Predicting Heart Disease with Personal Health Indicators

By: James Makanas

Background Information

- 2020 CDC Health Status Survey of 401,958 adults for the Behavioral Risk Factor Surveillance System (BRFSS)
  - Originally had 279 columns

- Dataset trimmed to 391,795 rows with 18 columns potentially related to heart disease

Data Description

- Dependent variable: Reported having coronary heart disease or myocardial infarction
  - ❑ Binary
  - ❑ Categorical

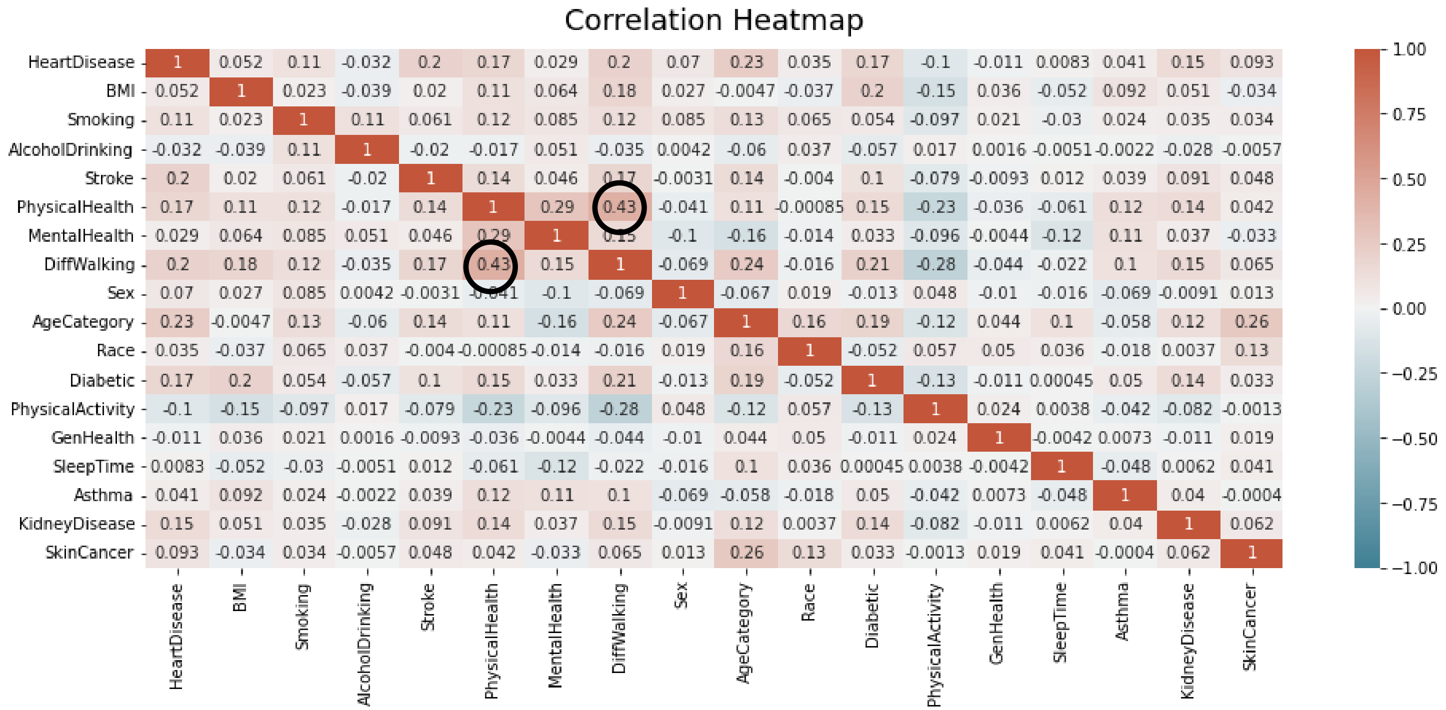
- Independent variables
  - ❑ BMI (continuous)
  - ❑ Smoking (categorical)
  - ❑ Alcohol Drinking (categorical)
  - ❑ Stroke (categorical)
  - ❑ Physical Health (continuous)
  - ❑ Mental Health (continuous)
  - ❑ Difficulty Walking (categorical)
  - ❑ Sex (categorical)
  - ❑ Age Category (categorical)
  - ❑ Diabetic (categorical)
  - ❑ Physical Activity (categorical)
  - ❑ General Health (categorical)
  - ❑ Sleep Time (continuous)
  - ❑ Asthma (categorical)
  - ❑ Kidney Disease (categorical)
  - ❑ Skin Cancer (categorical)

Primary Objectives

- Utilize the relevant health status indicators to create a binary classification model that predicts heart disease
- Go through the necessary preprocessing steps to prepare the data and then the model selection, tuning, and validation process

Methodology: Data Preprocessing

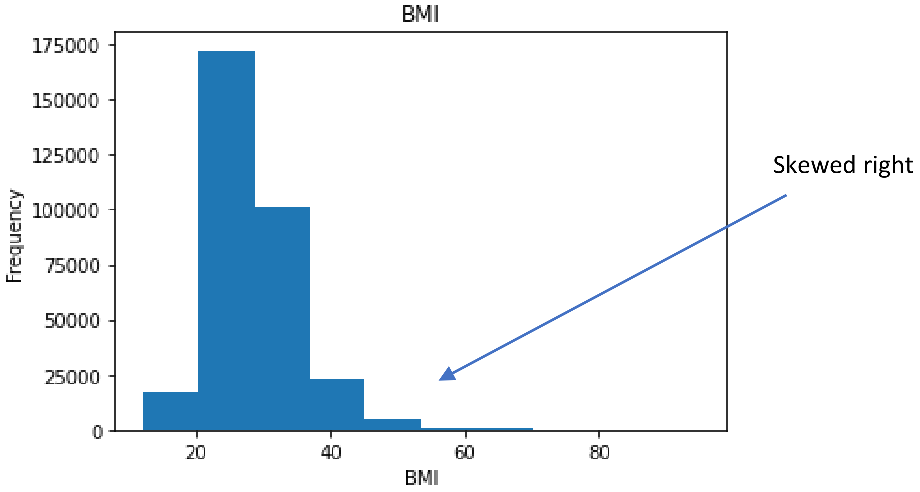
- No missing values
- No near-zero variance predictors out of numerical variables
- No collinearity
  - ❑ Highest correlation coefficient: .43



Methodology: Transforming Skewed Numerical Variables

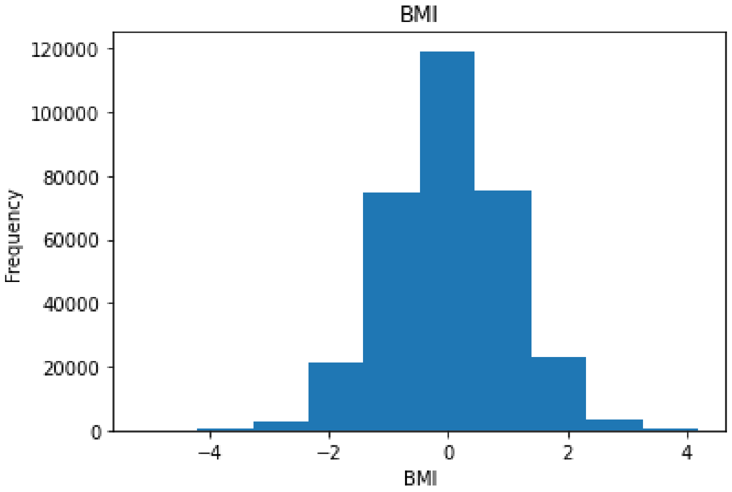
Before Yeo-Johnson

Variable	Skewness Coefficient
BMI:	1.33
PhysicalHealth:	2.60
MentalHealth:	2.33
SleepTime:	0.68



After Yeo-Johnson

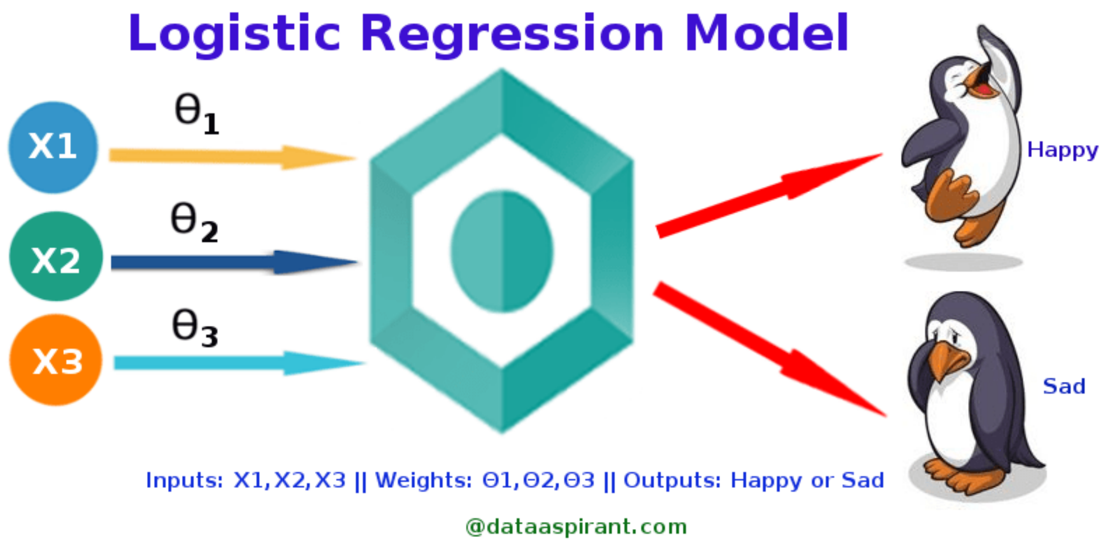
Variable	Skewness Coefficient
BMI:	-0.01
PhysicalHealth:	1.00
MentalHealth:	0.72
SleepTime:	0.68



Methodology: Model Selection

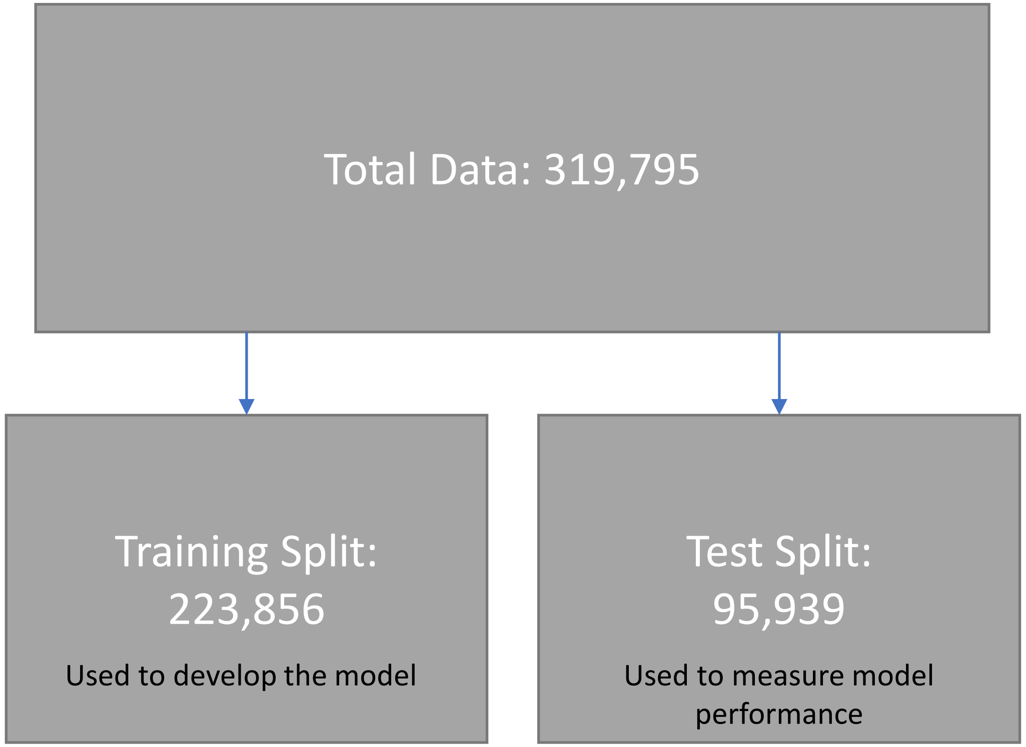
Model Type Selected: Logistic Regression

- Good for binary classification
  - ❑ 1: Heart disease present
  - ❑ 0: No heart disease present
- Estimates the relationship between the binary dependent variable and the independent features selected



(Source: http://dataaspirant.com/2017/03/02/how-logistic-regression-model-works/)

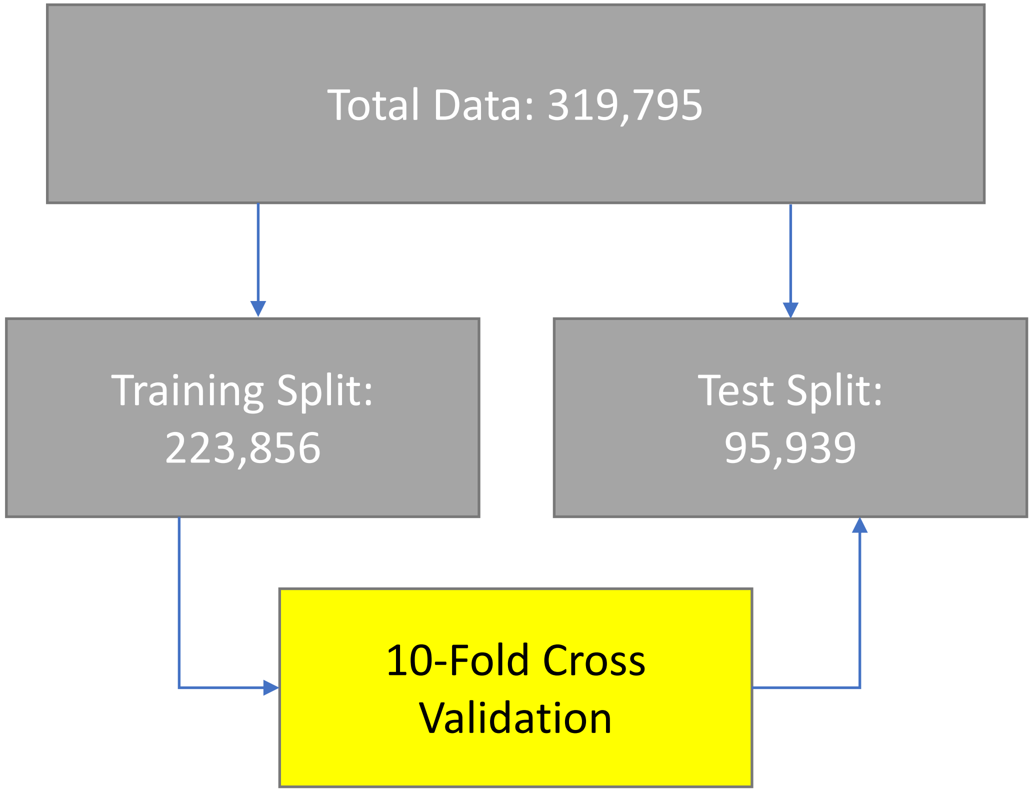
Methodology: Model Training and Validation



Data Split

- 70% of the data will be used within model development.
- 30% of the data will be used to measure model performance after the model was developed.
- Data split allows for a large sample for both training and testing. This also allows for additional model validation prior to evaluating model performance on test data.

Methodology: Model Training and Validation



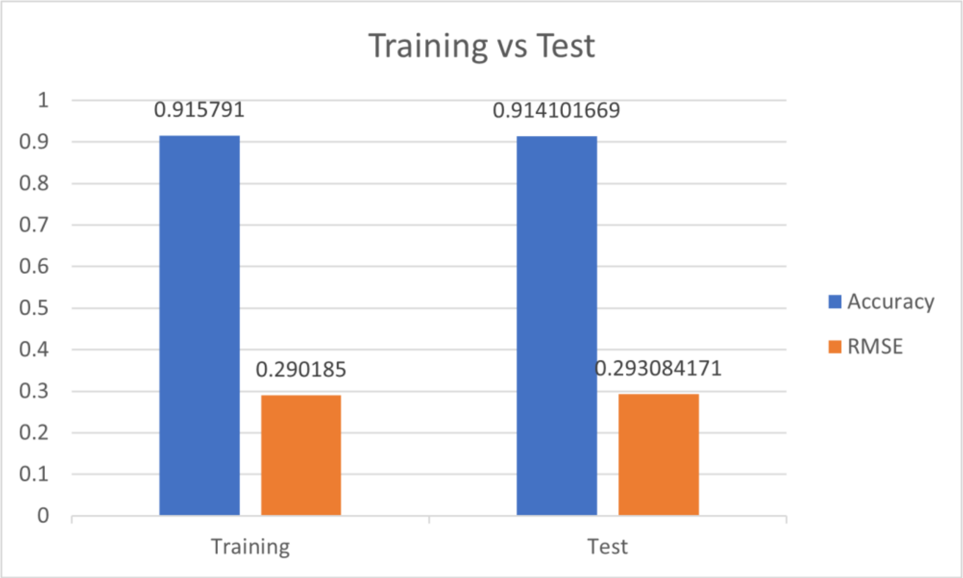
Model Validation

- Prior to testing, the model was validated with 10-fold cross validation on a portion of the training split not used to develop the model.
- Accuracy: 91.58%

Results: Model Performance

Model Validation

- After validating the model with a portion of the training data not used in developing the model, the model was then evaluated on the test split data.
- Training Model Validation
  - ❑ Accuracy: 91.58%
  - ❑ Root Mean Squared Error: .2901
- Test Model Performance
  - ❑ Accuracy: 91.41%
  - ❑ RMSE: .2931
- Strong performance in predicting different samples of the data shows that the model will perform well on brand new data



Potential Uses

- With this logistic regression model, we can potentially predict the presence of coronary heart disease or myocardial infarction at a large scale given that the people provide their health status indicators related to heart disease
- Health organizations
- At home use

References

https://www.kaggle.com/datasets/kamilpytlak/personal-key-indicators-of-heart-disease
https://dataaspirant.com/how-logistic-regression-model-works/