# BA 723

# Business Analytics Capstone

## “A Predictive Modeling for Identifying Individuals at High Risk of Heart Disease”

**Prepared**

**By:**

**Jason S. Yap**

**301293413**

# Executive Summary

## 0.1. Executive Introduction

Knowledge is power. By understanding the risks of heart disease, individuals can take proactive steps to improve their health. Analyzing how various indicators contribute to heart disease can help identify key factors that could reduce the risk. Heart disease remains a leading global cause of mortality, claiming an estimated over 17 million lives annually (World Health Organization, 2024). Identifying at-risk individuals is crucial for prevention and early intervention.

This document comprehensively analyzes a heart disease prediction model developed using a balanced dataset. The model leverages demographic and medical factors to estimate and identify heart disease risk.

## 0.2. Executive Objective

The project aims to create a predictive model for identifying individuals at high risk of heart disease. By analyzing the key health indicators to build a model capable of predicting heart disease risk.

The model demonstrates strong predictive capabilities for individuals with heart disease risk and holds promise for clinical application following further validation. Insights into the importance of features can inform the general communities and healthcare practitioners about key risk factors.

Inspired by the Behavioral Risk Factor Surveillance System (BRFSS) in the United States, this project aims to create a similar but more localized and digitized survey system. By utilizing simple questionnaires to identify individuals at risk of heart disease, we can mimic the BRFSS model. Initially, the heart disease indicator dataset from BRFSS will serve as a foundation for predictive modeling. Ultimately, this system can be used by communities, non-profit organizations, and healthcare institutions.

## 0.3. Executive Model Description

Several models were developed and simulated using a down sampled dataset to balance class distribution. Key features such as age, income, and other health metrics were incorporated after rigorous feature selection and engineering. The following tasks were

performed to identify the most significant features. Once identified, these significant features were fed into various models or simulations to determine the best predictors of individual heart disease risk.

**High Level Tasks**

**1. Data Cleaning and Preparation**

- Handle missing values through imputation as needed (**Not applicable, will be discussed in the exclusion section**).
- Address inconsistencies or outliers using capping and flooring if necessary.
- Convert categorical variables into a numerical format using one-hot or label encoding.
- Normalize or standardize numerical features to ensure consistent scale.

**2. Exploratory Data Analysis**

- Calculate descriptive statistics (mean, median, standard deviation) for each lifestyle factor.
- Visualize feature distributions and relationships with heart disease risk using skewness, histograms, and box plots.

**3. Correlation and Feature Selection**

- Conduct correlation analysis to identify potential relationships between features.
- Employ a Random Forest classifier to identify significant features for modeling.

**4. Predictive Modeling**

- Build multiple models (Decision Trees, Logistic Regression, Neural Networks) using selected features.
- Evaluate model performance using ROC AUC, accuracy, recall, precision, and F1 score.
- Apply cross-validation to assess model robustness and generalizability.

**5. Model Selection and Interpretation**

- Select the best-performing model based on evaluation metrics.
- Interpret model results to identify key indicators of heart disease risk.

**6. Recommendations and Insights**

- Utilize insights from the best indicators to inform heart disease risk identification and prevention.
- Create visualizations to effectively communicate findings.
- Prepare a comprehensive report summarizing the analysis, results, and recommendations.

## 0.4. Executive Recommendations

The Neural Neuter model demonstrates strong predictive capabilities for heart disease and holds promise for clinical application following further validation. Insights into feature importance can inform healthcare practitioners about key risk factors. The model is beneficial to different industries such as insurance companies, and hospitals or clinics, especially non-governmental or non-profit organizations, and a small community. The insights from this model and its leverage with analytics could improve health outcomes by identifying key lifestyle changes that have significant factor to heart disease risk, and healthcare providers or communities can offer tailored preventive care initiatives to mitigate the onset of heart disease.

# Introduction

## 1.0. Background

As mentioned in the executive summary, heart disease remains a leading cause of mortality worldwide claiming an estimated over 17 million lives annually (World Health Organization 2024)", and identifying at-risk individuals is crucial for prevention and early intervention. Below are common reasons why identifying heart disease risks are important:

- **Early Intervention and Prevention:** Early detection of heart disease risk indicators can significantly reduce the likelihood of devastating consequences such as heart attacks and strokes.
- **Personalized Healthcare:** By effectively managing and mitigating risk factors, individuals can significantly enhance their quality of life by preventing or delaying the onset of debilitating advanced heart disease.
- **Economic Benefits:** A healthy population is a productive population. By reducing the prevalence of heart disease, we can minimize productivity losses due to illness and disability.
- **Empowerment and Awareness:** Empowering individuals with knowledge about heart disease risks is essential for promoting proactive health management. Increased health literacy encourages individuals to take control of their well-being**.**

## 2.0. Problem Statement

The model aims to create a predictive model that can be utilized for a straightforward and convenient survey questionnaire to identify the likelihood of heart disease using and analyzing various demographic and medical indicators.

## 3.0. Objectives & Measurement

- **Objective:** Predict the likelihood of individual risk of heart disease.
- **Measurement:**
    - ROC**:** The best ability to distinguish between classes
    - Accuracy: Shows the overall correctness of the model

- Precision: (Positive Predictive Value): It tells how many of the predicted positive (With Heart Disease Risk) cases were actually positive.
- Recall: Actual positive cases were correctly identified by the model.
- F1 Score: A harmonic mean of precision and recall, and gives a balance between the two, especially useful when the classes are imbalanced.

## 4.0. Assumptions and Limitations

- Assumptions:
  - The data is accurate, assuming that respondents accurately report their health behaviors, conditions, and demographic details. Misreporting or recall bias could affect the validity of the data.
  - The sample is represented & consistent with the population: The sample is representative and consistent with the general population.
  - Independence of Observations: One's data does not influence another's within the dataset.
  - Balance of the Target Variable: Using down-sampling techniques.
  - Association vs. Causation: correlation-related (e.g., if ice cream sales and drowning incident increases, it means both tend to increase).
- Limitations:
  - May not capture complex non-linear relationships.
  - Confounding Factor: A variable that influences both the dependent variable and independent variable, leading to a spurious association. In this context, confounding factors are those not captured in the dataset due to the time constraints of telephone surveys.

## Data Sources

### 5.0. Data Set Introduction

A 2015 dataset that was originally from the Behavioral Risk Factor Surveillance System (BRFSS), was processed and posted on Kaggle by the Centers for Disease Control (CDC) and Prevention and a collaborator which was enhanced by Teboul, A. The Behavioral Risk Factor Surveillance System (BRFSS) is a comprehensive, ongoing data collection program that gathers information about health-related risk behaviors, chronic health conditions, and the use of preventive services among U.S. residents. For more information about the dataset, please visit the link below.

*https://www.kaggle.com/datasets/alexteboul/heart-disease-health-indicators-dataset*.

### 6.0. Exclusions

Records with missing or incomplete data were already excluded by the Behavioral Risk Factor Surveillance System (BRFSS) to maintain data quality by the time used by the Centers for Disease Control and Prevention.

### 6.1 Preparation

Performed simple data preparation such as Data transformation:

1. Applying a Downsampling technique to balance the data and be able to:
    a. Improved Model Performance on Minority Class.
    b. Better Model Generalization.
    c. More Reliable Evaluation Metrics (Accuracy).
    d. Balanced Data Solution (Meaningful metrics: precision, recall, and F1 score).
2. Feature engineering by data sub-setting or data filtering and creating a new feature by categorizing BMI variables.
    a. Ensuring subset classes are correct like with heart disease is 1 and else is 0. Categorizing BMI integers into 'Underweight', 'Normal Weight', 'Overweight', 'Obesity 1', 'Obesity 2', 'Obesity 3'.

Image 1: Codes for the Feature Engineering

Setting Values

```
[ ]  with_heart_risk = df[df['HeartDiseaseorAttack'] == 1]
     without_heart_risk = df[df['HeartDiseaseorAttack'] == 0]
```

```
     # BMI

     # Define the bins and corresponding labels
     bins = [0, 18.5, 24.9, 29.9, 34.9, 40, float('inf')]
     labels = ['Underweight', 'Normal Weight', 'Overweight', 'Obesity 1', 'Obesity 2', 'Obesity 3']

     # Check if 'BMI' column exists
     if 'BMI' in df.columns:
         # Categorize Cholesterol levels
         df['BMI_Category'] = pd.cut(df['BMI'], bins=bins, labels=labels, right=False)

         # Create a new column 'Age_Numeric' with numeric representation of categories
         # Use the codes attribute to get numeric representation of the categories
         df['BMI_Numeric'] = df['BMI_Category'].cat.codes

         df=df.drop(['BMI_Category','BMI'], axis=1)
     else:
         print("The 'BMI' column is not in the DataFrame.")
```

The code allows to ensure 1 and 0 are equal to Yes and No respectively. Consequently, the following code categorizes diverse BMI open continuous values according to the definition or category by Statistics of the Government of Canada and creates a new column with applied categories. Then the new column is transformed into numeric values for ease of modeling.

### 7.0. Data Dictionary

The table below provides a brief description of the data used in the dataset to give an overview of the data's content. For further clarification or understanding, please refer to the "2020-calculated-variables-version4-508" file attached with the capstone documentation. Simply search for the data variable using the find feature to locate specific details.

## Table 1: Heart Disease Data Dictionary

| Category | Data Variables | Survey Questions/ Context | Answer |
|---|---|---|---|
| Output | HeartDiseaseorAttack | With heart disease or experienced heart attack? | 1: Yes, 0: No |
| Health Details | HighBP | Is high blood confirmed by a medical professional? | 1: Yes, 0: No |
| Health Details | HighChol | Is high cholesterol confirmed by a medical professional? | 1: Yes, 0: No |
| Health Details | CholCheck | Chol check-up w/in 5yrs | 1: Yes, 0: No |
| Health Details | BMI | Confirmed BMI | 1: Underweight to 5: Obese 3 |
| Health Details | Smoker | Smoked 100 cigars? | 1: Yes, 0: No |
| Health Details | Stroke | Experienced Stroke? | 1: Yes, 0: No |
| Health Details | Diabetes | Has diabetes? | 1: Yes, 0: No |
| Lifestyle | PhysActivity | Active w/in 30 days? Except work | 1: Yes, 0: No |
| Lifestyle | Fruits | Eat 1 or more/ day | 1: Yes, 0: No |
| Lifestyle | Veggies | Eat 1 or more/ day | 1: Yes, 0: No |
| Lifestyle | HvyAlcoholConsump | Over 14 drinks/ week? (7 women) | 1: Yes, 0: No |
| Access to Healthcare | AnyHealthcare | Healthcare? | 1: Yes, 0: No |
| Access to Healthcare | NoDocbcCost | Avoid medical attention due to cost? | 1: Yes, 0: No |
| Health Status | GenHlth | Rate general health | 1: Excellent to 5: Poor |
| Health Status | MentHlth | Number of days of experiencing mental health problems within one month. | Open-ended in numbers |
| Health Status | PhysHlth | Number of days of experiencing physical health problems within one month. | Open-ended in numbers |
| Health Status | DiffWalk | Difficulty walking | 1: Yes, 0: No |
| Personal Details | Sex | Male/ Female | 1: Yes, 0: No |
| Personal Details | Age | The number of years belongs to | 1: 18-24, **(Lowest)** till 13: Over 80 **(Highest)** |
| Socioeconomic | Education | Educational Level | 1: Only kindergarten, 2-4: Grade - Highschool, 5 : Diploma/ Vocational, 6: Bachelor's Degree |
| Socioeconomic | Income | Income Level (Annually) | 1: Less than $10k, 2: $10k - $15k... 8 : over $75,000, With $5,000 gap. |

## Data Exploration

Data exploration is a crucial initial step in any data science project. It provides a foundation for understanding the dataset, uncovering patterns, identifying anomalies, and guiding subsequent modeling and analysis.

Data exploration allows the proponent to:

- Understand data structure and distribution: Grasp how data is organized and how values are spread across variables.
- Identify issues: Detect problems like missing values, outliers, and inconsistencies that could impact model performance.
- Inform feature engineering: Decide which variables to create, transform, or combine based on their characteristics and relationships.

By addressing these aspects, data exploration helps ensure that subsequent analyses are based on reliable and informative data.

### 8.0. Data Exploration Techniques

The proponent performs several data exploration techniques, including but not limited to “df.shape” for verifying observations and features, and “df.types” for identifying data types. However, the proponent chooses to highlight major exploration techniques that provide meaningful insights during the exploration. The following page shows the visualization or exploration techniques used for the data exploration and why it was used.

### Skewness

Describes the asymmetry of a dataset's distribution. A distribution is skewed when its data points are unevenly spread around the central value. This imbalance can be either positive (right-skewed), with a tail extending towards higher values, or negative (left-skewed), with a tail extending towards lower values.

Image 4: Skewness (Original Dataset)

Skewness of numerical variables:

| | Skewness | Is_Skewed |
|---|---|---|
| HeartDiseaseorAttack | 0.000000 | No |
| HighBP | -0.282878 | No |
| HighChol | -0.191647 | No |
| CholCheck | -5.948153 | Yes |
| Smoker | -0.081768 | No |
| Stroke | 2.740602 | Yes |
| Diabetes | 1.237548 | Yes |
| PhysActivity | -0.910341 | No |
| Fruits | -0.489044 | No |
| Veggies | -1.433359 | Yes |
| HvyAlcoholConsump | 4.237350 | Yes |
| AnyHealthcare | -4.449994 | Yes |
| NoDocbcCost | 2.747977 | Yes |
| GenHlth | 0.155068 | No |
| MentHlth | 2.336729 | Yes |
| PhysHlth | 1.492429 | Yes |
| DiffWalk | 0.991248 | No |
| Sex | -0.005106 | No |
| Age | -0.646947 | No |
| Education | -0.670350 | No |
| Income | -0.608604 | No |

The skewness image displays the skewness value in the second column. The third column indicates whether the values are skewed or not for easier visualization. The proponent decided not to transform the data because most variables are binary. Only Age, Education, Income, General Health, and BMI are categorical and not skewed. On the other hand, Physical Health and Mental Health are integers and were not categorized due to the varied number of days. The proponent also intends to examine the unique values after modeling.

Image 4: Skewness (Balanced Dataset)

Skewness of numerical variables:

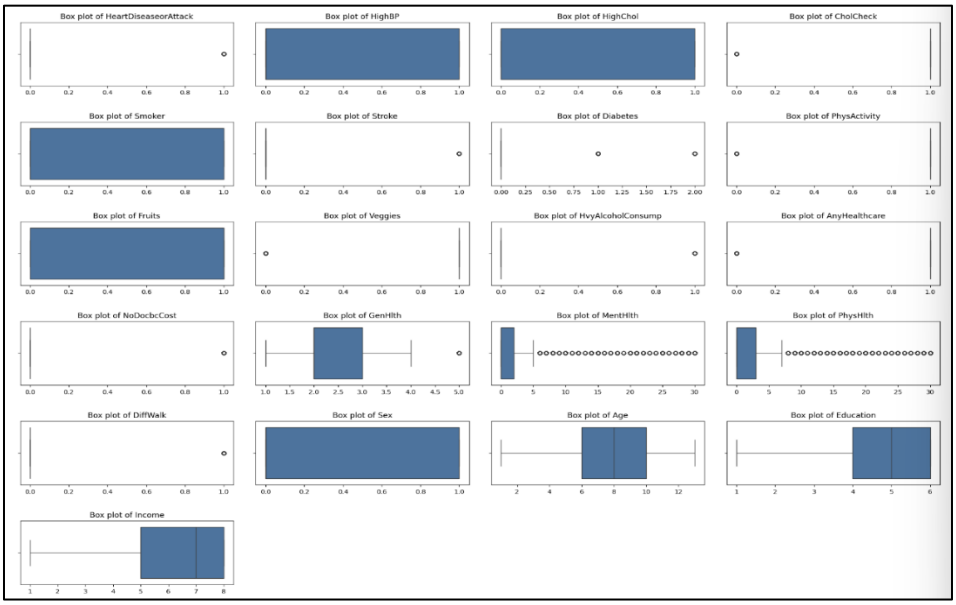
| | Skewness | Is_Skewed |
|---|---|---|
| HeartDiseaseorAttack | 0.000000 | No |
| HighBP | -0.282878 | No |
| HighChol | -0.191647 | No |
| CholCheck | -5.948153 | Yes |
| Smoker | -0.081768 | No |
| Stroke | 2.740602 | Yes |
| Diabetes | 1.237548 | Yes |
| PhysActivity | -0.910341 | No |
| Fruits | -0.489044 | No |
| Veggies | -1.433359 | Yes |
| HvyAlcoholConsump | 4.237350 | Yes |
| AnyHealthcare | -4.449994 | Yes |
| NoDocbcCost | 2.747977 | Yes |
| GenHlth | 0.155068 | No |
| MentHlth | 2.336729 | Yes |
| PhysHlth | 1.492429 | Yes |
| DiffWalk | 0.991248 | No |
| Sex | -0.005106 | No |
| Age | -0.646947 | No |
| Education | -0.670350 | No |
| Income | -0.608604 | No |

After the Downsampling technique was applied only Physical Health and Mental Health were the integers that remained skewed, still as mentioned earlier the data were retained due to a diverse number of days and the proponent believes that the result after modeling will be insightful.

**Boxplot**

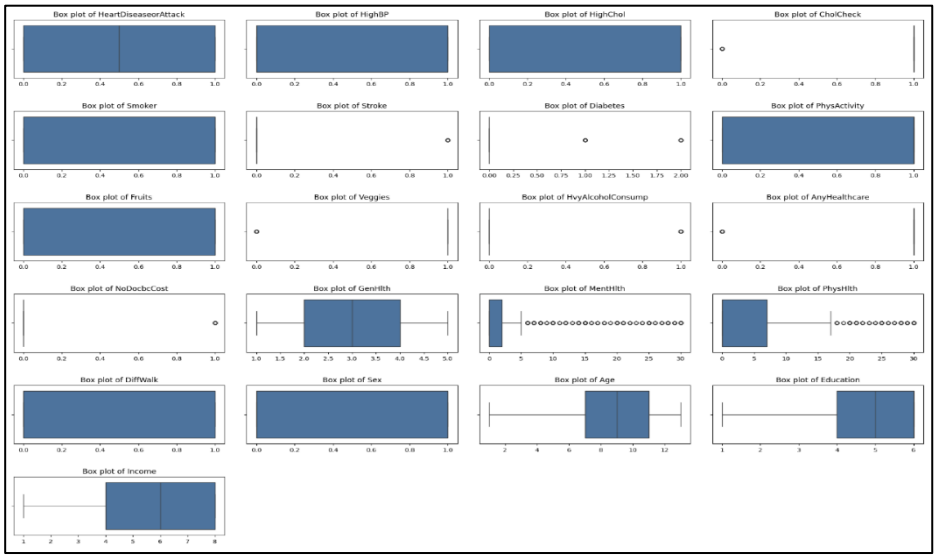
Boxplots: Displaying five-number summary distributions of each variable, helping the proponent to identify outliers, understand data distribution, compare multiple data sets, visualize skewness, and identify data spread.

Image 2: Original Dataset




The first boxplot image displays the original dataset, providing insights into the overall population distribution. As indicated by the skewness analysis, Physical Health, and Mental Health also exhibit outliers, which were retained due to their diverse range as depicted in the boxplots.

Image 3: Balanced Dataset

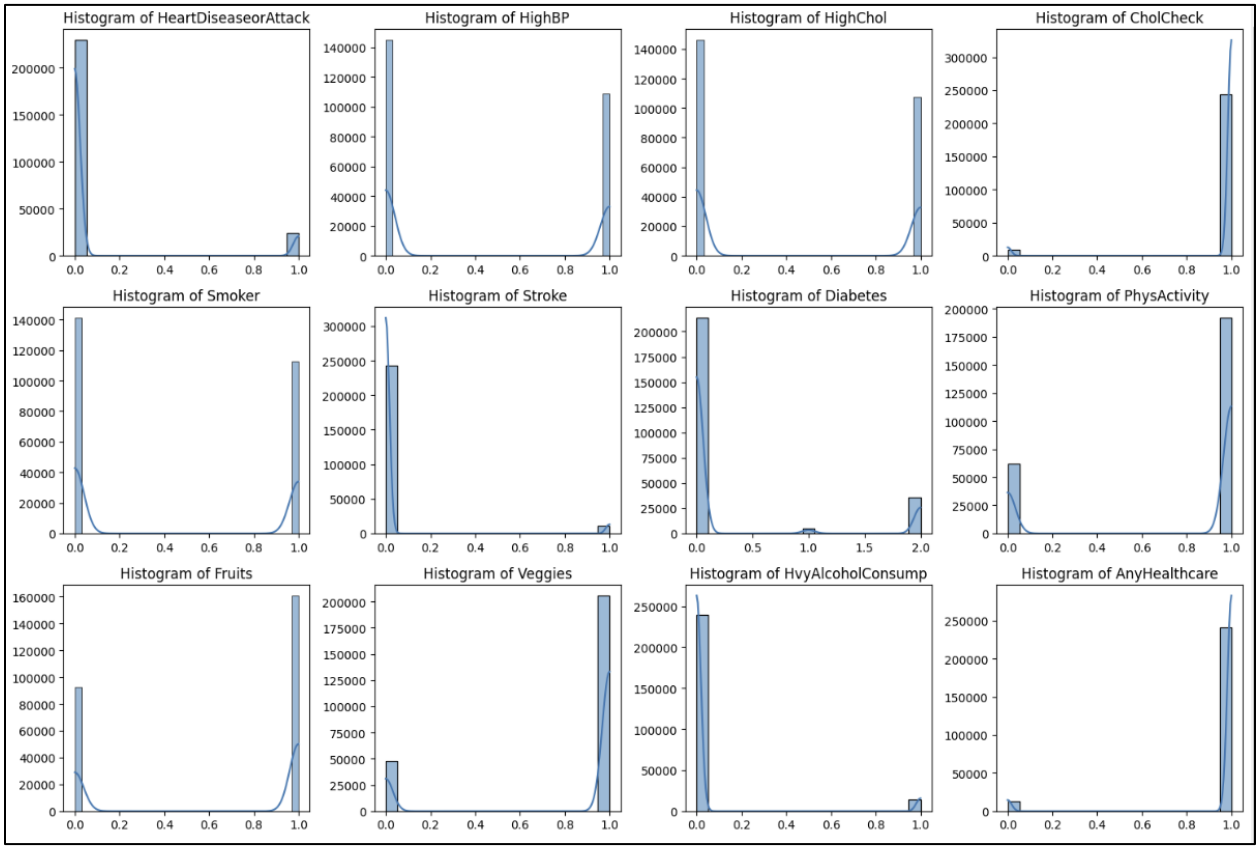


The second boxplot image visualizes how the distribution changed after balancing the dataset. It demonstrates a normal distribution for most variables, except for Physical Health and Mental Health, which the proponent decided to retain.

**Histograms**

Histograms are visual tools that reveal the distribution of numerical data. They help identify data patterns like normal distribution or skewness, spot outliers, and compare different data groups. Essentially, histograms offer a quick way to understand the shape and characteristics of the data. The proponent utilizes histograms to complement the observation of boxplots and skewness.

Image 4: Original Dataset



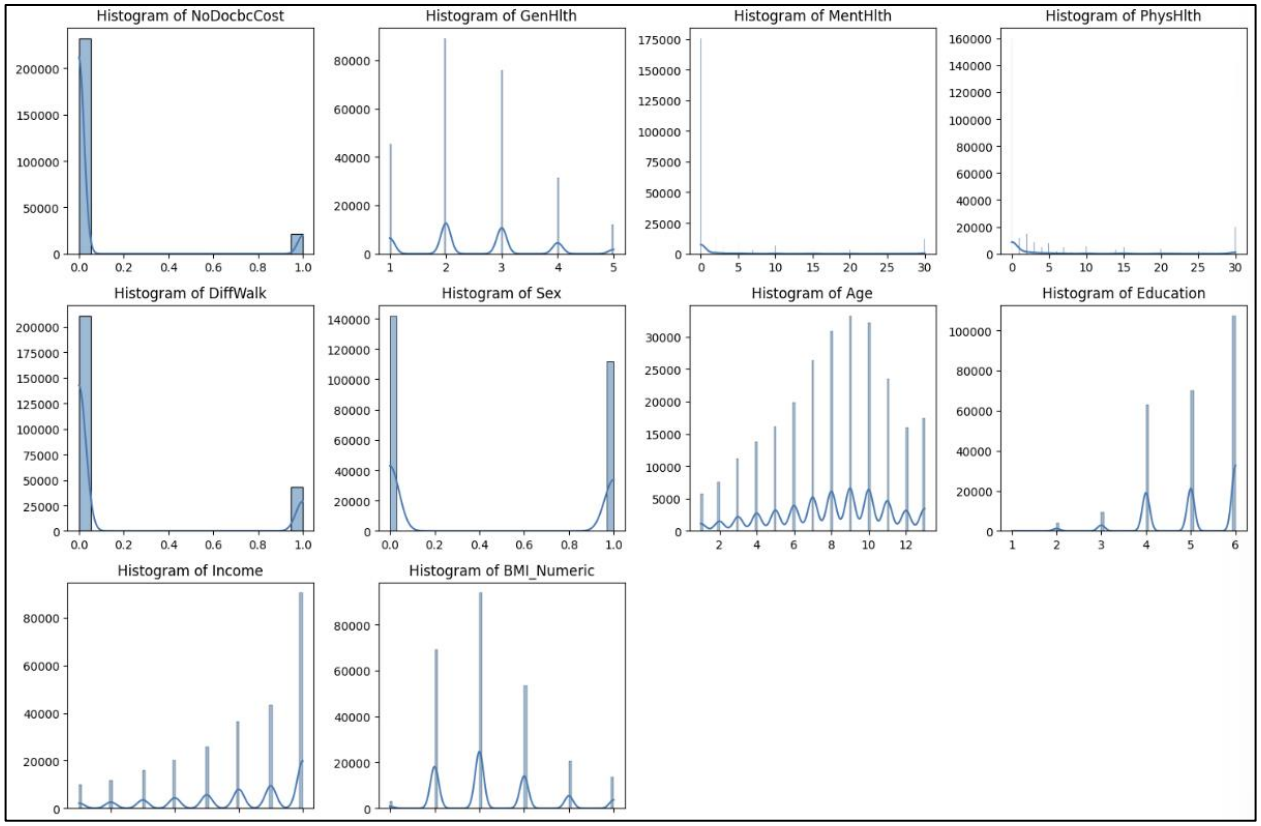
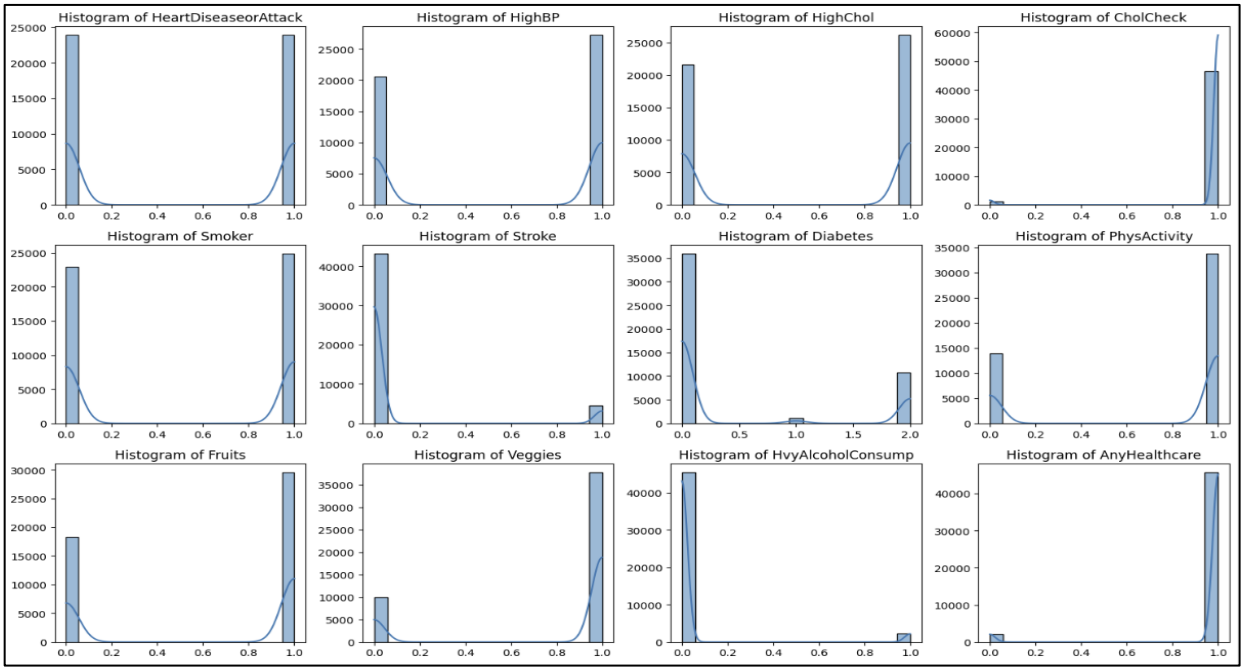
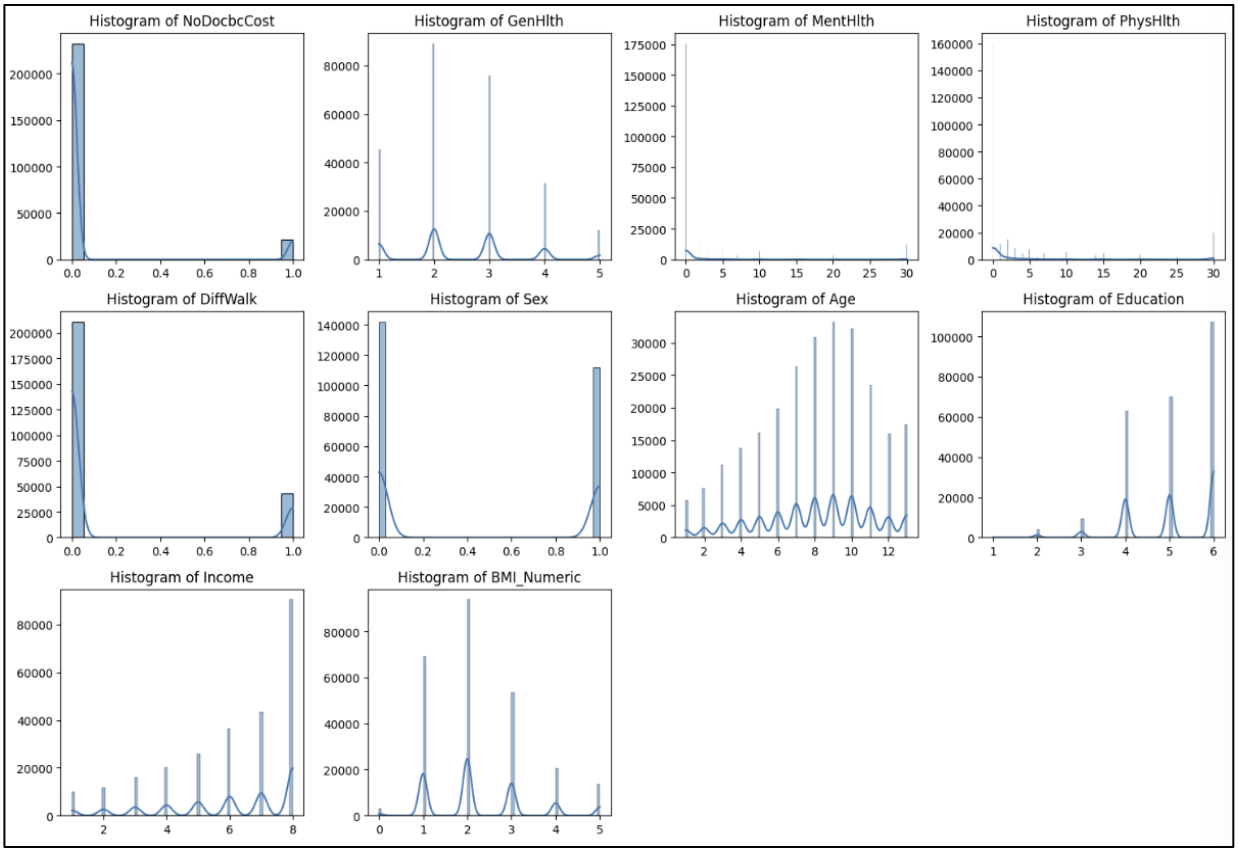



Image 5: Balanced Dataset

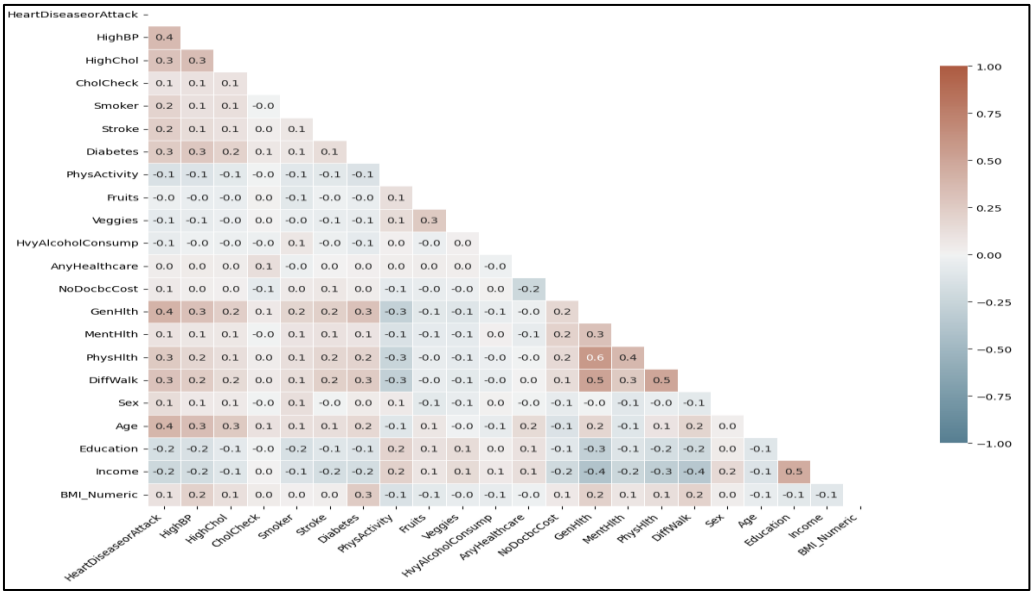





## Correlation Analysis

Heatmaps were used to visualize correlations between variables.

Image 6: Heatmap



The correlation heatmap indicates a lack of significant or extreme correlations (below 0.7) among the variables. The highest correlation of 0.6 suggests that a balanced dataset is likely to perform well in a model due to minimal predictor redundancy. This simplifies feature selection, leading to better interpretability and robust insights.

**9.0. Data Cleansing**

Skewness or outlier detection: Skewness/ Boxplot/ Histograms analysis identified outliers in BMI, Mental Health (MentHlth), and Physical Health (PhytHlth). BMI was categorized into Underweight (1) to Obese (5) to mitigate the impact of outliers. However, PhytHlth and MentHlth were retained due to their varying number of days, making categorization difficult. The proponent believes retaining the raw days significantly influences the decision tree model. For visualization, Image 1 in the preparation section demonstrates how BMI was categorized using a code.

**10.0. Summary (Original Dataset)**

- The data comprises 253,680 observations with 22 features.
- Of the 253,680 respondents, 23,893 have experienced heart attacks or heart disease.
- The community is generally healthy, with most people being active and eating healthily.
- Most do not smoke or consume alcohol heavily.
- Stroke incidence is low, and diabetes is uncommon.
- However, a significant number of individuals do not eat healthily, drink heavily, or smoke.
- More respondents do not eat fruits compared to those eating vegetables, suggesting a preference for vegetables over fruits.
- The community exhibits diversity in age and education levels, with a concentration in higher education.
- Income levels are distributed across multiple categories and show distinct peaks.
- While most people have an average BMI, some outliers exist who are underweight or obese.

- The some of the variables on the original dataset are too skewed which can disproportionately influence the results, leading to misleading insights or predictions.

### 10.1. Summary (Balance Dataset)

The insights from the original dataset and the balanced data are consistent, indicating that Downsampling effectively preserved the necessary insights and information for achieving accurate results or predictions without introducing significant bias or distortion.

## Data Preparation and Feature Engineering

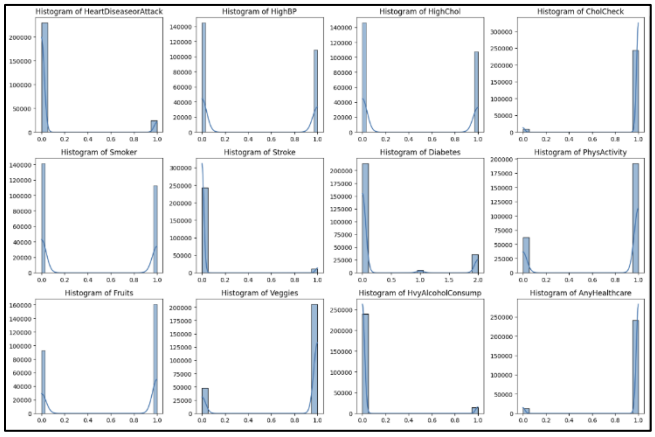
### 11.0. Data Preparation Needs

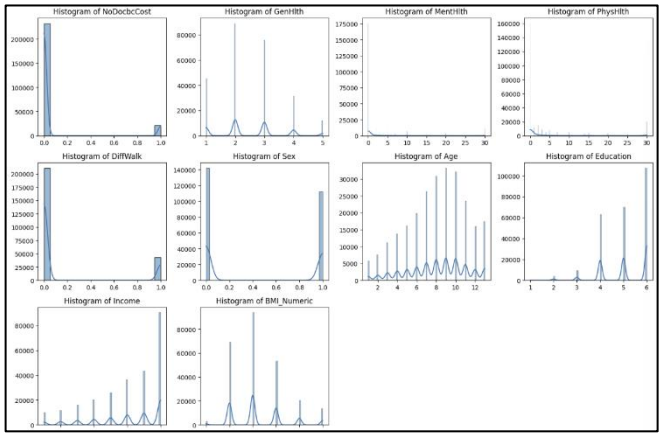
Data preparation is a critical initial step in any data analysis or machine learning endeavor. To ensure data is consistent and suitable for modeling, several key processes are performed:

### 11.1. Applied Downsampling Techniques

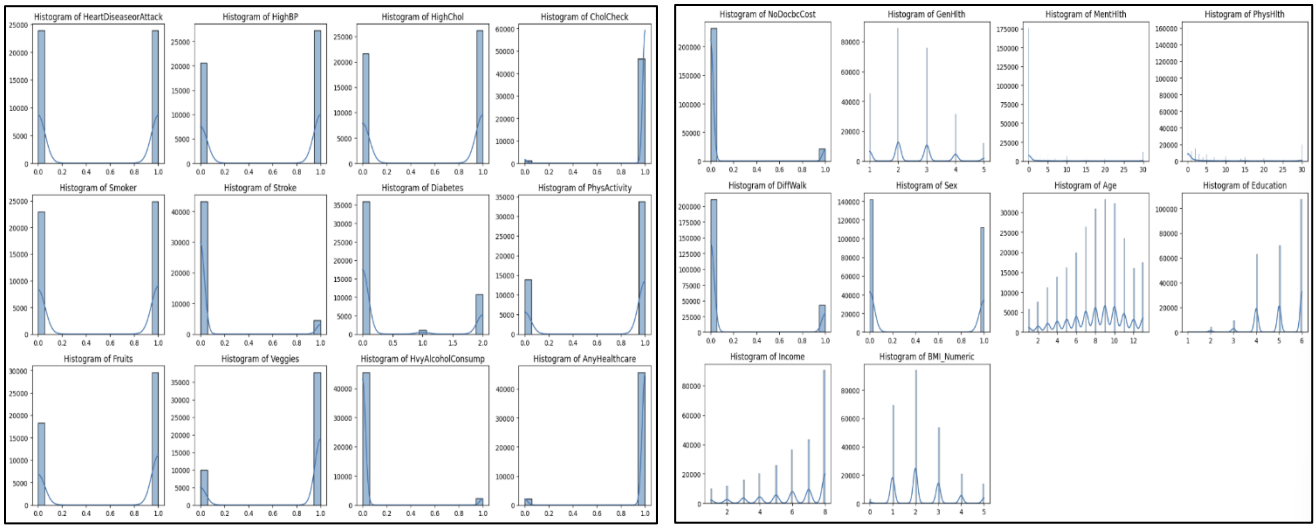
Image 7: Dataset Comparison

Original Dataset



Balanced Dataset




Images show significant changes while retaining the insights of the original as mentioned in the data exploration section.

### 11.2. Feature Engineering and Data Subsetting

Ensured that the classes are defined using the code below:

```
with_heart_risk = df[df['HeartDiseaseorAttack'] == 1]
without_heart_risk = df[df['HeartDiseaseorAttack'] == 0]
```

Drop the original "BMI" column and replace it with the "BMI_Numeric" column.

Image 8: BMI metrics table (Government of Canada)

**Health Risk Classification According to Body Mass Index (BMI)**

| Classification | BMI Category ($kg/m^2$) | Risk of developing health problems |
|---|---|---|
| Underweight | < 18.5 | Increased |
| Normal Weight | 18.5 - 24.9 | Least |
| Overweight | 25.0 - 29.9 | Increased |
| Obese class I | 30.0 - 34.9 | High |
| Obese class II | 35.0 - 39.9 | Very high |
| Obese class III | >= 40.0 | Extremely high |

Image 9: BMI column Comparison (Original, Engineered)

```
df.head(3)
```

| | HeartDiseaseorAttack | HighBP | HighChol | CholCheck | BMI |
|---|---|---|---|---|---|
| 0 | 0 | 1 | 1 | 1 | 40 |
| 1 | 0 | 0 | 0 | 0 | 25 |
| 2 | 0 | 1 | 1 | 1 | 28 |

| Sex | Age | Education | Income | BMI_Numeric |
|---|---|---|---|---|
| 1 | 11 | 5 | 6 | 2 |
| 0 | 10 | 4 | 7 | 1 |
| 0 | 9 | 5 | 2 | 2 |
| 0 | 12 | 5 | 4 | 1 |
| 0 | 10 | 6 | 7 | 2 |

** The BMI values (BMI column) and category numbers (BMI_Numeric column) are inconsistent because they originate from different data sources.

## Model Exploration

Selecting the optimal model for a given problem is crucial. It involves understanding model behavior, optimizing performance, mitigating risks, and gaining insights. By comparing multiple models and fine-tuning parameters, you can develop robust, fair, and efficient solutions that meet stakeholder needs. The model with the highest percentage or score across all specified metrics below will be considered the best model.

1. ROC-AUC: Best ability to distinguish between classes.
2. Accuracy: Measures the correct proportion of predictions among all instances.
3. Recall: Very good at identifying true positive cases.
4. Precision: Measuring the proportion of correct predictions among all instances.
5. F1 Score: Balances precision and recall, providing a comprehensive evaluation metric.

### 13.0. Modeling Approach/Introduction

Instead of independently feeding the balanced dataset to different models, the proponent employed a filtering approach. The balanced dataset was initially processed through a random forest model. Subsequently, the significant variables identified from this model were input into decision trees, logistic regression, and neural network models. By doing so, only the most significant variables remain.

## 14.0. Random Forest

Image 10: Feature Importance with Standard Deviation

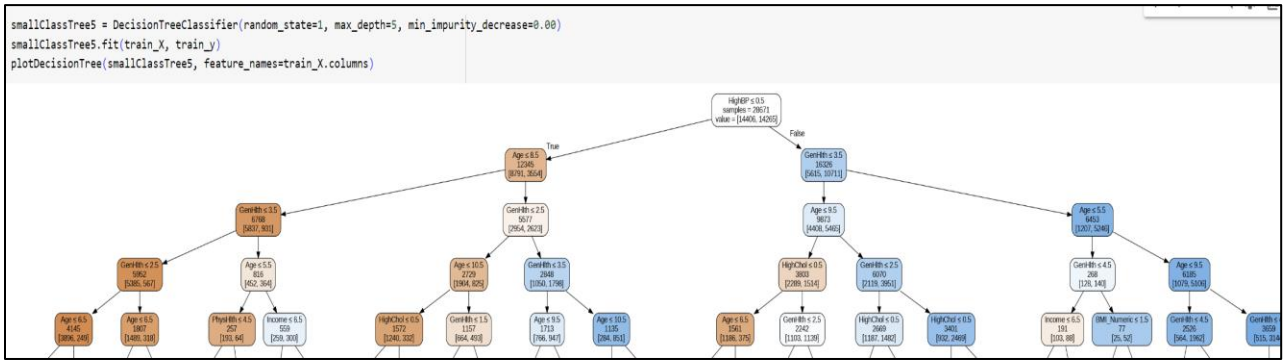
| | feature | importance | std |
|---|---|---|---|
| 17 | Age | 0.164292 | 0.022443 |
| 12 | GenHlth | 0.105131 | 0.034418 |
| 19 | Income | 0.090959 | 0.006581 |
| 14 | PhysHlth | 0.075790 | 0.014060 |
| 20 | BMI_Numeric | 0.072842 | 0.004563 |
| 0 | HighBP | 0.064715 | 0.038310 |
| 18 | Education | 0.060235 | 0.004254 |
| 13 | MentHlth | 0.052450 | 0.003021 |
| 1 | HighChol | 0.046289 | 0.021898 |
| 15 | DiffWalk | 0.039326 | 0.026379 |
| 16 | Sex | 0.032603 | 0.004288 |
| 5 | Diabetes | 0.032081 | 0.010827 |
| 7 | Fruits | 0.029544 | 0.002372 |
| 3 | Smoker | 0.026571 | 0.004276 |
| 4 | Stroke | 0.026209 | 0.008421 |
| 6 | PhysActivity | 0.024416 | 0.003326 |
| 8 | Veggies | 0.022670 | 0.002516 |
| 11 | NoDocbcCost | 0.012180 | 0.001298 |
| 9 | HvyAlcoholConsump | 0.009637 | 0.001235 |
| 10 | AnyHealthcare | 0.007503 | 0.000976 |
| 2 | CholCheck | 0.004555 | 0.000792 |

The image displays the results of the Random Forest model, ranking variables from highest to lowest importance. Variables from Age to HighChol were selected for subsequent models.

## 14.1. Significant Tree Model Insight

Among the Tree models, the most significant or with the highest percentage or score across ROC-AUC, Accuracy, Recall, Precision, and F1 Score is the SmallClassTree 5.

Image 11: Small Class Tree 5 Decision Tree Model

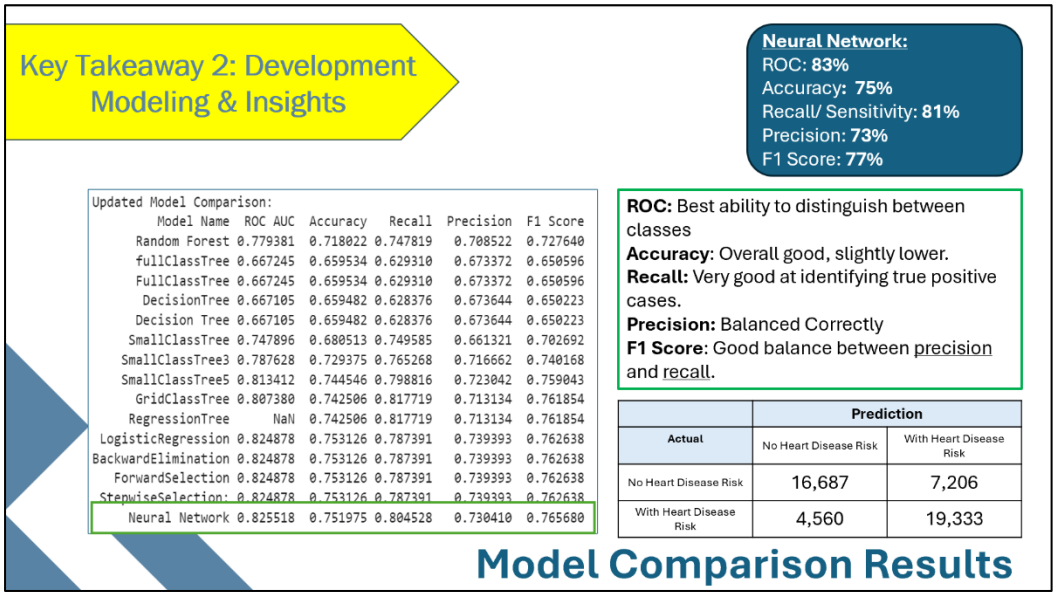



The tree model indicates that individuals under 60 without a confirmed diagnosis of high blood pressure, who have not undergone regular medical check-ups, and who perceive themselves as healthy, are at a higher risk of heart disease. In a personal interview with

a nurse professional, the results reflect real-life situations as the nurse reported from a 2015 case of a 7-year-old patient diagnosed with heart disease, emphasizing the unexpectedness of the condition, and the child was hospitalized due to severe chest pain. The nurse also noted that the child appeared healthy and active for their age. This case highlights the often-silent nature of heart disease, emphasizing the importance of regular medical check-ups for everyone.

### 15.0. Model Comparison and Recommendation

Image 12: Model Comparison and Insights




The image reveals that the neural network model significantly outperforms other models based on key performance metrics. The model correctly predicted heart disease risk for over 19,000 individuals and accurately identified over 16,000 individuals as being at low risk.

### 15.1. Odd Ratio Results

The odds ratio is a statistical tool used to compare the likelihood of an event happening between two variables or features. Also, it's a way to measure how much more (or less) likely something is to occur in one group compared to another.

Image 13: Odd Ratio Table

| No. | variable | coef | odds | Interpretation |
|---|---|---|---|---|
| **0** | HighChol | 0.710112 | 2.03422 | Individuals with confirmed high cholesterol are 2x more likely to have heart disease risk |
| **1** | HighBP | 0.636853 | 1.89052 | Followed by an 89% probability for individuals with confirmed high blood pressure. |
| **2** | GenHlth | 0.618677 | 1.85647 | Individuals with deteriorating health are 85% more likely to have heart disease risk. |
| **3** | Age | 0.284642 | 1.32929 | Every time the age increases, the probability of heart disease risk increase by 33%. |
| **4** | BMI_Numeric | 0.04611 | 1.04719 | Every time the BMI increases, the probability of heart disease risk increases by 5%. |
| **5** | PhysHlth | 0.007677 | 1.00771 | Less significant (1%) |
| **6** | MentHlth | 0.003311 | 1.00332 | Less significant (0.01%) |
| **7** | Education | -0.00456 | 0.99545 | Less significant (-1%) |
| **8** | Income | -0.04848 | 0.95268 | Less significant (-0.5%) |

The odd ratio table and logistic regression analysis consistently identified high cholesterol, high blood pressure, and general health as the primary predictors of heart disease risk. Education was found to be a less significant factor and was excluded from the final models.

## References


*World Health Organization. (2024). Cardiovascular diseases: https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1Kaggle. (2022). Heart Disease Health Indicators Dataset: https://www.kaggle.com/datasets/alexteboul/heart-disease-health-indicators-dataset American Heart Association. (2024). Understand Your Risks to Prevent a Heart Attack: https://www.heart.org/en/health-topics/heart-attack/understand-your-risks-to-prevent-a-heart-attack#:~:text=An%20inactive%20lifestyle%20is%20a,blood%20pressure%20in%20some%20people. National Library of Medicine. (Circa). Risk Factors for Coronary Artery Disease: Historical Perspectives: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5686931/*

*Parent, D. (2023). Business Analytics and Insights notes. Centennial College.*