldwide. Machine learning can be a powerful tool in detecting a potential heart disease diagnosis, which is critical in the treatment and management of cardiovascular diseases.
To identify patterns and trends, machine learning algorithms can analyse massive amounts of data from a variety of sources. Heart disease is a complicated condition with many risk factors, such as age, gender, blood pressure, cholesterol levels, and family history. All of these risk factors can be combined by machine learning models to generate a personalised risk score for each individual, which can aid in the early detection and prevention of heart disease. The model can update its predictions as more data becomes available, improving its accuracy over time. This adaptive learning approach can assist healthcare providers in staying current on the latest research and best practises, allowing them to provide better care to their patients. Machine learning algorithms can assist healthcare providers in making better decisions. Machine learning models can help healthcare providers prioritise patients for further testing or treatment by providing accurate and timely predictions, resulting in better patient outcomes and more efficient use of healthcare resources.
To summarise, machine learning is a useful tool in the prediction of heart disease. Machine learning algorithms have the potential to significantly improve the early detection and prevention of heart disease by analysing vast amounts of data, learning and adapting over time, providing new insights, and assisting healthcare providers in making more informed decisions.

2 LITERATURE SURVEY

2.1 Existing problem

Here are some of the articles which are published on existing solutions.

1."A comparative study of machine learning algorithms for heart disease prediction" by Raza et al. (2019)
The authors compare and contrast various machine learning models such as decision trees, random forests, support vector machines, and neural networks. In terms of accuracy and sensitivity, they discover that random forests and neural networks outperform other models. They also discover that incorporating feature selection techniques, such as the Chi-Square test, can improve machine learning model performance.
According to the findings, machine learning algorithms can accurately predict the risk of heart disease, and random forests and neural networks are the most effective models for this task. These models, according to the authors, can be used to identify patients at high risk of heart disease and develop targeted interventions to prevent it. They also recommend that future research look into the use of other feature selection techniques and evaluate the performance of these models in real-world clinical settings.

2."Predicting coronary heart disease using machine learning techniques" by Rana et al. (2021)
The study makes use of a patient database that contains demographic data, medical histories, and test results from clinical trials.
The authors assess a number of machine learning models, such as support vector machines, decision trees, random forests, and logistic regression. They discover that the random forest model performs better than other models in terms of sensitivity and accuracy.
Age, sex, smoking status, and cholesterol levels are just a few of the risk factors the study assesses for predicting coronary heart disease. The most significant risk variables for predicting coronary heart disease, according to the experts, are age and smoking habits.
According to the study's findings, machine learning algorithms can correctly estimate the likelihood of coronary heart disease and pinpoint significant risk factors for the condition. According to the authors, these models can be used to identify patients at high risk of coronary heart disease and develop targeted interventions to prevent the disease. They also note the importance of further research to evaluate the performance of these models in diverse populations and clinical settings.

3. "A machine learning approach for the prediction of cardiovascular disease based on risk factors," Kamble et al. (2021)
T A sample of 6,911 people aged 40 to 79 years old from the National Health and Nutrition Examination Survey (NHANES) were used in the study. The authors used nine risk factors, including age, sex, smoking status, systolic blood pressure, total cholesterol, HDL cholesterol, diabetes status, body mass index, and race/ethnicity, to predict the 10-year risk of CVD. These risk factors included age, sex, smoking status, systolic blood pressure, total cholesterol, HDL cholesterol, diabetes status, and body mass index.

All four machine learning algorithms had strong predictive accuracy, with the Random Forest algorithm doing the best, according to the study's findings. The three most significant risk factors for predicting CVD, according to the authors, were age, systolic blood pressure, and smoking status.
The authors suggest that their machine learning approach could be used to identify individuals at high risk of CVD and to provide targeted interventions to reduce their risk. They also acknowledge that their study has some limitations, such as the use of cross-sectional data and the lack of external validation of their model.

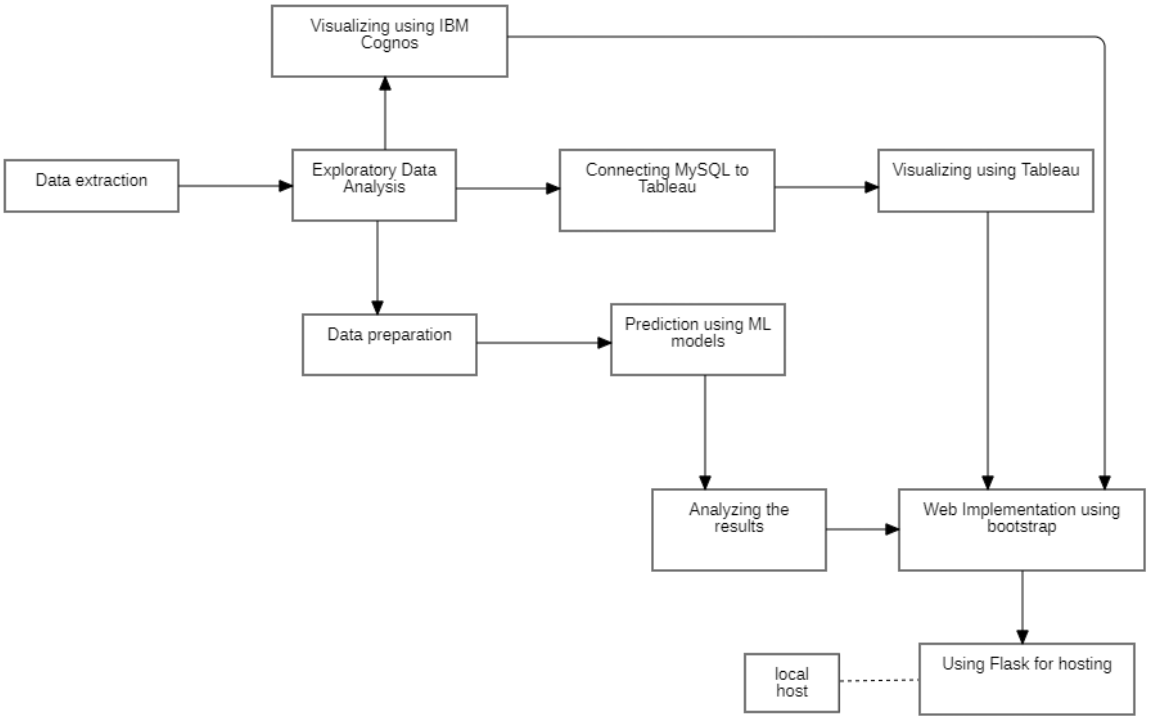
4.“An Intelligent Heart Disease Prediction System Using K-Means Clustering and Naïve Bayes Algorithm”
The Cleveland Heart Disease dataset, which has 303 records and 14 attributes, is used by the authors. They handle missing values, normalise the data, and choose the best features as part of the preprocessing. The data is then clustered into four categories using K-means clustering based on the similarity of the attributes. Then, each cluster is categorised using the Naive Bayes method into one of two categories: heart disease presence or absence.
The findings demonstrate that the suggested intelligent heart disease prediction system outperforms previous classification algorithms with an accuracy of 93.07%. The authors further examine the significance of the features in predicting the likelihood of developing heart disease and discover that the most crucial features are the type of chest discomfort, the highest heart rate attained, and number of major vessels.
The authors suggest that their proposed system can be used for early diagnosis and prevention of heart disease. The system is easy to use and can be integrated into existing healthcare systems for efficient and accurate prediction of heart disease.

2.2 Proposed solution

The solution we are suggesting is the use of tableau and IBM Cognos for better understanding of the dataset and improve the accuracy of the ML solution used for prediction. We are using the voting system for predicting heart disease, the use of regular ML algorithms will give some generic predictions with good but not the best accuracy using a voting system among all these various machine learning models will marginally improve the accuracy.

3 THEORITICAL ANALYSIS

3.1 Block diagram



3.2 Hardware / Software designing

Hardware Requirements:-

1. Intel i5 10th and above processor
2. 8GB ram

Software Requirements:-

1. Python
2. Pandas(pyhton package)
3. Tableau
4. IBM Cognos
5. Tabpy

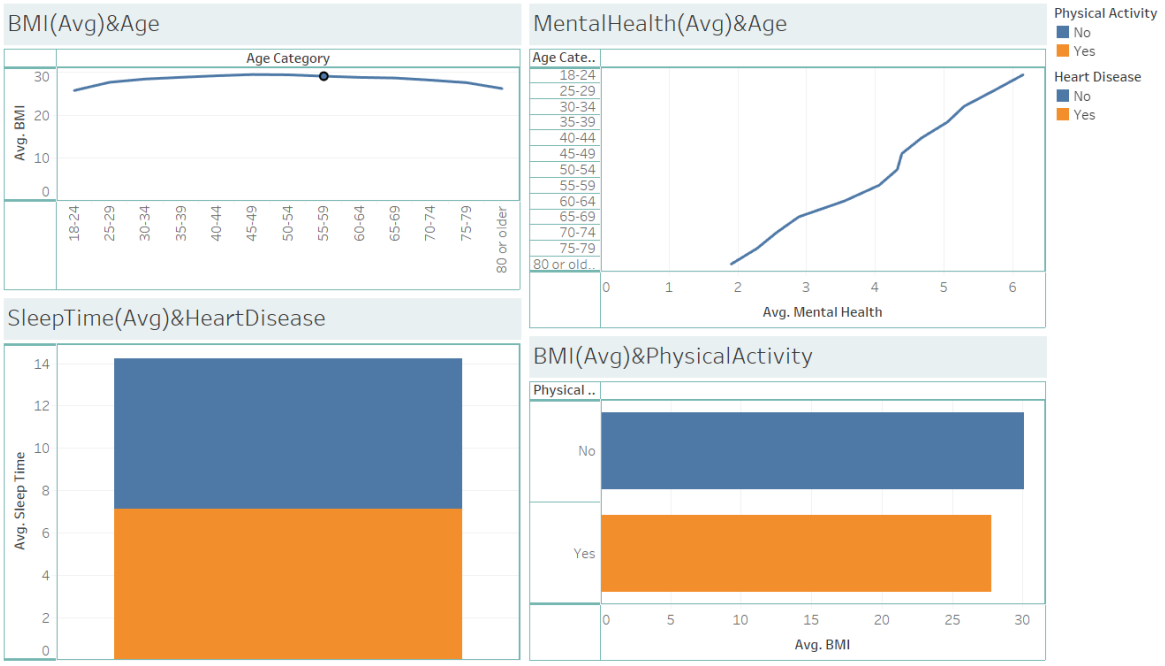
4 EXPERIMENTAL INVESTIGATIONS

We have done the analysis on the dataset using tableau and IBM Cognos below we are attaching all the dashboards and explanation of all the dashboards.

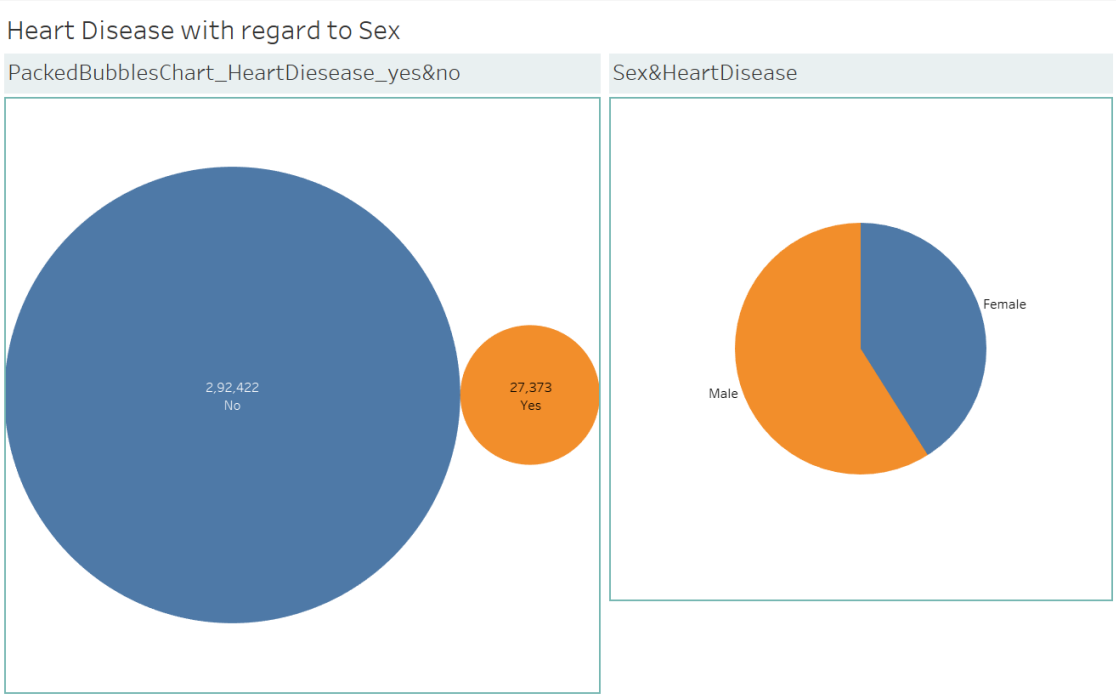
Dashboards using Tableau:

We have implemented a total of 7 dashboards using tableau and we have integrated tableau with MySQL to import the data from the dataset.

The above dashboard show the relationship between Heart Disease and four other major factors that can lead to heart disease we have visualized the number of people who were suffering with heart disease and have the habit of smoking and alcohol consumption, we have also checked for people who were suffering with both asthma and heart disease and skin cancer and heart disease.

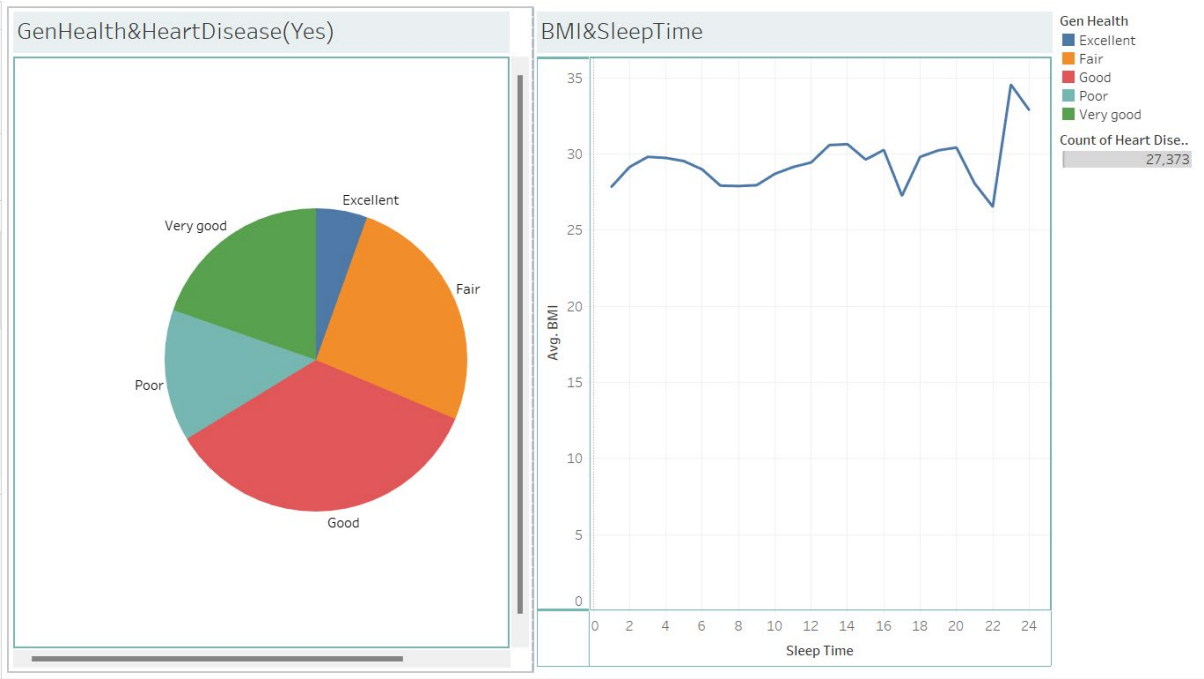


In this above dashboard we have taken two line and two bar visualizations the line visualizations are between the average BMI and age in which we can see a increase in average BMI with age in the first half and then a slight decrease in the curve in the second half of the graph. The second line graph is between age and average mental health in which we can see that with the increase in case the average mental health is decreasing. In the first bar graph which is between average sleep time and heart disease we can see that the average sleep time of people with and without heart disease is almost same so sleep time isn't a major factor effecting heart disease. And in the second bar graph between average BMI and physical activity we can see that people who are having physical activity regularly have average BMI between 25-30 which is healthy.

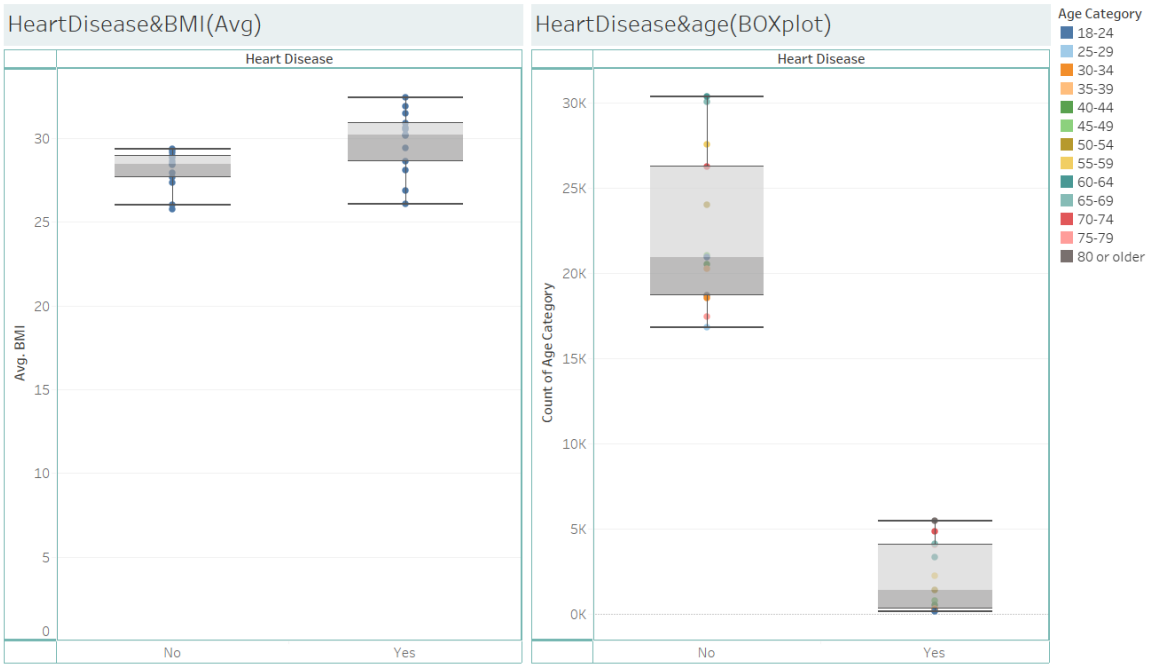


In this particular dashboard we can see that the dataset contains more than 90% of the entries with no heart disease therefore we can say that the dataset is not balanced and biased towards the no class. And in the pie chart we can see that heart disease in men is more common than in women.

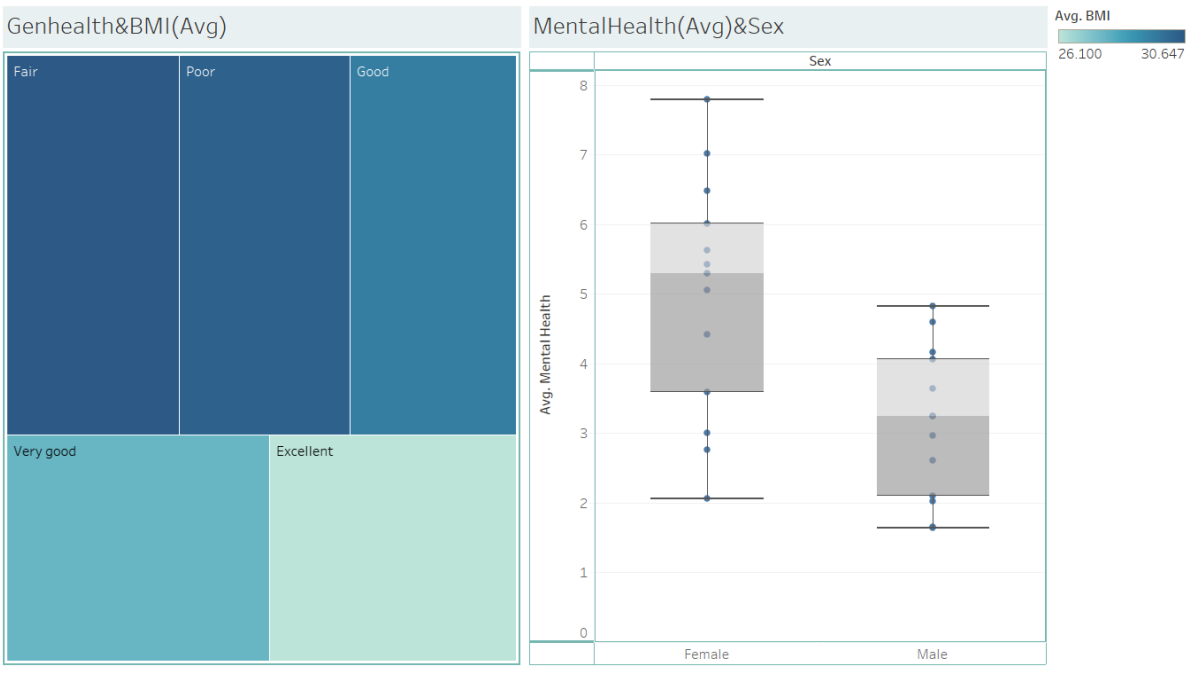
In this dashboard we can see the relation between heart disease and other factors like average mental health, stroke, kidney disease and race. As we know attributes like race do not effect heart disease this attribute can be dropped during the final prediction. From the other bar charts we can come to a conclusion that heart disease is more common among people with better average mental health and other two bar charts who have stroke and heart disease and kidney disease and heart disease.



From the above dashboard we can come to a conclusion that people with excellent general health are the least affected ones with heart disease and from the line graph between Average BMI and sleep time which says when the sleep time is more than 20 average BMI is more than 30 which means the person is obese.



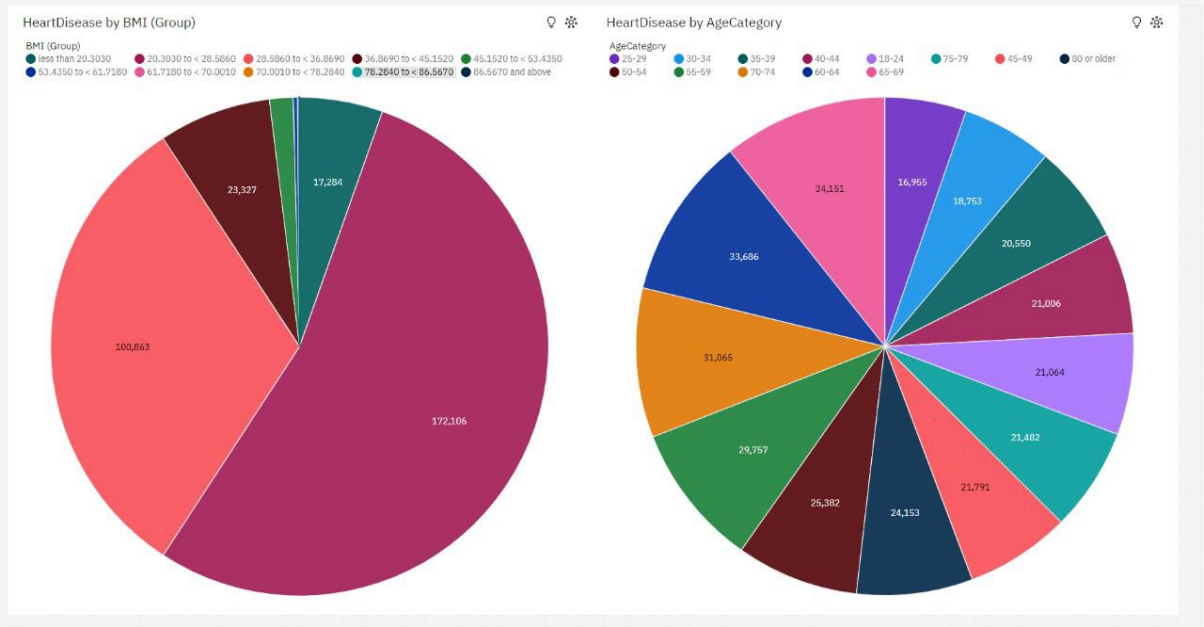
In this above dashboard we have used box plot to visualize the relation between heart disease with average BMI and heart disease with age it provides the maximum and minimum point of the data along with its upper and lower quartile and median.



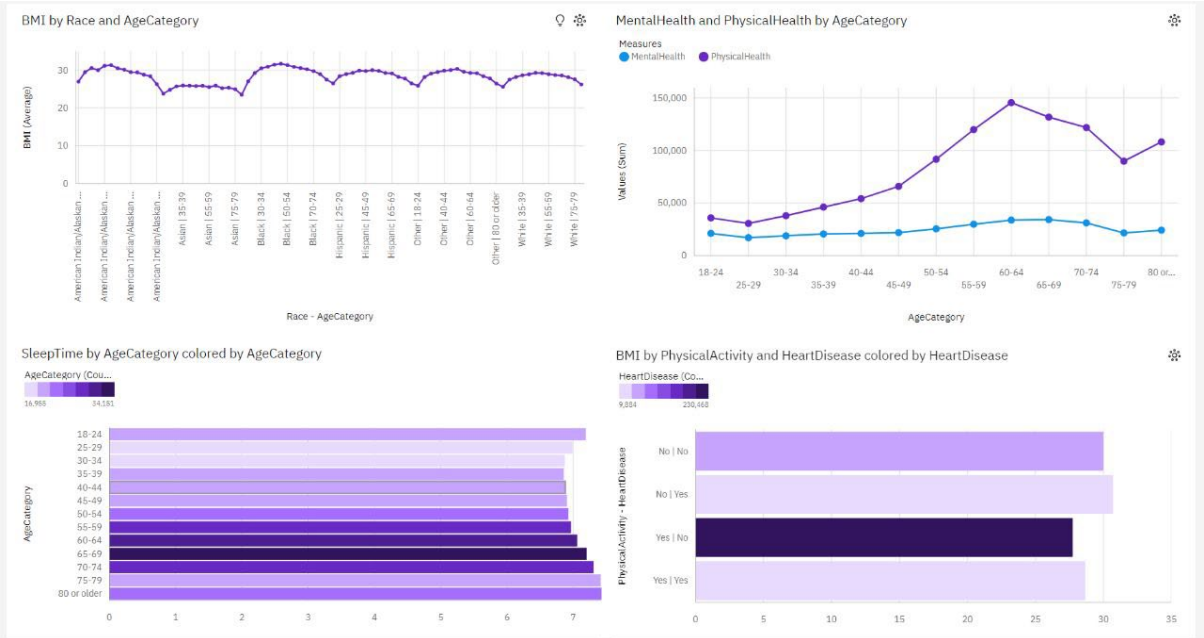
In this final dashboard from tableau we have used heap map to visualize the relationship between general health and Average BMI and we can see that excellent general health has the lowest average BMI nearly 27 which means the person is healthy. In the box plot we can see the maximum and minimum values of mental health and we can see upper and lower quartile and also the median value.

Dashboards using IBM Cognos:-

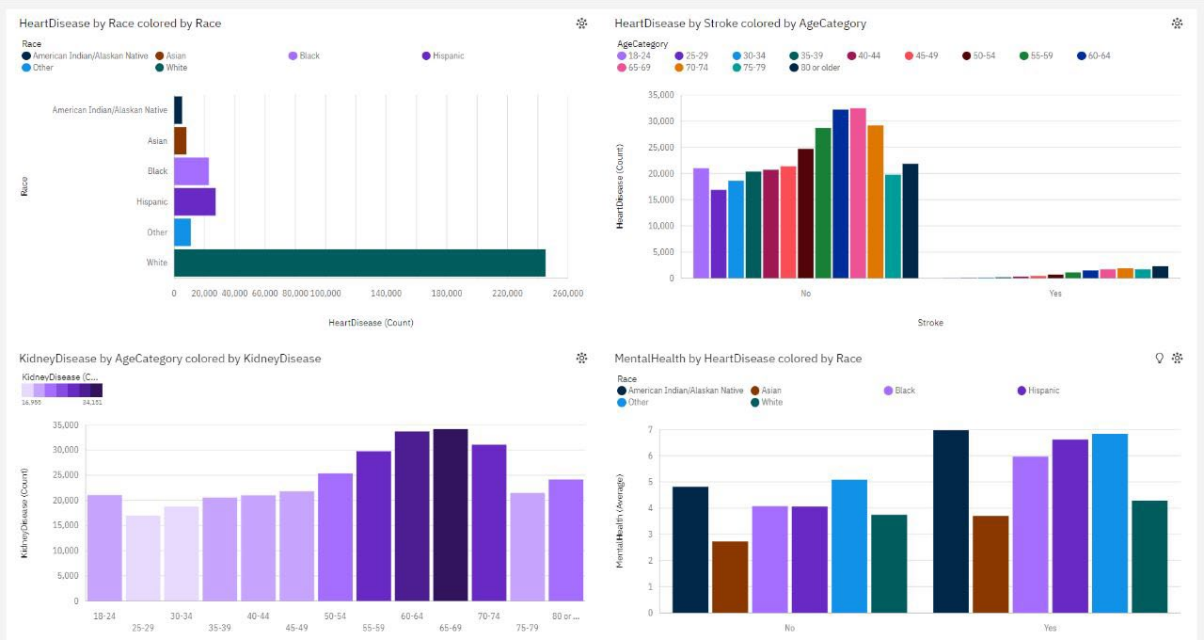
The Dashboards above gives us four data visualizations of different categories helping us see what the dataset tries to convey ultimately. The first visualization tries to tells us how alcohol Drinking has affected different age groups and how it has led then into having heart diseases, also the overall number of results for Heart Disease is nearly 320 thousand. The second visualization tells us how walking regularly decreases the probability of heart diseases. The Third Visualization conveys how Diabetes affects a person with regards to Heart Disease and the fourth Visualization helps us see how Skin Cancer is connected to Heart Disease for Different Age Categories.



This interactive Dashboard has two visualizations where both are pie charts each conveying a different meaning to the dataset available to us. The first pie Chart displays how different categories of BMI's are affected by heart Diseases. And the second Pie chart tells us about the Age Categories and the number of cases in each category.

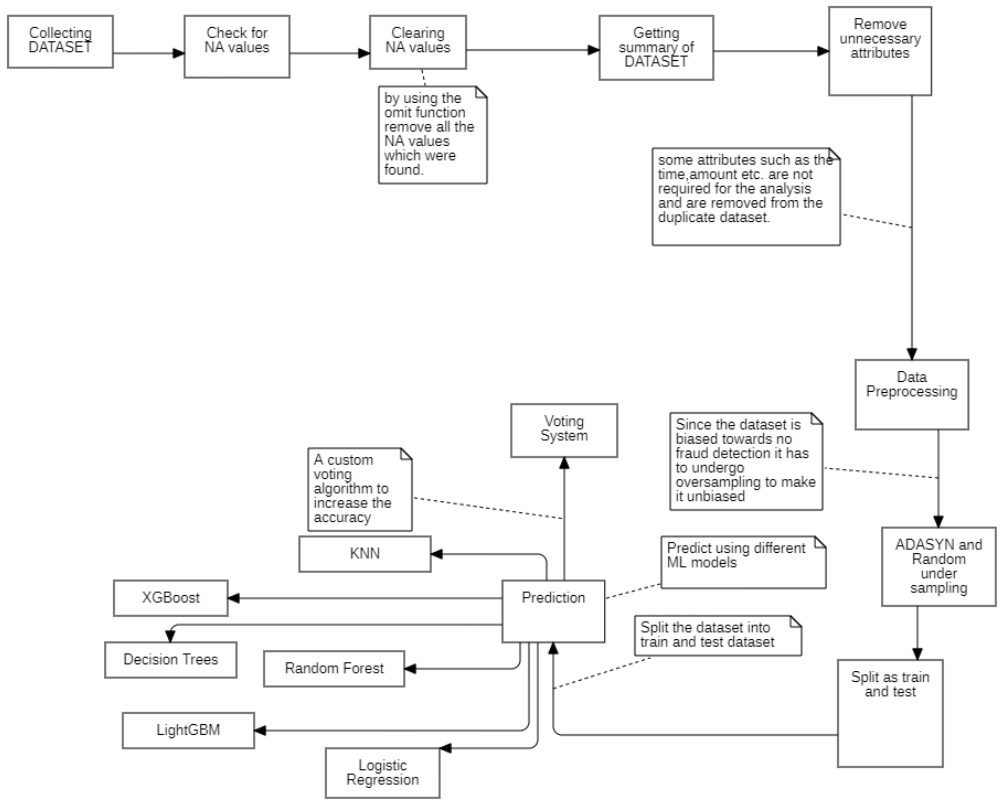


The above Dashboard has four visualizations to see. The first visualization gives us a line graph which tells us the average BMI for different races and age categories of those races. The second Visualization tells us how the physical and mental health varies for different age categories. The third visualization tells us about the average sleep time of different age categories and the fourth visualization tells us about the relation between heart Diseases, physical Activity and BMI.



This Dashboard has four visualizations to see. The first one is about Heart Diseases are affecting different races. The second one is about how stroke occur with respect to heart diseases for different age categories. The Third is about how common is Kidney diseases for different age categories and the fourth one is about how Different age categories are affected by heart disease due to their mental Health.

5 FLOWCHART



6 RESULT

we have used some of the machine learning algorithms on the ADASYN pre-processed such as Decision trees using both Gini Index and Entropy, Random Forest using Both Gini index and Entropy, KNN(K nearest neighbor), XGBoost, Logistic Regression, Light Gradient Boost, then we have taken the models which gave us high accuracy to run a voting system for example if most of the models predict true than the final output of the voting model will be yes and if most of the models predict false than the final output of the voting model will be no this ensures that even if any 1 or 2 models predict the wrong output the final output of the voting model will still be the correct one.
This helped us get a better accuracy than the machine learning models the maximum accuracy we got from the machine models is 88.2% where as the model proposed by us gave 88.3%.

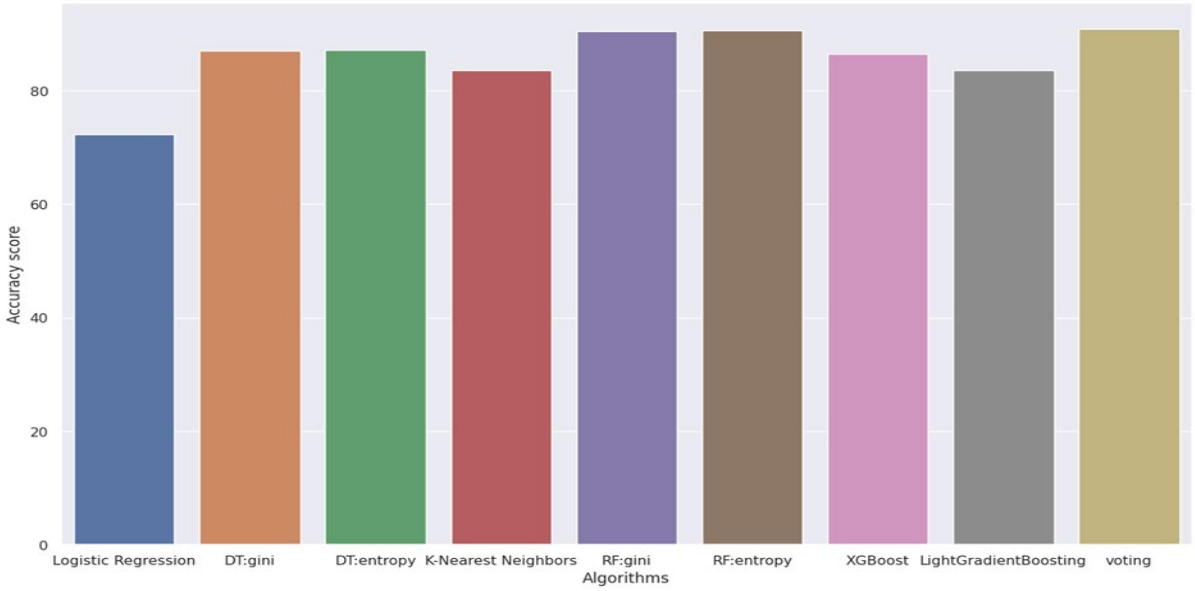


Fig 5: Result Comparison ADASYN

In the voting system for ADASYN preprocessed dataset we have taken XGBoost, Random forest both Gini index and entropy, Decision Tree using both Gini Index and Entropy.

Accuracy obtained from ADASYN:-

Model	**Accuracy**
Logistic Regression	72.57%
Decision Tree(Gini)	84.22%
Decision Tree(Entropy)	84.3%
Random Forest(Gini)	88.21%
Random Forest(Entropy)	88.1%
KNN	83.34%
XGBoost	82.18%
Light Gradient Boosting	77.17%
Voting	88.3%

7 ADVANTAGES & DISADVANTAGES

Advantages:

- Use of over sampling helps improve accuracy.
- Oversampling method leads to no information loss.
- Outperforms under sampling.

Disadvantages:

- Takes a lot of time and resources.
- Might lead to overfitting of the data.

8 APPLICATIONS

Classification Models: ML algorithms can be trained on labeled datasets to classify patients into different categories based on the presence or absence of heart disease. These models use a combination of features such as patient demographics, medical history, and laboratory results to make predictions. Popular algorithms for classification include logistic regression, support vector machines (SVM), random forests, and neural networks.

Risk Stratification: ML models can help stratify patients into different risk categories based on their likelihood of developing heart disease. By analyzing a wide range of patient characteristics, ML algorithms can provide a more personalized risk assessment than traditional risk scores. This can aid in determining appropriate preventive measures and interventions.

Feature Selection: ML techniques can be utilized to identify the most relevant features or risk factors contributing to heart disease. By analyzing large datasets, ML algorithms can automatically select the most informative features, helping clinicians and researchers understand the underlying factors that contribute to heart disease development and progression.

Predictive Modeling: ML techniques can be used to develop predictive models that estimate the likelihood of future cardiac events, such as heart attacks or heart failure, based on a patient's clinical data. These models can assist in identifying high-risk individuals who may benefit from targeted interventions or monitoring.

9 CONCLUSION

From the results we can conclude that the proposed model is giving better accuracy than normal Decision Tree, Random Forest, XGBoost, Light GBM, KNN and logistic Regression. We can also say that using Random under sampling is causing lot of information loss and decreasing the accuracy of the model. ADASYN provides better accuracy of nearly 91% and is better in this cause than SMOTE because SMOTE might lead to overfitting of the dataset.

10 FUTURE SCOPE

The future work of the project include implementing mode preprocessing techniques and more machine learning models which give better accuracy and add them to the voting model to increase the overall accuracy of the proposed architecture.

11 BIBILOGRAPHY

1.Abdar, M., Książek, W., Acharya, U. R., Tan, R. S., Makarenkov, V., & Pławiak, P. (2019). A new machine learning technique for an accurate diagnosis of coronary artery disease. Computer methods and programs in biomedicine, 179, 104992.

2.Hsich, E., Gorodeski, E. Z., Blackstone, E. H., Ishwaran, H., & Lauer, M. S. (2011). Identifying important risk factors for survival in patient with systolic heart failure using random survival forests. Circulation: Cardiovascular Quality and Outcomes, 4(1), 39-45.

3.Alsharif, M. H., Kelechi, A. H., Yahya, K., & Chaudhry, S. A. (2020). Machine learning algorithms for smart data analysis in internet of things environment: taxonomies and research trends. Symmetry, 12(1), 88.

4.Kumar, N., & Kumar, D. (2021, August). Machine learning based heart disease diagnosis using non-invasive methods: a review. In Journal of Physics: Conference Series (Vol. 1950, No. 1, p. 012081). IOP Publishing.

5.Ramachandran, A., & Karuppiah, A. (2021, July). A survey on recent advances in machine learning based sleep apnea detection systems. In Healthcare (Vol. 9, No. 7, p. 914). MDPI.

6.Ahsan, M. M., Luna, S. A., & Siddique, Z. (2022, March). Machine-learning-based disease diagnosis: A comprehensive review. In Healthcare (Vol. 10, No. 3, p. 541). MDPI.

7.Nabeel, M., Majeed, S., Awan, M. J., Muslih-ud-Din, H., Wasique, M., & Nasir, R. (2021). Review on Effective Disease Prediction through Data Mining Techniques. International Journal on Electrical Engineering & Informatics, 13(3).

8.Arif, H., Siddique, M., Aslam, N., Pervez, M. T., & Khan, M. K. (2022). Early-Stage Heart Disease Prediction using supervised Machine Learning Algorithms.

9.Gupta, A., Kumar, L., Jain, R., & Nagrath, P. (2020). Heart disease prediction using classification (naive bayes). In Proceedings of First International Conference on Computing, Communications, and Cyber-Security (IC4S 2019) (pp. 561-573). Springer Singapore.

APPENDIX

Collab Notebook for python code:
https://colab.research.google.com/drive/1BkgvgqK85I66_rubL7XGtdQ-RHz58yZh?usp=sharing

Demo video:-
https://drive.google.com/file/d/14XD78dQgZn25FMNBwoFhfu-Mt-NS-oiE/view?usp=sharing

Bootstrap page:-
https://drive.google.com/file/d/1KERjl6P9PTG9VF7xg-8NM0TtESA_YEWx/view?usp=sharing

Web Assets:-
https://drive.google.com/file/d/1dA-nvevazZkr7v1p7FroHcFmez2K9TmN/view?usp=sharing

Changes made:-
https://drive.google.com/file/d/13zkhsDyIDHY7gzmdEqTyMuTnQQMUp9Bi/view?usp=sharing

Integrations:-
https://drive.google.com/file/d/16j3qRVlhwhWpzYguk97dqD4Oe-j7niMa/view?usp=sharing

Flask Implementation:-
https://drive.google.com/file/d/1HwmTngabilvGB64YmD200rosmIvagM6G/view?usp=sharing

IBM Cognos Dashboard-1:-
https://drive.google.com/file/d/1IaT3TR7IrL8zI2SezF0erA4SGS2qZwSS/view?usp=sharing

IBM Cognos Dashboard-2:-
https://drive.google.com/file/d/1_Zy7j7SPxGsbeZoTWYOYF34p3uXRs8JN/view?usp=sharing

IBM Cognos Dashboard-3:-
https://drive.google.com/file/d/1aiDj7a6jbMcQAslDB0oTivxaiNDaqu2A/view?usp=sharing

IBM Cognos Dashboard-4:-
https://drive.google.com/file/d/1YpOyQ34Mph0b0INxkG6W90yTH3_1MjoF/view?usp=sharing

Tableau - MySQL connection:-
https://drive.google.com/file/d/1Y54GV4xHxtpzvQkKCyBGD94IGmOsqODF/view?usp=sharing

Tableau Dashboard-1:-
https://drive.google.com/file/d/1tO3DTTbpXyOeZznfzWApmY6DhhQzGDyS/view?usp=sharing

Tableau Dashboard-2:-
https://drive.google.com/file/d/1_9uG8eqRc8j7L_c85Hc6Q-3xFyugXiKj/view?usp=sharing

Tableau Dashboard-3:-
https://drive.google.com/file/d/1g2c2gnPEKNowYUcSHp-SO4xKpEvX6FlK/view?usp=sharing

Tableau Dashboard-4:-
https://drive.google.com/file/d/1EMr_aOvWBdE_ID0Z6fyTXQGl9GLlt_U-/view?usp=drive_link

Tableau Dashboard-5:-
https://drive.google.com/file/d/1CMUssFWslHoXvCngaGl3RvtkFmpu7tfQ/view?usp=drive_link

Tableau Dashboard-6:-
https://drive.google.com/file/d/1xTFuFkdxYNcDIcrITzYnoWlVnWAxsOJj/view?usp=drive_link

Tableau Dashboard-7:-
https://drive.google.com/file/d/1BXxBLpEuR9xcd7GO1758CYRX6LLzKQzu/view?usp=drive_link

Prediction:-
https://drive.google.com/file/d/1cOmEfSslHAfB0NeHa97TeyzjGckgN4OZ/view?usp=sharing