Risk Factors of Coronary Heart Disease Among the General United States Population

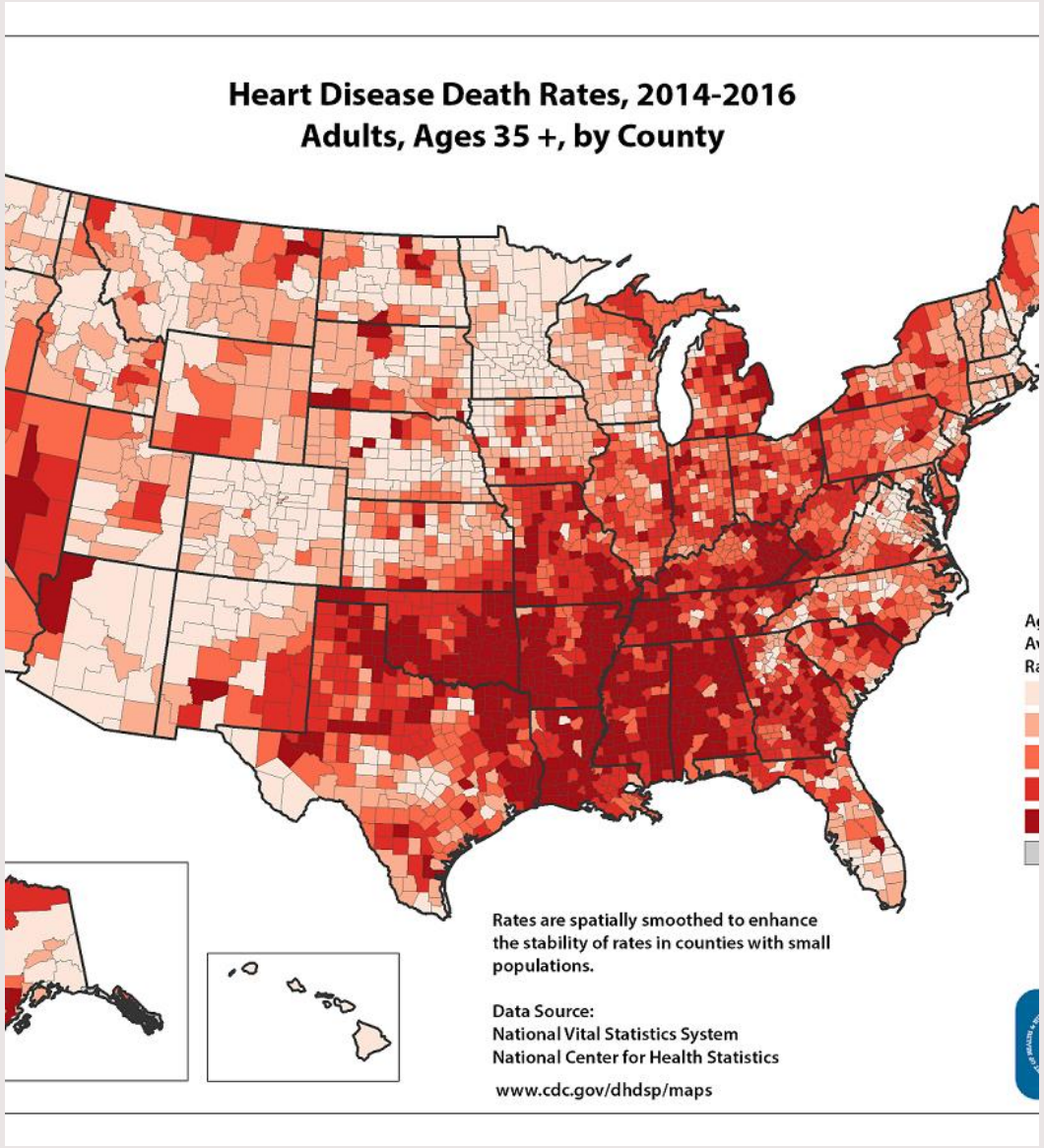
Katie Adams

DSC 630 – Predictive Analytics

Spring 2022

Heart Disease Risk in the United States

- "Heart disease is **the leading cause of death** for men, women, and people of most racial and ethnic groups in the United States.
- **One person dies every 36 seconds** in the United States from cardiovascular disease.
- About **659,000 people in the United States die from heart disease** each year
  - that's 1 in every 4 deaths
- **Heart disease costs the United States about $363 billion each year** from 2016 to 2017." (CDC, 2022).



Predictive or Related Factors of Heart Disease

- Interested parties **would be interested in understanding what factors are most predictive of heart disease** to decrease patient mortality and save costs
- Doctors and healthcare professionals would be interested to be able **to better advise patients on heart disease risk factors and reduce heart disease events** (coronary heart disease and myocardial infarction)
- Life & health insurance companies would also be interested in these heart disease predictive features to **better understand how risky a particular customer** would be to insure

Predictive or Related Factors of Heart Disease

- To understand **what patient lifestyle and health factors are most related to or most predictive of heart disease** (single characteristics and interactions) an open-source dataset from Kaggle was used with 319,795 rows and 18 columns of US patient data
- Each dataset **column represents a factor or feature that** represents a measurable piece of data that can be used in this analysis

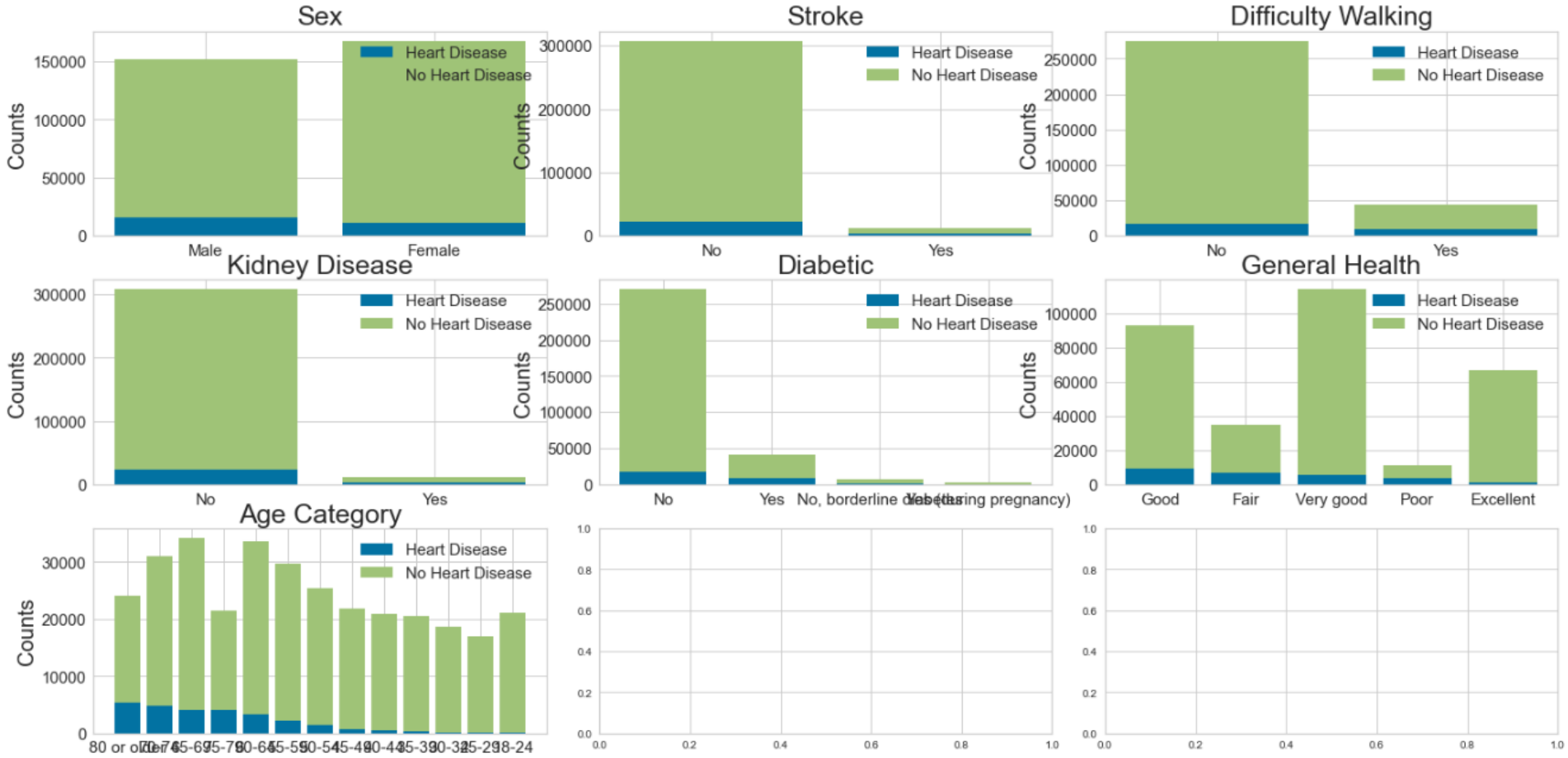
Dataset Features

- **Heart Disease:** If subject has experienced heart disease (Yes or No)
- **BMI (Body Mass Index):** Subject's current BMI (numerical value)
- **Smoking:** If subject has smoked more than 100 cigarettes (Yes or No)
- **Alcohol Drinking:** If males subjects have more than 14 drinks/week and if females have more than 7/week (Yes or No)
- **Stroke:** If subject has experienced at stroke (Yes or No)
- **Physical Health:** Subject's daily exercise count in last 30 days (numerical value)
- **Mental Health:** Subject's poor mental health days count in last 30 days (numerical value)
- **Difficulty Walking:** If subject has difficulty walking or climbing stairs (Yes or No)
- **Sex:** If subject is male or female
- **Age Category:** The age category of the subject (14 categories)
- **Race:** The race and/or ethnicity of the subject
- **Diabetic:** If subject has experienced diabetes (Yes, Yes – Pregnancy, No, No- borderline diabetes)
- **Physical Activity:** If subject had exercised in the last month outside of their job (Yes or No)
- **General Health:** Subject's general health (Excellent, Very Good, Good, Fair, Poor)
- **Sleep Time:** Subject's average hourly sleep at night (numerical value)
- **Asthma:** If subject has experienced asthma (Yes or No)
- **Kidney Disease:** If subject has experienced kidney disease (Yes or No)
- **Skin Cancer:** If subject has experienced skin cancer (Yes or No)

Discovery

- **Most people in the dataset do not have** heart disease, stroke, difficulty walking, kidney disease, or diabetes
- Heart disease rates are **relatively the same in both men and women** (however women have slightly lower heart disease rate)
- **General Health and Age appear to be a significant factor** in predicting heart disease
- **Correlations**
  - The dataset patient features most correlated to heart disease are: **Physical Health, Stroke, Difficulty Walking Up Stairs, Age 80 or older, Diabetes, Fair & Poor General Health, and Kidney Disease**

Discovery

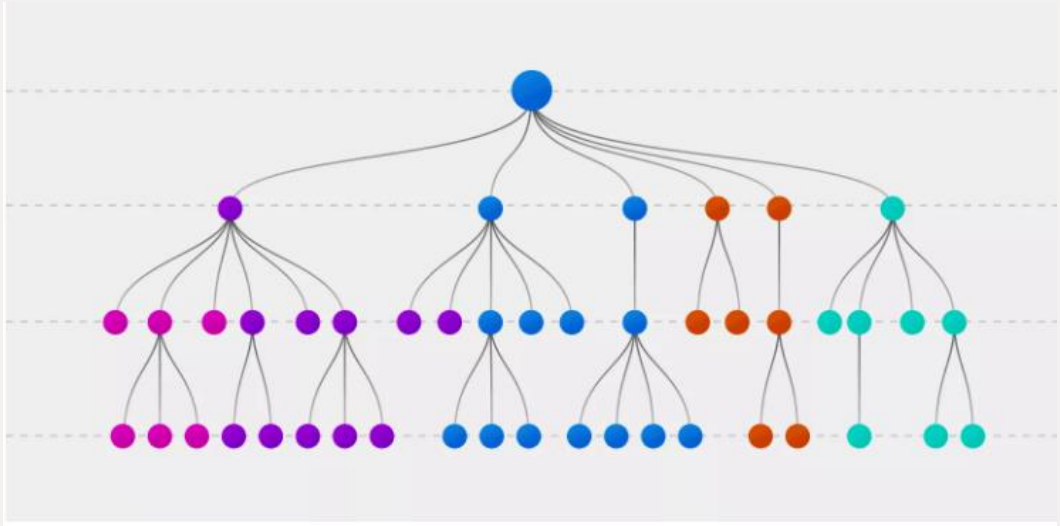


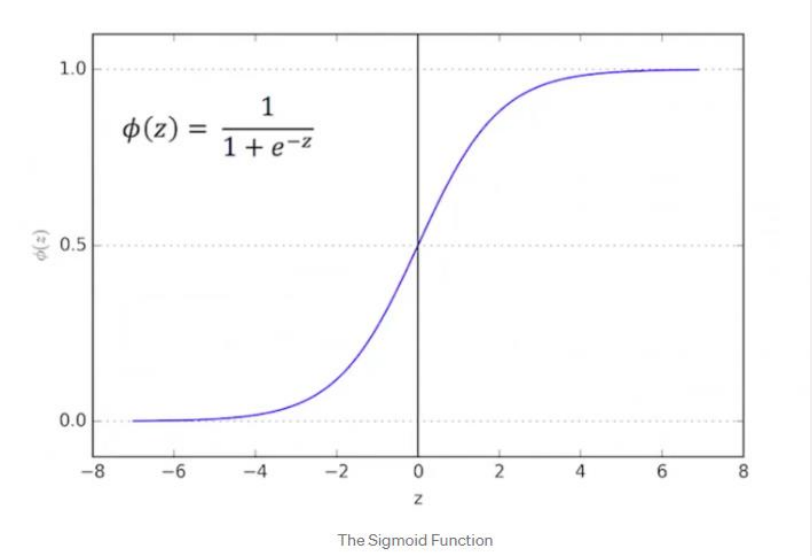
Observations

- Most people in the dataset do not have heart disease, stroke, difficulty walking, kidney disease, or diabetes
- Heart disease rates are relatively the same in both men and women (however women have slightly less occurences of heart disease)
- General Health and Age appear to be a significant factor in predicting heart disease

Methods

- To determine which patient features are most predictive to heart disease, **a logistic regression model was created** as the prediction is either "Heart Disease" or "No Heart Disease"



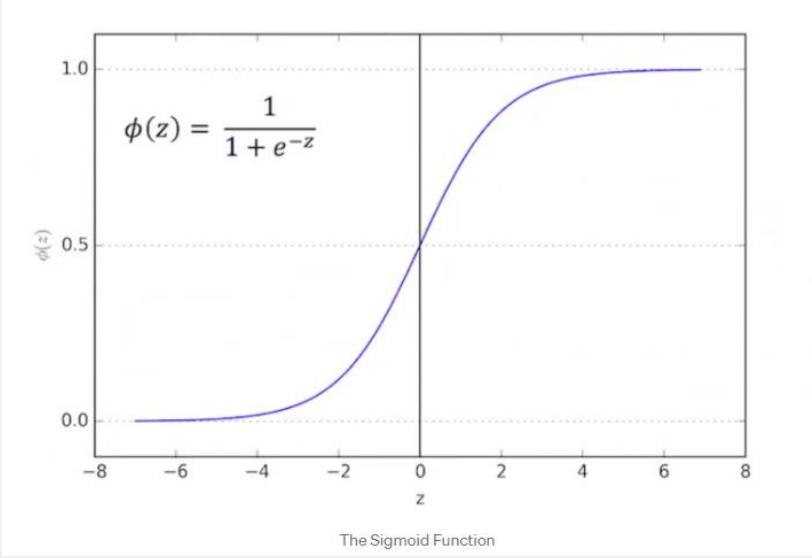


The Sigmoid Function

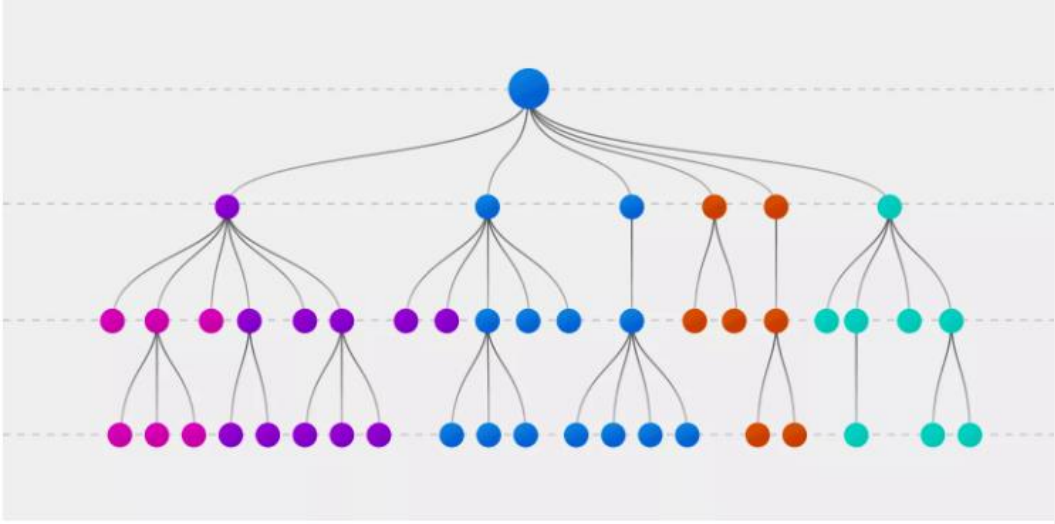
- In addition to the logistic regression model, a **decision tree model** was created to also determine which features are most predictive of heart disease
- Both models were implemented **to find out which model has better performance of predicting whether a patient will have heart disease based on the patient features**

Methods

- **Logistic Regression Models:** 7 models using a combination of **all features, 5 best features from X2, and PCA** and automatic and manual weighting of the target class (Heart Disease) was since there are **10 times as many subjects with no history of heart disease** compared to those with heart disease



The Sigmoid Function



- **Decision Tree:** 2 models using a combination of **all features and 5 best features from X2**
- **Performance Metrics:** The model accuracy, precision, recall, F1 and AUC scores were calculated to determine which model best predicts heart disease using the features

Methods

- **Accuracy:** “Accuracy represents the number of correctly classified data instances over the total number of data instances” (B, H. N., 2020)
- **Precision:** “Precision talks about how precise/accurate your model is out of those predicted positive, how many of them are actual positive.” (B, H. N., 2020)
- **Recall:** “Recall actually calculates how many of the Actual Positives our model capture through labeling it as Positive (True Positive).” (B, H. N., 2020)
- **F1 Score:** “F1-score is a metric which takes into account both precision and recall.” (B, H. N., 2020)
- **AUC Score:** “"So, what area under the ROC curve describes good discrimination? Unfortunately there is no "magic" number, only general guidelines. In general, we use the following rule of thumb:

- 0.5 = This suggests no discrimination, so we might as well flip a coin.
- 0.5-0.7 = We consider this poor discrimination, not much better than a coin toss.
- 0.7-0.8 = Acceptable discrimination
- 0.8-0.9= Excellent discrimination
- >0.9 = Outstanding discrimination" (Dixon, 2020)

Methods

- **Chi- Squared (X2):** When two features are independent, the observed count is close to the expected count, thus we will have smaller Chi-Square value. So high Chi-Square value indicates that the hypothesis of independence is incorrect. In simple words, higher the Chi-Square value the feature is more dependent on the response and it can be selected for model training.

  The 5 Best Features with the highest X2 value are as follows:
  PhysicalHealth
  Stroke_Yes
  DiffWalking_Yes
  Diabetic_Yes
  GenHealth_Poor

Results

- **The 2 Decision Tree models with all 37 features and the 5 best features should not be selected as our model** as their AUC scores (0.59) was less than that of the Logistic Regression Models, and overall, is not much better than random chance (0.5)
- **The 3 Logistic Regression Models with no class balancing should not be selected as our model** because their F1 score for minority class (Yes - heart disease) is low due to the class imbalance of the target variable
- **The 1 Logistic Regression Model with automatic class balancing should not be selected as our model** because the accuracy lowered to 75%
- **The 3 Logistic Regression Models with manual class balancing should be selected as our model** because their accuracy, precision, recall, F1 and AUC scores were all acceptable, including the F1 score for the minority class (Yes - heart disease)
- Although the manual class balancing Logistic Regression Models had relatively the same performance, the fact that the **5 best features from X2 did relatively the same as both all 37 features and PCA (34 features) lends itself to show that these 5 best features (PhysicalHealth, Stroke_Yes, DiffWalking_Yes, Diabetic_Yes, GenHealth_Poor) have a bigger impact on the target variable (HeartDisease)** than any of the other variables combined
- **The Logistic Regression Model with manual classing balancing using 5 best features from X2 should be selected as our model of choice**

Results

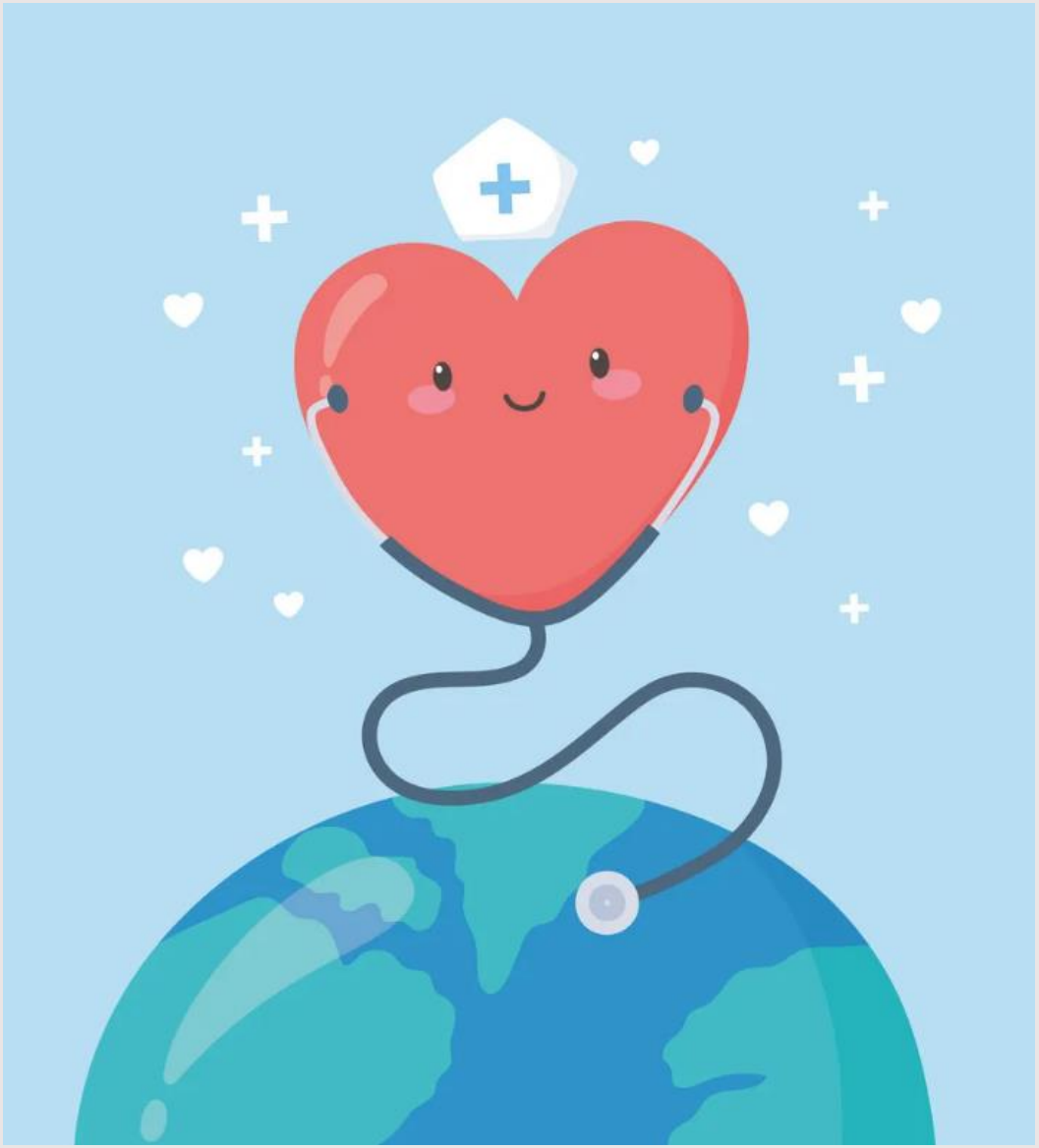
Model	Accuracy	F1 Score (Yes-Heart Disease)	F1 Score (No-Heart Disease)	AUC Score
Decision Tree Classifier - All 37 features	86.44%	0.250	0.925	0.59
Decision Tree Classifier - 5 Best Features using X2	91.5%	0.060	0.955	0.59
Logisitic Regression - No class weighting using PCA (34 features)	91.53%	0.171	0.955	0.83
Logistic Regression - No class weighing using all 37 features	91.71%	0.187	0.956	0.85
Logistic Regression - No class weighting using 5 best features using X2	91.46%	0.102	0.955	0.72
Logistic Regression - Automatic class weighting of target variable with 37 features	75.12%	0.353	0.846	0.85
Logistic Regression - Manual class weighting of target variable with PCA (34 features)	87.01%	0.380	0.927	0.83
Logistic Regression - Manual class weighting of target variable with 37 features	87%	0.402	0.927	0.85
Logistic Regression - Manual class weighting of target variable with 5 best features using X2	88.61%	0.312	0.938	0.72

Recommendations

- As a recommendation, the X2 5 best features are easier to implement in a health care environment/setting than all 37 (i.e., it's easier to inform healthcare professionals on 5 features that impact or are most related to heart disease, rather than 37 features). **The 5 best features are:**
  - **PhysicalHealth:** Subject's report of how many days in the past 30 days they have exercised (integer)
  - **Stroke_Yes:** Subject has been told or has known that they have had a stroke
  - **DiffWalking_Yes:** Subject's opinion is that they have difficulty walking or climbing stairs
  - **Diabetic_Yes:** Subject has had or currently has diabetes
  - **GenHealth_Poor:** Subject's opinion of their general health is poor

Recommendations

- When **healthcare professionals are advising patients on heart health**, these are the 5 recommendations to provide:
  1. You should have as much **daily exercise** in a month as possible
  2. You should avoid anything that can cause or increase your **risk of stroke** (e.g., high blood pressure, high cholesterol)
  3. You should keep up and maintain your **ability to walk up or climb stairs easily**
  4. You should avoid anything that can cause or increase your **risk of diabetes** (e.g., body mass index over 25, inactivity, high blood pressure)
  5. You should have a **good routine related to your own general health** (e.g., exercise and nutritional habits)



Recommendations

- When **advising life & health insurance companies** on how to assess how risky customers would be to insure, these are the recommendations to provide:
  1. Customers with **high daily exercise, low incident of stroke, high ability to climb stairs, low incident of diabetes, and good general health are at less risk of heart disease** and heart disease treatments, surgeries, and death
  2. These **particular customers should be provided better insurance rates** as their likelihood of needing to submit insurance claims related to heart disease is lower than that of customers who don't exercise, have incident of stroke, low ability to climb stairs, high incident of diabetes, and poor general health
  3. Customer incentives for decreasing the risk of heart disease could include: **free access to dieticians, activity trackers, digital scales, healthy eating cookbooks, health routine logging**

Next Steps

- The **logistic regression model with the 5 best features isn't necessarily ready for deployment**, but it is ready to be beta tested with real-life data from area hospitals.
- This will **ensure that more potential real-world scenarios** that relate or do not relate to heart disease are captured in the dataset.
- To improve the model, **doctors and healthcare staff will also be interviewed on data collection process and possible methods for improvement**.
- Examples of feedback include **different data repositories, new features/measurements from patients to the dataset (e.g., type of diet), possible anecdotal features that doctors** and healthcare professionals have observed in the field.