GROUP NO.: 09

**Title: Predictive Modeling of Heart Disease Risk Assessment**

**Abstract**

One of the most common chronic illnesses in the US, heart disease affects millions of people annually and has a large financial impact on the nation's economy. About 647,000 people die from heart disease in the United States alone each year, making it the leading cause of mortality. Heart disease is caused by and has risk factors such as diabetes, high blood pressure, chronic inflammation, aging-related molecular changes, plaque accumulation inside bigger coronary arteries, and high blood pressure.

Although there are various forms of coronary heart disease, most people don't realize they have it until they have symptoms like chest pain, a heart attack, or unexpected cardiac arrest. This information emphasizes the value of screenings and preventative interventions that can reliably identify heart disease in the general population before adverse events such as myocardial infarctions (heart attacks) occur.

Three major risk factors for heart disease have been recognized by the Centers for Disease Control and Prevention: smoking, high blood pressure, and high blood cholesterol. These three risk factors are present in about half of the American population. The National Heart, Lung, and Blood Institute provides physicians with a more comprehensive list of variables to consider when diagnosing coronary heart disease, including age, sex, race or ethnicity, environment and occupation, family history and genetics, lifestyle habits, and other medical conditions. An initial assessment of these common risk factors, followed by blood testing and other tests, typically serves as the basis for the diagnosis.

This study is to investigate the viability of employing survey responses for heart disease risk prediction by utilizing the Behavioral Risk Factor Surveillance System (BRFSS) 2015 dataset. Using 253,680 cleaned survey responses—23,893 of whom had heart disease—the study aims to evaluate the predictive ability of BRFSS data and its potential applicability for screening for preventive health issues.

In conclusion, utilizing large-scale health surveys like the BRFSS offers a priceless chance to expand on our knowledge of the risk factors for heart disease and create efficient preventative and early detection plans. We can enhance public health initiatives targeted at lowering the prevalence of heart disease and enhancing cardiovascular outcomes for people and communities by utilizing data analytics and predictive modeling.

**Introduction**

Heart disease, which includes a variety of disorders that impact the heart and blood arteries, is still a major global public health concern. It can show up as a variety of symptoms, such as congenital cardiac abnormalities, arrhythmias, heart failure, and coronary artery disease. Heart disease continues to be a major global source of morbidity and death, despite advances in medical research and healthcare.

Strategies for early detection and effective prevention are essential to reducing the burden of heart disease. Preventive measures frequently center on altering risk factors and encouraging heart-healthy habits like consistent exercise, eating a balanced diet, abstaining from tobacco use, and stress management. Early identification lowers the risk of unfavorable outcomes such heart attacks, strokes, and sudden cardiac death by enabling prompt intervention and treatment.

Numerous health-related variables, including heart disease risk factors, can be found in abundance in large-scale health surveys such as the Behavioral Risk Factor Surveillance System (BRFSS). By gathering data from a variety of populations, these surveys provide important insights into the distribution and prevalence of risk factors among various demographic groups and geographical areas.

Researchers can examine the intricate interactions between various factors that contribute to heart disease by utilizing the vast amounts of data gathered from surveys such as the BRFSS. This covers both established risk factors like smoking, high blood pressure, and high cholesterol as well as newer ones like socioeconomic status, healthcare access, and lifestyle choices.  
  
Predictive models for heart disease risk assessment that are based on survey data have the potential to enhance population health management and preventive intervention. Healthcare professionals can reduce risk and enhance outcomes by identifying patients who are more likely to develop heart disease and implementing tailored treatment plans, targeted therapies, and lifestyle changes.

**Related Work**

Previous studies on cardiovascular health have looked closely at the connection between a number of risk factors and the onset of heart disease. For a more thorough explanation, see this:  
  
**High blood pressure, or hypertension:** Heart disease is known to be associated with hypertension. High blood pressure increases the risk of diseases like coronary artery disease, heart failure, and stroke because it strains the heart and blood vessels. Several research works have indicated a robust correlation between high blood pressure and unfavorable cardiovascular consequences, underscoring the significance of blood pressure control in averting heart disease.

**High cholesterol (hyperlipidemia):** High cholesterol, especially low-density lipoprotein (LDL) cholesterol, is linked to atherosclerosis, the accumulation of plaque in the arteries. Heart attacks and coronary artery disease can result from the arteries' hardening and constriction, which reduces blood flow to the heart. The necessity of controlling cholesterol in order to prevent heart disease is highlighted by the numerous studies that have demonstrated the connection between elevated cholesterol levels and an increased risk of cardiovascular events.

**Smoking:** Using tobacco products, either by smoking cigarettes themselves or by being around secondhand smoke, increases the risk of heart disease. Cardiovascular disorders are largely caused by the toxic compounds in tobacco smoke, which also damage blood vessels, cause inflammation, and raise the risk of blood clot formation. Several epidemiological studies have demonstrated that smoking is a major preventable cause of heart disease, highlighting the significance of quitting smoking in lowering the risk of cardiovascular disease.

Researchers can learn more about the extent of these risk factors, their prevalence in various groups, and their relative effects on cardiovascular health by looking through the literature and databases that are already available. Studies also examine the efficacy of interventions including medication treatments, public health campaigns, and lifestyle changes that try to lower these risk factors.   
  
By placing the current study in the larger context of research, scientists can fill in knowledge gaps, advance our understanding of heart disease risk assessment and prevention, and develop new perspectives. Additionally, by combining information from many sources, scientists can create more thorough models for estimating the risk of heart disease and guiding focused actions to enhance cardiovascular outcomes.

**Methods**   
  
The CDC gathers information about health-related telephone surveys called the Behavioral Risk Factor Surveillance System (BRFSS) once a year. Over 400,000 Americans participate in the annual survey, which gathers information on health-related risk behaviors, chronic illnesses, and the use of preventative services. Since 1984, it has been held annually. To begin this project, I downloaded the 2015 Kaggle dataset in csv format. This original dataset has 330 features and responses from 441,455 people. These characteristics are variables that are computed based on responses from specific participants, or they are questions that are posed to participants directly.

This dataset, which is mainly intended for use in the binary categorization of heart disease, includes 253,680 survey responses from the cleaned BRFSS 2015. This dataset does not exhibit a significant class imbalance. 23,893 respondents had heart disease, compared to 229,787 respondents who do not have or have not had heart disease.

1 binary tаrgеt variable: target аnd 21 fеаturе variables that аrе either bіnаrу оr оrdіnаl.

* HіghBP: Adultѕ whо hаvе been tоld thеу have high blood рrеѕѕurе bу a dосtоr, nurѕе, or оthеr health professionals
* HighChol: Have you EVER bееn told by a dосtоr, nurѕе, or оthеr health рrоfеѕѕіоnаlѕ thаt уоur blооd cholesterol is hіgh?
* ChоlChесk: Chоlеѕtеrоl check wіthіn раѕt five уеаrѕ.
* BMI: Body Mаѕѕ Indеx (BMI)
* Smоkеr: Hаvе уоu ѕmоkеd аt least 100 сіgаrеttеѕ in уоur еntіrе lіfе? [Nоtе: 5 расkѕ = 100 сіgаrеttеѕ]
* Strоkе: (Evеr tоld) уоu hаd a ѕtrоkе.
* Dіаbеtеѕ: 0 is no dіаbеtеѕ, 1 іѕ рrе-dіаbеtеѕ, and 2 іѕ diabetes.
* PhуѕAсtіvіtу: Adults who reported dоіng physical асtіvіtу or exercise during the past 30 days оthеr than thеіr rеgulаr jоb.
* Fruits: Consume Fruіt 1 or mоrе tіmеѕ реr dау
* Vеggіеѕ: Cоnѕumе Vеgеtаblеѕ 1 оr more tіmеѕ реr day
* HvуAlсоhоlCоnѕumр: Hеаvу drinkers (adult men hаvіng mоrе than 14 drіnkѕ реr week аnd аdult wоmеn hаvіng mоrе thаn 7 drіnkѕ реr week)
* AnуHеаlthсаrе: Dо уоu hаvе аnу kind of health саrе coverage, including hеаlth іnѕurаnсе, prepaid plans ѕuсh as HMOѕ, оr government рlаnѕ ѕuсh as Mеdісаrе, оr Indian Hеаlth Sеrvісе?
* NоDосbсCоѕt: Wаѕ there a time іn thе раѕt 12 mоnthѕ whеn уоu needed tо ѕее a dосtоr but could nоt because оf cost?
* GеnHlth: Would уоu say thаt іn general, уоur hеаlth іѕ:
* MentHlth: Nоw thinking аbоut your mental hеаlth, whісh іnсludеѕ stress, dерrеѕѕіоn, аnd problems wіth emotions, fоr hоw mаnу days durіng thе past 30 dауѕ wаѕ уоur mental hеаlth not gооd?
* PhуѕHlth: Nоw thіnkіng аbоut уоur рhуѕісаl hеаlth, which іnсludеѕ рhуѕісаl illness аnd іnjurу, fоr hоw mаnу dауѕ during the раѕt 30 dауѕ wаѕ уоur physical hеаlth nоt good?
* DіffWаlk: Dо уоu hаvе ѕеrіоuѕ difficulty wаlkіng оr сlіmbіng ѕtаіrѕ?
* Sеx: Indicate ѕеx.
* Agе: Fourteen-level аgе саtеgоrу
* Education: What іѕ thе highest grаdе оr уеаr of school уоu соmрlеtеd?
* Inсоmе: Iѕ your аnnuаl household income frоm аll sources: (If thе раtіеnt refuses аt аnу іnсоmе lеvеl, соdе "Refused.")

1. **Exploratory data analysis, or EDA**: it is a vital stage in deciphering the dataset's properties and spotting trends or connections between different variables. EDA was used in this work to investigate correlations between predictor variables and the target variable (heart disease status), as well as to look at the distribution of the variables, identify outliers, and evaluate missing values.

* **Variable Distribution:** For numerical variables, descriptive statistics including mean, median, standard deviation, and range were computed; for categorical variables, frequency tables were produced.
* **Outlier Detection**: Statistical and graphical approaches, such as boxplots, were used to identify outliers, or data points that differ considerably from the rest of the dataset (e.g., z-score or interquartile range).
* **Missing Value Analysis:** The dataset's missing values were evaluated, and methods for dealing with them—such as imputation or deletion—were taken into consideration.
* **Relationship Analysis:** To investigate the links between predictor factors (such as high blood pressure and cholesterol) and the target variable (heart disease status), correlation analysis, scatter plots, and other visualization techniques were used.

A graph of numbers and a number of numbers

Description automatically generated with medium confidence

* This visualization provides insights into the relationships between different variables in the dataset, helping to identify patterns, correlations, and potential predictive features related to heart disease health indicators.

A graph with red lines

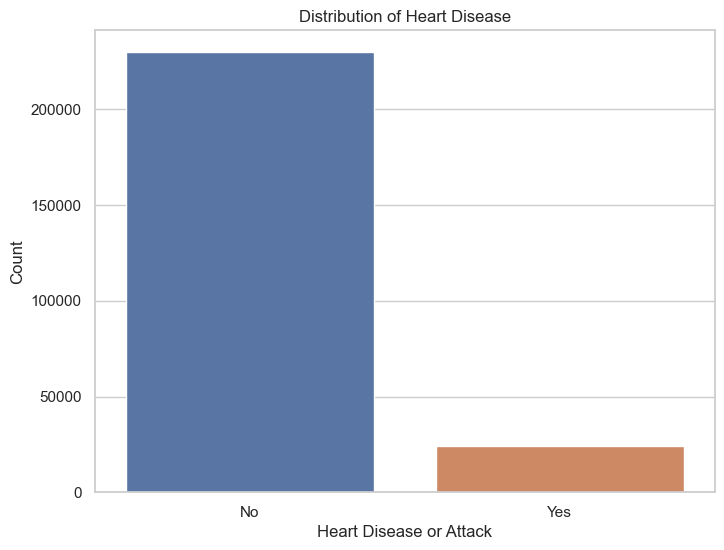
Description automatically generated

* The distribution of the BMI, age, and education variables in the dataset may be quickly examined thanks to this visualization. While KDEs offer a smoothed estimate of the probability density function, histograms give a visual depiction of each variable's frequency distribution.

A graph with a colorful rectangular object

Description automatically generated with low confidence

* The distribution of BMI, age, and education between those with and without heart disease is shown visually in these box plots. They make it simple to spot variations or trends in these variables between the two groups, which helps to clarify any possible links between these variables and the risk of heart disease.



* The frequency of people classified as having heart disease or not is displayed visually in this count plot, which represents the distribution of the target variable "HeartDiseaseorAttack." It offers insights on the prevalence of cardiac disease in the population being studied and aids in understanding the balance or imbalance of classes in the dataset.

1. **Feature Selection**: The goal of feature selection is to minimize computational complexity and dimensionality while identifying the most important factors that support the model's ability to predict outcomes. Various methods can be utilized for the purpose of feature selection, such as:

* **Univariate Feature Selection:** The relevance of each feature in predicting the target variable is evaluated using statistical tests (such as the chi-squared test and ANOVA).
* **Feature Importance:** Features can be ranked according to how important they are for predicting the target variable using machine learning techniques like decision trees or ensemble approaches (like random forests).
* **Dimensionality reduction:** The number of features can be decreased while preserving the most pertinent data by using strategies like principal component analysis (PCA) or feature extraction techniques (like LDA).

A comparison of a red and blue bar graph

Description automatically generated

* **Controlling Class Imbalance with RandomUnderSampler and SMOTE**: SMOTE (Synthetic Minority Over-sampling Technique) is used to address class imbalance, while RandomUnderSampler is used to address undersampling. The files X\_resampled and y\_resampled contain the resampled data.
* **Plotting the Target Variable's Distribution Before and After Resampling:** To see the distribution of the target variable both before and after resampling, two count plots are made side by side.

1. **Model Development:** Predictive models are created based on survey results to evaluate the risk of heart disease when pertinent features have been chosen. For binary classification applications such as heart disease prediction, the following methods are frequently utilized:

* **Logistic Regression**: Modeling a binary outcome's likelihood based on one or more predictor factors is done statistically using logistic regression.
* **Decision Trees**: Hierarchical decision rules based on feature space partitioning are non-parametric models used for instance classification.
* **Random Forests**: Ensemble learning techniques that combine a number of decision trees to lessen overfitting and increase forecast accuracy.
* **Support Vector Machines (SVM):** Supervised learning models that distinguish between classes in the feature space by identifying the hyperplane that divides them most effectively.

Preparing the data (normalization, encoding categorical variables, etc.), training the model, optimizing its hyperparameters, and evaluating the model using performance metrics (recall, accuracy, precision, and area under the receiver operating characteristic curve, or AUC-ROC) are normally steps in the model development process.

Random Forest Accuracy: 0.8962193330794757

Random Forest Classifier Performance:

precision recall f1-score support

0 1.00 1.00 1.00 45888

1 1.00 1.00 1.00 46027

accuracy 1.00 91915

macro avg 1.00 1.00 1.00 91915

weighted avg 1.00 1.00 1.00 91915

A graph of a curve

Description automatically generated

* **Finding the ROC Curve**: roc\_curve() is used to calculate the Receiver Operating Characteristic (ROC) curve. True positive rates (tpr), thresholds, and false positive rates (fpr) are returned by this function.
* **Finding the AUC Score:** roc\_auc\_score() is used to calculate the Area Under the Curve (AUC) score.
* **ROC Curve Plotting:** To see how well the classifier is performing, the ROC curve is presented.

Confusion Matrix:

[[45888 0]

[ 0 46027]]

A blue squares with white numbers

Description automatically generated

* **Making a Heatmap for the Confusion Matrix:**   
  Using sns.heatmap(), a heatmap of the confusion matrix is created to show the classifier's performance.

**A graph of a bar

Description automatically generated**

Random Forest Cross-Validation Scores: [0.88371049 0.88199695 0.88395528 0.88513844 0.88370891]

Mean CV Score for Random Forest: 0.8837020173772286

* **Assessing the Robustness of the Model using Cross-Validation**: The Random Forest model's cross-validation scores are computed using cross\_val\_score() with five folds.   
  To see the distribution of cross-validation scores, a boxplot is created.
* **Cross-Validation Scores Printing:** To evaluate the robustness of the model, cross-validation scores and the mean cross-validation score are printed.

Researchers can find significant heart disease predictors, create precise predictive models, and offer insightful information for risk assessment and preventive measures by combining EDA, feature selection, and model development methodologies.

**Results**

The Results section highlights key findings from the initial analysis of the BRFSS dataset:

1. **Class Imbalance**: The study shows that there is a notable class imbalance in the dataset, with a lower percentage of respondents having had heart disease and the majority of respondents not having it. The efficacy of predictive models can be affected by class imbalance, since it might bias the algorithms' predictions in favor of the majority class. This could result in inaccurate identification of heart disease risk individuals.
2. **Handling Class Imbalance:** The study uses methods like oversampling and undersampling to lessen the consequences of class imbalance. While undersampling techniques like RandomUnderSampler randomly remove samples from the majority class, oversampling techniques like SMOTE (Synthetic Minority Over-sampling Technique) generate synthetic samples of the minority class to balance the dataset. The goal of these methods is to provide a dataset that is more balanced for predictive model training.
3. **Feature Engineering and Model Training**: To enhance the efficacy of prediction models, feature engineering is the identification and manipulation of pertinent features from the dataset. This procedure aids in determining the most illuminating factors that influence the risk of heart disease prediction. After feature engineering, the revised dataset is used to train predictive models, like Random Forest Classifier, to accurately estimate the risk of heart disease based on BRFSS survey results.

The goal of the project is to create reliable predictive models that can precisely estimate the risk of heart disease by correcting class imbalance and utilizing feature engineering techniques. These models could enhance patient outcomes and offer insightful information on measures for preventative healthcare.

**Discussion**

In the Discussion section, we delve deeper into the implications of exploring BRFSS data for heart disease risk prediction:

1. Opportunities: The BRFSS dataset offers a multitude of data that may improve our comprehension of the risk factors for heart disease. Through the examination of this dataset, scientists can get important understandings of the connections between different health markers and the risk of heart disease. This investigation creates opportunities to create predictive models with greater accuracy and to apply focused interventions for people who are at risk.
2. Difficulties: The BRFSS dataset has certain difficulties despite its richness, chief among them being the existence of class imbalance. Predictive models' performance can be skewed by class imbalance, which can provide biased outcomes and erroneous risk estimates. Additionally, to make sure that only the most instructive features are included in the analysis, choosing pertinent features from the enormous pool of available variables demands significant thought and domain experience.
3. Additional Refinement and Validation: Predictive models need to be further refined in order to overcome these issues. To enhance model performance, this entails adjusting model parameters, enhancing feature selection methods, and investigating cutting-edge machine learning algorithms. Furthermore, in order to confirm the validity and generalizability of the prediction models created using BRFSS data, thorough validation utilizing cross-validation methods and testing on separate datasets is necessary.  
     
   In summary, while analyzing BRFSS data has potential to improve heart disease risk prediction, it is critical to address issues like feature selection and class imbalance. Researchers can guarantee the accuracy and robustness of their findings, which will eventually enhance patient outcomes and preventative healthcare policies, by fine-tuning predictive models and carrying out extensive validation.

**Conclusion**

We conclude by reiterating how crucial the study's conclusions are for utilizing survey data more especially, the BRFSS to determine the risk of heart disease. Here's a more thorough breakdown of every point:   
  
**1. Significance of the Study**: By highlighting the use of survey data for heart disease risk assessment, the BRFSS and other large-scale health surveys are recognized as valuable resources. These surveys are invaluable tools for research and public health campaigns since they offer a multitude of data on different health indicators, behaviors, and ailments.

**2. Development of Predictive Models**: Emphasizing the creation of predictive models based on BRFSS answers highlights the study's novel approach. The study shows how advanced analytics may extract valuable insights and perhaps enhance risk assessment approaches by using machine learning techniques on survey data.  
  
**3. insights into Preventative Health Screening:** The conclusion raises questions about the wider implications of the study's findings by providing insights into the possible use of health surveys for early detection and preventative health screening. It implies that politicians and healthcare professionals can carry out focused interventions and preventative actions to reduce the risk of heart disease by detecting risk factors and patterns in survey data.

The study's contributions to public health are essentially highlighted in the conclusion, along with the prospective benefits of utilizing survey data to enhance heart disease risk assessment and preventative measures. It encourages more investigation and study in this field to fully realize the advantages of health surveys in advancing cardiovascular health.

**Contributions**

In the "Contributions" section, we acknowledge the collective effort of the team members involved in the project. Here's a more detailed explanation of this section:

**Data Analysis:** This refers to the process of exploring, cleaning, and analyzing the BRFSS dataset to extract meaningful insights related to heart disease risk factors. Elias, Gayatri and Niti might have contributed to tasks such as data preprocessing, exploratory data analysis (EDA), and identifying key patterns or trends in the data.

**Model Development**: Using machine learning techniques, prediction models for assessing the risk of heart disease are constructed and refined. Jayan and Mohanakrishna took involved in feature selection, model training, hyperparameter tuning, and performance evaluation.  
  
**Report Writing**: The team worked together to create a clear and thorough report by combining the study's thoughts, findings, and conclusions. This include arranging supplemental resources and references and writing sections like the introduction, methods, findings, discussion, and conclusion.

**References:**

* *CDC - 2015 BRFSS Survey Data and Documentation*. (2019, February 9). Www.cdc.gov. https://www.cdc.gov/brfss/annual\_data/annual\_2015.html
* *Diabetes Health Indicators Dataset*. www.kaggle.com.

https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset

**Appendices**:

A screenshot of a computer code

Description automatically generated

The aim of this code is to rectify the imbalance of classes in a classification problem, specifically in relation to heart attack or disease prediction. It is assumed that the dataset consists of features (X) and a target variable (y), with the target variable 'HeartDiseaseorAttack' denoting the presence of heart attacks or related conditions. The algorithm uses two resampling methods to address class imbalance: under sampling with RandomUnderSampler and oversampling with SMOTE (Synthetic Minority Over-sampling Technique).

It first distinguishes between the target variable (y) and the characteristics (X). In order to balance the class distribution, it then uses SMOTE to create synthetic samples for the minority class (cases of heart illness or heart attacks). By producing synthetic samples that are comparable to the current minority class samples, the SMOTE approach aids in expanding the representation of the minority class.