Description of Business

Problem Statement

* Globally rising incidences of skin cancer, heart disease, diabetes, and stroke present a pressing public health challenge. The lack of efficient methods for early identification of these illnesses causes delayed diagnosis and higher medical expenses.

Hypothesis

* There is a positive correlation between a higher patient's BMI (Body Mass Index) and their likelihood of developing heart disease.

Projected Outcome of the analysis

* Identify which individuals who have not been diagnosed with heart disease has a high likelihood of contracting heart disease.
* Identify individual demographics that increase the likelihood of them contracting heart disease.
* Help HCA Healthcare increase the number of successfully treated patients by 10% by predicting the risk of patients contracting heart disease for early treatment, leading to better treatment outcomes and prevention of heart disease progression.

User Stories

1. As a middle-aged individual concerned about my health, I want to use a heart disease prediction model to assess my risk, so that I can take necessary preventive measures and maintain my health.

* The prediction model should allow me to input relevant health data such as age, gender, past medical records, and BMI.
* The model should provide a risk assessment based on the entered data, indicating the likelihood of developing heart disease.
* The results should be presented in an easily understandable format, such as a percentage or risk category.

1. As a healthcare provider, I want to leverage a heart disease prediction model to assist me in identifying high-risk patients, so that I can provide personalized care to high-risk patients.

* The prediction model should integrate with our electronic health record system, allowing easy access to patient data.
* The model should analyse patient data, including demographics, and medical history, to generate a classification.
* The results should be presented in a clear and concise manner, providing an indication of the patient's risk level.

User Journey Map

A screenshot of a diagram

Description automatically generated

Details of Dataset

Scope of datasets

* The Heart Disease dataset has been chosen for further analysis due to its strong predictive performance. Logistic regression without tuning has already achieved a commendable F1 score of 0.68, indicating good accuracy. Moreover, the dataset is highly versatile and can be utilized for multiple mining goals. It contains a sufficient number of records, ensuring robustness in the analysis. Given its relevance to the healthcare industry, the dataset holds significant potential for extracting valuable insights and facilitating advancements in healthcare research and practices.

Dataset Information

* Description: Key indicators of heart disease. Data collected from 2020 annual CDC survey data of 400k adults related to their health status.
* Source of the data: <https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease>
* Type of dataset: CSV and shared with all members.
* How will the dataset be used: All columns will be used.

Individual Contribution

* Dataset used is the heart disease dataset from Kaggle as specified in the dataset information.
* Data Cleaning and transformation done (please refer to the Jupyter notebook):

1. Removed duplicates in the dataset.
2. Removed outliers in the numerical columns – SleepTime and BMI. There are records with unlikely values like BMI 90 and SleepTime of 24 hours.
3. Replace values in categorical columns.

* The binary categorical values have values of ‘Yes’ and ‘No’ rather than 0s and 1s. I convert binary categories to 0s and 1s.
* Diabetic column has 4 unique values 'Yes' 'No' 'No, borderline diabetes' 'Yes (during pregnancy)'. I convert ‘No, borderline diabetes’ to ‘No’ and ‘Yes (during pregnancy)’ to ‘Yes’ to reduce the unique values.

1. Performed one hot encoding on multiclass columns - Sex, AgeCategory, Race, GenHealth

* I use one hot encoding instead of label encoding for easily interpretability for the users when they view the results.

1. Performed two types of data balancing – Undersampling on the target column – HeartDisease as the HeartDisease column after cleaning is unbalanced, with 9% of 1s and 81% of 0s. Performed data balancing such that 1s contribute 40% and 0s contribute 60%. This ensures that the 0 class retains majority of the original data without the model being biased toward the 0 class.

* How is the dataset related to the problem/ hypothesis?
* This dataset contains individual’s health records which helps us gather insights on what possible demographics/factors that could increase the likelihood of an individual contracting heart disease.
* It also allows healthcare providers to gather insights into patterns/combinations of characteristics that lead to an increased chance of heart disease, allowing them to better plan treatment plans for individuals, reducing time to diagnosis and less medical expenses.

Exploratory Data Analysis (You can view the codes in the attached Jupyter notebook/ HTML file).

1. No Null values in the dataset

A screenshot of a computer

Description automatically generated

1. There are a total of 18078 duplicate records. They will be removed to reduce noise in the dataset.



1. Looking at the continuous columns, there are outliers in all columns. Apart from Meantal and PhysicalHealth column, I will remove outliers in BMI and SleepTime as they are not logical and are likely outliers (e.g., 24 hours sleep and BMI of 90) which could affect model prediction performance.

A group of graphs showing different types of health

Description automatically generated

1. Some categorical column values are not in the correct format. “Yes”/”No” will be converted to 1/0 respectively. The diabetic column also has interesting values. Yes during pregnancy means having diabetes so I convert it to 'yes'. No, borderline diabetes means the individual does not have diabetes so 'no'. The other multiclass categorical column like Sex, AgeCategory, GenHealth, and Race will be cleaned with one hot encoding.

A screenshot of a computer

Description automatically generated

1. Looking at distribution of binary value columns, majority of individuals are fairly unbalanced.

A diagram of different types of health

Description automatically generated A white background with black dots

Description automatically generated

1. Let’s take a closer look at the target column: HeartDisease. The distribution is heavily unbalanced with the 1 class contributing only 9% of the total records. Data balancing is required to prevent the model from being biased toward the 0 class.

A blue and orange pie chart

Description automatically generated

1. Looking at the occurrence of each demographic in the dataset, majority of the individuals are of the White race, aged above 50, and have good general health statuses. This might result in the model being biased toward the white race.

A group of blue bars

Description automatically generated

1. Looking at the percentage of heart disease in each class, more individuals with an age above 50 have heart disease as compared to the younger individuals. Although the general health of patients are fair/good/verygood, majority of them still have heart diseases.

A group of green and orange bars

Description automatically generated

1. Checking correlation among columns after cleaning and encoding. There are no highly correlated features in the dataset. Therefore, I will not remove any columns.

A green square with yellow and blue squares

Description automatically generated

Data Catalogue

Description of the fields in the data set:

* HeartDisease: Respondents that have ever reported having coronary heart disease (CHD) or myocardial infarction (MI)
* BMI: Body Mass Index (BMI)
* Smoking: Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes]
* AlcoholDrinking: Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week
* Stroke: (Ever told) (you had) a stroke?
* PhysicalHealth: Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days was your physical health not good? (0-30 days)
* MentalHealth: How many days during the past 30 days was your mental health not good? (0-30 days)
* DiffWalking: Do you have serious difficulty walking or climbing stairs?
* Sex: Are you male or female?
* AgeCategory: Fourteen-level age category
* Race: Imputed race/ethnicity value
* Diabetic: (Ever told) (you had) diabetes?
* PhysicalActivity: Adults who reported doing physical activity or exercise during the past 30 days other than their regular job
* GenHealth: Would you say that in general your health is...
* SleepTime: On average, how many hours of sleep do you get in a 24-hour period?
* Asthma: (Ever told) (you had) asthma?
* KidneyDisease: Not including kidney stones, bladder infection or incontinence, were you ever told you had kidney disease?
* SkinCancer: (Ever told) (you had) skin cancer?