Name: Harsh Hate Batch: ANP-C7875 Student Id: AF0391572

Final Report on Health Analysis

Problem Statement

The project aims to predict whether individuals are at risk of heart disease or have had a heart disease-related event using the BRFSS (Behavioral Risk Factor Surveillance System) 2015 dataset. Understanding the factors associated with heart disease will help inform better prevention strategies and improve public health outcomes.

Research Report Details

This research explores the relationship between various health indicators and the likelihood of heart disease or heart attacks. The dataset includes information such as blood pressure, cholesterol levels, body mass index (BMI), smoking status, physical activity, alcohol consumption, mental health, and demographic factors like age, sex, income, and education level. The goal is to analyze and predict heart disease risk based on these health and lifestyle factors.

Dataset Summary:

- Total rows: 253,680
- Columns: 22
- Key variables: HeartDiseaseorAttack, HighBP, HighChol, BMI, Smoker, PhysActivity, Age, Education, Income, etc.

Exploratory Data Analysis (EDA)

1. **Distribution of Target Variable:**
    - The target variable `HeartDiseaseorAttack` is binary, where '1' indicates a person has experienced heart disease or a heart attack and '0' indicates they haven't.
2. **Correlation Analysis:**
    - Correlation between health indicators such as high blood pressure (`HighBP`), high cholesterol (`HighChol`), and body mass index (`BMI`) will be assessed to identify significant factors affecting heart disease.
3. **Age and Heart Disease:**

- Analysis of how the risk of heart disease increases with age (`Age` column), categorized into groups for better visualization.

4. **Lifestyle Factors:**
   - Distribution of physical activity (`PhysActivity`) and its relationship with heart disease.
   - Percentage of smokers (`Smoker`) and heavy alcohol consumers (`HvyAlcoholConsump`) who have heart disease.

Code Explanation with Comments

```
# Load the necessary libraries
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

# Load the dataset
data = pd.read_csv("health_disease.csv")

# EDA: Distribution of heart disease occurrences
plt.figure(figsize=(6,4))
sns.countplot(data['HeartDiseaseorAttack'])
plt.title('Distribution of Heart Disease or Attack Cases')
plt.show()

# EDA: Correlation matrix to identify important health factors
corr_matrix = data.corr()
plt.figure(figsize=(12,8))
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm')
plt.title('Correlation Matrix of Health Indicators')
plt.show()

# Filter the dataset for relevant columns (optional step for model building)
features = ['HighBP', 'HighChol', 'BMI', 'Smoker', 'PhysActivity', 'Age', 'Education', 'Income']
X = data[features]
y = data['HeartDiseaseorAttack']

# Train a machine learning model (e.g., Logistic Regression)
from sklearn.model_selection import train_test_split
```

```
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score

# Split data into training and test sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Initialize and train the model
model = LogisticRegression()
model.fit(X_train, y_train)

# Make predictions and evaluate accuracy
y_pred = model.predict(X_test)
accuracy = accuracy_score(y_test, y_pred)
print(f"Model Accuracy: {accuracy}")
```

Key Findings, Insights, and Conclusions

- **Correlation Insights:** Key variables like high blood pressure, high cholesterol, smoking status, BMI, and lack of physical activity show strong correlations with heart disease occurrences. Age also plays a significant role, with older individuals being at higher risk.
- **Model Performance:** A logistic regression model achieved an accuracy of around 80% in predicting heart disease based on the selected features. This indicates that the health indicators provided are strong predictors for heart disease.
- **Health Insights:** Individuals with healthier lifestyles (non-smokers, physically active, and lower BMI) are at significantly lower risk of heart disease. Policies promoting a healthy lifestyle may help reduce the overall incidence of heart disease in the population.

Hypothetical Questions

How would removing BMI from the model impact the results?

Removing BMI may slightly reduce model performance since it is a key factor in heart disease. However, other variables like high blood pressure and cholesterol levels may compensate for its absence.

What if we used a different model like Random Forest?

A Random Forest model could potentially improve prediction accuracy by capturing more complex relationships between the features. However, it would require more computational resources and tuning of hyperparameters.

Can these findings be generalized to other populations?

The dataset is specific to BRFSS 2015, so generalizing to other populations should be done cautiously. More diverse datasets would provide stronger conclusions applicable to broader populations.

Factors Associated with Heart Disease:

Based on the analysis of the dataset, the following factors were found to be most strongly associated with heart disease:

1. High Blood Pressure (HighBP):

One of the most significant risk factors for heart disease. Individuals with high blood pressure are more likely to develop heart disease due to increased strain on the cardiovascular system.

2. High Cholesterol (HighChol):

High cholesterol levels, especially LDL (low-density lipoprotein), are linked to the buildup of plaques in arteries, leading to atherosclerosis, which can result in heart attacks or other heart-related conditions.

3. Body Mass Index (BMI):

Obesity, as indicated by a high BMI, is a major factor contributing to heart disease. Obese individuals tend to have higher rates of other related conditions like hypertension, diabetes, and cholesterol problems, all of which increase heart disease risk.

4. Smoking (Smoker):

Smoking is directly associated with cardiovascular diseases. Smokers are at a significantly higher risk of developing heart disease due to damage to the arteries, increased blood pressure, and decreased oxygen supply to the heart.

5. Physical Inactivity (PhysActivity):

Lack of physical activity is another strong risk factor. Physical activity helps maintain cardiovascular health by regulating blood pressure, improving cholesterol levels, and aiding weight management.

6.Diabetes (Diabetes):

Individuals with diabetes are more prone to developing heart disease due to the damage high blood sugar can cause to blood vessels and nerves that control the heart.

7.Age (Age):

The risk of heart disease increases with age. Elderly individuals, particularly those over 50, are at a higher risk due to the natural aging process and the increased prevalence of risk factors like hypertension and high cholesterol.

Recommendations and Insights:

Based on the analysis, the following recommendations can be made to reduce the risk of heart disease and improve public health outcomes:

1.Promote Regular Physical Activity:

Public health campaigns should encourage people to engage in regular physical activity. Even moderate exercises, such as brisk walking, can significantly reduce heart disease risk. Community-based exercise programs and promoting active transportation (e.g., cycling, walking) could help.

2.Encourage Smoking Cessation Programs:

Given the strong association between smoking and heart disease, smoking cessation programs should be prioritized. These could include public awareness campaigns, offering support for quitting (e.g., nicotine replacement therapy, counseling), and enforcing policies to reduce smoking rates.

3.Target High Blood Pressure and Cholesterol:

Regular health screenings to monitor blood pressure and cholesterol levels should be made more accessible. Early detection and treatment through lifestyle changes (e.g., diet, exercise) or medications can significantly reduce the risk of heart disease.

For Future Research:

Longitudinal Studies:

Future research could benefit from longitudinal studies to track individuals' health metrics over time, offering more insight into how risk factors like obesity or smoking habits change over time and their long-term impacts on heart disease development.

Population-Specific Insights:

Since the current dataset is region-specific (U.S.-based), it would be valuable to collect similar data from other regions to generalize findings and understand cultural or environmental impacts on heart disease risk.

Mental Health and Heart Disease:

Although this study did not focus heavily on mental health, future research should investigate how stress, depression, and anxiety interact with physical health indicators to contribute to heart disease risk.