greatlearning

PGPDSE FT Capstone Project – Final Report

Topic: Empowering Early Detection of Heart Attack Risks with Machine Learning

Batch Details	PGP-DSE MAR'24
Team Members	Aarti Gupta, Baljit Singh Karnavat, Nikhil Vinod, Raghavendra S, Shashwat Girish Irny, Vergin J
Domain of Project	Predictive Analysis
Proposed Project Title	Empowering Early Detection of Heart Attack Risks with Machine Learning
Group Number	1
Team Leader	Raghavendra S
Mentor Name	Mr. Ankush Bansal

Industry Review:

Heart disease remains one of the leading causes of death worldwide. Early detection and prediction are crucial in preventing heart-related complications and improving patient outcomes. Over the last few decades, the healthcare industry has increasingly turned to

machine learning (ML) and data analytics to support clinical decision-making, improve diagnostic accuracy, and optimize healthcare delivery. In particular, heart disease prediction has emerged as a key application area where advanced data techniques are helping to revolutionize patient care.

1. Current State of Heart Disease Prediction in the Industry

In the healthcare industry, early diagnosis of heart disease is critical in reducing mortality rates and improving quality of life. Traditionally, heart disease diagnoses have relied on clinical tests, such as ECG (electrocardiograms), blood tests, X-rays, and stress tests. However, with the advancement of data science, machine learning algorithms are now increasingly being adopted to automate diagnosis
and predict the risk of heart disease.

- Data-Driven Healthcare: The rise of electronic health records (EHRs) and other digital health technologies has generated vast amounts of data, creating an opportunity to apply machine learning to predict heart disease risk. These records often include patient demographics, lab test results, vital signs, and medical histories, which serve as inputs for predictive models.
- Use of Predictive ML Models: Several machine learning models, such as logistic regression, decision trees and ensemble methods (e.g., Random Forests, Gradient Boosting Machines), are being used by healthcare providers and medical research centers to predict heart disease. These models analyze a variety of factors—such as cholesterol levels, blood pressure, age, smoking habits, and family history of heart disease—to assess a patient's risk and recommend preventative actions.

2. Industry Applications and Machine Learning Adoption

- Wearables and Remote Monitoring: The use of wearable devices like Fitbit, Apple Watch, and Garmin has also made a significant impact on heart disease prediction. These devices continuously monitor heart rate, activity levels, blood pressure, and even ECG, feeding data into cloud-based platforms that use machine learning to identify any abnormalities or risks associated with heart disease. This data is then sent to healthcare providers for analysis, facilitating early intervention.
- Clinical Decision Support Systems (CDSS): Many hospitals and healthcare organizations have started adopting clinical decision support systems powered by machine learning. These systems assist physicians by analyzing patient data and recommending appropriate diagnostic tests, lifestyle changes, or treatments. For example, ML-based systems can alert doctors when a patient's vital signs indicate the possibility of heart disease, enabling timely intervention.
- Insurance and Risk Assessment: The insurance industry has also adopted machine learning techniques to assess heart disease risk when underwriting health insurance policies. Predictive models are used to analyze an individual's health data and generate more accurate risk assessments, allowing insurance companies to provide personalized

insurance premiums.

- Predictive Analytics Software: Companies like Health Catalyst and Cerner are offering healthcare analytics platforms that incorporate machine learning models to predict a range of health outcomes, including heart disease. These platforms are designed to help healthcare providers monitor patient health over time, identify at-risk individuals, and deliver personalized healthcare plans.

3. Challenges in the Industry

- Data Quality and Privacy: Medical data is often incomplete, inconsistent, and sometimes biased, which can negatively affect the accuracy of machine learning models. Additionally, patient data privacy remains a significant concern in healthcare, and strict regulations like the Health Insurance Portability and Accountability Act (HIPAA) in the U.S. govern how patient data can be used and shared.

- Model Interpretability: Many of the machine learning models used in heart disease prediction are often seen as "black boxes." This lack of interpretability makes it difficult for healthcare professionals to understand how decisions are being made, which can be a barrier to adoption. The ability to explain model predictions is crucial for gaining trust from medical practitioners and patients.

- Generalization: Models that perform well in specific datasets may struggle to generalize to broader, real-world populations. Ensuring that machine learning models are trained on diverse, representative datasets is crucial for ensuring that they provide accurate and reliable predictions across different patient groups.

4. Future Directions in Heart Disease Prediction

As technology evolves, the future of heart disease prediction looks promising, with several exciting trends on the horizon:

- Integration with Genetic and Omics Data: Advances in genomics and precision medicine may lead to more personalized heart disease prediction models that incorporate genetic data. By combining traditional medical data with genomic information, machine learning models can offer more precise and tailored risk assessments.

- Real-Time Data Integration: Real-time monitoring of patients via connected devices will become increasingly integrated into heart disease prediction models. As wearable technologies continue to improve, real-time data on heart rate, blood pressure, and activity levels can be used to dynamically adjust predictions and alert healthcare professionals to emerging risks.

- Telemedicine and Remote Health Monitoring: With the growing adoption of telemedicine and remote health monitoring, machine learning models will play an even greater role in providing ongoing heart disease risk assessments. This will allow healthcare providers to deliver timely interventions without needing patients to visit the clinic physically.

Literature Review:

- Research on heart disease prediction has extensively explored machine learning models like Logistic Regression, Decision Trees, Random Forests, and Support Vector Machines, demonstrating their ability to identify at-risk individuals based on health and lifestyle factors.

- These models are favoured for their interpretability and effectiveness in handling structured datasets such as electronic health records. Numerous studies have validated their application in predicting cardiovascular events and informing preventive care.

- This project builds on such research, focusing exclusively on machine learning techniques to develop a reliable and interpretable model for heart attack prediction, addressing challenges like data imbalance and missing values.

Dataset and Domain

Data Dictionary

Data Type: Integer

Attribute Name	Description	Notes
1. PhysicalHealthDays	The number of days in the past month the individual experienced physical health issues.	Physical health issues can strain the heart, increasing the likelihood of heart attacks.
2. MentalHealthDays	The number of days in the past month the individual experienced mental health issues.	Stress and mental health conditions are associated with inflammation and increased cardiac risk.
3. SleepHours	Average number of hours the individual sleeps each night.	Insufficient (<6 hours) or excessive sleep (>9 hours) is linked to poor cardiovascular health.
4. HeightInMeters	The height of the individual in meters.	These contribute to BMI, which directly impacts

		cardiovascular health.
5. WeightInKilograms	The weight of the individual in kilograms.	These contribute to BMI, which directly impacts cardiovascular health.
6. BMI	The body mass index (BMI) of the individual.	Overweight or obesity is a well-documented risk factor for heart attacks due to associated conditions like hypertension and diabetes.

Data Type: Categorical

7. State	The state in which the individual resides ('California', 'Texas', etc.).	State wise air quality, weather and climate are impacting on heart.
8. Sex	Gender of the individual ('Male', 'Female').	Men have a higher risk at an earlier age, but the risk for women increases after menopause.

9. GeneralHealth	Self-reported general health status ('Excellent', 'Good', 'Fair', 'Poor').	Poor general health correlates with higher risks of cardiovascular events.
10. LastCheckupTime	The last time the individual had a health checkup ('1 year ago', '2 years ago').	Regular checkups is mandatory.
11. PhysicalActivities	Whether the individual engages in physical activities ('Yes', 'No').	Regular exercise lowers the risk by improving blood circulation and heart function.

12. RemovedTeeth	Whether the individual has had teeth removed ('Yes', 'No').	Tooth loss is often a result of periodontal disease (gum disease), which is a chronic inflammatory condition
13. HadHeartAttack	Whether the individual has had a heart attack (*target variable*, 'Yes', 'No').	Target Variable.
14. HadAngina	Whether the individual has a history of angina ('Yes', 'No').	Angina is a symptom of coronary artery disease and a warning sign of heart attacks.
15. HadStroke	Whether the individual has had a stroke ('Yes', 'No').	A history of stroke indicates compromised cardiovascular health, increasing heart attack risk.
16. HadAsthma	Whether the individual has a history of asthma ('Yes', 'No').	Severe asthma can strain the cardiovascular system, particularly in uncontrolled cases.
17. HadSkinCancer	Whether the individual has a history of skin cancer ('Yes', 'No').	The relationship between skin cancer and heart attacks is an area of growing research interest. While there isn't direct evidence that having skin cancer causes heart attacks

18. HadCOPD	Whether the individual has chronic obstructive pulmonary disease ('Yes', 'No').	Chronic obstructive pulmonary disease is linked to heart problems due to

		poor oxygenation.
19. HadDepressiveDisorder	Whether the individual has had a depressive disorder ('Yes', 'No').	Depression is linked to unhealthy lifestyles (e.g., poor diet, inactivity), which increase cardiac risk.
20. HadKidneyDisease	Whether the individual has had kidney disease ('Yes', 'No').	Kidney disease is associated with high blood pressure and poor heart health
21. HadArthritis	Whether the individual has had arthritis ('Yes', 'No').	Chronic inflammation in arthritis can damage blood vessels, indirectly affecting heart health.
22. HadDiabetes	Whether the individual has had diabetes ('Yes', 'No', “No, pre diabetes or borderline diabetes”, “Yes, but only during pregnancy (female)”)	Diabetes accelerates arterial plaque buildup and increases cardiovascular risk.
23. DeafOrHardOfHearing	Whether the individual is deaf or has hearing difficulties ('Yes', 'No').	May indicate comorbid conditions or barriers to accessing preventive care.
24. BlindOrVisionDifficulty	Whether the individual is blind or has vision difficulties ('Yes', 'No').	May indicate comorbid conditions or barriers to accessing preventive care.
25. DifficultyConcentrating	Whether the individual experiences difficulty concentrating ('Yes', 'No').	Chronic illnesses or neurological conditions can indirectly impact cardiovascular health.

26. DifficultyWalking	Whether the individual has difficulty walking ('Yes', 'No').	May reflect physical inactivity, a strong contributor to cardiovascular risk.
27. DifficultyDressingBathing	Whether the individual has difficulty dressing or bathing ('Yes', 'No').	Conditions like stroke, Parkinson's disease, or neuropathy can make dressing and bathing difficult and impacting on heart attacks.

28. DifficultyErrands	Whether the individual has difficulty running errands ('Yes', 'No').	Disease like Parkinson's are impacting on capability of doing work and Calculations.
29. SmokerStatus	Whether the individual is a smoker ('Yes', 'No').	Smoking is a direct risk factor for heart disease and can lead to arterial damage.
30. ECigaretteUsage	Whether the individual uses e cigarettes ('Yes', 'No').	While less studied than traditional smoking, vaping may still contribute to cardiovascular risks.
31. ChestScan	Whether the individual has had a chest scan ('Yes', 'No').	If an individual has undergone chest scans, it could signal previous heart or lung issues.
32. RaceEthnicityCategory	The individual's race or ethnicity category.	Certain ethnic groups may have predispositions to heart disease due to genetic or lifestyle

		factors.
33. AgeCategory	The individual's age group.	Risk increases significantly with age due to arterial plaque buildup and decreased cardiovascular health. Example: Older individuals (e.g., 45+) have a higher likelihood of heart attacks.
34. AlcoholDrinkers	Whether the individual drinks alcohol	Excessive drinking increases blood pressure and cholesterol
35. HIVTesting	Whether the individual has undergone HIV testing	Indirectly relevant; individuals with untreated HIV may have higher inflammation levels affecting cardiovascular health.
36. FluVaxLast12	Whether the individual had a flu vaccination in the past 12.	Receiving vaccines can indicate health awareness, possibly correlating with better prevention efforts.

37. PneumoVaxEver	Whether the individual has ever received a pneumonia vaccination.	Receiving vaccines can indicate health awareness, possibly correlating with better prevention efforts.
38. TetanusLast10Tdap	Whether the individual has received a tetanus vaccination in the last 10 years.	Similar to flu vaccines, it indirectly indicates health seeking behavior.

39. HighRiskLastYear	Whether the individual was considered high-risk for health issues in the last year.	Activities like extreme stress or substance abuse can increase the likelihood of cardiac events.
40. CovidPos	Whether the individual has tested positive for COVID-19.	Post-COVID, individuals are at a higher risk of cardiovascular complications, including heart attacks.

Variable categorization:

- Total Numeric Variables: 6
- Total Categorical Variables: 34

Pre-Processing Data Analysis:

Duplicates:

In the context of a heart disease dataset, duplicates can skew the analysis and lead to overfitting, where the model overly learns repetitive patterns.

There are 157 duplicate records. Therefore we are removing duplicate data from this dataset. Handle Missing Values:

Only State and Sex column have no missing value otherwise all attributes have some amount of missing values. Many machine learning algorithms cannot handle null values directly and fail or produce incorrect results.

Algorithms like Linear Regression and Logistic Regression typically require complete data. Tree based models (e.g., Random Forest, XGBoost) can handle nulls but may still perform sub optimally.

Missing values decreasing interpretability and leads to inaccurate predictions therefore we need to treat missing values.

Treatments:

- Check row wise missing values those are containing 75% of null values, consider >25% of data are missing removed those rows.
- Columns wise missing values
    - Logical and Mode Imputation for Categorical variable.

- ➢ 'PhysicalActivities' imputing with 'No' when the patient has arthritis, difficulty in dressing, bathing and walking.
- ➢ 'FluVaxLast12' imputing 'fulvax last 12' with 'Yes' when the patient age is 65-69 or older (assuming all the older peoples are vaccine).
- ➢ 'RaceEthnicityCategory' column, states with a higher population of a particular race or ethnicity have been imputed accordingly.

- Knn Imputation for Numerical variable (except 'BMI').
- For 'BMI' imputation done by formula (BMI = weight in (kg) / height in (M)**2).
- Conditional mode imputation or Group based mode imputation.

Presence of outliers and its treatment:

- All the numerical columns contain outliers.
- The 'SleepHours' column has outliers on both the lower and upper ends.
- The 'HeightInMeters', 'WeightInKilograms', and 'BMI' columns exhibit a significant number of outliers on both ends.

There are outliers in the dataset but no treatment was done because the variables are significant and Standard scaling was done.

Project Justification:

1. Project Statement:

Heart disease is a leading cause of global mortality, and early prediction of heart attacks can save countless lives. This project focuses on developing a machine-learning model to predict the likelihood of heart attacks using various health-related factors and medical history. By identifying high-risk individuals, timely interventions and preventive measures can be implemented, ultimately improving cardiovascular health outcomes.

2. Complexity Involved:

- Data Preprocessing: Managing a large dataset with missing values, and unbalanced target classes requires robust data cleaning, imputation, and feature engineering techniques.
- Feature Selection: Identifying the most relevant health and lifestyle factors from a diverse range of variables is essential for creating an effective predictive model.
- Model Development: Implementing and tuning machine learning models like Logistic Regression, Decision Tree, and Random Forest to achieve high accuracy without overfitting demands expertise.

3. Project Outcome:

- Social Value: This project has immense societal impact by enabling healthcare providers to identify and support at-risk individuals, reducing mortality and improving quality of life.
- Commercial Value: Hospitals, insurance companies, and wellness organizations can adopt the predictive model to streamline patient care, reduce costs, and offer targeted health

interventions.

- Academic Value: The project contributes to research in machine learning applications in healthcare, setting a benchmark for future studies and innovations in predictive health analytics.

Data Exploration (EDA):

Statistics Summary:

The describe() function was applied to analyze key statistics like mean, median, standard deviation, and quartiles for numerical columns. This step provided insights into the data distribution and variability.

	count	mean	std	min	25%	50%	75%	max
Physical_HealthDays	414430.0	4.255228	8.603966	0.00	0.00	0.00	3.00	30.00
Mental_HealthDays	414430.0	4.310335	8.318915	0.00	0.00	0.00	4.00	30.00
Sleep_Hours	414430.0	7.022981	1.478186	1.00	6.00	7.00	8.00	24.00
Height_In_Meters	414430.0	1.702857	0.105886	0.91	1.63	1.70	1.78	2.41
Weight_In_Kilograms	414430.0	82.973876	20.878367	22.68	68.04	80.29	92.99	292.57
BMI	414430.0	28.465196	6.339031	12.02	24.37	27.40	31.32	99.64

- The majority of individuals report 0 days of physical or mental health issues. However, there are outliers, with a small portion of the population reporting up to 30 days of health problems, suggesting chronic conditions or data irregularities.

- Most individuals report a healthy sleep duration of 6-8 hours per night. However, extreme values of 1 hour and 24 hours, as they may indicate data entry errors or unusual behaviour.

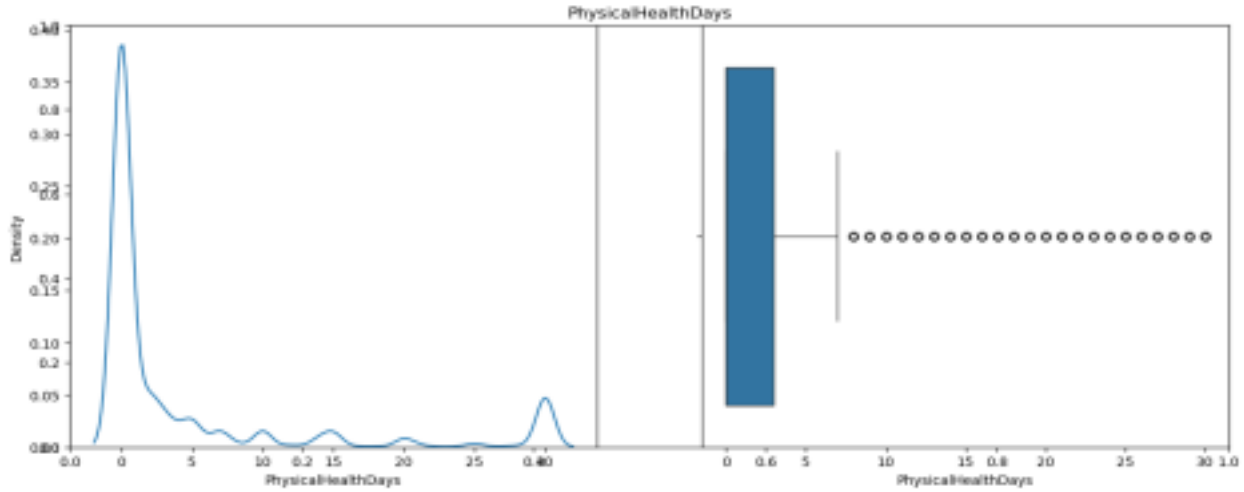
- The average ‘BMI’ suggests that the population is mostly in the overweight category, which could be a public health concern. Extreme values, both low (12.02) and high (99.64), likely indicate data entry issues or outliers.

- There are significant outliers in both height and weight, including extreme values like 0.91 meters for height and 292.57 kg for weight.

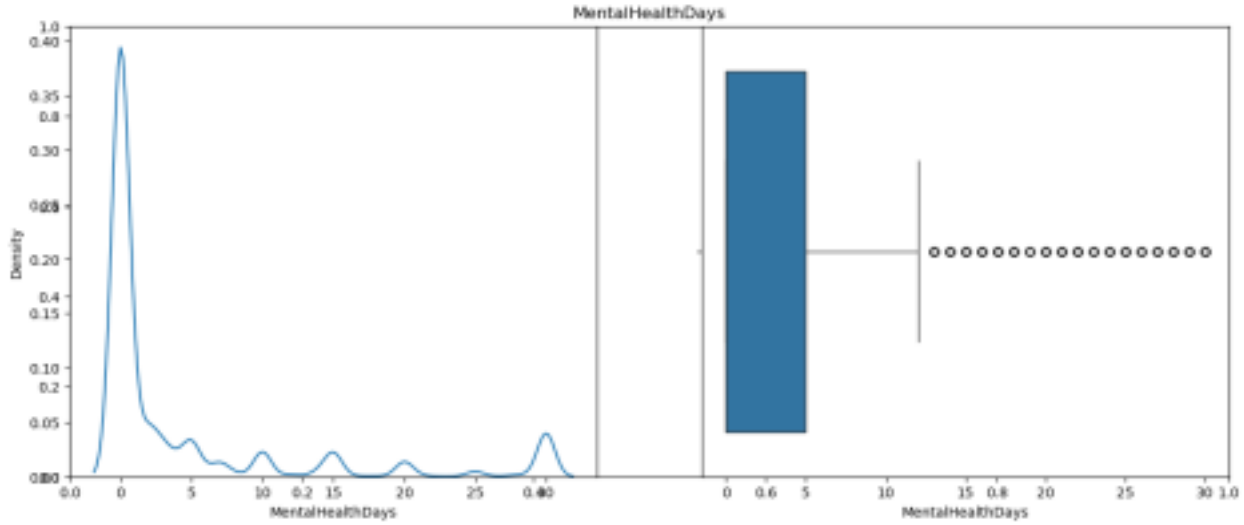
Univariate Analysis:

Conducted for individual columns to understand their distributions, including histograms, kde plot for numerical variables and bar plot, count plot for categorical variables.

Numerical Variable:

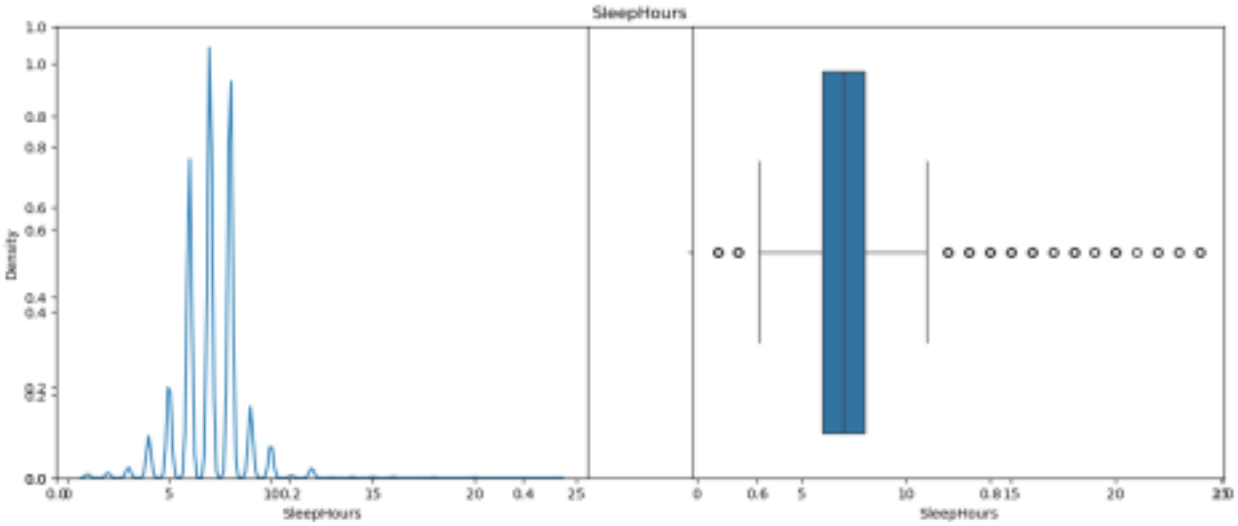


Skewness: 2.1792691574342538
Kurtosis: 3.4250618380994986
Skewness (2.179): indicates that the data is positively skewed, This suggests that there are a few observations with significantly higher values compared to the rest of the data.
Kurtosis (3.425): indicates that the distribution is leptokurtic, as it is slightly above 3 (the kurtosis value of a normal distribution). This means the data has heavier tails and a sharper peak than a normal distribution suggesting the presence of outliers or extreme values.



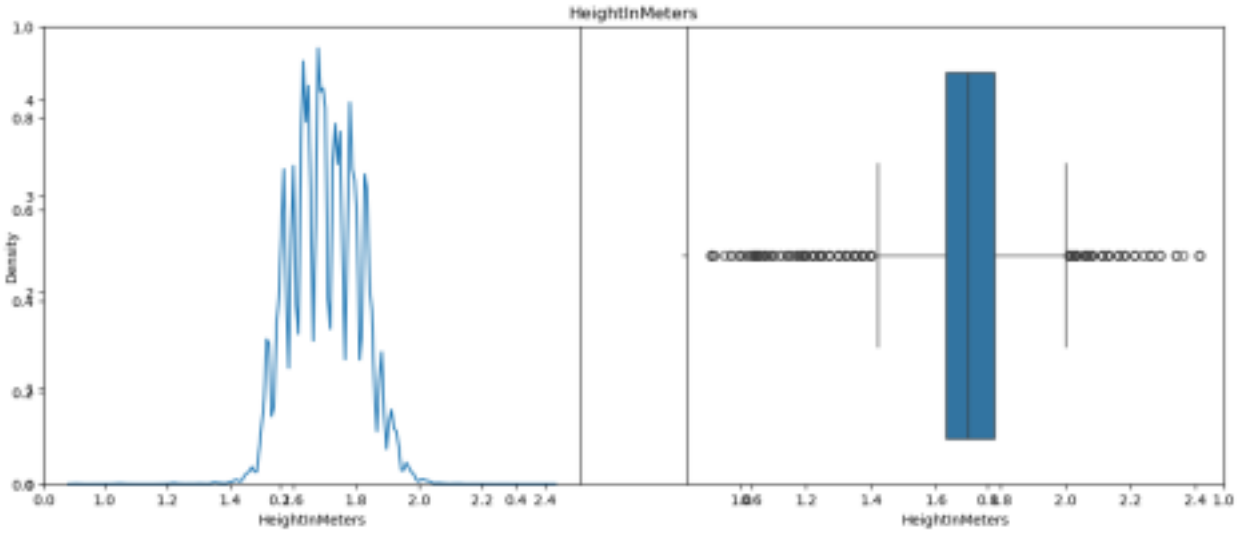
Skewness: 2.12265370351103
Kurtosis: 3.356652764476231

The data is positively skewed, indicating that higher values dominate the right tail of the distribution. Additionally, the slightly leptokurtic nature of the data suggests a sharper peak and the potential presence of outliers, making the distribution deviate from normality.



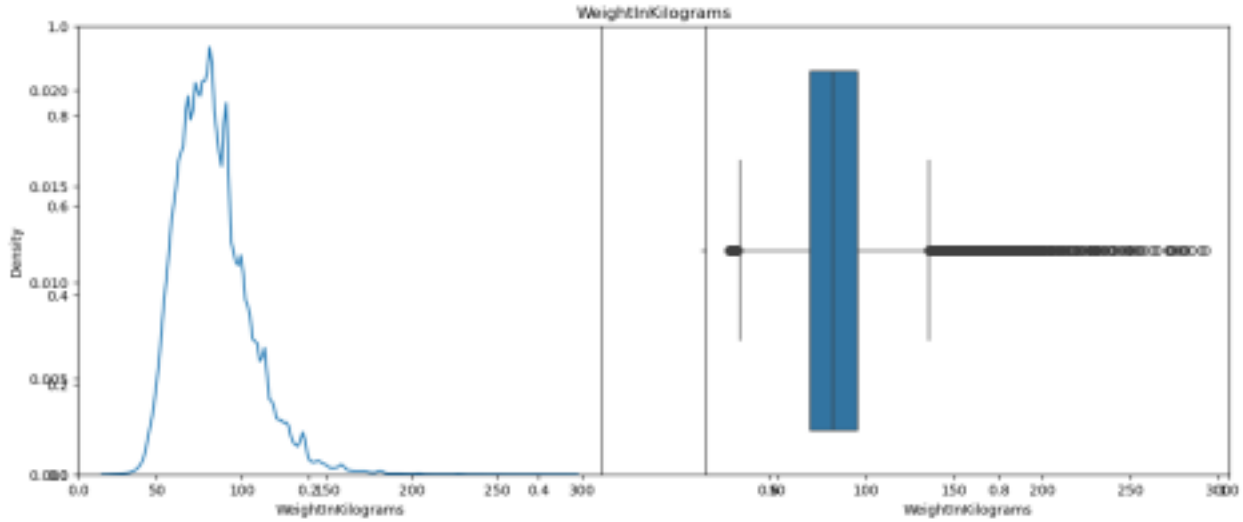
Skewness: 0.7647202620733116
Kurtosis: 8.739265011746813

The data shows a moderate positive skew, suggesting that higher values slightly dominate. However, the very high kurtosis indicates the presence of extreme outliers or a high concentration of data around the central peak making the distribution heavily deviate from normality.



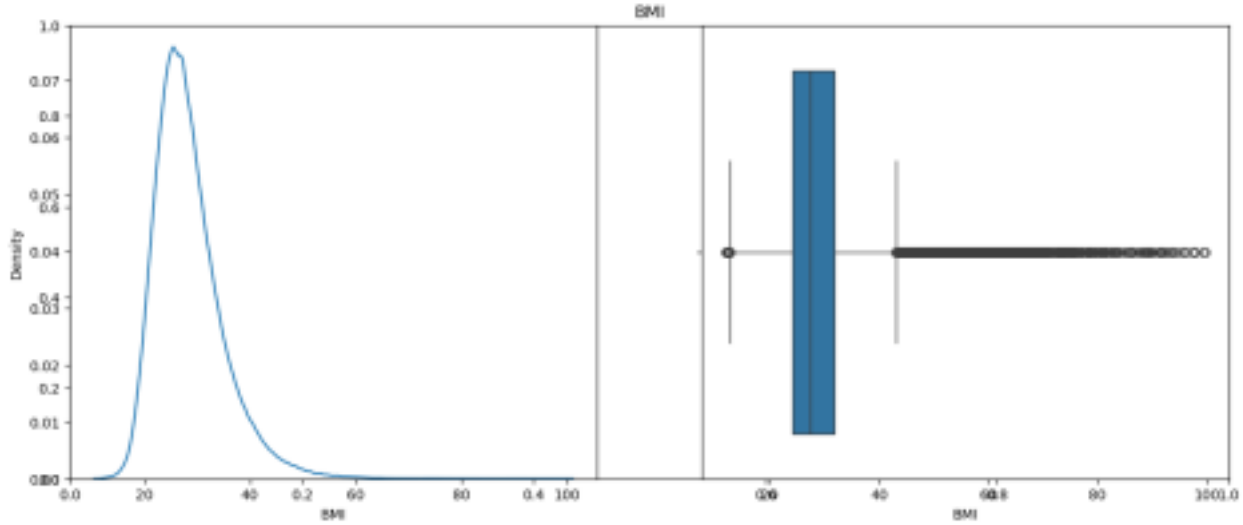
Skewness: 0.0288869956004554
Kurtosis: 0.18230087077564594

The data distribution is almost symmetric, with no notable skewness. The slightly platykurtic nature suggests the distribution is relatively flat and has fewer extreme values or outliers compared to a normal distribution. The data appears well-behaved and close to normal in shape.



Skewness: 1.0756226917604688
Kurtosis: 2.7389750517766167

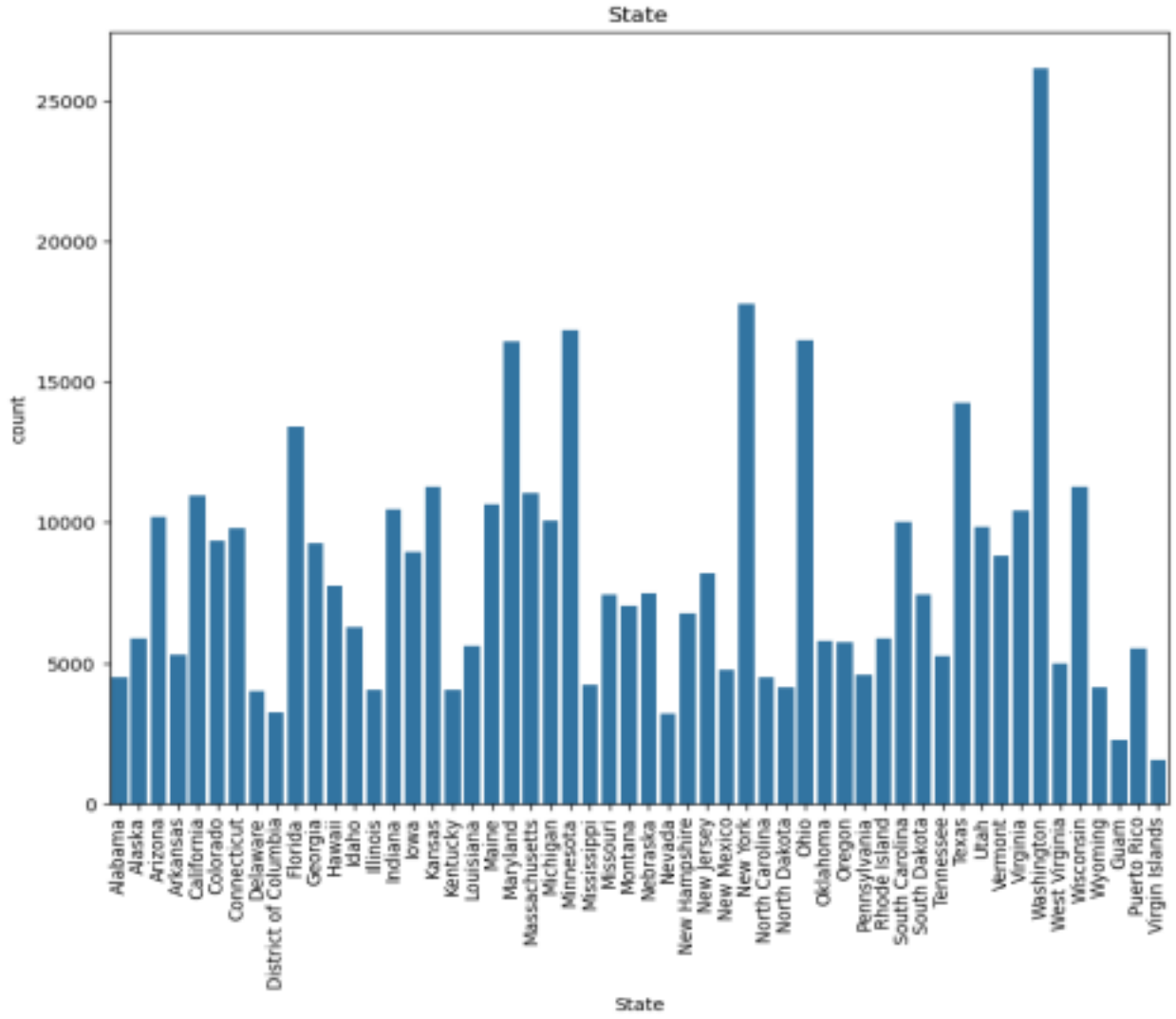
The data distribution is positively skewed, with a tendency for values to cluster toward the lower end while a few higher values stretch the distribution. The slightly platykurtic nature indicates a relatively flat peak, with fewer extreme values than a normal distribution would have.



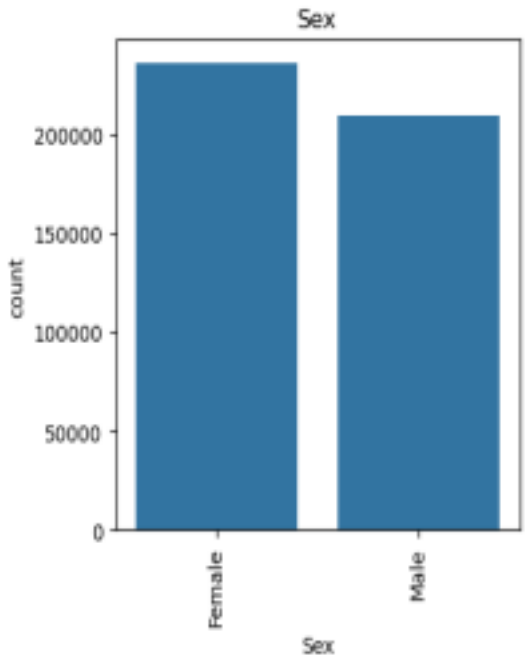
Skewness: 1.3877487186852828
Kurtosis: 4.428355854190881

The data distribution is positively skewed, with most values concentrated toward the lower range while a few higher values stretch the distribution. The leptokurtic nature of the distribution indicates that there are more extreme outliers, making the data more prone to rare but significant deviations.

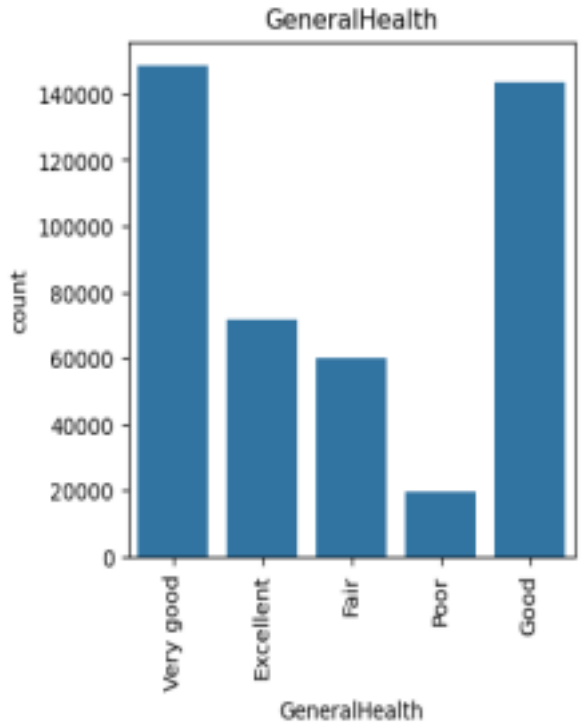
Categorial Variable:



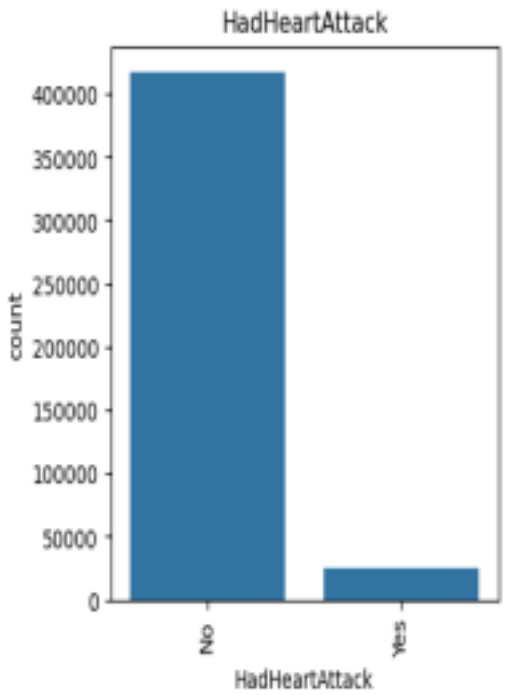
From the above chart it indicates the distribution of person who are from different stats and different regions.



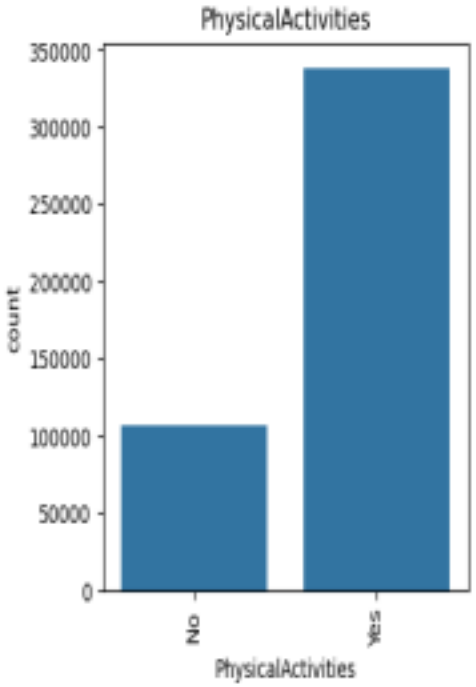
Above chart is showing Sex ratio from the data.



From the above chart we can infer that health category of the person.

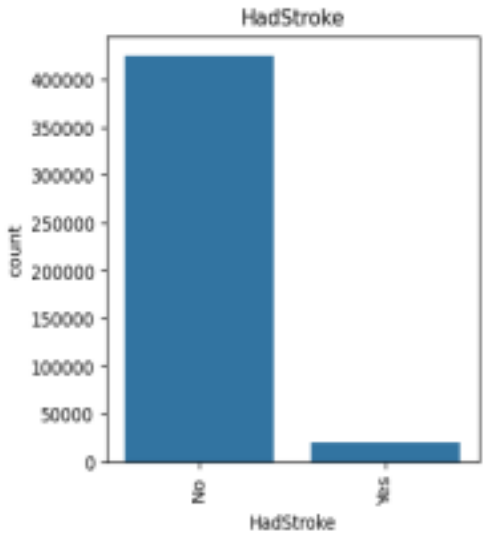


From the above chart showing the heart attack ratio from the

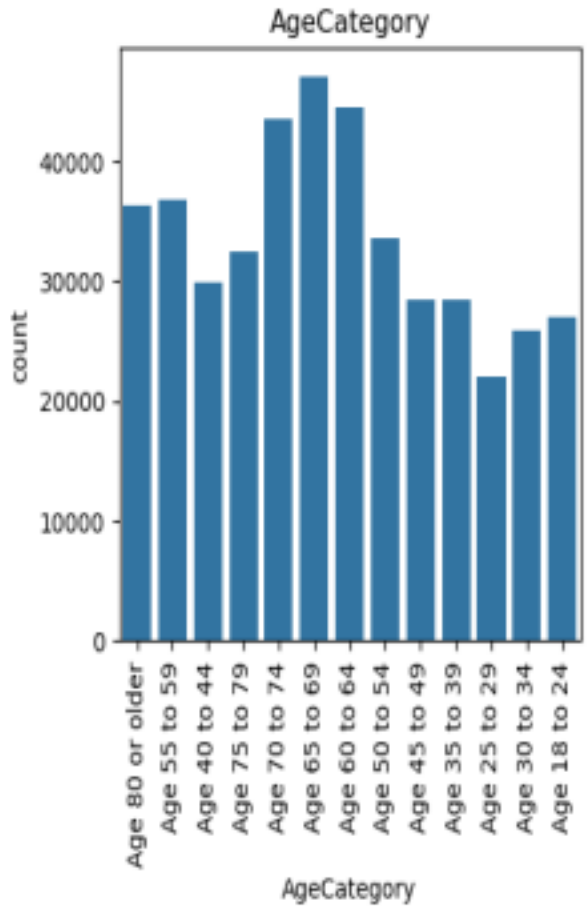


data.

The above chart showing Physical activity of the person.



Above chart is for what is the ratio of stroke in the data.

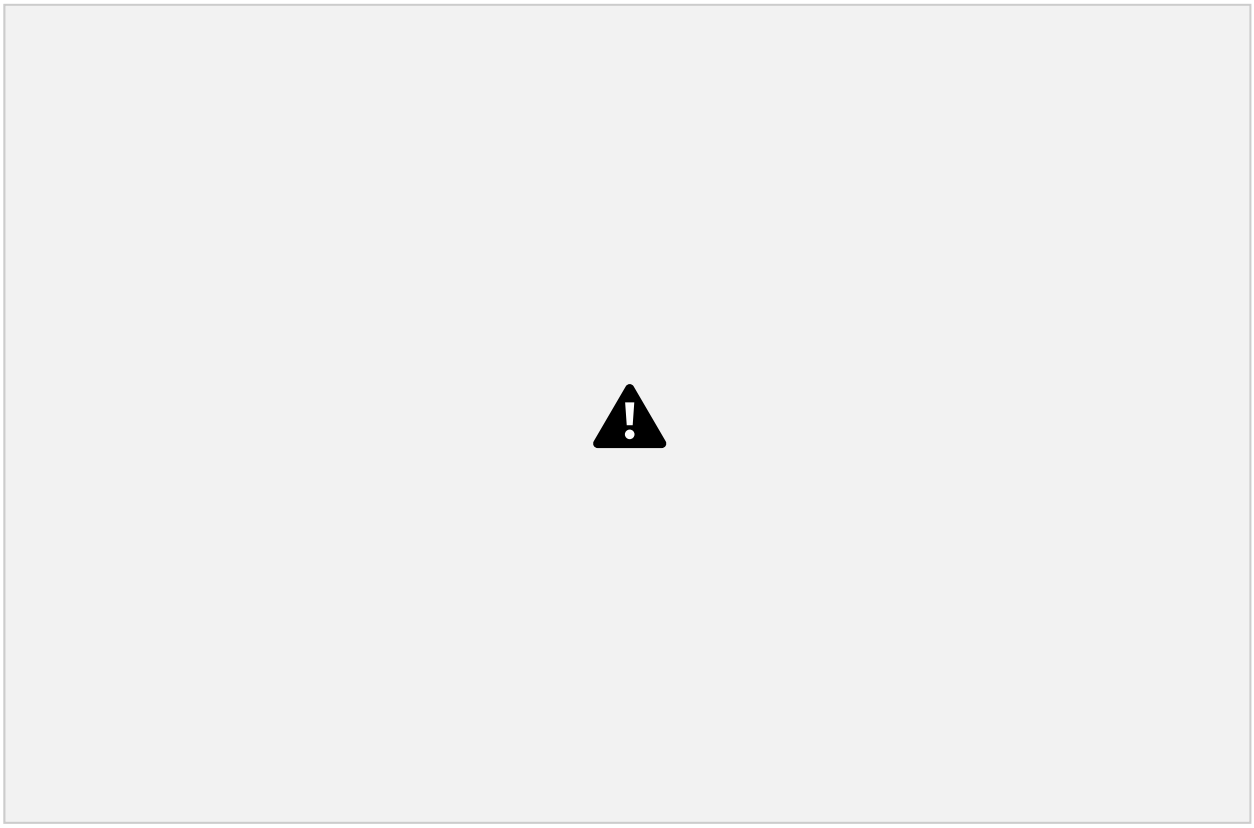


Above chart says that the age category from the data, data are available teenager to senior citizens.

Bivariate Analysis:

Numerical Variable:

Explored relationships between the target column and other variables.



- Physical Health Days:
  * The median and mean values for both groups are relatively low, suggesting that a majority of individuals, regardless of heart attack history, experience a limited number of physically healthy days.
  * The distribution for the "Yes" group appears slightly more skewed to the right compared to the "No" group, suggesting that individuals who have experienced a heart attack may have, on average, fewer physically healthy days.

- Mental Health Days:
  * Similar to physical health, the median and mean values for mental health days are also low for both groups.
  * The distribution for the "Yes" group is again slightly more skewed to the right, indicating that individuals with a history of heart attack may experience fewer mentally healthy days.
- Sleep Hours:
  * The distributions for both groups are highly skewed to the right, with a peak around 7-8 hours of sleep.
  * The "Yes" group appears to have a slightly wider distribution compared to the "No" group, suggesting greater variability in sleep duration among individuals who have experienced a heart attack.

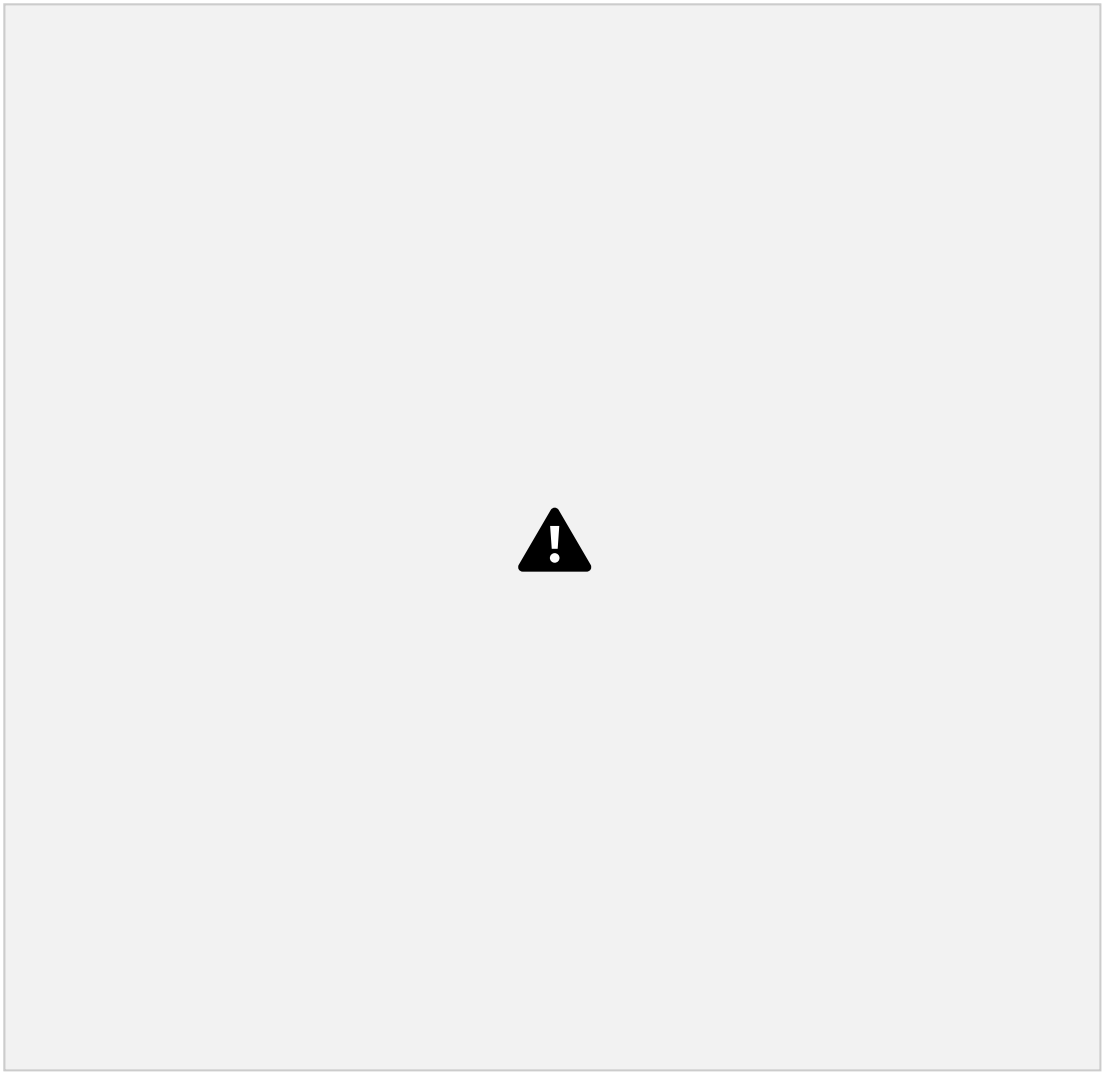
* Overall, the graph suggests that individuals who have experienced a heart attack may experience a lower number of physically and mentally healthy days compared to those who have not. Additionally, their sleep patterns might exhibit greater variability.

- Height (HeightinMeters):
  * The median and mean values for both groups are relatively similar, suggesting that height may not be a significant factor in differentiating between individuals with and without a history of heart attack.

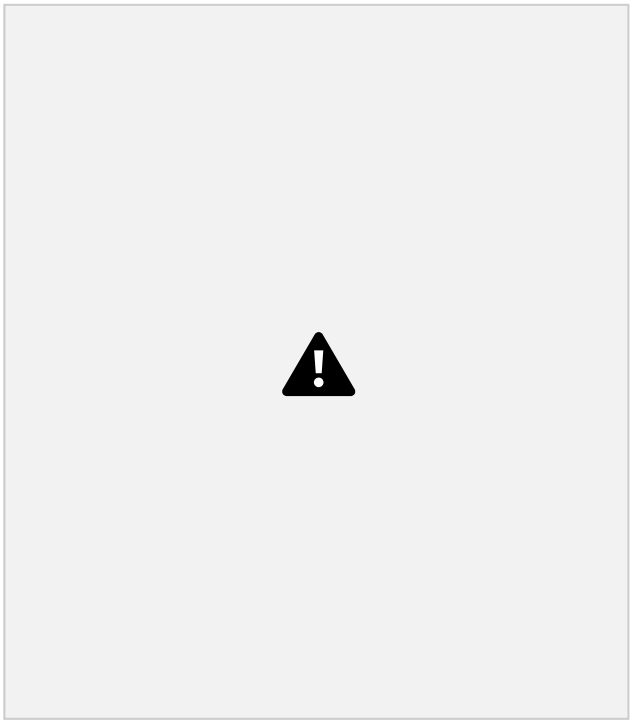
- Weight (WeightinKilograms):
  * The median and mean values for the "Yes" group are slightly higher compared to the "No" group, suggesting that individuals who have experienced a heart attack may have a higher average weight.
  * The distribution for the "Yes" group is also slightly wider, indicating greater variability in weight among individuals with a history of heart attack.

- BMI (Body Mass Index):
  * The median and mean values for the "Yes" group are also slightly higher compared to the "No" group, suggesting a potential association between higher BMI and heart attack risk. * The distribution for the "Yes" group is again slightly wider, indicating greater variability in BMI among individuals with a history of heart attack.
  * Overall, the graph suggests a potential association between higher weight and BMI and an increased risk of heart attack. However, the differences between the groups are relatively small.

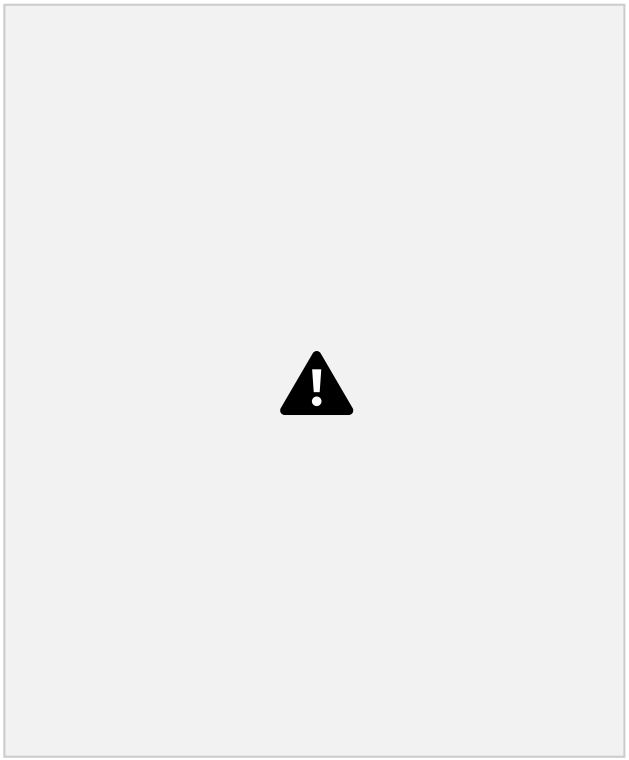
Catagorical Variable:



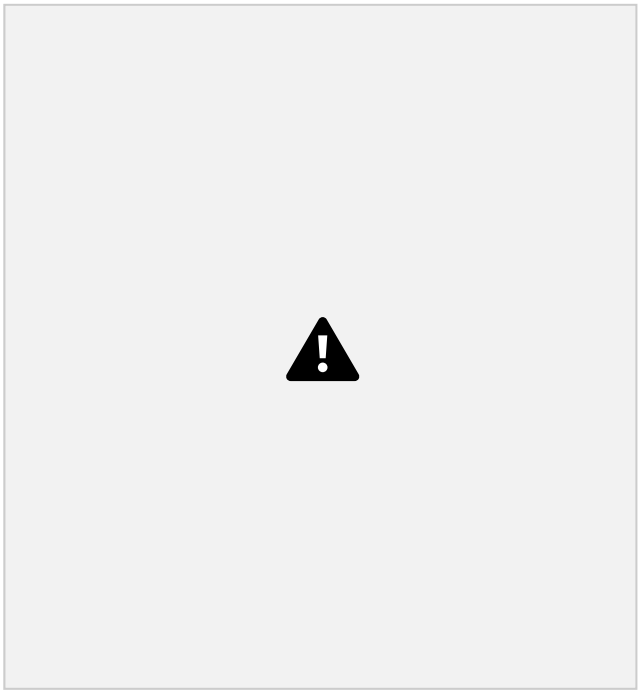
The chart highlights the state-wise distribution of heart attack or heart disease cases across various states in the USA.



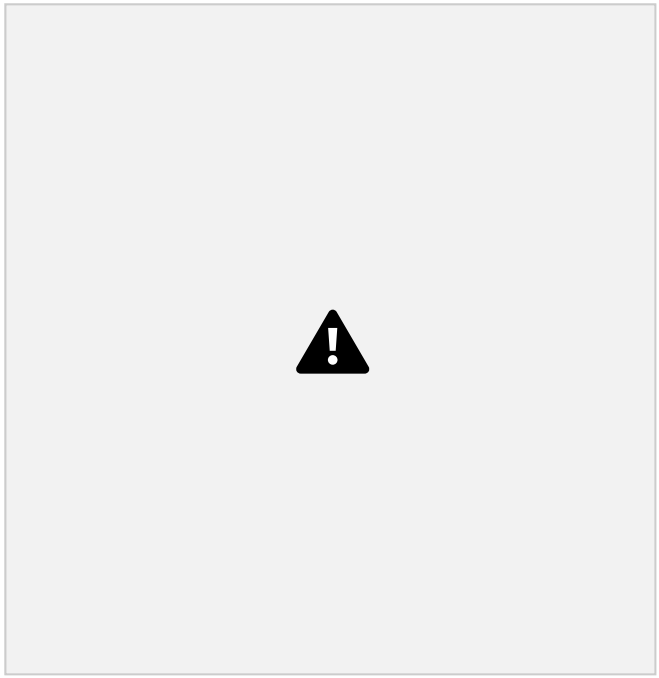
Above chart is for count of Heart Attack cases of the categorised by Male and Female it shows the chance is more for man as compare to womens.



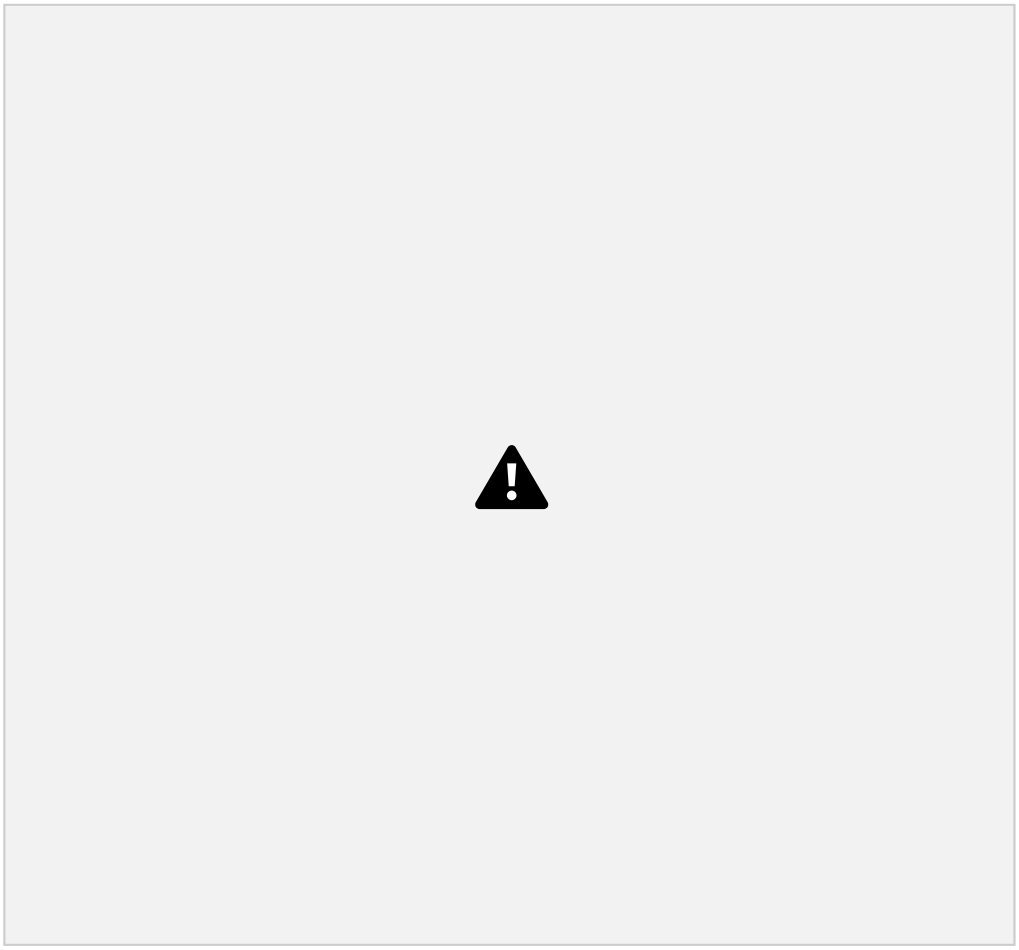
From the above chart it shows the persons general health are impacting on Heart health if the persons health is good still have chance for having Heart Attack.



Above chart showing that the persons physical activity are effective on Heart Health physically active person have less chance to have Heart Disease.



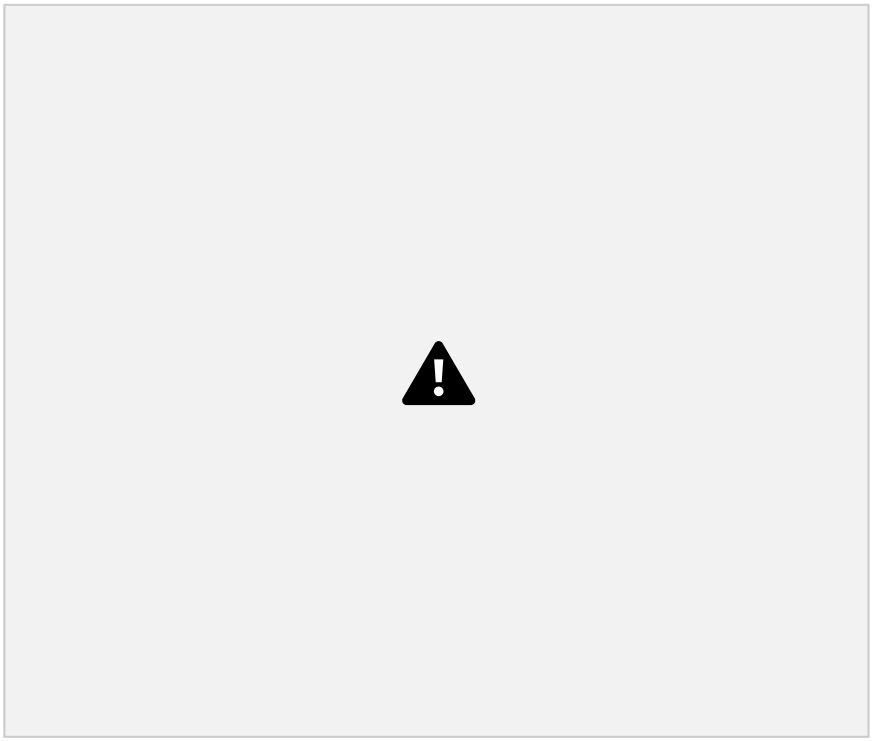
The chart indicates that individuals who have experienced a stroke are at a higher risk of having a heart attack.


The chart indicates that age has a significant impact on heart health, with the likelihood of experiencing a heart attack increasing as age advances.

Multivariate Analysis:

Examined interactions between multiple variables to uncover deeper

insights.



- Example: Relationship among ‘BMI’, ‘GeneralHealth’ and target column.
- Techniques like pair plots and heatmaps were used for visual exploration.

Feature Engineering:

Enhancing the accuracy, features for better and understandable model used Feature

Engineering. Standardization:

- Simplified ‘LastCheckupTime’ into <1 Year, 1-2 Years, 2-5 Years, and >=5 Years.
- Mapped ‘HadDiabetes’ and ‘CovidPos’ values to binary categories.

Column Extraction:

- Extracted Race and Ethnicity from ‘RaceEthnicityCategory’.

Age Grouping:

- Merged ‘AgeCategory’ 5-year intervals into broader 10-year intervals for simplified analysis.

State to Regional Mapping:

- Mapped individual states into Northeast, Midwest, South, West, and Territories under the Region column.

After separation the attributes ‘RaceEthnicityCategory’, ’Age’ and ‘State’ needs to drop.

Encoding:

In this dataset maximum important variables are in categorical variable, the Machine Learning model are working only in numerical data therefore needs to encode categorical variable into numerical variable.

- Label Encoding for all the variables that has exactly two categories.
- Ordinal encoding to the columns that exhibit a hierarchical structure.
- Dummy encoding for other columns and dropping the first column.

Checking Multi-Collinearity:

Checking for multicollinearity in regression analysis to ensure the reliability and interpretability of our model.

Using Variance Inflation Factor method for Checking multicollinearity and got some attribute they have huge amount of collinearity as per VIF criteria.

- WeightInKilograms - 77.822766
- BMI - 71.964811

Class imbalance and its treatment:

There is an imbalance in the dataset.

HadHeartAttack:

- Yes - 5.681 %
- No – 94.318 %

Train-Test Split:

The dataset was divided into training and testing subsets, with 70% of the data allocated for model training and 30% reserved for testing and evaluation purposes.

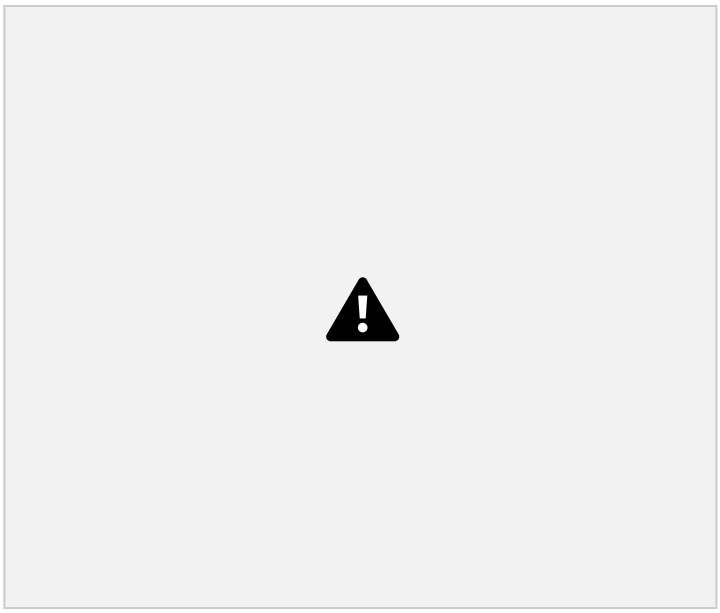
Imbalance Treatment:

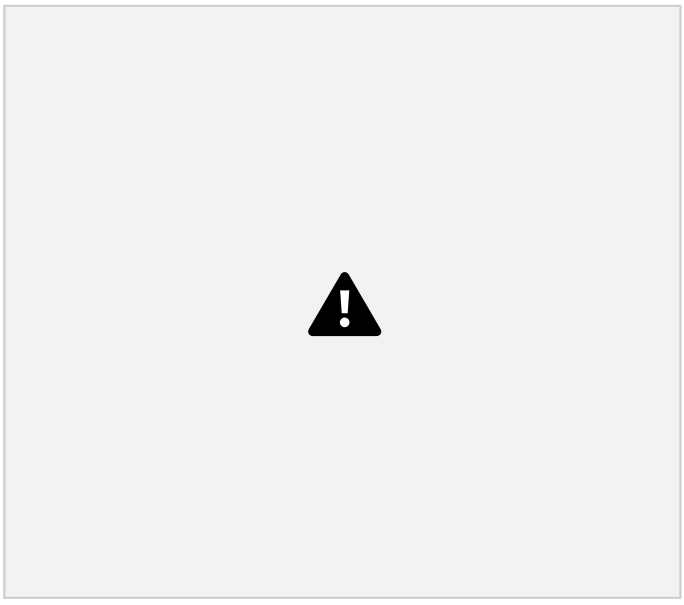
For treating imbalanced data, ‘Undersampling’ or ‘Class Weights’ while training the model can be used.

Model Building:

Several machine learning models were built to classify Had Heart Attack or not, including Logistic Regression, Decision Tree,Naïve Bayes Classifier,KNN Classifier Random Forest, Gradient boosting, XGboost, Adaboost. Each model was evaluated based on accuracy, precision, recall and F1 score.

Logistic Regression:





Recall [class-1]: 0.79 it means Out of all the people who actually have a heart attack, the model correctly identified 79%.The model is good at catching most of the true positives (TP), meaning it can identify the majority of individuals at risk.

Precision[class-1]: 0.24 out of all the people the model predicted as having a heart attack, only

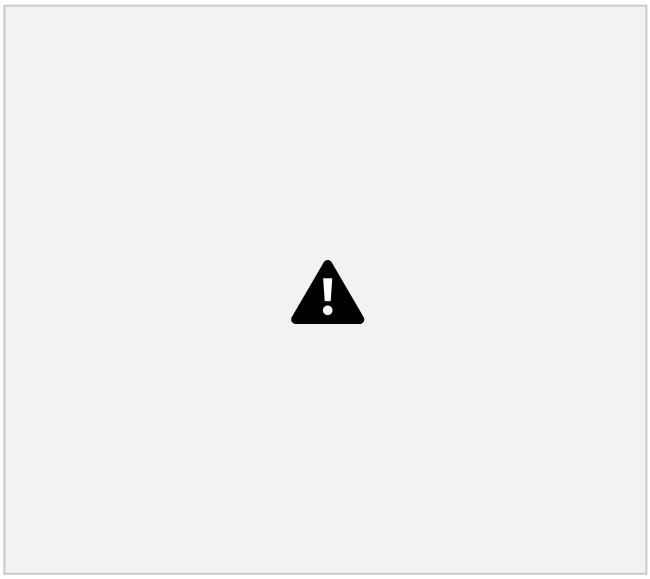
24% actually have the condition.

F1 Score[class-1]: 0.37 it Means the harmonic mean of precision and recall is low, indicating an imbalance between the two

Recall [class-0]: 0.85 the model correctly identifies 85% of the actual Class 0 cases (no heart attack). This suggests the model misses 15% of the actual no-heart-attack cases (false negatives). Precision[class-0]: 0.98 Among all predictions made for Class 0 (no heart attack), 98% are correct.

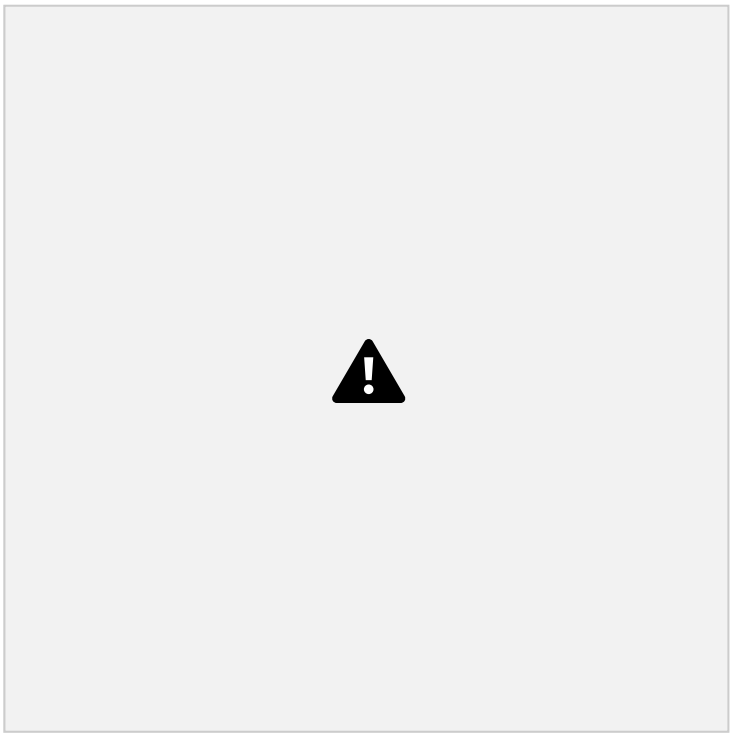
This indicates the model is very good at avoiding false positives (incorrectly predicting no heart attack when there is one).

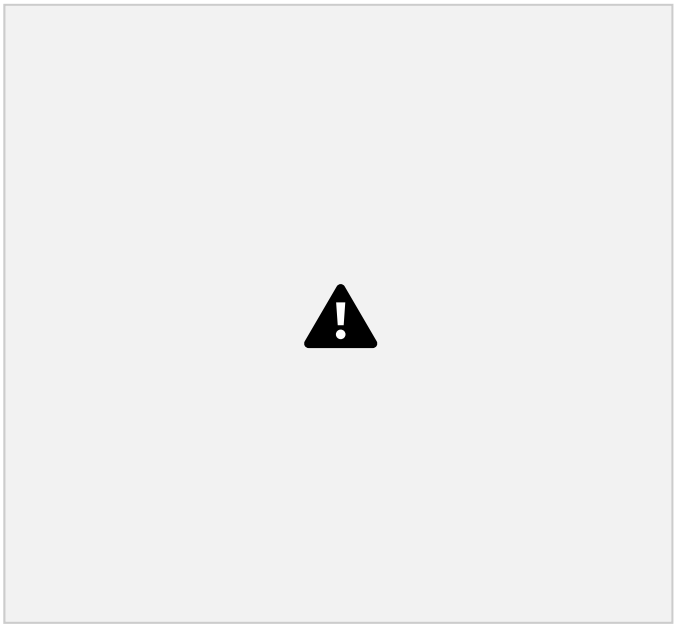
F1 Score[class-0]: 0.91 the harmonic mean of precision and recall indicates a good balance between avoiding false positives and false negatives.



The ROC AUC score is 0.900 (Test), which indicates the model has good discriminatory power between the two classes. The curve's shape shows that the model can balance sensitivity (true positive rate) and specificity (false positive rate) effectively.

KNN Classifier(after tunning):





Recall [class-1]: 0.75 it means Out of all the people who actually have a heart attack, the model correctly identified 75%.The model is good at catching most of the true positives (TP)

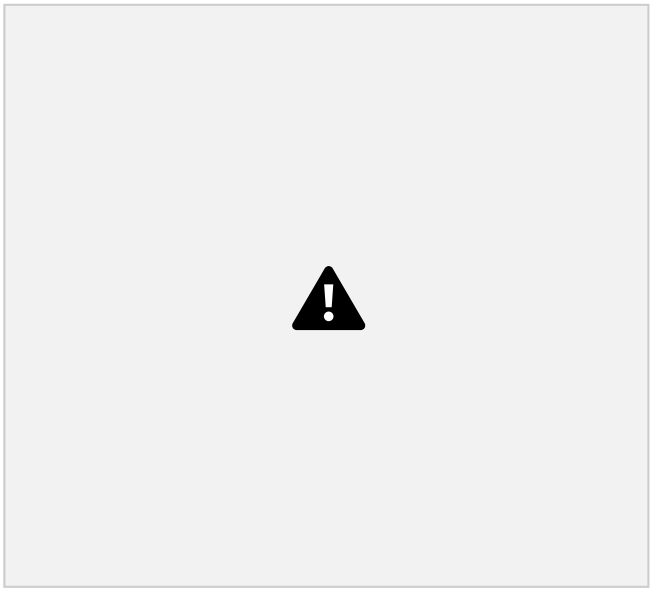
Precision[class-1]: 0.20 out of all the people the model predicted as having a heart attack, only 20% actually have the condition.

F1 Score[class-1]: 0.32 it Means the harmonic mean of precision and recall is low, indicating an imbalance between the two

Recall [class-0]: 0.82 the model correctly identifies 82% of the actual Class 0 cases (no heart attack). This suggests the model misses 15% of the actual no-heart-attack cases (false negatives). Precision[class-0]: 0.98 Among all predictions made for Class 0 (no heart attack), 98% are correct.

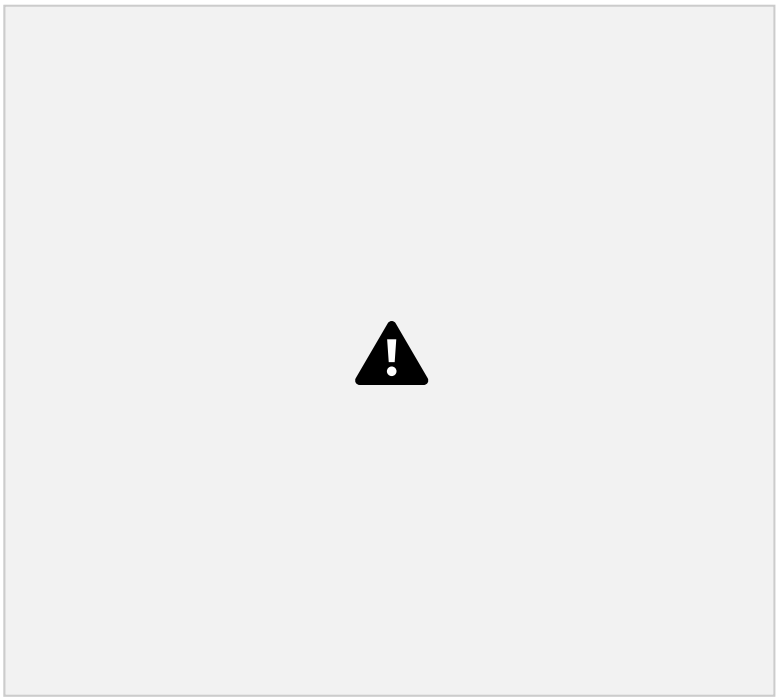
This indicates the model is very good at avoiding false positives (incorrectly predicting no heart attack when there is one).

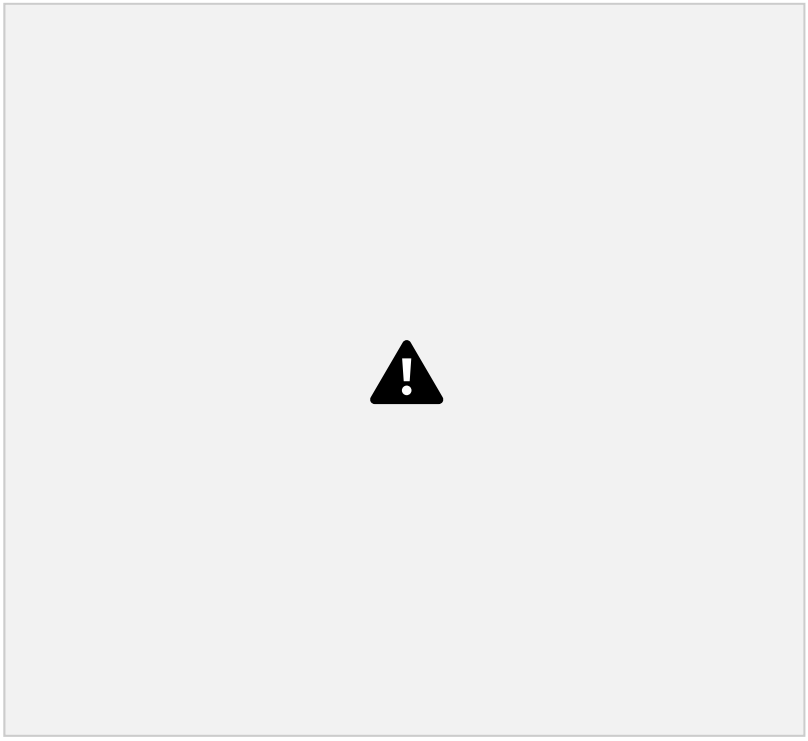
F1 Score[class-0]: 0.89 the harmonic mean of precision and recall indicates a good balance between avoiding false positives and false negatives.



The ROC AUC score is 0.863 (Test), which indicates the model has good discriminatory power between the two classes. The curve's shape shows that the model can balance sensitivity (true positive rate) and specificity (false positive rate) effectively.

Naive Bayes Classifier:





Recall [class-1]: 0.79 it means Out of all the people who actually have a heart attack, the model correctly identified 79%.The model is good at catching most of the true positives (TP), meaning it can identify the majority of individuals at risk.

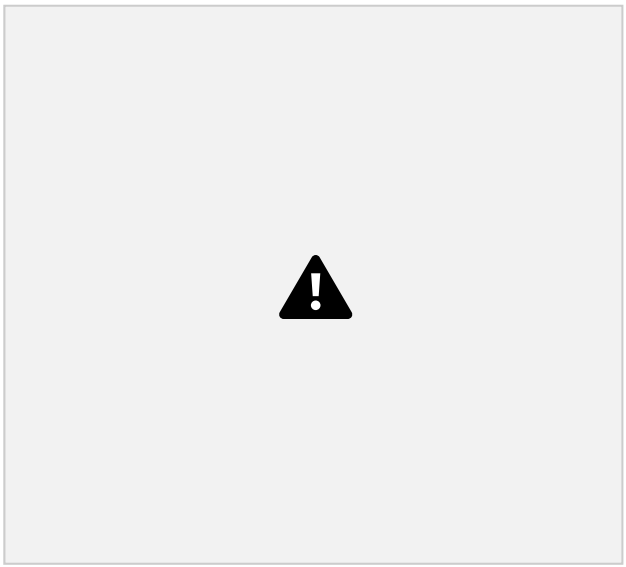
Precision[class-1]: 0.19 out of all the people the model predicted as having a heart attack, only 19% actually have the condition.

F1 Score[class-1]: 0.30 it Means the harmonic mean of precision and recall is low, indicating an imbalance between the two

Recall [class-0]: 0.79 the model correctly identifies 79 % of the actual Class 0 cases (no heart attack). Precision[class-0]: 0.98 Among all predictions made for Class 0 (no heart attack), 98% are correct.

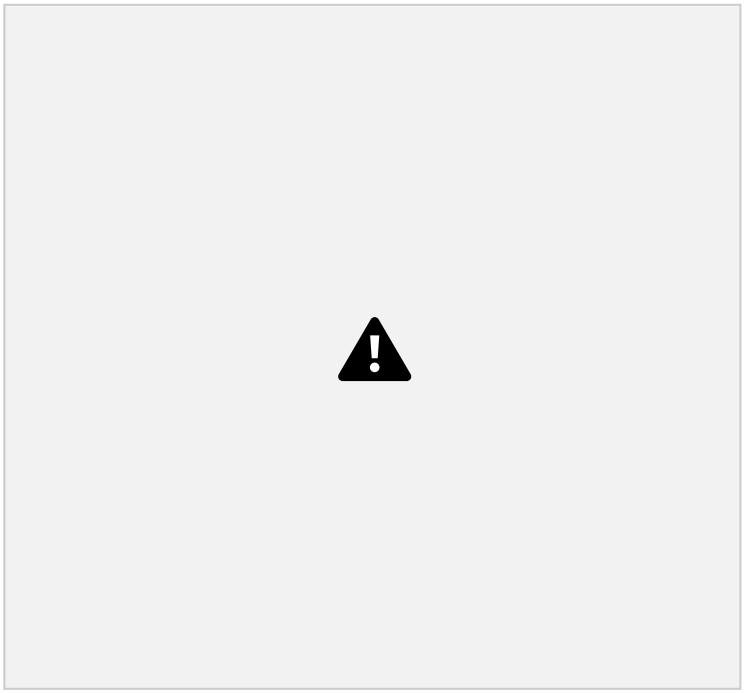
This indicates the model is very good at avoiding false positives (incorrectly predicting no heart attack when there is one).

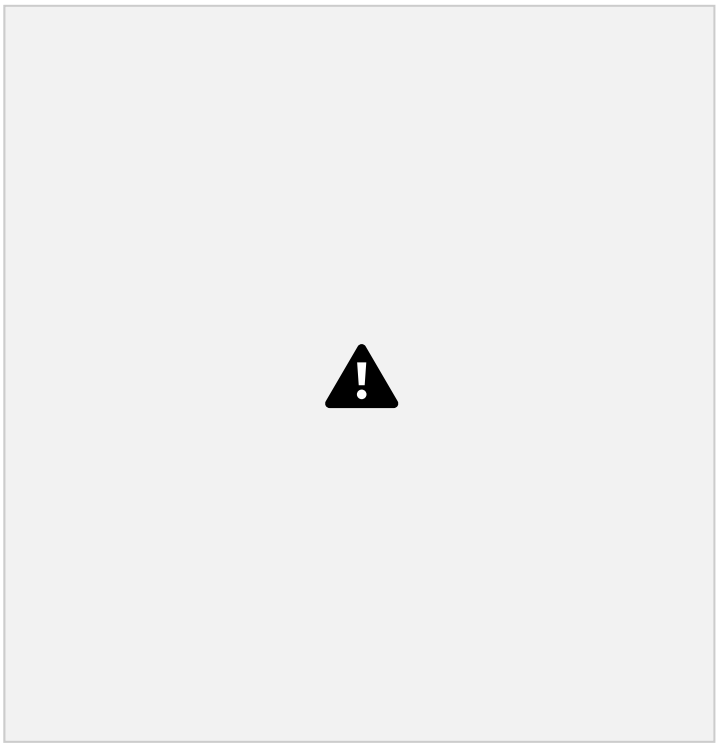
F1 Score[class-0]: 0.88 the harmonic mean of precision and recall indicates a good balance between avoiding false positives and false negatives.



The ROC AUC score is 0.861(Test), which indicates the model has good discriminatory power between the two classes. The curve's shape shows that the model can balance sensitivity (true positive rate) and specificity (false positive rate) effectively.

Decision Tree Classifier(after tunning):





Recall [class-1]: 0.81 it means Out of all the people who actually have a heart attack, the model correctly identified 81%.The model is good at catching most of the true positives (TP), meaning it can identify the majority of individuals at risk.

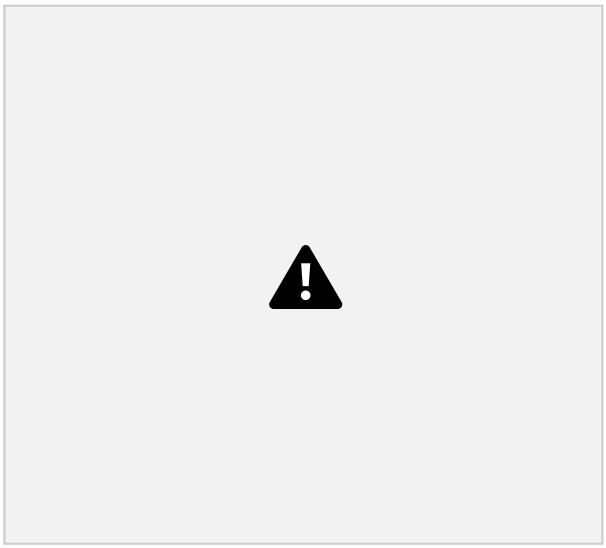
Precision[class-1]: 0.20 out of all the people the model predicted as having a heart attack, only 20% actually have the condition.

F1 Score[class-1]: 0.33 it Means the harmonic mean of precision and recall is low, indicating an imbalance between the two

Recall [class-0]: 0.81 the model correctly identifies 81% of the actual Class 0 cases (no heart attack). Precision[class-0]: 0.99 Among all predictions made for Class 0 (no heart attack), 99% are correct.

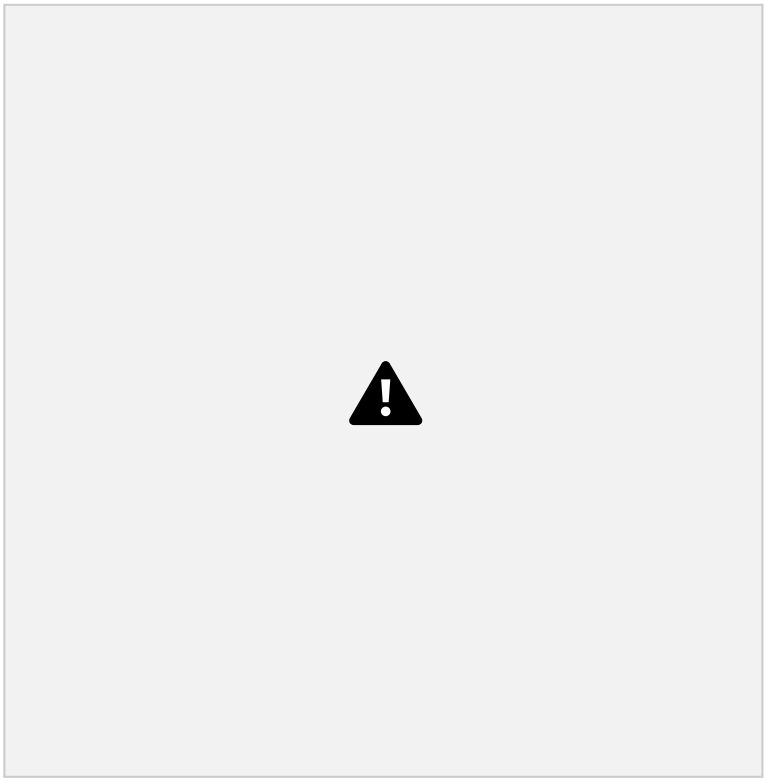
This indicates the model is very good at avoiding false positives (incorrectly predicting no heart attack when there is one).

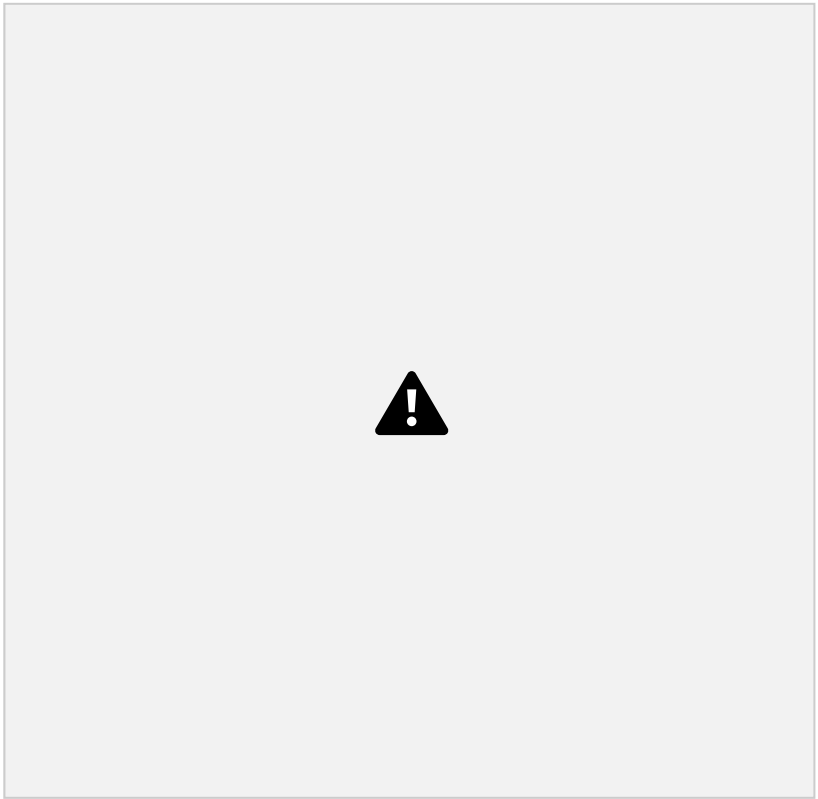
F1 Score[class-0]: 0.89 the harmonic mean of precision and recall indicates a good balance between avoiding false positives and false negatives.



The ROC AUC score is 0.893 (Test), which indicates the model has good discriminatory power between the two classes. The curve's shape shows that the model can balance sensitivity (true positive rate) and specificity (false positive rate) effectively.

Random Forest Classifier(after tunning):





Recall [class-1]: 0.80 it means Out of all the people who actually have a heart attack, the model correctly identified 80%.The model is good at catching most of the true positives (TP), meaning it can identify the majority of individuals at risk.

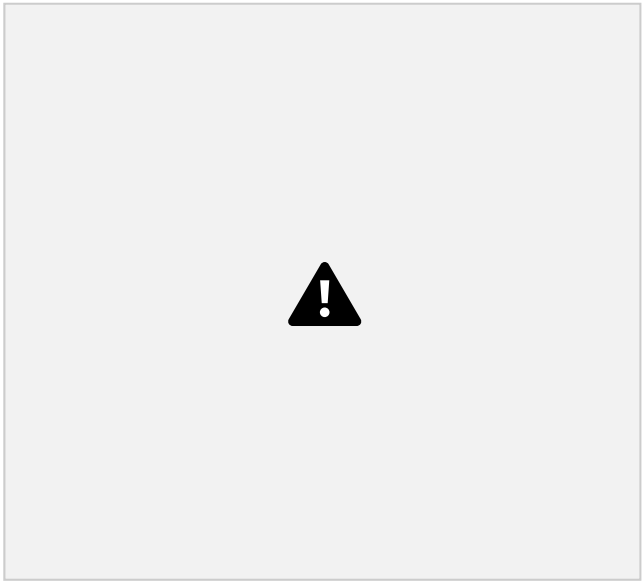
Precision[class-1]: 0.23 out of all the people the model predicted as having a heart attack, only 23% actually have the condition.

F1 Score[class-1]: 0.35 it Means the harmonic mean of precision and recall is low, indicating an imbalance between the two

Recall [class-0]: 0.83 the model correctly identifies 83% of the actual Class 0 cases (no heart attack). Precision[class-0]: 0.99 Among all predictions made for Class 0 (no heart attack), 99% are correct.

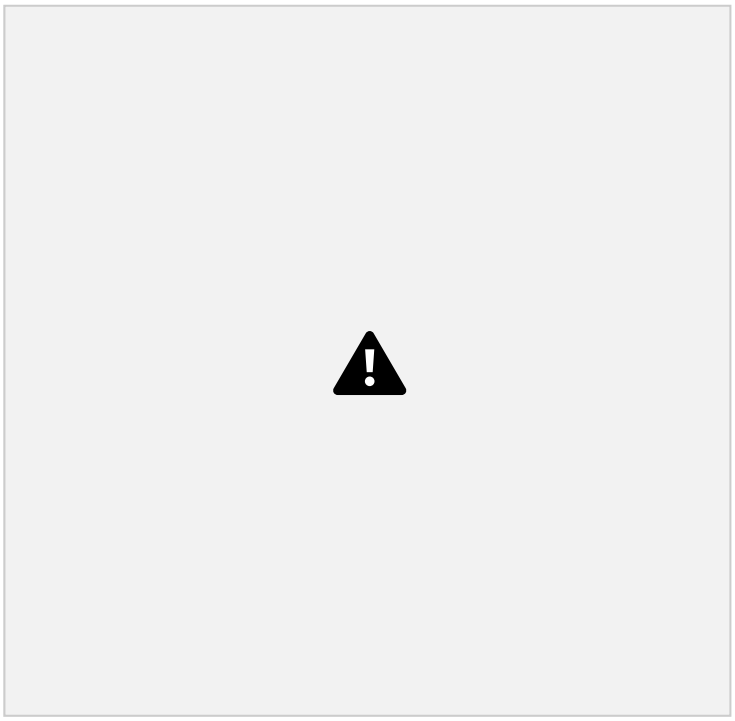
This indicates the model is very good at avoiding false positives (incorrectly predicting no heart attack when there is one).

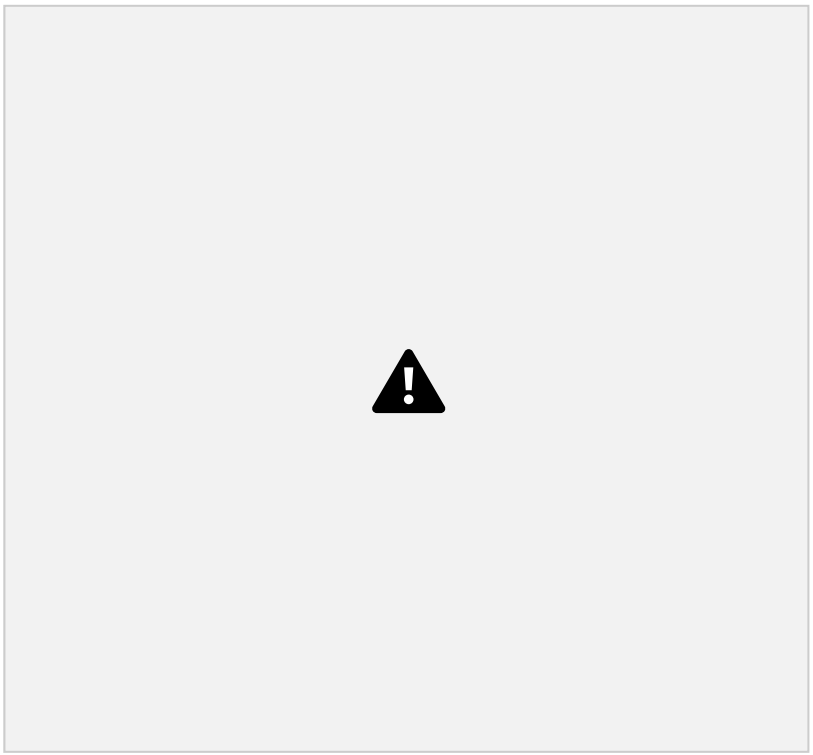
F1 Score[class-0]: 0.90 the harmonic mean of precision and recall indicates a good balance between avoiding false positives and false negatives.



The ROC AUC score is 0.901 (Test), which indicates the model has good discriminatory power between the two classes. The curve's shape shows that the model can balance sensitivity (true positive rate) and specificity (false positive rate) effectively.

Ada Boost(after tunning):





Recall [class-1]: 0.78 it means Out of all the people who actually have a heart attack, the model correctly identified 78%.The model is good at catching most of the true positives (TP), meaning it can identify the majority of individuals at risk.

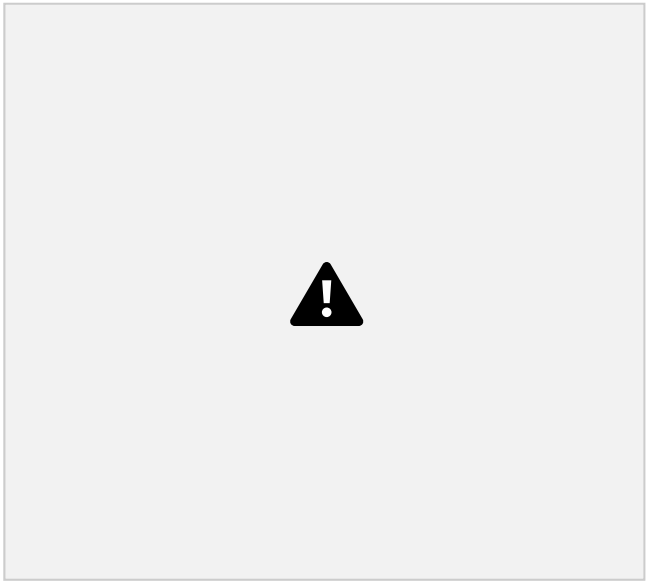
Precision[class-1]: 0.24 out of all the people the model predicted as having a heart attack, only 24% actually have the condition.

F1 Score[class-1]: 0.37 it Means the harmonic mean of precision and recall is low, indicating an imbalance between the two

Recall [class-0]: 0.85 the model correctly identifies 85% of the actual Class 0 cases (no heart attack). Precision[class-0]: 0.98 Among all predictions made for Class 0 (no heart attack), 98% are correct.

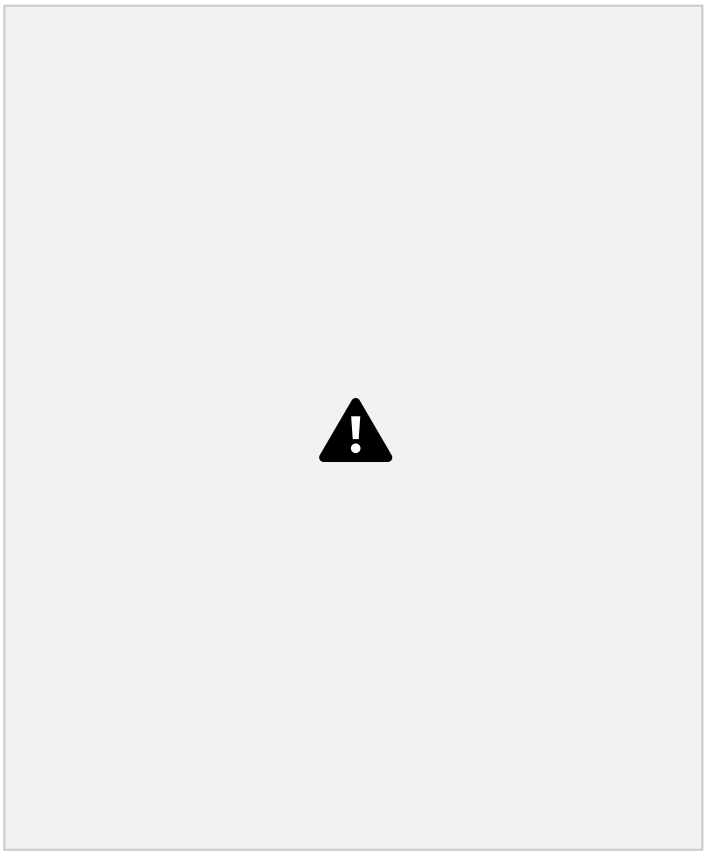
This indicates the model is very good at avoiding false positives (incorrectly predicting no heart attack when there is one).

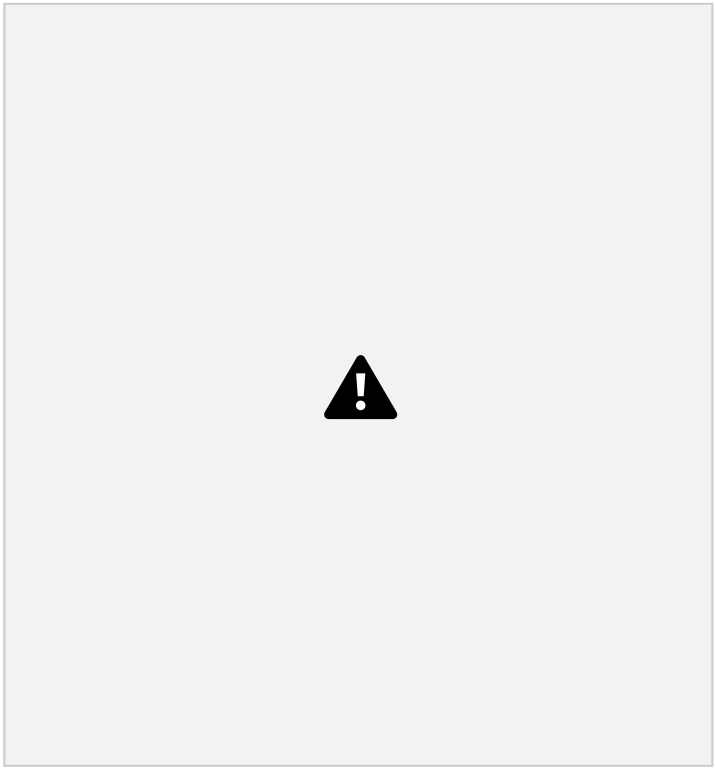
F1 Score[class-0]: 0.91 the harmonic mean of precision and recall indicates a good balance between avoiding false positives and false negatives.



The ROC AUC score is 0.9(Test), which indicates the model has good discriminatory power between the two classes. The curve's shape shows that the model can balance sensitivity (true positive rate) and specificity (false positive rate) effectively.

Gradient Boost(after tunning):





Recall [class-1]: 0.81 it means Out of all the people who actually have a heart attack, the model

correctly identified 81%.The model is good at catching most of the true positives (TP), meaning it can identify the majority of individuals at risk.

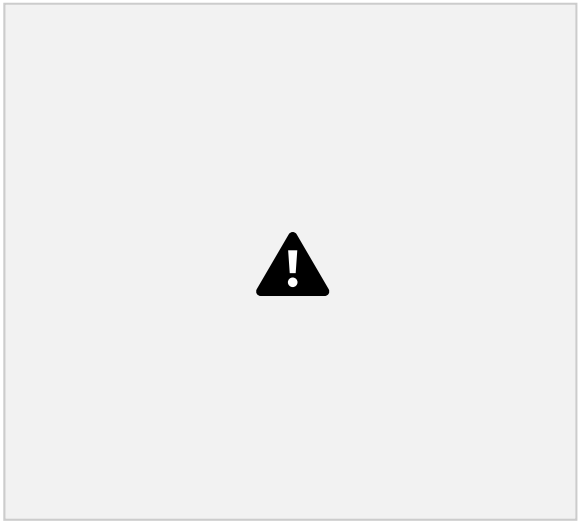
Precision[class-1]: 0.23 out of all the people the model predicted as having a heart attack, only 23% actually have the condition.

F1 Score[class-1]: 0.36 it Means the harmonic mean of precision and recall is low, indicating an imbalance between the two

Recall [class-0]: 0.84 the model correctly identifies 84% of the actual Class 0 cases (no heart attack). Precision[class-0]: 0.99 Among all predictions made for Class 0 (no heart attack), 99% are correct.

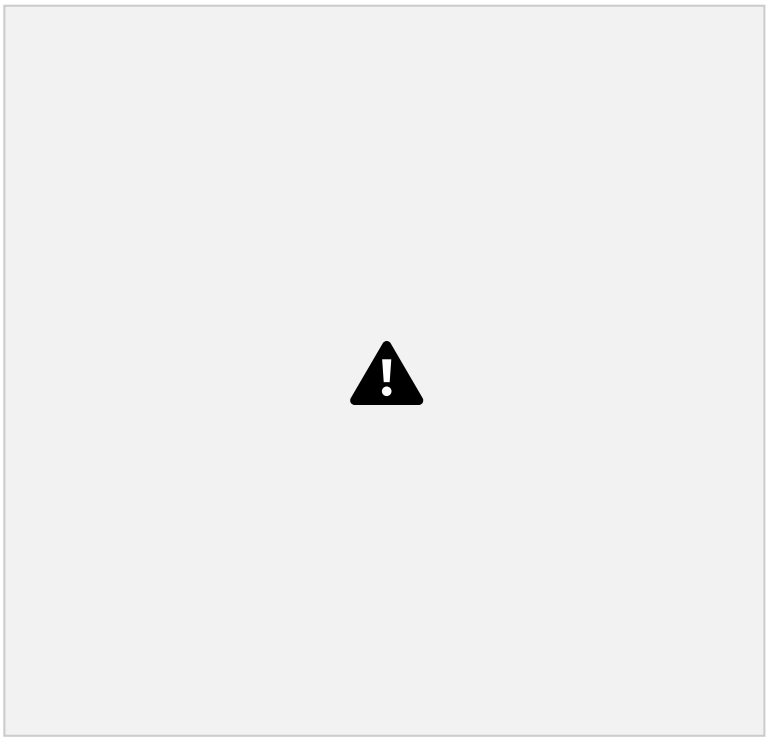
This indicates the model is very good at avoiding false positives (incorrectly predicting no heart attack when there is one).

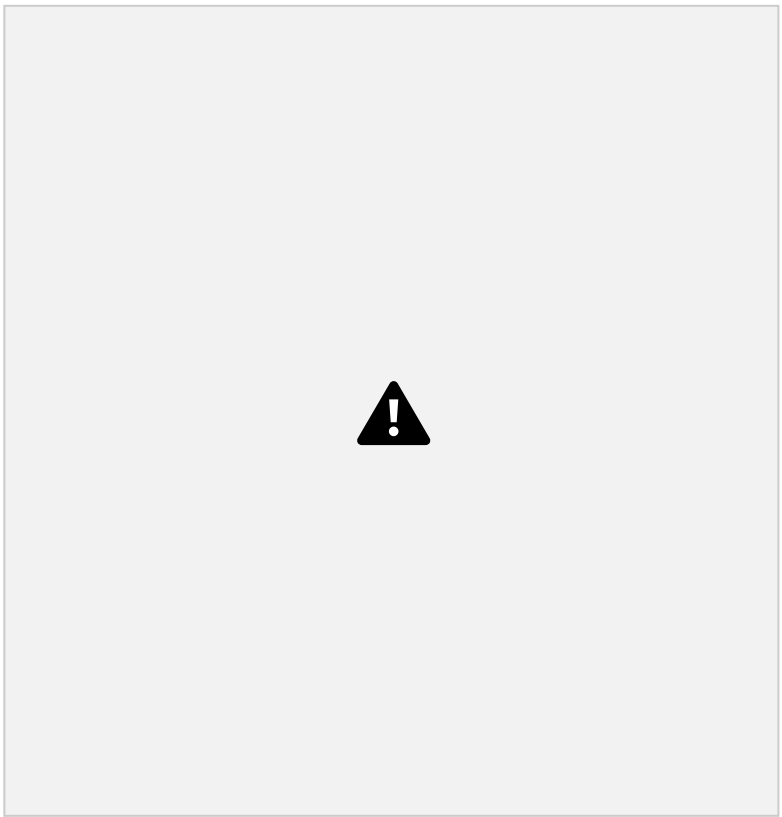
F1 Score[class-0]: 0.91 the harmonic mean of precision and recall indicates a good balance between avoiding false positives and false negatives.



The ROC AUC score is 0.905 (Test), which indicates the model has good discriminatory power between the two classes. The curve's shape shows that the model can balance sensitivity (true positive rate) and specificity (false positive rate) effectively.

XGBoost(after tunning):





Recall [class-1]: 0.81 it means Out of all the people who actually have a heart attack, the model correctly identified 81%.The model is good at catching most of the true positives (TP), meaning it can identify the majority of individuals at risk.

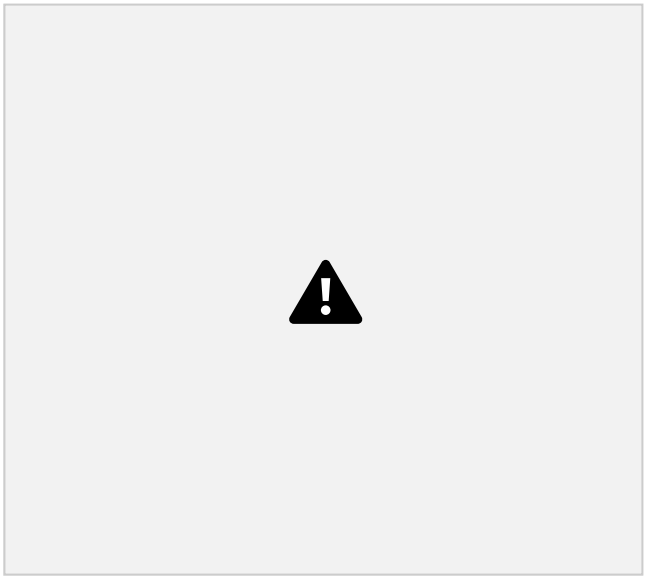
Precision[class-1]: 0.23 out of all the people the model predicted as having a heart attack, only 23% actually have the condition.

F1 Score[class-1]: 0.36 it Means the harmonic mean of precision and recall is low, indicating an imbalance between the two

Recall [class-0]: 0.83 the model correctly identifies 83% of the actual Class 0 cases (no heart attack). Precision[class-0]: 0.99 Among all predictions made for Class 0 (no heart attack), 99% are correct.

This indicates the model is very good at avoiding false positives (incorrectly predicting no heart attack when there is one).

F1 Score[class-0]: 0.90 the harmonic mean of precision and recall indicates a good balance between avoiding false positives and false negatives.



The ROC AUC score is 0.905 (Test), which indicates the model has good discriminatory power between the two classes. The curve's shape shows that the model can balance sensitivity (true positive rate) and specificity (false positive rate) effectively.

Comparison of Model Performance Metrics:

Model	Accuracy		Precision		Recall		F1 score	
	Train	Test	Train	Test	Train	Test	Train	Test
Logistic Regression	0.84	0.84	0.23	0.24	0.79	0.79	0.36	0.37
KNN classifier	0.82	0.82	0.84	0.20	0.80	0.75	0.82	0.32

Naïve Bayes Classifier	0.79	0.79	0.18	0.19	0.79	0.79	0.30	0.30
Decision Tree Classifier	0.81	0.81	0.20	0.20	0.83	0.81	0.33	0.33
Random Forest Classifier	0.83	0.83	0.23	0.23	0.81	0.80	0.35	0.35
Ada Boost	0.82	0.85	0.84	0.24	0.79	0.78	0.81	0.37
Gradient Boost	0.83	0.84	0.84	0.23	0.81	0.81	0.82	0.36
XGBoost	0.84	0.83	0.84	0.23	0.83	0.81	0.83	0.36

As per the performance metrics, XGBoost is giving the good results. All the boosting algorithms are stable and consistent. The best model for detecting Heart Attack is XGBoost (After Tunning). This model achieves a recall of 0.81 for classifying Had Heart Attack (class 1), meaning it successfully identifies 81% of the actual Heart patient.

Limitations of the Solution:

Class Imbalance:

While our dataset showed imbalance between "Yes-5.6%" and "No-94.3%", we adopted to apply resampling techniques (e.g., class weight or under sampling), which worked reasonably well.

Impact of External Factors:

Predicting Had Heart Attack is a multifaceted problem influenced by many external factors that is not in aur dataset, including Poor Diet, Cholesterol, High Blood Pressure, Air Pollution, High stress jobs, Drug Abuse, Cold weather or sudden temperature drops.

Model Interpretability:

The XGBoost model, used for predicting the likelihood of heart attacks, provides interpretability through its ability to highlight key relationships between features and outcomes. This interpretability aligns with the business goal of reducing heart attack risks by offering healthcare providers actionable insights.

Enhancements to Improve the Solution:

While the current model based on XGBoost demonstrates promising results with a recall of 0.81 and accuracy of 0.83, there are several avenues for further improvement. These enhancements focus on refining the model's predictive capabilities and ensuring that it performs well across different aspects of heart attack prediction.

1.Model Ensemble:

While XGBoost is performing well, combining multiple models in an ensemble could boost performance. A Stacking Classifier that combines the predictions of different models, such as Random Forest, Logistic Regression, and XGBoost, can potentially improve both recall and accuracy. This would help leverage the strengths of multiple algorithms and enhance the overall prediction capability.

2.Model Interpretability and Explainability:

Building a model that can explain its predictions is crucial, especially in healthcare applications. Incorporating SHAP (Shapley Additive Explanations) values can provide transparency into which features are driving the predictions, helping healthcare professionals trust and understand the results. This could also assist in identifying which factors most contribute to heart attack risk, leading to better preventive measures.

3.External Data Integration:

Integrating external datasets or publicly available health records with your existing dataset could potentially improve model performance. This would allow the model to learn from a larger, more diverse set of data, leading to better generalization. Additionally, considering factors like lifestyle and socio-economic status, which can influence heart disease, could enhance the model's predictions.

4.Continuous Monitoring and Model Updating:

Heart disease predictions may change over time due to shifts in population health or changes in medical practices. Implementing a system for continuous learning where the model is periodically retrained on newer data can keep the model's performance up-to-date and ensure its continued effectiveness in predicting heart attacks.

Closing Reflections:

This project provided a deep understanding of predictive analysis, model development, and the critical role data-driven insights play in solving real-world Health care problems in the Healthcare industry. Here are the key lessons and potential improvements for the future:

Lessons Learned:

1. Data Exploration and Feature Engineering:

Effective exploratory data analysis (EDA) is crucial to understanding the data and identifying relationships between variables. The univariate and bivariate analysis in this project revealed meaningful patterns, such as the impact of physical activity and age on the Heart health, which guided model building.

2. Handling Imbalanced Data:

While our dataset showed a 70:30 imbalance between "Yes" and "No" Heart Attack, we adopted to apply resampling techniques (e.g., class weight or under sampling), which worked reasonably well. However, it also highlighted the importance of carefully analysing the impact of class imbalance on model performance.

3. Model Selection and Evaluation:

Experimenting with various machine learning algorithms helped me gain insights into their strengths and weaknesses. Ensemble methods, such as Random Forest and XGBoost, demonstrated strong performance, while simpler models like logistic regression provided interpretability. The ability to compare and optimize multiple models is a key skill developed through this project.

4. Overfitting and Cross-Validation:

The decision tree model performed well, but overfitting was an area of concern. Incorporating hyperparameter tuning improved generalization, using techniques like bagging (e.g., Random Forests) or boosting (e.g., XGBoost) to reduce overfitting strategies in preventing overfitting.

5. Real-World Challenges:

Predicting Had Heart Attack is a multifaceted problem influenced by many external factors that is not in aur dataset, including Poor Diet, Cholesterol, High Blood Pressure, Air Pollution, High stress jobs, Drug Abuse, Cold weather or sudden temperature drops. While we achieved high performance in model testing, the unpredictable nature of human behaviour means that real-world performance may not always align perfectly with our predictions.

Base Model:

Assumptions:

Logistic Regression

Since the target variable is binary (categorical), including predicting the presence of heart disease we chose the base model as Logistic Regression effectively.

For best prediction we are focusing on Recall and Accuracy.

References:

Pytlak, K. (2021). *Personal Key Indicators of Heart Disease* [Dataset]. Kaggle.

Retrieved from

https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-diseas

e