Katie Adams
DSC 630 – Predictive Analytics
Spring 2022

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

Introduction

Many Americans die of heart disease every year, "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. This includes the cost of health care services, medicines, and lost productivity due to death." (CDC, 2022). Understanding what factors contribute to heart disease could save lives and healthcare costs. In this project, the goal will be to understand what lifestyle and health features (i.e., risk factors) are most related to or most predictive of heart disease (single characteristics and interactions) using an open-source dataset. This dataset was obtained from Kaggle (https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease?resource=download), has a large volume of records (319,795 rows) and 18 columns or features (9 booleans, 5 strings and 4 decimals), which are:

- **HeartDisease:** Respondents that have ever reported having coronary heart disease (CHD) or myocardial infarction (MI) - either yes or no
- **BMI:** Body Mass Index (BMI) - float
- **Smoking:** Have you smoked at least 100 cigarettes in your entire life? [Note: 5 packs = 100 cigarettes] - either yes or no
- **AlcoholDrinking:** Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week - either yes or no

- **Stroke:** (Ever told) (you had) a stroke? - either yes or no
- **PhysicalHealth:** Now thinking about your physical health, which includes physical illness and injury, for how many days during the past 30 days have you exercised - integer
- **MentalHealth:** Thinking about your mental health, for how many days during the past 30 days was your mental health not good? - integer
- **DiffWalking:** Do you have serious difficulty walking or climbing stairs? - either yes or no
- **Sex:** Are you male or female? - string
- **AgeCategory:** Fourteen-level age category - string
- **Race:** Imputed race/ethnicity value- string
- **Diabetic:** (Ever told) (you had) diabetes? - either yes, no, no-borderline diabetes, or yes pregnancy
- **PhysicalActivity:** Adults who reported doing physical activity or exercise during the past 30 days other than their regular job - either yes or no
- **GenHealth:** Would you say that in general your health is... - Excellent, Very Good, Good, Fair, Poor
- **SleepTime:** On average, how many hours of sleep do you get in a 24-hour period? - integer
- **Asthma:** (Ever told) (you had) asthma? - either yes or no
- **KidneyDisease:** Not including kidney stones, bladder infection or incontinence, were you ever told you had kidney disease? - either yes or no
- **SkinCancer:** (Ever told) (you had) skin cancer? - either yes or no

This dataset references individual lifestyle and other disease information, which is useful as they could have an impact on heart disease risk potential. It is the goal of this project to determine which features in the dataset, if any, are predictive of heart disease.

Doctors, healthcare professionals, patients, and life & health insurance companies would all be interested in understanding what factors are most predictive of heart disease. Life & health insurance companies would be interested to better understand how risky a particular customer would be to insure. 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 (CHD) or myocardial infarction (MI)).

Methods

Multiple Logistic Regression and Decision Tree Classifier models with an 80% train and 20% test sets were used to determine which heart disease dataset features are most related to or most predictive of heart disease, where the "HeartDisease" feature is the target. This was a supervised learning model where "the supervisor is the target variable, a column in the data representing values to predict from other columns in the data. The target variable is chosen to represent the answer to a question the organization would like to answer or a value unknown at the time the model is used that would help in decisions. Sometimes supervised learning is also called predictive modeling. The primary predictive modeling algorithms are classification for categorical target variables or regression for continuous target variables." (Abbott, 2014, p. 5).

In addition to the logistic and decision tree classifier models with 1 target and 17 features (37 features when dummy features were created for categorical features), multiple logistic models were

used with the 5 best features only derived with highest chi-squared statistics, and features extracted from Principal Component Analysis (PCA).

Logistic models are not perfect and have their disadvantages, “Logistic Regression is a statistical analysis model that attempts to predict precise probabilistic outcomes based on independent features. On high dimensional datasets, this may lead to the model being over-fit on the training set, which means overstating the accuracy of predictions on the training set and thus the model may not be able to predict accurate results on the test set. This usually happens in the case when the model is trained on little training data with lots of features. So on high dimensional datasets, Regularization techniques should be considered to avoid over-fitting (but this makes the model complex). Very high regularization factors may even lead to the model being under-fit on the training data.” (Grover, 2020). If overfitting is identified as an issue via cross-validation, two types of regularization techniques will be used: Lasso (L1 Regularization) or Ridge (L2 Regularization), where λ is called the regularization parameter and controls the trade-off between fitting the training data well and keeping the parameters small to avoid overfitting.

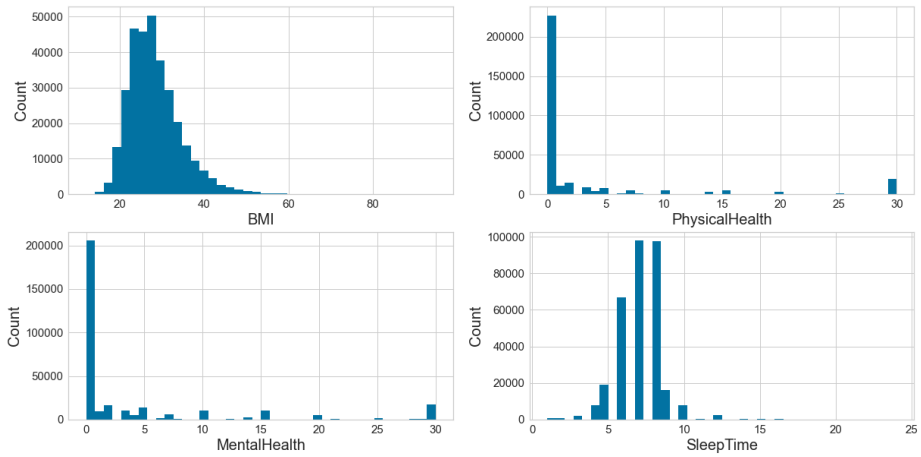
To measure the models’ performance, the accuracy, precision, recall and F1 score will be calculated for the logistic and decision tree models for both the 17 features, best 5 features, and PCA features, as well as visualize a confusion matrix and an ROC curve where "The area under the ROC curve (AUC) results were considered excellent for AUC values between 0.9-1, good for AUC values between 0.8-0.9, fair for AUC values between 0.7-0.8, poor for AUC values between 0.6-0.7 and failed for AUC values between 0.5-0.6." (El Khouli, 2009). All these together were used to evaluate the results of the logistic regression models. The definitions of accuracy, precision, recall, F1 and AUC scores can be found below:

- **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.
  - **Precision:** positive value
  - **Recall:** true positive rate” (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)

The data was prepared and cleaned to support the creation of several models and evaluation of their performance and results for interpretation. With these results, a conclusion and recommendations in reference to the original question and problem will be detailed.

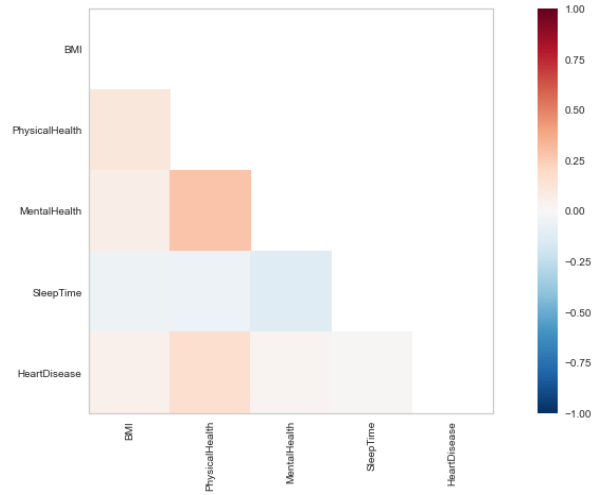
Katie Adams
DSC 630 – Predictive Analytics
Spring 2022

Histograms, bar charts, Pearson's correlation matrix (numerical data) and Spearman's correlation matrix (categorial & numerical data) were created to describe the data.

- **Numerical Data**
  - **Histogram:** To show the distribution of the numerical data (BMI, PhysicalHealth, MentalHealth, and SleepTime), histograms were used, and observations were captured:
    - Most people in the dataset have a BMI between 20 and 40
    - Most people in the dataset exercise between 0-5 days in the past 30 days
    - Most people in the dataset have 0-5 bad mental health days in the past 30 days
    - Most people in the dataset get between 5-9 hours of sleep a night
  - 

  - A **Pearson's Correlation Matrix** was then created to understand the correlation between 'HeartDisease" and the numerical features (BMI, PhysicalHealth, MentalHealth, SleepTime) and observations were captured:
    - The numerical feature most highly correlated with 'HeartDisease' is 'PhysicalHealth'

- There is not significant collinearity between the features (Rule of thumb: If the correlation > 0.8 then severe multicollinearity may be present)



Observations

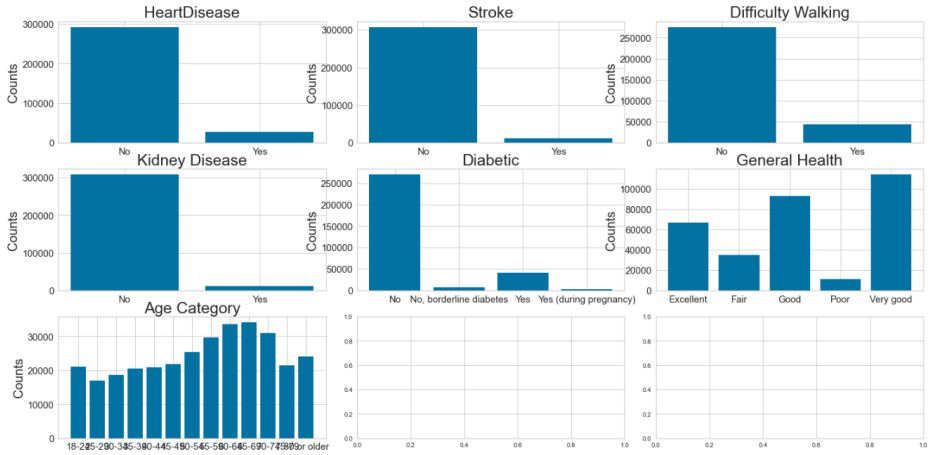
- The numerical feature most highly correlated with 'HeartDisease' is 'PhysicalHealth'
- There is not significant collinearity between the features (Rule of thumb: If the correlation > 0.8 then severe multicollinearity may be present)

- **Categorial Data**
  - **Spearman's Correlation Matrix:** Created to find the features most correlated to the target variable, 'HeartDisease' and observations were captured:
    - The categorical features most highly correlated with 'HeartDisease' are 'Stroke_Yes', 'DiffWalking_Yes', 'AgeCategory_80 or older', 'Diabetic_Yes', 'GenHealth_Fair', 'GenHealth_Poor', 'KidneyDisease_Yes'
    - There is not significant collinearity between the features (Rule of thumb: If the correlation > 0.8 then severe multicollinearity may be present)

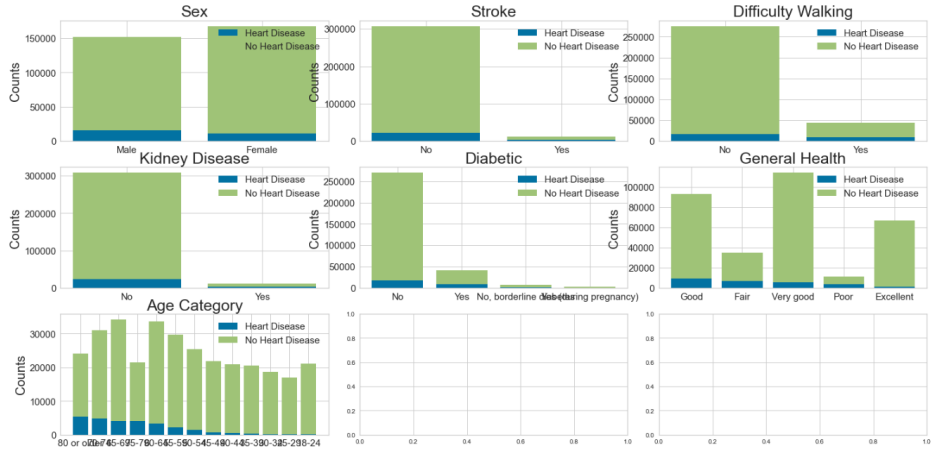
- 

	HeartDisease	BMI	PhysicalHealth	MentalHealth	SleepTime	Smoking_No	Smoking_Yes	AlcoholDrinking_No	AlcoholDrinking_Yes	Stroke_No	Stroke_Yes	DiffWalking_No	DiffWalking_Yes	Sex_Female	Sex_Male	AgeCategory_18-24	AgeCategory_25-29	AgeCategory_30-34	A
HeartDisease	1.000000	0.057355	0.143084	-0.003444	0.007521	-0.107764	0.107764	0.032080	-0.032080	-0.196835	0.196835	-0.201258	0.201258	-0.070040	0.070040	-0.073385	-0.065759	-0.065611	
BMI	0.057355	1.000000	0.093134	0.024011	-0.064533	-0.028788	0.028788	0.009295	-0.009295	-0.021597	0.021597	-0.154506	0.154506	-0.064332	0.064332	-0.125673	-0.034811	-0.004537	
PhysicalHealth	0.143064	0.093134	1.000000	0.276983	-0.079916	-0.092659	0.092659	0.012853	-0.012853	-0.116252	0.116252	-0.355052	0.355052	0.062744	-0.062744	-0.024341	-0.027438	-0.025344	
MentalHealth	-0.003444	0.024011	0.276983	1.000000	-0.132638	-0.050658	0.050658	-0.056523	0.056523	-0.026438	0.026438	-0.107651	0.107651	0.137340	-0.137340	0.119557	0.084244	0.068923	
SleepTime	0.007521	-0.064533	-0.079916	-0.132638	1.000000	0.035726	-0.035726	0.003927	-0.003927	-0.009227	0.009227	0.032894	-0.032894	0.017853	-0.017853	0.018542	-0.019822	-0.041772	
Smoking_No	-0.107764	-0.028788	-0.092659	-0.050658	0.035726	1.000000	-1.000000	0.111768	-0.111768	0.061226	-0.061226	0.120074	-0.120074	0.085052	-0.085052	0.138397	0.052149	0.015226	
Smoking_Yes	0.107764	0.028788	0.092659	0.050658	0.035726	-1.000000	1.000000	-0.111768	0.111768	-0.061226	0.061226	-0.120074	0.120074	-0.085052	0.085052	-0.138397	-0.052149	-0.015226	
AlcoholDrinking_No	0.032080	0.009295	0.012853	-0.056523	0.003927	0.111768	-0.111768	1.000000	-1.000000	-0.019858	0.019858	-0.035328	0.035328	0.004200	-0.004200	-0.004334	-0.023069	-0.015902	
AlcoholDrinking_Yes	-0.032080	-0.009295	-0.012853	0.056523	-0.003927	-0.111768	0.111768	-1.000000	1.000000	0.019858	-0.019858	0.035328	-0.035328	-0.004200	0.004200	0.004334	0.023069	0.015902	
Stroke_No	-0.196835	-0.021597	-0.116252	-0.026438	-0.009227	0.061226	-0.061226	-0.019858	0.019858	1.000000	-1.000000	0.174143	-0.174143	-0.003091	0.003091	0.048552	0.040268	0.040279	
Stroke_Yes	0.196835	0.021597	0.116252	0.026438	0.009227	-0.061226	0.061226	0.019858	-0.019858	-1.000000	1.000000	-0.174143	0.174143	0.003091	-0.003091	-0.048552	-0.040268	-0.040279	
DiffWalking_No	-0.201258	-0.154506	-0.355052	-0.107651	0.032894	0.120074	-0.120074	-0.035328	0.035328	0.174143	-0.174143	1.000000	-1.000000	-0.068860	0.068860	0.094715	0.080330	0.077637	
DiffWalking_Yes	0.201258	0.154506	0.355052	0.107651	-0.032894	-0.120074	0.120074	0.035328	-0.035328	-0.174143	0.174143	-1.000000	1.000000	0.068860	-0.068860	-0.094715	-0.080330	-0.077637	
Sex_Female	-0.070040	-0.064332	0.062744	0.137340	0.017853	0.085052	-0.085052	0.004200	-0.004200	-0.003091	0.003091	-0.068860	0.068860	1.000000	-1.000000	-0.040865	-0.026289	-0.018928	
Sex_Male	0.070040	0.064332	-0.062744	-0.137340	-0.017853	-0.085052	0.085052	-0.004200	0.004200	0.003091	-0.003091	0.068860	-0.068860	-1.000000	1.000000	0.040865	0.026289	0.018928	
AgeCategory_18-24	-0.073385	-0.125673	-0.024341	0.119557	0.018542	0.138397	-0.138397	-0.004334	0.004334	0.048552	-0.048552	0.094715	-0.094715	-0.040865	0.040865	1.000000	-0.062831	-0.046275	
AgeCategory_25-29	-0.065759	-0.034811	-0.027438	0.084244	-0.019822	0.052149	-0.052149	-0.023069	0.023069	0.040268	-0.040268	0.080330	-0.080330	-0.026289	0.026289	-0.062831	1.000000	-0.059056	
AgeCategory_30-34	-0.065611	-0.004537	-0.025344	0.068923	-0.041772	0.015226	-0.015226	-0.015902	0.015902	0.040279	-0.040279	0.077637	-0.077637	-0.018928	0.018928	-0.046275	-0.059056	1.000000	
AgeCategory_35-39	-0.066685	0.014903	-0.023281	0.056060	-0.049279	-0.004290	0.004290	-0.021545	0.021545	0.038985	-0.038985	0.069503	-0.069503	-0.008302	0.008302	-0.069586	-0.052006	-0.046406	
AgeCategory_40-44	-0.059196	0.032061	-0.019066	0.036191	-0.044768	-0.010680	0.010680	-0.018818	0.018818	0.033103	-0.033103	0.058260	-0.058260	-0.001780	0.001780	-0.070408	-0.062738	-0.066178	
AgeCategory_45-49	-0.049733	0.047434	-0.008614	0.026357	-0.040506	0.009657	-0.009657	-0.009857	0.009857	0.025682	-0.025682	0.040195	-0.040195	0.000091	-0.000091	-0.071806	-0.063984	-0.067492	
AgeCategory_50-54	-0.032648	0.053273	0.006397	0.013537	-0.039621	0.011667	-0.011667	-0.010588	0.010588	0.016567	-0.016567	0.013943	-0.013943	0.000380	-0.000380	-0.077968	-0.069475	-0.073284	
AgeCategory_55-59	-0.013276	0.043527	0.019442	0.000321	-0.033470	-0.009701	0.009701	-0.008189	0.008189	0.001244	-0.001244	-0.016113	0.016113	-0.002744	0.002744	-0.085055	-0.075790	-0.079945	
AgeCategory_60-64	0.016152	0.033760	0.027004	-0.029892	-0.010693	-0.031892	0.031892	-0.001621	0.001621	-0.011316	0.011316	-0.038116	0.038116	-0.002120	0.002120	-0.091115	-0.081190	-0.085641	
AgeCategory_65-69	0.042626	0.027822	0.011022	-0.062354	0.031262	-0.030367	0.030367	0.009026	-0.009026	-0.022110	0.022110	-0.037431	0.037431	0.002596	-0.002596	-0.091816	-0.081815	-0.086300	
AgeCategory_70-74	0.082578	0.004235	0.011868	-0.076074	0.054879	-0.045288	0.045288	0.021730	-0.021730	-0.039817	0.039817	-0.057190	0.057190	0.016669	-0.016669	-0.087100	-0.077613	-0.081868	
AgeCategory_75-79	0.098690	-0.021306	0.013039	-0.078905	0.063770	-0.048040	0.048040	0.027861	-0.027861	-0.058948	0.058948	-0.074948	0.074948	0.022959	-0.022959	-0.071258	-0.063496	-0.066977	
AgeCategory_80 or older	0.143041	-0.095123	0.018517	-0.102625	0.094476	-0.013569	0.013569	0.045226	-0.045226	-0.086099	0.086099	-0.156933	0.156933	0.048396	-0.048396	-0.075898	-0.067631	-0.071339	
Race_American Indian/Alaskan Native	0.008547	0.026708	0.019108	0.014166	-0.006002	-0.033667	0.033667	0.004243	-0.004243	-0.014357	0.014357	-0.023205	0.023205	0.003434	-0.003434	0.002726	0.000397	0.006306	

- Observations
  - The categorical features most highly correlated with 'HeartDisease' are 'Stroke_Yes', 'DiffWalking_Yes', 'AgeCategory_80 or older', 'Diabetic_Yes', 'GenHealth_Fair', 'GenHealth_Poor', 'KidneyDisease_Yes'
  - There is not significant collinearity between the features (Rule of thumb: If the correlation > 0.8 then severe multicollinearity may be present)
- **Histogram:** Created for only the most correlated features to 'HeartDisease' ('Stroke, 'DiffWalking', 'AgeCategory', 'Diabetic, 'GenHealth' 'KidneyDisease') to show their distributions and observations were captured:
  - Most people in the dataset do not develop heart disease
  - Most people in the dataset do not develop stroke
  - Most people in the dataset do not have difficulty walking
  - Most people in the dataset do not have kidney disease
  - Most people in the dataset do not have diabetes, pre-diabetes or gestational diabetes
  - Most people in the dataset have Excellent, Good or Fair general health
  - The dataset age categories span from the lowest at 16955 (ages 25-29) and highest at 34151 (ages 65-59)



- **Stacked Bar Chart:** Created to compare heart disease and no heart disease by the most correlated features found in the Spearman's Correlation Matrix, and observations were captured:
  - 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



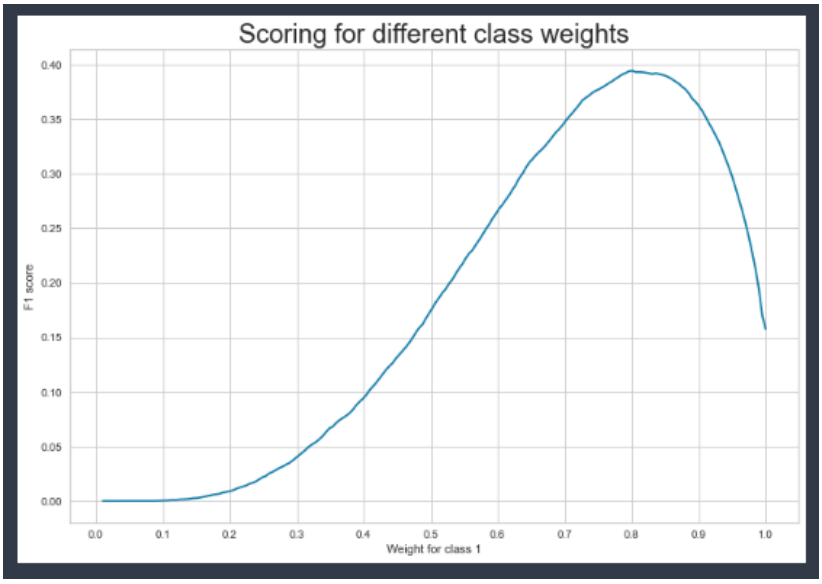
- **Train and Test Sets**
  - **Train and Test Sets:** To support the training and testing of the models, the dataset was split data into training (80%) and test (20%) sets with Heart Disease feature (binary Yes/No converted to 1/0) as the target
  - **Missing Data:** To ensure that any missing data is captured and then managed, isna() was applied to the training and test sets. It was found that there is no missing data in the dataset.
  - **Dummy Variables:** To ensure the data can be processed by the models, dummy variables were created in the dataframe for the categorical data which increased the features from 17 to 37 due to Yes/No options for categorical features.
  - **5 Best Features:** To find the best 5 features, X2 was used, and additional train & test sets were created using the chi2 function.
    - The 5 best features from X2 are as follows: PhysicalHealth, Stroke_Yes, DiffWalking_Yes, Diabetic_Yes, GenHealth_Poor

- **PCA Features:** To find the best 5 features, PCA was used and additional train & test sets were created using the PCA function. .95 was used for the number of components parameter. This means that scikit-learn choose the minimum number of principal components such that 95% of the variance is retained.
    - Of the 37 features with dummy variables, 34 features were identified using PCA

- **Automatic & Manual Class Balancing of the Target (HeartDisease)**
    - **Class Balancing:** "A classification problem in machine learning is where we have given some input (independent variables), and we have to predict a discrete target. It is highly possible that the distribution of discrete values will be very different. Due to this difference in each class, the algorithms tend to get biased towards the majority values present and don't perform well on the minority values. This difference in class frequencies affects the overall predictability of the model." (Singh, 2021). In this dataset, the distribution of the target variable HeartDisease is very different (ten times as many with HeartDisease_No, count of 58497, compared to HeartDisease_Yes, count of 5462)
    - To remedy this class imbalance, two methods were used: automatic and manual class balancing. In automatic class balance logistic model "We have added the class_weight parameter to our logistic regression algorithm and the value we have passed is 'balanced'." (Singh, 2021). In manual class balance logistic regression "we are trying to find optimal weights with the highest score using grid search. We will search for weights between 0 to 1. The idea is, if we are giving n as the weight for the minority class, the majority class will get 1-n as the weights. Here, the

magnitude of the weights is not very large but the ratio of weights between majority and minority class will be very high." (Singh, 2021).

- In this dataset, it was found through the graph below that the highest value for the minority class is peaking at about 0.81 class weight. Therefore, the minority class is 0.19. (class_weight={0: 0.19, 1: 0.81}).



- **Automatic class balance** will be implemented into the logistic regression models using solver='newton-cg', class_weight='balanced'
- **Manual class balance** will be implemented into the logistic regression models using solver='newton-cg', class_weight={0: 0.19, 1: 0.81}

Results

- **Models:** A total of 2 Decision Tree and 7 Logistic Regression models were created to find the model with the best performance using the heart disease dataset. The Decision Tree Classifier used all 37 features and 5 best features using chi2 (X2), and the Logistic Regression models used a combination of all 37 features, 5 best features from X2, and PCA in addition to no weighting of

the target class, automatic weighting of the target class, and manual weighting of the target class. The precision and recall were also calculated, but because the F1 score is a combination of these two, the F1 score is the metric used below.

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	87.01%	0.380	0.927	0.83

- Manual class weighting of target variable with PCA (34 features)				
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

- **Model Findings**
    - **Decision Tree Classifier:** The Decision Tree models with all 37 features and the 5 best features should not be used to find the features that impact heart disease the most as its AUC score was less than that of the Logistic Regression Models, and overall, not much better than random chance (0.5)
    - **Logistic Regression Models:** 3 logistic regression models were created to find the best model performance *without class balancing*
        - All 3 logistic regression models created without class balancing (all 37 features, PCA with 34 features, and 5 best features from X2) had relatively the same performance (91% accuracy, 0.1 F1 score for minority class (Yes - heart disease), 0.95 F1 score for majority class (No – heart disease), and 0.7-0.8 AUC score.

- In all 3 regression models, the F1 score for minority class (Yes - heart disease) is low due to the class imbalance of the target variable (heart disease), and therefore the class should be balanced for the target variable, and the logistic regression model run again (see the next 4 models created below)

- **Logistic Regression Models:** 4 logistic regression models were created to find the best model performance *with class balancing*
  - In this dataset, the distribution of the target variable HeartDisease is very different (many more HeartDisease_No compared to HeartDisease_Yes), and therefore the previous 3 logistic regression models were biased towards the majority class (HeartDisease_No). This is shown in the previous 3 logistic regression model performance metrics where the F1 score for HeartDisease_No is > 0.9, and HeartDisease_Yes is < 0.2
  - All 4 logistic regression models created with class balancing (automatic with all 37 features, manual PCA with 34 features, manual with all 37 features, and manual with 5 best features from X2) had relatively the same performance (75-88% accuracy, 0.3-0.4 F1 score for minority class (Yes - heart disease), 0.84-0.93 F1 score for majority class (No – heart disease), and 0.7-0.85 AUC score.
  - In all 4 regression models with class balancing, the F1 score for minority class (Yes - heart disease) improved from the 3 regression models without class balancing. These performance metrics prove that the class balancing efforts did what they needed to do in terms of improving the model performance, specifically F1 on the minority class.

- 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

Conclusion

Multiple Logistic Regression and Decision Tree Classifier models were used on the dataset to determine which features are most related to or most predictive of heart disease, where the HeartDisease feature was the target.

- **Matrices & Tests used to evaluate features in relation to the target variable, HeartDisease**

Matrices & Tests	Features
Pearson's & Spearman's Correlation Matrix - Feature correlation to Heart Disease target variable	• **Pearson's Correlation Matrix (numeric features)** o 'PhysicalHealth' • **Spearman's Correlation Matrix (categorical features)** o 'Stroke_Yes' o 'DiffWalking_Yes' o 'AgeCategory_80 or older' o 'Diabetic_Yes' o 'GenHealth_Fair' o 'GenHealth_Poor' o 'KidneyDisease_Yes'

Chi-Square (X2) Test **- 5 best features** **correlated to Heart** **Disease**	• PhysicalHealth • Stroke_Yes, • DiffWalking_Yes, • Diabetic_Yes • GenHealth_Poor

- **Findings:** The Pearson's & Spearman's Correlation Matrix features and the Chi-Squared (X2) features complement each other (with the exceptions of 'AgeCategory_80 or older', 'GenHealth_Poor' and 'KidneyDisease_Yes'), providing more evidence that the X2 5 best features are accurate
- **Recommendations:**
  - The logistic regression with class balancing using X2 5 best features had about the model performance as all 37 features in the same type of model (logistic regression with class balancing). 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)
  - When advising healthcare professionals and patients on what features are most related to or most predictive of heart disease, they are the 5 best features extrapolated from X2 with good logistic regression modeling performance:
    1. **PhysicalHealth:** Subject's report of how many days in the past 30 days they have exercised (integer)
    2. **Stroke_Yes:** Subject has been told or has known that they have had a stroke

3. **DiffWalking_Yes:** Subject's opinion is that they have difficulty walking or climbing stairs
4. **Diabetic_Yes:** Subject has had or currently has diabetes
5. **GenHealth_Poor:** Subject's opinion of their general health is poor

- 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)

- 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

- **Model Deployment**
    - The model has been trained and then the testing set used to obtain the model accuracy, precision, and F1 scores. Using the Logisitic Regression model that has been class weighted and X2 best features, the next step would be to do a real-life test with patient data from area hospitals.
    - The model 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 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.

References

1. CDC. (2022, February 7). *Heart disease facts*. Centers for Disease Control and Prevention. Retrieved March 21, 2022, from https://www.cdc.gov/heartdisease/facts.htm
2. Abbott, D. (2014). *Applied predictive analytics: Principles and techniques for the professional data analyst*. Wiley.
3. El Khouli, R. H., Macura, K. J., Barker, P. B., Habba, M. R., Jacobs, M. A., & Bluemke, D. A. (2009, November). Relationship of temporal resolution to diagnostic performance for dynamic contrast

enhanced MRI of the breast. Journal of magnetic resonance imaging : JMRI. Retrieved February 20, 2022, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2935260/

4. Grover, K. (2020, June 23). *Advantages and disadvantages of logistic regression*. OpenGenus IQ: Computing Expertise & Legacy. Retrieved March 21, 2022, from https://iq.opengenus.org/advantages-and-disadvantages-of-logistic-regression/
5. Mehta, S. (2022, March 16). *A beginner's guide to chi-square test in python from scratch*. Analytics India Magazine. Retrieved May 14, 2022, from https://analyticsindiamag.com/a-beginners-guide-to-chi-square-test-in-python-from-scratch/
6. Dixon, Chris. (2020). Re: What is the value of the area under the roc curve (AUC) to conclude that a classifier is excellent?. Retrieved from: https://www.researchgate.net/post/What-is-the-value-of-the-area-under-the-roc-curve-AUC-to-conclude-that-a-classifier-is-excellent/5eb02bdf39db6760dc56c904/citation/download.
7. B, H. N. (2020, June 1). *Confusion matrix, accuracy, precision, recall, F1 score*. Medium. Retrieved May 14, 2022, from https://medium.com/analytics-vidhya/confusion-matrix-accuracy-precision-recall-f1-score-ade299cf63cd#:~:text=F1%20Score%20becomes%201%20only,a%20better%20measure%20than%20accuracy.
8. Singh. (2021, January 6). *How to dealing with imbalanced classes in machine learning*. Analytics Vidhya. Retrieved May 15, 2022, from https://www.analyticsvidhya.com/blog/2020/10/improve-class-imbalance-class-weights/