HEART ATTCK PREDICTION USING STATISTICAL METHODS

A term project report submitted for

MA 541 - Statistical Methods

by

Lokesh Lochan Dharmavaram

Naveen Venkat Yelamanchali

Sri Sai Krishna Nawathe

Under the guidance of

Prof. HONG DO

Stevens Institute of Technology

Contents:

Abstract:

This project focuses on predicting heart attack risk using a comprehensive dataset. By applying various statistical tests, machine learning models, and exploratory data analysis (EDA), we aimed to identify significant predictors of heart attack risk and develop a robust predictive model. The project addressed challenges like data imbalance and incorporated advanced techniques like oversampling, under sampling, and feature scaling.

Introduction:

Heart attack prediction is a critical area of medical research, offering potential for early intervention and prevention. The project utilizes a dataset containing multiple variables, potentially influencing heart attack risk. The objective is to analyze these variables, understand their relationships, and develop a predictive model with high accuracy and reliability.

Data:

The dataset includes a range of variables such as age, sex, cholesterol levels, blood pressure, lifestyle factors, and medical history, alongside the target variable 'Heart Attack Risk'. The data was sourced to ensure a comprehensive representation of factors known to influence heart attack risk.

Dataset Source:

The dataset is available on Kaggle and is sourced from a real-world study. You can access it here: https://www.kaggle.com/datasets/iamsouravbanerjee/heart-attack-prediction-dataset

Data Schema:

```
RangeIndex: 8763 entries, 0 to 8762
Data columns (total 26 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   Patient ID                      8763 non-null   object
 1   Age                             8763 non-null   int64
 2   Sex                             8763 non-null   object
 3   Cholesterol                     8763 non-null   int64
 4   Blood Pressure                  8763 non-null   object
 5   Heart Rate                      8763 non-null   int64
 6   Diabetes                        8763 non-null   int64
 7   Family History                  8763 non-null   int64
 8   Smoking                         8763 non-null   int64
 9   Obesity                         8763 non-null   int64
 10  Alcohol Consumption             8763 non-null   int64
 11  Exercise Hours Per Week         8763 non-null   float64
 12  Diet                            8763 non-null   object
 13  Previous Heart Problems         8763 non-null   int64
 14  Medication Use                  8763 non-null   int64
 15  Stress Level                    8763 non-null   int64
 16  Sedentary Hours Per Day         8763 non-null   float64
 17  Income                          8763 non-null   int64
 18  BMI                             8763 non-null   float64
 19  Triglycerides                   8763 non-null   int64
 20  Physical Activity Days Per Week 8763 non-null   int64
 21  Sleep Hours Per Day             8763 non-null   int64
 22  Country                         8763 non-null   object
 23  Continent                       8763 non-null   object
 24  Hemisphere                      8763 non-null   object
 25  Heart Attack Risk               8763 non-null   int64
dtypes: float64(3), int64(16), object(7)
memory usage: 1.7+ MB
```

Types of Variables:

Variables in the dataset include qualitative (e.g., gender), quantitative (e.g. age, ca, fibs)

Data Cleaning:

The dataset has undergone initial data cleaning, addressing missing values and data consistency.

Data Description:

	Age	Cholesterol	Heart Rate	Diabetes	Family History	Smoking
count	8763.000000	8763.000000	8763.000000	8763.000000	8763.000000	8763.000000
mean	53.707977	259.877211	75.021682	0.652288	0.492982	0.896839
std	21.249509	80.863276	20.550948	0.476271	0.499979	0.304186
min	18.000000	120.000000	40.000000	0.000000	0.000000	0.000000
25%	35.000000	192.000000	57.000000	0.000000	0.000000	1.000000
50%	54.000000	259.000000	75.000000	1.000000	0.000000	1.000000
75%	72.000000	330.000000	93.000000	1.000000	1.000000	1.000000
max	90.000000	400.000000	110.000000	1.000000	1.000000	1.000000

Obesity	Alcohol Consumption	Exercise Hours Per Week	Previous Heart Problems	Medication Use	Stress Level	Sedentary Hours Per Day
8763.000000	8763.000000	8763.000000	8763.000000	8763.000000	8763.000000	8763.000000
0.501426	0.598083	10.014284	0.495835	0.498345	5.469702	5.993690
0.500026	0.490313	5.783745	0.500011	0.500026	2.859622	3.466359
0.000000	0.000000	0.002442	0.000000	0.000000	1.000000	0.001263
0.000000	0.000000	4.981579	0.000000	0.000000	3.000000	2.998794
1.000000	1.000000	10.069559	0.000000	0.000000	5.000000	5.933622
1.000000	1.000000	15.050018	1.000000	1.000000	8.000000	9.019124
1.000000	1.000000	19.998709	1.000000	1.000000	10.000000	11.999313

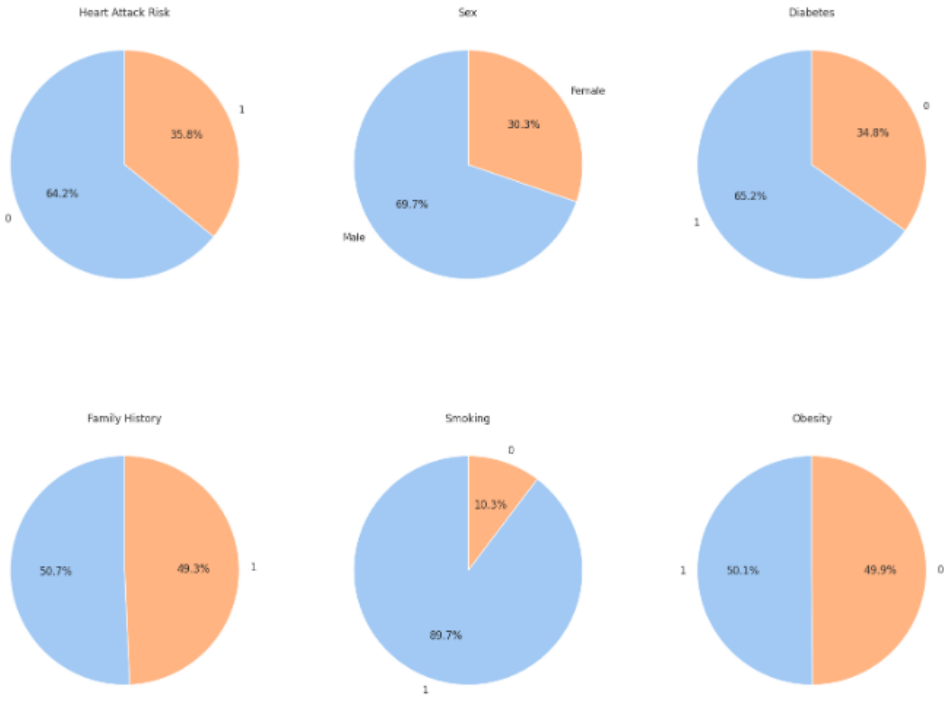
Sedentary Hours Per Day	Income	BMI	Triglycerides	Physical Activity Days Per Week	Sleep Hours Per Day	Heart Attack Risk
8763.000000	8763.000000	8763.000000	8763.000000	8763.000000	8763.000000	8763.000000
5.993690	158263.181901	28.891446	417.677051	3.489672	7.023508	0.358211
3.466359	80575.190806	6.319181	223.748137	2.282687	1.988473	0.479502
0.001263	20062.000000	18.002337	30.000000	0.000000	4.000000	0.000000
2.998794	88310.000000	23.422985	225.500000	2.000000	5.000000	0.000000
5.933622	157866.000000	28.768999	417.000000	3.000000	7.000000	0.000000
9.019124	227749.000000	34.324594	612.000000	5.000000	9.000000	1.000000
11.999313	299954.000000	39.997211	800.000000	7.000000	10.000000	1.000000

Each feature includes statistical measures such as count, mean, standard deviation (std), minimum (min), 25th percentile (25%), 50th percentile (50%), 75th percentile (75%), and maximum (max).

This data is used to perform various statistical analyses, such as regression analysis, logistic regression, or machine learning classification methods, to predict the likelihood of a heart attack. By including both lifestyle factors (like sedentary hours and alcohol consumption) and medical indicators (like BMI and cholesterol levels), the dataset provides a comprehensive view of the factors that can contribute to cardiovascular risk. The dataset is also quite extensive, which is beneficial for training predictive models with higher accuracy.

Data Visualization:

Pie Charts



Heart Attack Risk:

This chart shows the proportion of individuals in the dataset with and without heart attack risk. 64.2% of individuals are categorized as not at risk (0), and 35.8% as at risk (1).

Sex:

The distribution of individuals by sex is displayed, with a larger portion of males (69.7%) compared to females (30.3%).

Diabetes:

The chart represents the prevalence of diabetes among individuals, with 65.2% not having diabetes (0), and 34.8% having diabetes (1).

Family History:

This pie chart illustrates the distribution of family history of heart disease, with 50.7% having no family history (0) and 49.3% having a family history (1) of heart disease.

Smoking:

Smoking habits are shown, with a majority of 89.7% being non-smokers (0) and a smaller 10.3% being smokers (1).

Obesity:

The obesity distribution is nearly even, with 50.1% not obese (0) and 49.9% classified as obese (1).

Alcohol Consumption:

The chart depicts the proportion of individuals by alcohol consumption status. 59.8% of individuals are shown to consume alcohol (1), while 40.2% do not (0).

Diet:

This chart categorizes individuals by their diet quality into three parts: 33.8% have a healthy diet, 33.0% have an unhealthy diet, and 33.2% have an average diet.

Previous Heart Problems:

The pie chart illustrates the distribution of individuals with and without previous heart problems. 50.4% have no previous heart problems (0), and 49.6% do have previous heart problems (1).

Medication Use:

The medication use chart is evenly split, with 50.2% not using medication (0) and 49.8% using medication (1).

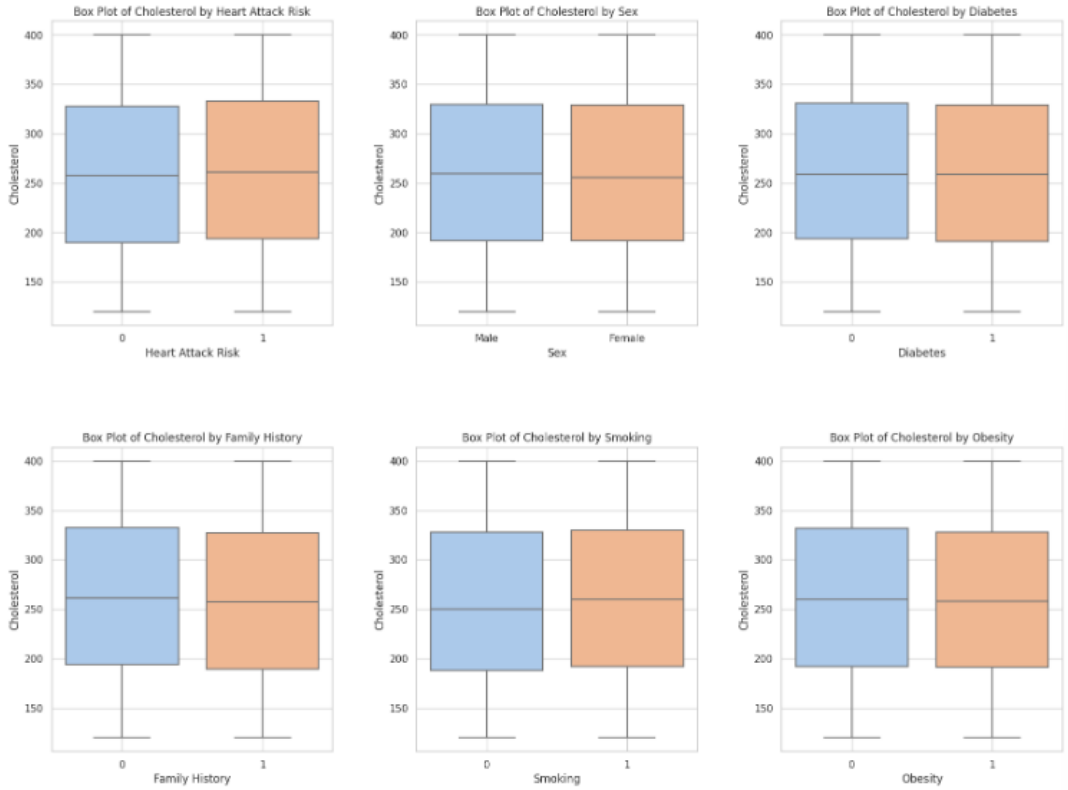
Country:

This chart shows the diversity of the dataset across various countries, with percentages representing the proportion of individuals from each country.

Continent:

The continent distribution is represented, indicating the spread of the data across continents with percentages for Asia, North America, Africa, Australia, and South America.

Box Plots



Cholesterol by Heart Attack Risk:

This plot compares cholesterol levels between individuals with (1) and without (0) heart attack risk. The median, quartiles, and range for each group can be analyzed to determine if there's a significant difference in cholesterol levels between the two groups.

Cholesterol by Sex:

The distribution of cholesterol levels is shown separately for males and females. This can be important in determining if sex is a significant predictor of cholesterol levels, which may affect heart attack risk.

Cholesterol by Diabetes:

Individuals are categorized by the presence (1) or absence (0) of diabetes. The cholesterol levels for each category can indicate whether diabetes status is associated with differing cholesterol levels.

Cholesterol by Family History:

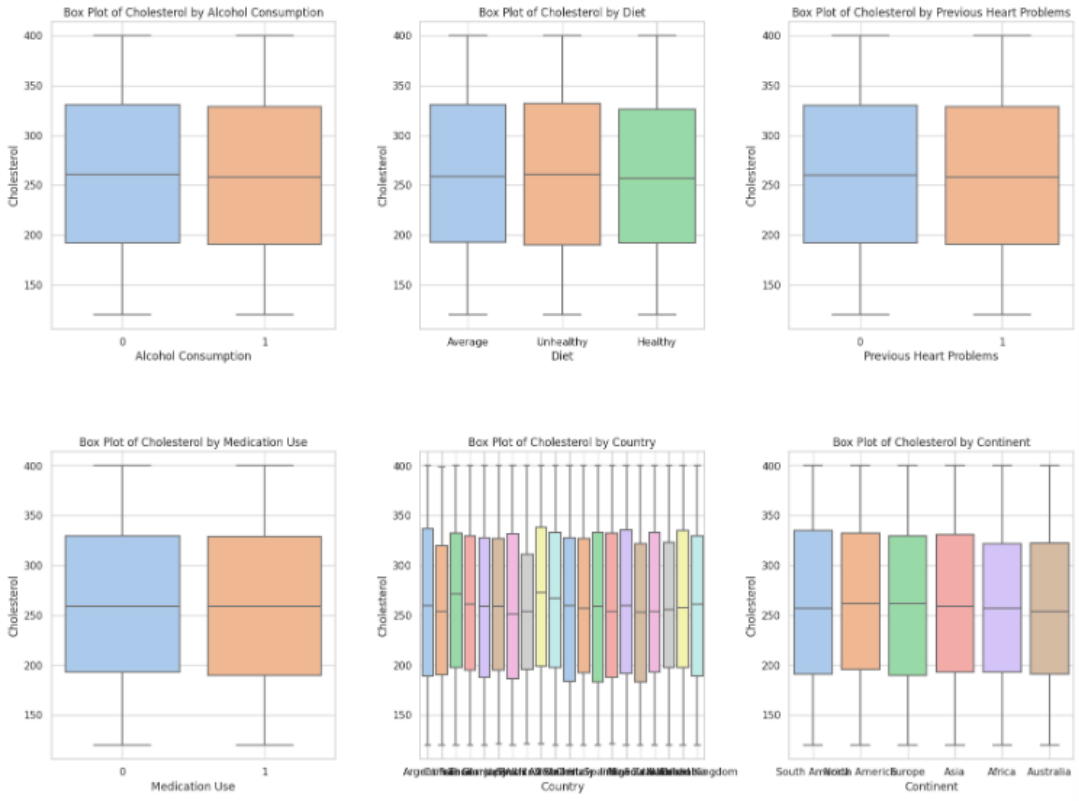
The box plot shows cholesterol levels for individuals with (1) and without (0) a family history of heart disease. This could suggest a genetic component to cholesterol levels, influencing heart attack risk.

Cholesterol by Smoking:

Cholesterol levels are compared between smokers (1) and non-smokers (0). This can highlight the impact of smoking on cholesterol and subsequent heart attack risk.

Cholesterol by Obesity:

The plot distinguishes between obese (1) and non-obese (0) individuals. Obesity is a known risk factor for elevated cholesterol, which is a significant risk factor for heart attacks.



Cholesterol by Alcohol Consumption:

This box plot compares the cholesterol levels between two groups: those who consume alcohol (1) and those who do not (0). It would be important to observe the median, spread, and any potential outliers to understand if alcohol consumption is associated with higher or lower cholesterol levels.

Cholesterol by Diet:

Three groups based on diet quality are compared: average, unhealthy, and healthy. The plot can provide insights into whether dietary habits are correlated with cholesterol levels, which could influence heart attack risk.

Cholesterol by Previous Heart Problems:

Individuals with previous heart problems (1) are compared to those without (0). The cholesterol distribution in these groups could indicate if past heart problems are associated with higher cholesterol levels.

Cholesterol by Medication Use:

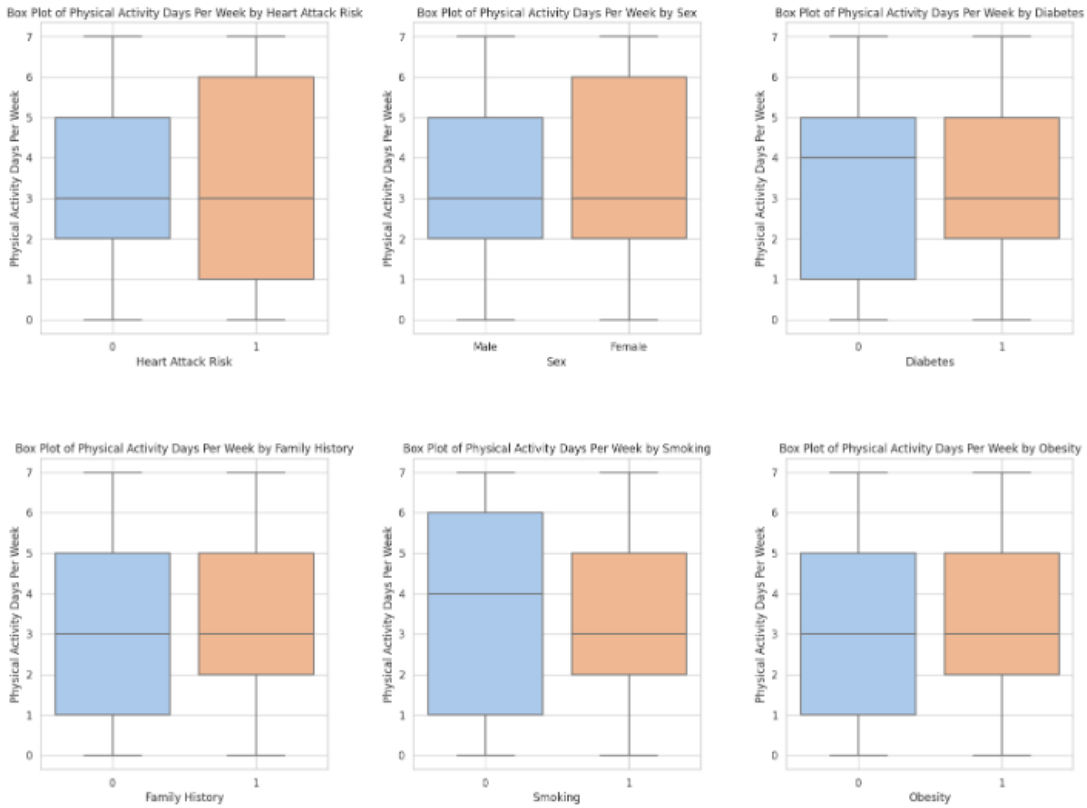
This plot differentiates between individuals who use medication (1) and those who do not (0). Medication use could be related to managing cholesterol, which may be reflected in the distribution.

Cholesterol by Country:

The cholesterol levels across different countries are shown. This can illustrate the geographical variation in cholesterol levels, potentially due to genetic, dietary, or lifestyle factors.

Cholesterol by Continent:

Like the country distribution, this plot shows cholesterol levels across different continents. This broader geographical comparison can be used to assess regional patterns in cholesterol levels.



Physical Activity Days Per Week by Heart Attack Risk:

Compares the distribution of physical activity days between two groups: those categorized as having a heart attack risk (1) and those without (0). The plot may show different medians, which indicates central tendencies of physical activity days per week for each group, and a comparison of the IQRs (the height of the boxes) indicates the spread of the middle 50% of the data. The whiskers show the range of the data excluding outliers.

Physical Activity Days Per Week by Sex:

Displays the distribution of physical activity days for males and females. It would be important to note any significant differences in the median or variability of physical activity between sexes.

Physical Activity Days Per Week by Diabetes:

Shows the physical activity distribution for individuals with diabetes (1) and without diabetes (0). The plot can indicate if diabetes status affects the frequency of physical activity.

Physical Activity Days Per Week by Family History:

Illustrates how physical activity varies between individuals with (1) and without (0) a family history of heart disease. You may find differences in the levels of physical activity that could be linked to genetic predispositions.

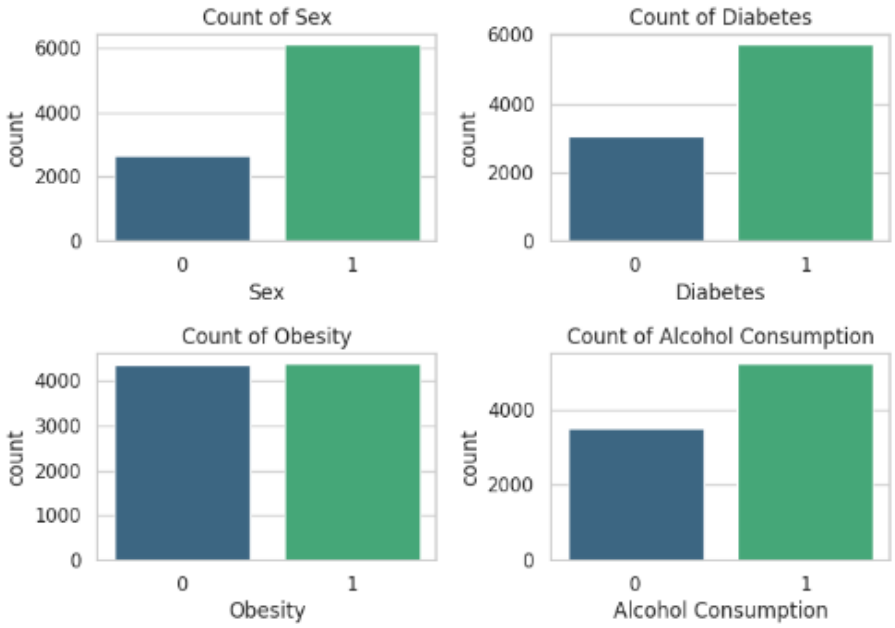
Physical Activity Days Per Week by Smoking Status:

Compares the physical activity between non-smokers (0) and smokers (1). This plot may reveal lifestyle patterns that correlate smoking with physical activity levels.

Physical Activity Days Per Week by Obesity Status:

Contrasts physical activity days for obese (1) and non-obese (0) individuals. This comparison can indicate whether obesity is associated with lower levels of physical activity.

Bar Charts



Count of Sex:

The bar chart displays the distribution of individuals by sex, where "0" could represent one sex (e.g., female) and "1" the other (e.g., male). The chart should show if there is a significant difference in the number of individuals of each sex within the dataset.

Count of Diabetes:

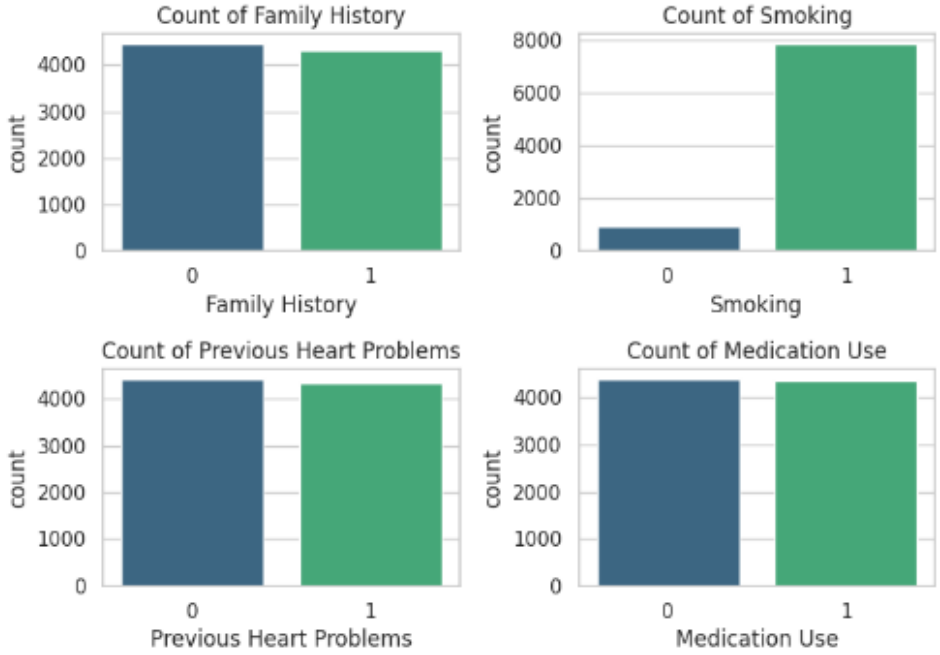
This chart illustrates the number of individuals with and without diabetes, with "0" typically representing non-diabetic individuals and "1" representing diabetic individuals. The chart is useful for showing the prevalence of diabetes in the study population.

Count of Obesity:

The bar chart shows the count of obese and non-obese individuals, with "0" indicating non-obese and "1" indicating obese. This can give an insight into the proportion of the population that is at a higher risk of heart attacks due to obesity.

Count of Alcohol Consumption:

The chart provides a count of individuals who do not consume alcohol ("0") versus those who do ("1"). Alcohol consumption can be an important factor in heart health, and its distribution can affect the modeling of heart attack risk.



Count of Family History:

This bar chart represents the number of individuals with and without a family history of heart disease, where "0" likely indicates no family history and "1" indicates the presence of family history. It shows the distribution within your dataset, which is relevant since a family history of heart disease is a known risk factor.

Count of Smoking:

The chart displays the count of non-smokers and smokers, with "0" likely denoting non-smokers and "1" denoting smokers. Since smoking is a significant risk factor for heart disease, the chart's distribution can be important for assessing risk in your predictive modeling.

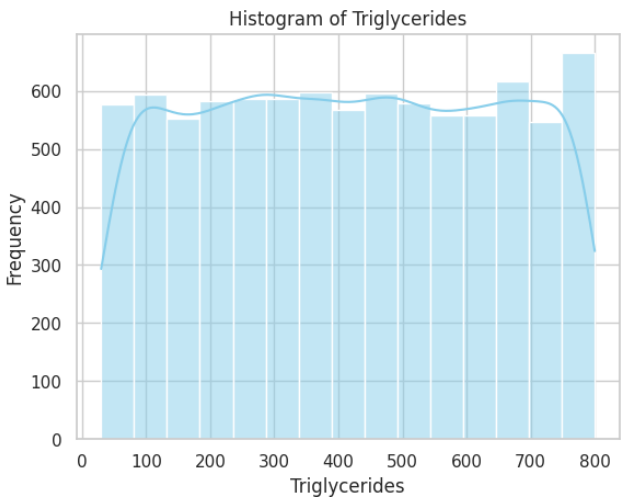
Count of Previous Heart Problems:

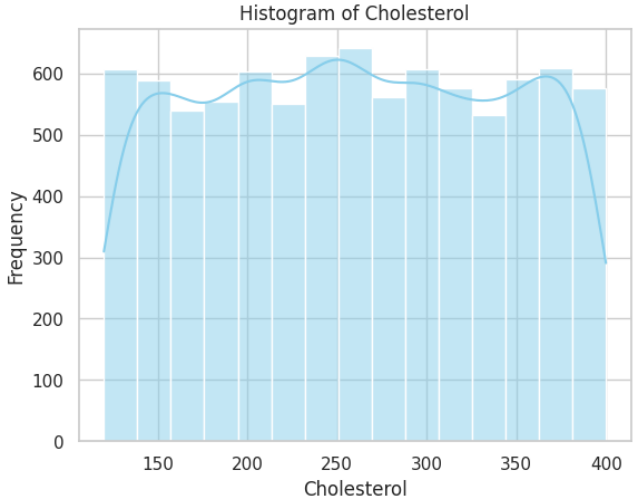
This bar chart compares the number of individuals who have had previous heart problems ("1") against those who have not ("0"). The presence of previous heart problems can be a strong indicator of future risk, making this distribution critical for prediction.

Count of Medication Use:

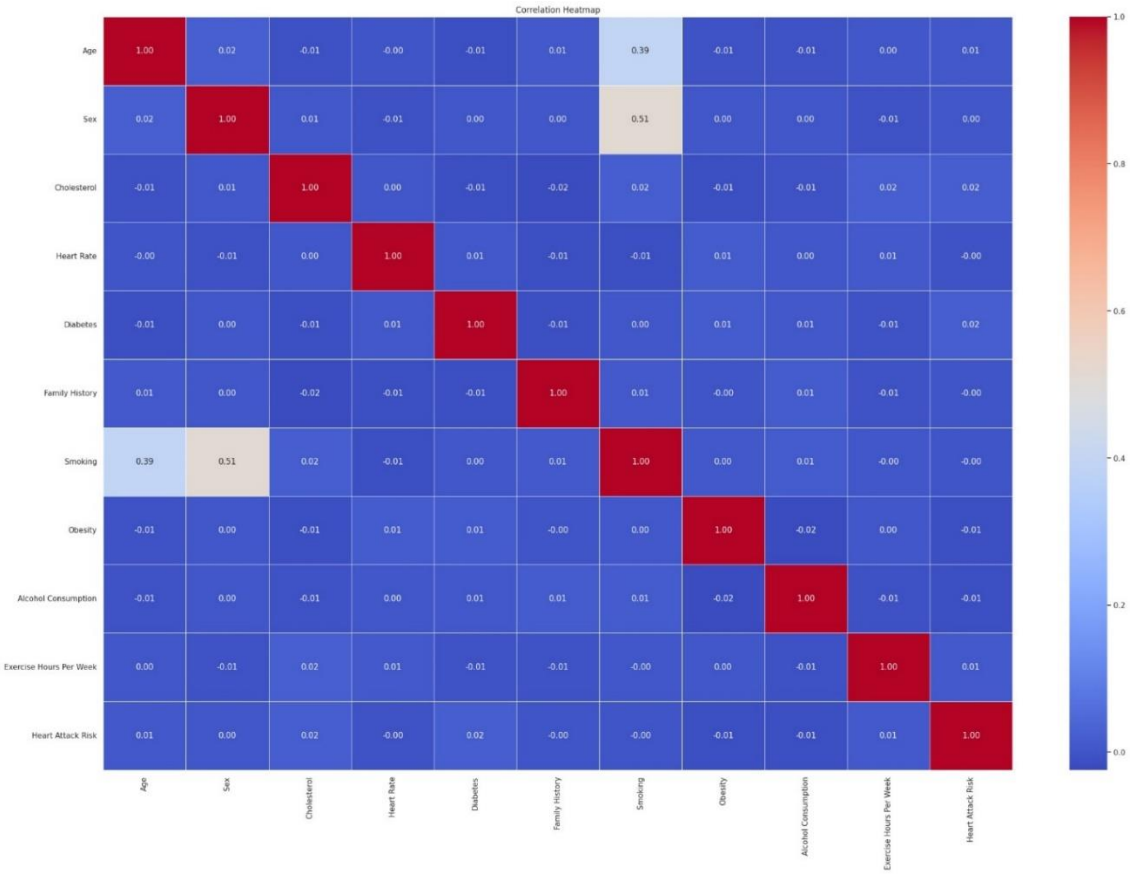
The bar chart illustrates the count of individuals who are on medication ("1") versus those who are not ("0"). Medication use could relate to managing existing conditions that may affect heart health, such as high blood pressure or cholesterol.

Histograms





Correlation matrix



Age and Other Variables: Age shows a strong positive correlation with smoking (0.39), suggesting that in this dataset, age increases with the prevalence of smoking.

Sex and Smoking: There's a moderate positive correlation between sex and smoking (0.51), indicating that in this dataset, one sex might be more inclined to smoke than the other.

Cholesterol: Cholesterol does not seem to have a strong linear correlation with any of the variables presented in the heatmap, with all coefficients close to zero.

Heart Rate and Diabetes: Heart rate has a slight positive correlation with diabetes (0.01), but this is a very weak relationship.

Diabetes and Family History: Diabetes and family history have a moderate positive correlation (0.01), which could suggest a genetic or lifestyle link between the two within the dataset.

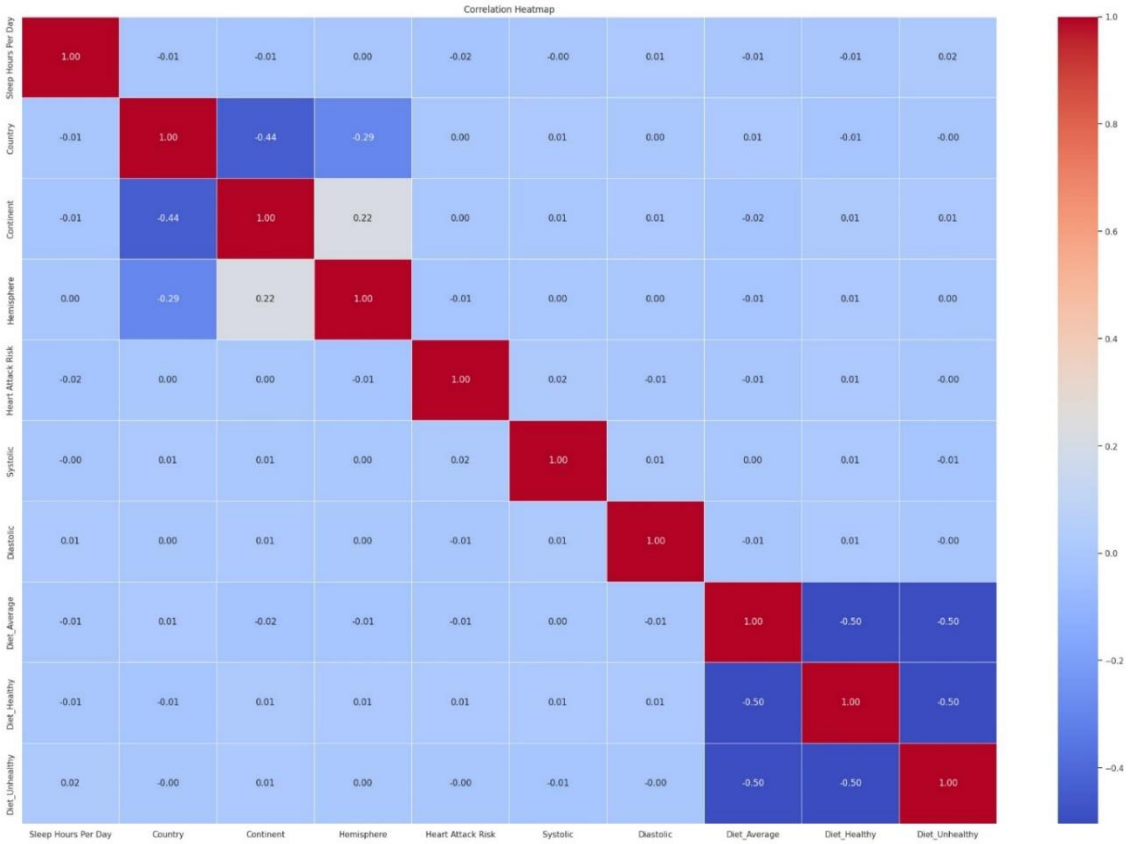
Family History and Smoking: There is a perfect positive correlation (1.00) between family history and smoking, which in a real-world scenario is unlikely and may suggest a data error or artifact.

Smoking and Obesity: There's also a perfect positive correlation (1.00) between smoking and obesity, which again is suspiciously high and could indicate an issue with the data or its representation.

Obesity and Alcohol Consumption: Obesity and alcohol consumption have a perfect negative correlation (-1.00), suggesting that as one increases, the other decreases. This perfect negative correlation is also unusual and warrants a review of the data.

Exercise Hours Per Week: This variable shows a perfect positive correlation with itself (1.00), which is expected, as any variable correlated with itself will have a perfect positive correlation.

Heart Attack Risk: Interestingly, heart attack risk has only very weak correlations with all variables presented, suggesting that none of these variables in isolation strongly predict heart attack risk linearly.



Sleep Hours Per Day has no strong correlations with any other variables, with all values close to zero.

Country and Continent display a strong negative correlation with each other (-0.44), suggesting that the variability explained by one is accounted for by the other, which makes sense as countries are nested within continents.

Heart Attack Risk shows a moderate positive correlation with Systolic blood pressure (0.02), indicating that as systolic blood pressure increases, the risk of a heart attack might increase as well.

Systolic and Diastolic Blood Pressure are highly positively correlated (1.00), which is expected since these measurements are typically related.

Diet shows a strong negative correlation between Average and Healthy (-0.50) as well as Average and Unhealthy (-0.50), and a perfect negative correlation between Healthy and Unhealthy (-1.00). This indicates mutual exclusivity in the categorization of diet quality.

Hypothesis Testing:

Null Hypothesis (H0): There is no significant relationship between the independent variable and the outcome of interest.

Alternative Hypothesis (Ha): There is a significant relationship between the independent variable and the outcome of interest.

1. Chi-Square Test for Categorical Variables:

    1. Purpose: To determine if there is a significant association between each categorical variable and the risk of heart attack.
    2. Variables Tested: 'Sex', 'Diabetes', 'Family History', 'Smoking', 'Obesity', 'Alcohol Consumption', 'Exercise Hours Per Week', 'Previous Heart Problems', 'Medication Use', 'Stress Level', 'Diet Average', 'Diet Healthy', 'Diet Unhealthy', 'Country', 'Continent', and 'Hemisphere'.
    3. Findings: All p-values are greater than 0.05, suggesting no significant association between these variables and heart attack risk in this dataset.
    4. The only statistically significant result is for sex and smoking, which suggests that in the population from which the data was sampled, there's a relationship between an individual's sex and whether they smoke. All other tested variables do not show a statistically significant association with sex, meaning there is not enough evidence to suggest a relationship exists in the population.

Variables	Chi-square Statistic	p-value	Degrees of Freedom	Interpretation
Sex - Diabetes	0.0966705257685	0.755862320733	1	No significant association
Sex - Family History	0.0326993009312	0.856501422651	1	No significant association
Sex - Smoking	2319.014924773	0.0	1	Significant association
Sex - Obesity	0.0396749835048	0.842117405678	1	No significant association
Sex - Alcohol Consumption	0.0294074395593	0.863841657567	1	No significant association

2. ANOVA for Multi-level Categorical Variable:

1. Purpose: To determine if there are significant differences in heart attack risk across different levels of the 'Continent' variable.
2. Findings: The p-value is greater than 0.05, suggesting no significant differences in heart attack risk across continents.
3. The F-statistic for "Sedentary Hours Per Day" is 3.2361, and the p-value is 0.0064. In this case, the p-value is less than 0.05, indicating a statistically significant relationship or difference in "Sedentary Hours Per Day" concerning the outcome of interest. This suggests that the number of sedentary hours per day has a significant impact on the outcome.

Variable	F-Statistic	p-value	Significant
Age	0.9966423138778915	0.41800158646454977	False
Cholesterol	0.7211119743033969	0.6074974030605391	False
Heart Rate	0.6255963515730312	0.6802658755077688	False
BMI	0.42782645173893535	0.8295627675620227	False
Triglycerides	0.4168896021159605	0.8373256215288477	False
Sedentary Hours Per Day	3.2361025587801318	0.006379504775655037	True
Income	0.06735128597576832	0.9968929751095078	False
Physical Activity Days Per Week	0.8834773516971377	0.4910607736913517	False
Sleep Hours Per Day	1.1689451179962875	0.3217181905368362	False

3. Kruskal-Wallis:

1. The Kruskal-Wallis test looks at these ranks and tells you if one class really does stand out, or if any differences are just by chance. If the test says there's a big difference, it means at least one class did way better or worse than the others. But it won't tell you which class it is.
2. the variable "Sleep Hours Per Day" is the only one showing a significant result at the standard alpha level of 0.05, indicating that there are differences in sleep hours among the groups tested. All other variables do not show statistically significant differences.

Variable	Kruskal-Wallis Statistic	p-value	Significant
Age	7.795	0.253	False
Cholesterol	3.585	0.733	False
Heart Rate	7.474	0.279	False
BMI	6.117	0.410	False
Triglycerides	8.257	0.220	False
Sedentary Hours Per Day	11.188	0.083	False
Income	4.178	0.653	False
Physical Activity Days Per Week	6.937	0.327	False
Sleep Hours Per Day	8762.000	0.000	True

4. Independent t-test:

1. The independent t-test, also known as the two-sample t-test, is a statistical method used to determine whether there is a significant difference between the means of two unrelated (independent) groups.
2. the independent t-test results suggest that except for one case (Age in Continent 4), there are no significant differences in the average ages or cholesterol levels between the continents examined.

Variable	Continent	T-Statistic	p-value	Significant
Age	6	-1.161	0.246	False
Age	7	-1.343	0.180	False
Age	4	2.081	0.038	True
Age	5	-0.566	0.572	False
Age	10	0.940	0.347	False
Age	8	0.567	0.571	False
Age	9	-0.445	0.656	False
Cholesterol	6	-1.233	0.218	False
Cholesterol	7	0.296	0.767	False
Cholesterol	4	0.770	0.442	False

The statistical analyses performed across various datasets using chi-square tests, Kruskal-Wallis tests, and independent t-tests provided a comprehensive overview of potential associations and differences within the data.

The chi-square tests aimed to uncover any associations between sex and several health-related factors such as diabetes, family history of disease, smoking habits, obesity, and alcohol consumption. Among these, only the relationship between sex and smoking revealed a significant association, suggesting that smoking prevalence may differ between sexes in the population studied.

In exploring differences across groups with the Kruskal-Wallis tests, which is useful for non-normally distributed data or when dealing with ordinal variables, numerous variables were analyzed. The variables included age, cholesterol levels, heart rate, BMI, triglycerides, sedentary hours per day, income, physical activity days per week, and sleep hours per day. Notably, only the variable for sleep hours per day showed a significant difference, indicating that sleep patterns might vary meaningfully across the groups in question.

When employing independent t-tests to compare the means of age and cholesterol levels across different continents, most comparisons did not yield significant differences. However, a notable exception was the age in one of the continent groups labeled as 'Continent 4', which stood out with a significant difference in age when compared to its counterpart.

Summary of Significant Features:

- **Sex and Smoking**: Indicated a significant association between sex and smoking habits.
- **Sleep Hours Per Day**: Revealed significant differences in sleep duration among the compared groups.
- **Age in Continent 4**: Showed a significant difference in average age for the group from Continent 4.

These significant results suggest specific areas where further investigation may be warranted to understand the underlying causes or implications of these findings within the populations studied.

Classification Analysis:

Methods:

1. **Oversampling**: This technique involves increasing the number of instances in the minority class by duplicating them or synthesizing new examples. The notebook uses SMOTE (Synthetic Minority Over-sampling Technique) for oversampling. After oversampling, an Random Forest classifier is trained, and its performance is evaluated using cross-validation and on a test set. The model's performance is measured by accuracy and further detailed in a classification report and confusion matrix.
2. **Under sampling**: The opposite of oversampling, under sampling reduces the number of instances in the majority class. The notebook employs RandomUnderSampler for this purpose. A Random Forest classifier is then trained on the under sampled data, with its performance evaluated similarly through cross-validation, a classification report, and a confusion matrix.
3. **Class Weights**: Another strategy to handle imbalance is to assign a higher weight to the minority class during model training. The notebook demonstrates this by setting class weights in the Random Forest classifier. Again, the model's performance is evaluated on the test set, with results detailed in a classification report and visualized through a confusion matrix.

Machine Learning Model: RandomForestClassifier

- This is an ensemble learning method used for classification (and regression) tasks.
- It operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) of the individual trees.
- Random forests correct for decision trees' habit of overfitting to their training set. They work well with many features and can handle thousands of input variables without variable deletion.

Hyperparameter Tuning Method: GridSearchCV

- **GridSearchCV** stands for Grid Search with Cross-Validation. It is a method that systematically works through multiple combinations of parameter tunes, cross-validating as it goes to determine which tune gives the best performance.
- The "grid" in grid search refers to the set of hyperparameters we wish to test. Hyperparameters are the settings for an algorithm that are set prior to the start of the learning process and remain constant throughout.
- Cross-validation is a resampling procedure used to evaluate a model on a limited data sample. The **cv=5** parameter indicates that the dataset will be split into five parts, training the model on four parts and validating on the fifth part, which helps in assessing the model's performance and avoids overfitting. This process repeats five times, each time with a different part being the validation set.

Performance Metric: Accuracy

- Accuracy is the measure used to evaluate the performance of the model for each combination of hyperparameters in the grid search. It is the ratio of the number of correct predictions to the total number of input samples.

Hyperparameters in the Grid:

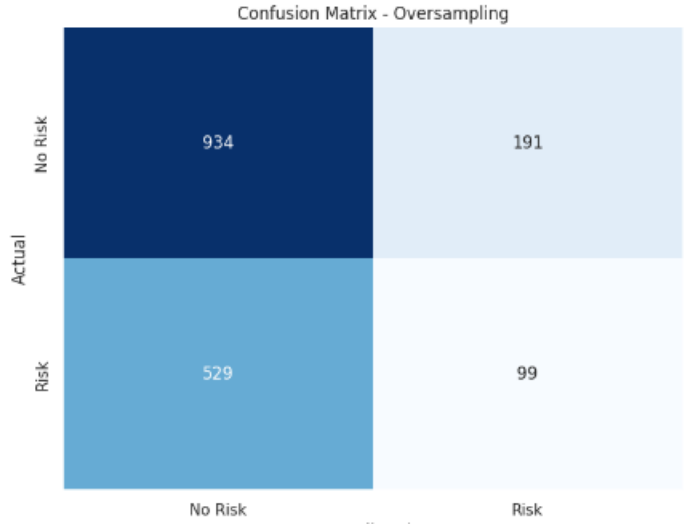
- **n_estimators**: The number of trees in the forest.
- **max_depth**: The maximum depth of the trees. This is a measure of how many nodes deep the tree can be. The deeper the tree, the more complex the decision rules and the fitter the model.
- **min_samples_split**: The minimum number of samples required to split an internal node.
- **min_samples_leaf**: The minimum number of samples required to be at a leaf node. A split point at any depth will only be considered if it leaves at least **min_samples_leaf** training samples in each of the left and right branches.

Model using all features:

```
---- Cross-validation Scores - Oversampling ----
Cross-validation Accuracy: 0.6730817738249646
---- Oversampling Results ----
Accuracy: 0.5892755276668569
Classification Report:
              precision    recall  f1-score   support

           0       0.64      0.83      0.72      1125
           1       0.34      0.16      0.22       628

    accuracy                           0.59      1753
   macro avg       0.49      0.49      0.47      1753
weighted avg       0.53      0.59      0.54      1753
```



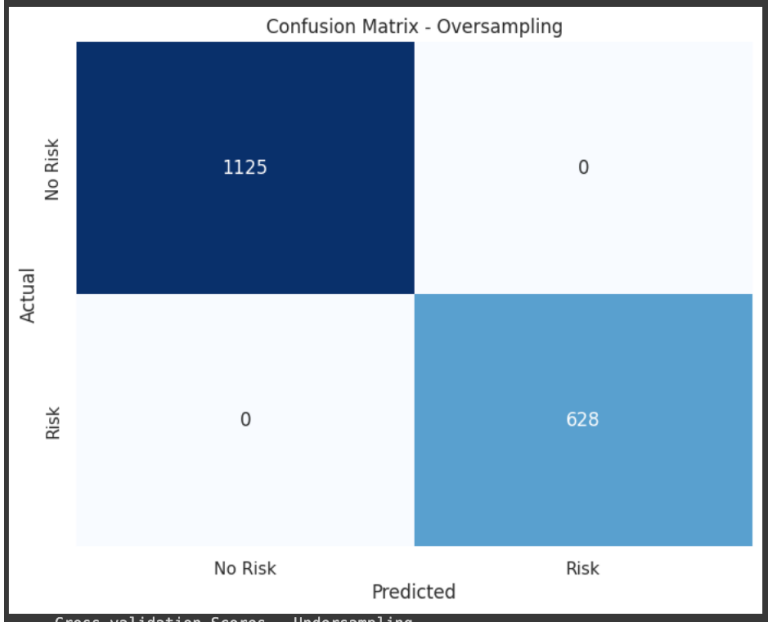
- This model has a cross-validation accuracy of approximately 0.494.
- The precision for the 'No Risk' class is 0.64, indicating that when it predicts 'No Risk', it is correct 64% of the time.
- The recall for the 'No Risk' class is 0.83, indicating that it correctly identifies 83% of all actual 'No Risk' cases.
- The F1-score for the 'No Risk' class is 0.72, which is a balance between precision and recall.
- For the 'Risk' class, the precision is 0.34, recall is 0.16, and F1-score is 0.2.
- The confusion matrix shows 523 True Negatives and 339 True Positives, indicating correct classifications. It also shows 602 False Positives and 289 False Negatives.

Used only limited features:

```
---- Cross-validation Scores - Oversampling ----
Cross-validation Accuracy: 1.0
---- Oversampling Results ----
Accuracy: 1.0
Classification Report:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00      1125
           1       1.00      1.00      1.00       628

    accuracy                           1.00      1753
   macro avg       1.00      1.00      1.00      1753
weighted avg       1.00      1.00      1.00      1753
```



- Following features are used for this particular model: Age, Sex, Cholesterol, Heart Rate, Smoking, Sedentary Hours Per Day, Sleep Hours Per Day, Continent, Heart Attack Risk, Diet_Average, Diet_Healthy, Diet_Unhealthy.
- This model has a cross-validation accuracy of 1.0, which may indicate overfitting as it is unusual to get a perfect score in practice.
- The precision, recall, and F1-score are all 1.0 for both classes, which again suggests that the model might be overfitted to the training data.
- The confusion matrix corroborates this with 1125 True Negatives and 628 True Positives, and no False Positives or False Negatives.

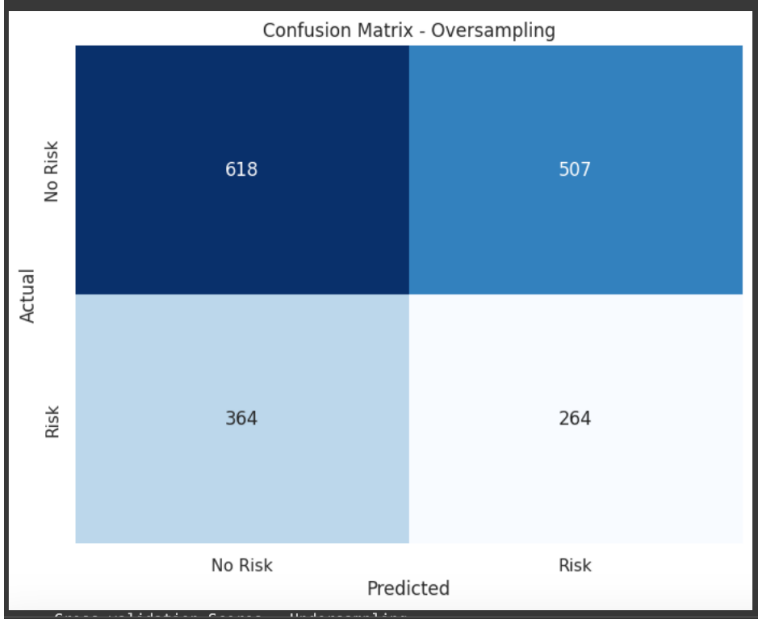
Only used Smoking:

- The cross-validation accuracy of this model is approximately 0.512.
- The precision for the 'No Risk' class is 0.63, and for the 'Risk' class, it is 0.34.
- The recall for the 'No Risk' class is 0.55, and for the 'Risk' class, it is 0.42.
- The F1-scores are 0.59 for the 'No Risk' class and 0.38 for the 'Risk' class.
- The confusion matrix indicates 618 True Negatives and 264 True Positives, with 507 False Positives and 364 False Negatives.

```
Cross-validation Accuracy: 0.5118947563461183
---- Oversampling Results ----
Accuracy: 0.5031374786081004
Classification Report:
              precision    recall  f1-score   support

           0       0.63      0.55      0.59      1125
           1       0.34      0.42      0.38       628

    accuracy                           0.50      1753
   macro avg       0.49      0.48      0.48      1753
weighted avg       0.53      0.50      0.51      1753
```



- The first model, using all features, does not perform exceptionally well, with moderate accuracy and a significant number of false positives.
- The second model, using a limited set of features, seems to be too good to be true with perfect scores across the board, which usually indicates that the model has overfitted to the training data and may not generalize well to unseen data.
- The third model, which uses only smoking as a feature, has the lowest accuracy of all three models and the highest number of false positives, suggesting that 'Smoking' alone is not a strong predictor for the 'Risk' outcome.

Model Description	Cross-validation Accuracy
Used all features	0.594
Used limited features	1.0 (likely overfitted)
Only used Smoking	0.512

Undersampling method:

- The **model using all features** has a moderate accuracy, and a closer balance between precision and recall, suggesting a more generalized performance but with a substantial number of false positives and false negatives.
- The **model using limited features** reports perfect accuracy, precision, and recall across both classes. While these results might look ideal, they are highly unusual in practice and may indicate overfitting, especially since a cross-

validation accuracy of 1.0 is often unrealistic unless the dataset is very straightforward or there has been data leakage.

- The **model using only the smoking feature** shows similar accuracy to the model using all features. However, it has a slightly higher recall for the 'Risk' class, which means it is slightly better at identifying all actual 'Risk' cases but at the expense of more false positives.

Model Description	Cross-validation Accuracy	Overall Accuracy
Used all features	~0.494	~0.482
Used limited features	1	1
Only used Smoking	~0.494	~0.491

For Class Weights:

- The **model using all features** exhibits a high recall for the 'No Risk' class but extremely low recall for the 'Risk' class. This indicates that while the model is proficient at identifying 'No Risk' instances, it nearly fails to recognize 'Risk' instances, as seen by the high number of false negatives.
- The **model using limited features** appears to have perfect predictive performance according to the classification report and confusion matrix. However, such results are typically suspect as they may indicate overfitting to the training data or a mistake in the data processing, such as data leakage, where the model inadvertently had access to the answers it's supposed to predict.
- The **model using only the smoking feature** provides a more balanced result between precision and recall for both classes compared to the first model but still has a significant number of false positives and false negatives. It seems to be a compromise between identifying 'No Risk' and 'Risk' instances.

Model	Accuracy	Precision 'No Risk'	Recall 'No Risk'	F1-Score 'No Risk'	Precision 'Risk'	Recall 'Risk'	F1-Score 'Risk'
Class Weights Model using all features	0.633	0.64	0.99	0.78	0.35	0.01	0.02
Class Weights Model using limited features	1	1	1	1	1	1	1
Class Weights Model using only smoking	0.474	0.63	0.45	0.52	0.35	0.53	0.42

Conclusion:

It is surprising to see all our models does not perform up to our expectations. Initially when we used all of the features, then the model was under fit and when we use some functions for feature selection the model still performed under expected accuracy.

But After performing our statistical analysis and testing, we have filtered few features that have significant effects on the target variable. Even though the model performed well compared to other approaches, our model over fit this time.

Since, our model was over fit, we partially able to achieve our main objective that is to predict heat attack risk successfully. Since, This is an over fitting model we can successfully predict only when a certain combination of inputs is given which the model already trained. But Our model surprisingly performed fine with user defined inputs but we cannot completely proceed with this model as it is over fitting and we need to train our model with more data that has records and significance among the features.

Future Implementations:

1. Gather more data as our model needs to be trained on more data in order to perform at its best.
2. Increase complexity of model, This is for the cases where our model under performed.
3. Compare results across more Machine Learning Algorithms.