**Fall 2022 DSGA – 1001 Introduction to Data Science**

Capstone Project – Team Skyways

**Heart Disease - Health Indicators**

**Sri Vikas Prathanapu – sp6904** **Kaushik Tummalapalli – kt2651**

**Prathyusha Kadali – pk2669**

## 1. Introduction

Heart disease is a common chronic condition in the United States, affecting many people and costing a lot of money. Heart disease is the leading cause of death in the United States, causing about 647,000 deaths each year. It can be caused by the build-up of plaques in the larger coronary arteries, changes in molecules due to aging, chronic inflammation, high blood pressure, and diabetes, which factors can also increase the risk of heart disease.

Most people with coronary heart disease do not know they have it until they experience symptoms like chest pain, heart attack, or sudden cardiac arrest. This emphasizes the need for preventive measures and accurate tests to predict heart disease in the population to avoid any heart diseases.

The Centres for Disease Control and Prevention (CDC) considers high blood pressure, high blood cholesterol, and smoking to be major risk factors for heart disease, approximately half of Americans have at least one of these. The National Heart, Lung, and Blood Institute (NHLBI) lists a range of factors for clinicians to consider when diagnosing coronary heart disease, including age, environment and occupation, family history and genetics, lifestyle habits, other medical conditions, race or ethnicity, and sex. Diagnosis typically begins with an assessment of these common risk factors, followed by blood tests and other tests.

Although, we have many existing models for heart diseases prediction. I find this issue important, as we need to get more health conscious given the recent pandemic situation and we don't know what more to come, so analysing such data and bringing awareness among the people would make us prepare for the worst. So, we here try to investigate the statistical significance of our hypotheses, we conduct hypothesis testing. The goal of our study was to develop diabetes predictive models using 2015 BRFSS data and machine learning techniques of Logistic Regression, Decision Tree Classifier, Random Forest Classifier, XG Boost.

### Data Preparation

The dataset for this project was obtained from Kaggle (Heart Disease Health Indicators Dataset). This project's dataset contains 21 features and 253680 participants (rows of data). 1 binary target variable: heart disease or Attack and 21 feature variables that are either binary or ordinal.

We did a check and found that there are **No missing values.**

We handled **the outliers** using turkeys' method and pandas' functions. (Turkey's method is used to find the appropriate quantiles which tells about the maximum and the minimum value in the box plot.)

| | **Column name** | **Explanation** |
|---|---|---|
| 1 | HeartDiseaseorAttack | If value is<br>1 - Participant has or had a myocardial infarction or heart disease.<br>0 – Otherwise |

| | | |
|---|---|---|
| 2 | HighBP | High blood pressure and if value is<br>1 - high blood pressure<br>0 - low/normal blood pressure. |
| 3 | HighChol | High cholesterol and If value is<br>1 - High cholesterol<br>0 - Low/Normal cholesterol |
| 4 | CholCheck | Cholesterol check and If value is<br>1 - Participant had a cholesterol test within the last 5 years<br>0 – Participant did not had a cholesterol test within the last 5 years |
| 5 | BMI | Body mass index |
| 6 | Smoker | If value is<br>1 - participant has smoked at least 100 cigarettes in his entire life<br>0 - participant has smoked less than 100 cigarettes in his entire lifetime |
| 7 | Stroke | If value is<br>1 - participants had a stroke<br>0 - Otherwise |
| 8 | Diabetes | Stages of diabetes |
| 9 | PhysActivity | Physicial activity, If value is<br>1 - Participant's involvement in any kind of physical activity apart from a job in the past 30 days<br>0 - Otherwise |
| 10 | Fruits | If value is<br>1 – Participant consumes fruit consumption being 1 or more times per day<br>0 - Otherwise |
| 11 | Veggies | If value is<br>1 – participant consumes vegetable consumption being 1 or more times per day |
| 12 | HvyAlcoholConsump | Heavy alcohol consumption, if value is<br>1 - Heavy drinkers (adult men having more than 14 drinks per week and adult women having more than 7 drinks per week)<br>0 - non-heavy drinkers |
| 13 | AnyHealthCare | If value is:<br>1 - The participants having health care coverage or health care insurance or prepaid plans<br>0 - Otherwise1 |
| 14 | NoDocbcCost | No Doctor because of cost, If value is<br>1 - The participant had to see a doctor in the past 12 months but could not because of the cost<br>0 - Otherwise |
| 15 | GenHlth | Self-rating by the participant of his/her general health on a scale of 5. |
| 16 | MenHlth | The number of days in the past 30 days during which the individual experienced poor mental health (stress, depression, and other mental health issues) was measured. |
| 17 | PhysHlth | The number of days in the past 30 days during which the individual experienced poor physical health (injuries and other physical health issues) was measured. |
| 18 | DiffWalk | If value is<br>1 - participant has difficulty in climbing stairs and walking<br>0 - Otherwise |
| 19 | Sex | If value is<br>1 - Female<br>0 - Male |

| | | |
|---|---|---|
| 20 | Age | The value indicates age range<br>1 – 18-24, 2 – 25-29, 3 – 30-34, 4 – 35-39, 5 – 40-44, 6 – 45-49, 7 – 50-54, 8 – 55-59, 9 – 60-64, 10 - 65-69, 11 – 70-74, 12 – 75-79, 13 – 80 or older |
| 21 | Education | The value indicates<br>1 - A participant who has never attended school or only kindergarten<br>2 - A participant who has attended grade 1 through 8 (Elementary)<br>3 - who has attended grade 9 through 11 (Some high schools)<br>4 - grade 12 or GED (High school graduate)<br>5 - college 1 year to 3 years (Some college or technical school)<br>6 - college 4 years or more (College graduate) |
| 22 | Income | The value indicates income range<br>1 - less than $10,000<br>8 - $75,000 or more |

## 2. HYPOTHESIS TESTING

Here we want to analyse the relationship between various features and stroke especially. Most interesting features for us to consider are mental Health, NoDocbcCost and Healthy habits. Healthy habits column is not present in original dataset. So, we combined few features to derive this column. Performing this will help us in understanding few relationships which we consider important in real life.

**Hypothesis Test 1: Does Mental health effect the chances of stoke in an individual?**

For that we divided the dataset into two parts based on whether patient had stroke or not. Among the two groups we considered the mental health column for analysis. Below you can see the distributions of mental health in both the groups.

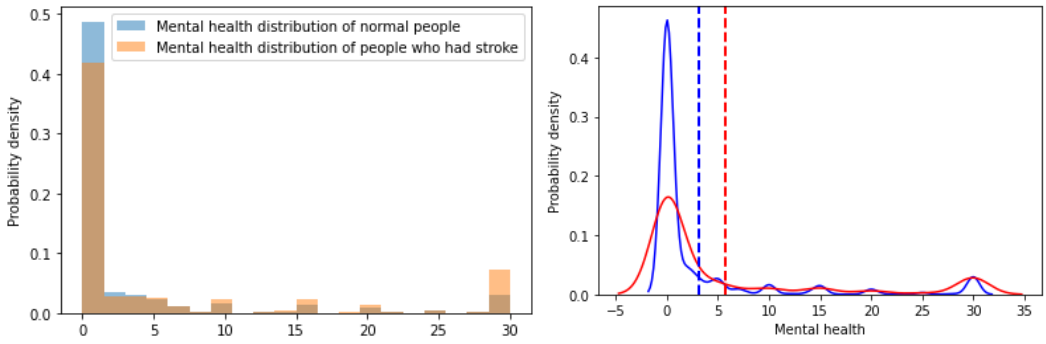



**Fig 0 and Fig 1**

Variance of the both groups are not similar and vary a lot. So, here we performed Welch's t-test is a statistical test. Here we use this test to know if there is any significant difference between the means of the both groups. Here we got p value of 2.85116793201112e-149.

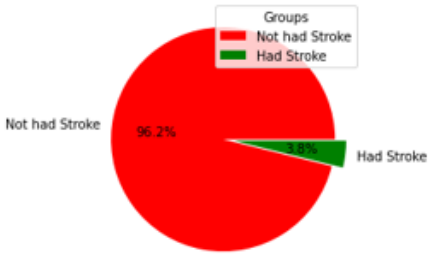
Here Null Hypothesis is “People with mental health issues in past 30 days are not highly prone to stroke”

Alternate Hypothesis: "People with mental health issues in past 30 days are highly prone to stroke"

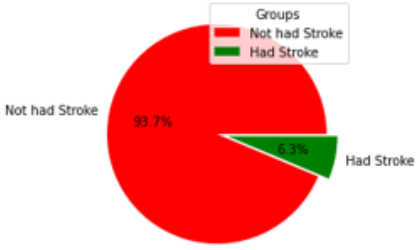
Here clearly $p < \alpha$. So, result is statistically significant and we reject our null hypothesis. So, we conclude stating that “People with mental health issues in past 30 days are highly prone to stroke”

**Hypothesis Test 2: Does people who had to see a doctor in the past 12 months but could not because of the cost are highly prone to strokes?**

Here we are considering columns stroke and NoDocbcCost. We need to perform chi square test to compare the observed and expected frequencies of the people based on two categories. Below you can see the ratio of stroke to no stroke for NoDocbcCost = 0 and NoDocbcCost = 1.




"NoDocbcCost"= 0




"NoDocbcCost"= 1

**Fig 2 and Fig 3**

For chi square test we got p value of 1.8750489229257827e-67.

Here Null Hypothesis is “The people who had to see a doctor in the past 12 months but could not because of the cost are not highly prone to strokes”

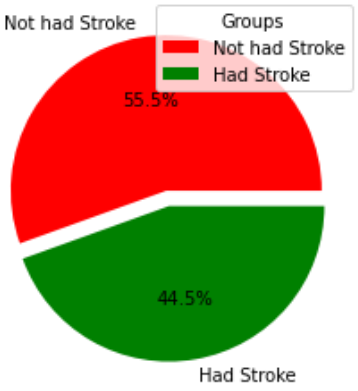
Alternate Hypothesis: " The people who had to see a doctor in the past 12 months but could not because of the cost are highly prone to strokes"

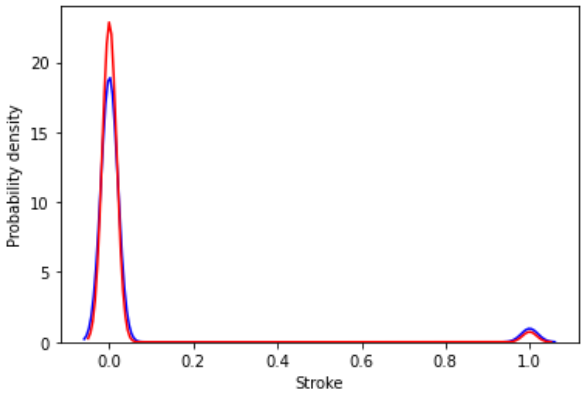
Here clearly $p < \alpha$. So, result is statistically significant and we reject our null hypothesis. So, we conclude stating that “The people who had to see a doctor in the past 12 months but could not because of the cost are highly prone to strokes”

**Hypothesis Test 3: Does peoples healthy life style like eating veggies, fruits regularly and participating in physical activity along with regular Cholesterol check effect the chances of stroke?**

Here we are creating new called healthy with combined values of columns 'CholCheck’, 'PhysActivity', 'Fruits', 'Veggies'. All these columns are numerical and has values 0 or 1. So we created new column value has 1 if people follow all above habits and those column values are 1s else 0.

After creating new column, we plotted chart for stroke to no stoke ratios in healthy category. It looks like below.




Probability density for strokes in people with healthy and not healthy life styles

**Fig 4 and Fig 5**

Here also we performed chi square test as both the columns are categorical and like above test. We got p values of 1.8426101845137378e-106.

Here Null Hypothesis is “The people with healthy life style considering given parameters like eating veggies, fruits, physical activity and cholesterol check are more prone to strokes”

Alternate Hypothesis: " The people with healthy life style considering given parameters like eating veggies, fruits, physical activity and cholesterol check are not more prone to strokes"

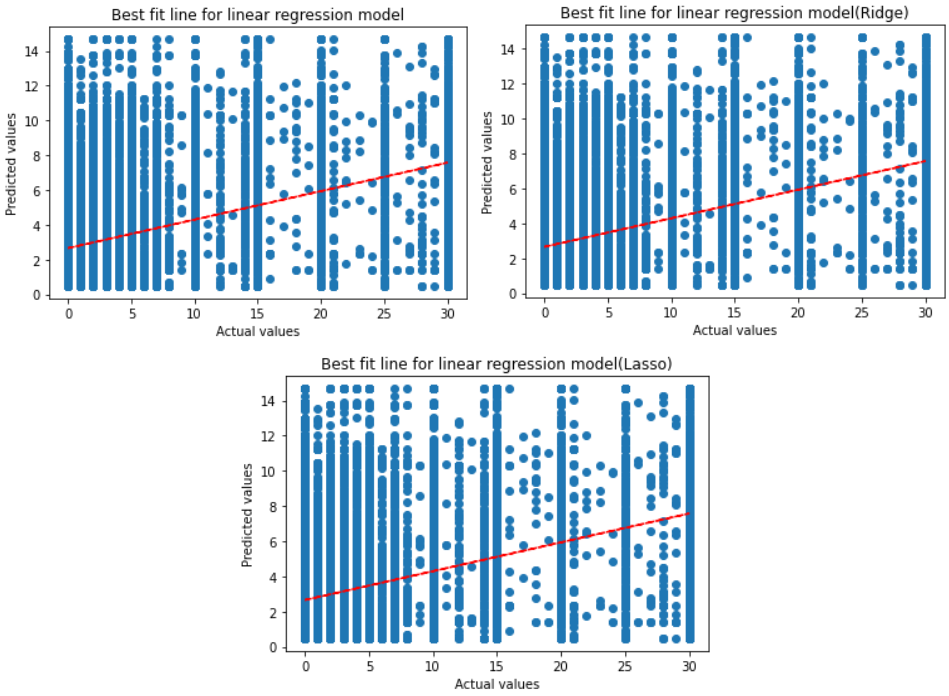
Here clearly $p < \alpha$. So, result is statistically significant and we reject our null hypothesis. So, we conclude stating that “The people with healthy life style considering given parameters like eating veggies, fruits, physical activity and cholesterol check are not more prone to strokes”

**3. Regression**

Here we used standard train test split of 80 to 20. First, we tried to figure out whether we can use few parameters from dataset to predict the mental health of a person. So, here we considered correlation among all features and choose four parameters with highest correlation. Chosen four features are "DiffWalk", "PhysHlth", "GenHlth", "NoDocbcCost". Then we used these features to predict the "MentHlth" of the patient. Luckily, we do not have any null values in our data and all categories are numerically encoded. After making sure that data is clean, we proceeded to regression.

Initially we performed multivariable regression model with above mentioned parameters to predict mental health. There we got R^2 score of 0.1610611046820506 and RMSE value of 6.770221659168112. Then we ran lasso and ridge regression for alpha values in the range of 1e-08 to 1e+08. Here we used gridSearchCV to choose best alpha value. For Ridge 'alpha': 21.412201548157228 and for lasso 'alpha': 1e-08. After running lasso and ridge regression for above alpha values R^2 and RMSE values very similar to the R^2 and RMSE values of regular regression.

But still regression models did not perform particularly well, as the R-squared values are relatively low and the RMSE values are relatively high. So, here we can say that chosen features are not good predictors of the dependent variable.



Best fit lines for all the three regression models.

**Fig 6, Fig 7, and Fig 8**

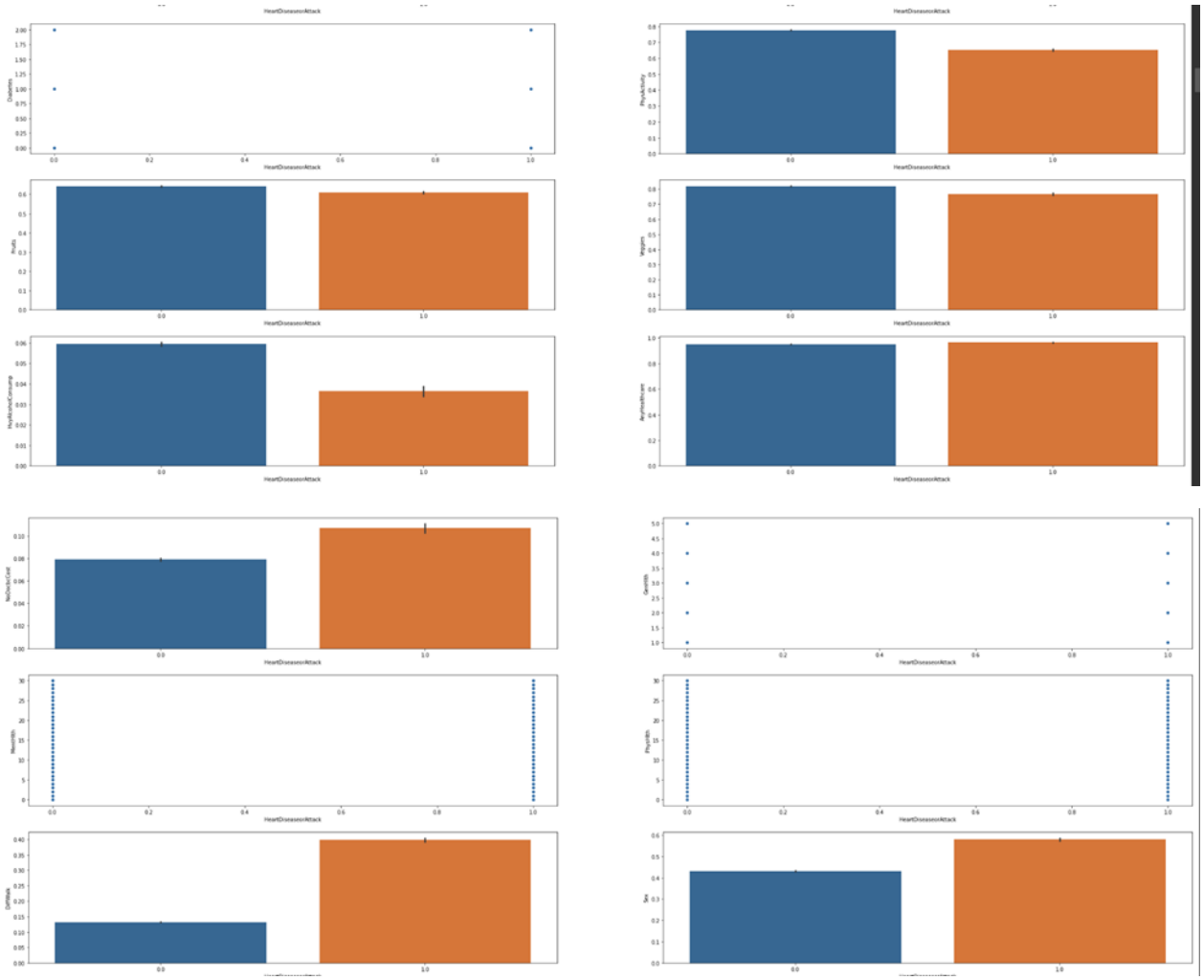
## 4. Classification

**How can we classify patients with a heart disease (Heart Disease Attack) with all the other parameters like HighBP, High Cholesterol, BMI, Smoker, Stroke, Diabetes etc?**

### Data Preparation Stage:

Data Preparation is one of the important stages in the complete data science cycle, as the data which is transformed, cleaned etc can be fit into the machine learning and deep learning algorithms to find out the relevant patterns. Some of the columns such as “BMI " have some extreme values also called the outliers, and can be controlled (removed), by using Tukey's method.

Tukey's method is used to find the appropriate quantiles which tells about the maximum and the minimum value in the box plot, and using pandas’ functions, all the outliers are controlled.



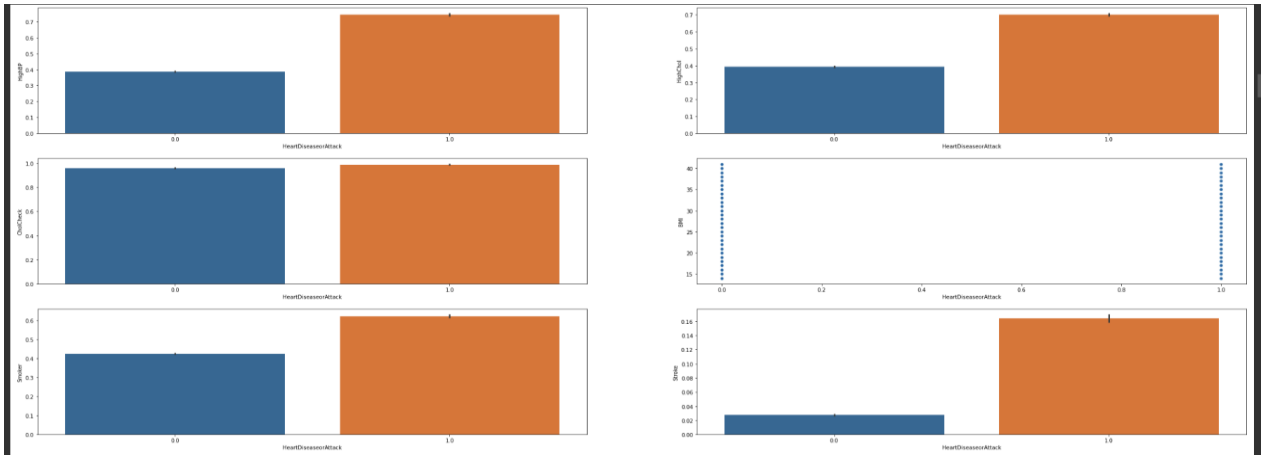


**Fig 9, Fig 10 and Fig 11**

As the data that we collected has some imbalance present to it, we will perform oversampling to improve the model performance i.e to improve the model generalization on future realistic data.

**Feature Engineering:**

To understand which feature is important, we will be performing various methods like the Chi-Square test, but this test failed to quantify the similarity between the features. We have also used to other methods like correlation matrix, where can understand how the columns i.e the features are correlated, i.e the Positive correlation and the negative correlation, and by combining the results of both correlation and the Chi-square test, we will be able to identify all the important features.

Features that are considered to for the model building are: ”HighBP”, “HighCol”, “Smoker”, “Stroke”, “Diabetes”, “GenHlth”, “PhysHlth”, “DiffWalk”, “Age”.

We will be splitting the data into two parts i.e Training data, Testing Data with a 80% - 20% split in the data set with the appropriate random state, where the training data is fed into the machine learning models and the test data is used to find out the scores and performance.

**Dimensionality Reduction:**

To get better results in data modeling, we will be using PCA method which is also called as Dimensionality Reduction Method, where we need to perform the Standard Scaling on the data before performing the PCA, as PCA converts the data from higher dimensionality to the lower dimensionality.

**Modeling and Evaluations:**

After performing Dimensionality reduction techniques, the results tend to be much better towards the data that was not performed PCA on, as it was examined on various Machine Learning algorithms like Logistic regression, Decision Tree Classifier, Random Forest Classifier, Boosting Algorithm(XGBoost). Except for the XGBoost algorithm, the data without PCA out performs the performance in all other algorithms when compared to the data that has been done PCA on.

These are the results with and without performing PCA:

**For Logistic Regression:**
(Without PCA)
Train Data Score: 0.18030755106724813
Test Data Score: 0.19105243857719106
Score: 0.9095289847642873

(With PCA)
Train Data Score: 0.39043362164044815
Test Data Score: 0.3973245175674637
Score: 0.8392560542990137

**For Decision Tree Classifier:**
(Without PCA)
Train Data Score: 0.3574281664733926
Test Data Score: 0.1858436766185844
Score: 0.900465478704861 9

(With PCA)
Train Data Score: 0.4879637938433927
Test Data Score: 0.36360794122244805
Score: 0.8330428363442491

**For Random Forest Classifier:**
(Without PCA)
Train Data Score: 0.38399143574751926
Test Data Score: 0.19842993151828964
Score: 0.9015932905448356

(With PCA)
Train Data Score: 0.48638580590652475
Test Data Score: 0.36600389863547755
Score: 0.8332683987122439

**For XG Boost:**
(Without PCA)
Train Data Score: 0.3465830405273941
Test Data Score: 0.34666787889808165
Score: 0.7052925133799496

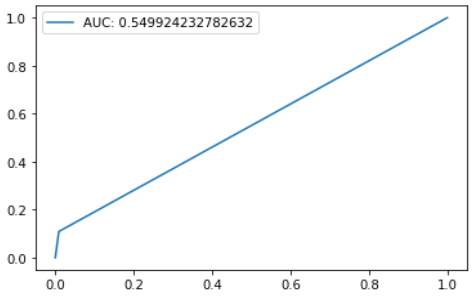
(With PCA)
Train Data Score: 0.48638580590652475
Test Data Score: 0.36600389863547755
Score: 0.833329915721697

**Fig 12**

**Hyper Parameter Tuning:** We perform this using GridSearchCV, which is used to improve the performance of the Machine learning models. AUC curves are used to represent the measurement of the classification model. Score metric is used to get an understanding on how well the model performs overall.

**For Logistic Regression,**
Beta: [[0.54559264 0.58305327 0.45105846 0.90408623 0.17786236 0.51232519
0.00208942 0.2210021 0.24481167]]
Params {'C': 0.01}

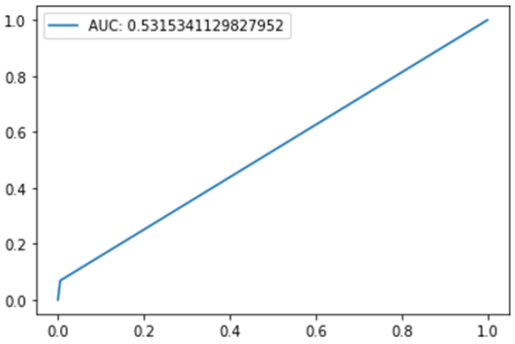
| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0.0 | 0.92 | 0.99 | 0.95 | 44254 |
| 1.0 | 0.55 | 0.11 | 0.18 | 4513 |
| accuracy | | | 0.91 | 48767 |
| macro avg | 0.73 | 0.55 | 0.57 | 48767 |
| weighted avg | 0.88 | 0.91 | 0.88 | 48767 |




For Decision Tree Classifier:
{'max_depth': 7, 'min_samples_leaf': 3, 'min_samples_split': 2}

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0.0 | 0.91 | 0.99 | 0.95 | 44254 |
| 1.0 | 0.56 | 0.07 | 0.12 | 4513 |
| accuracy | | | 0.91 | 48767 |
| macro avg | 0.74 | 0.53 | 0.54 | 48767 |
| weighted avg | 0.88 | 0.91 | 0.88 | 48767 |

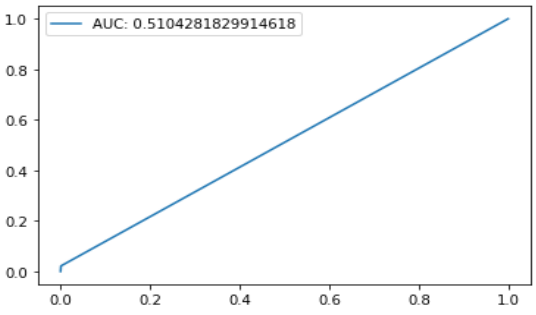
Score: 0.9088933089999385




For Random Forest Classifier:
{'max_depth': 5, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 10}

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0.0 | 0.91 | 1.00 | 0.95 | 44254 |
| 1.0 | 0.72 | 0.02 | 0.04 | 4513 |
| accuracy | | | 0.91 | 48767 |
| macro avg | 0.81 | 0.51 | 0.50 | 48767 |
| weighted avg | 0.89 | 0.91 | 0.87 | 48767 |

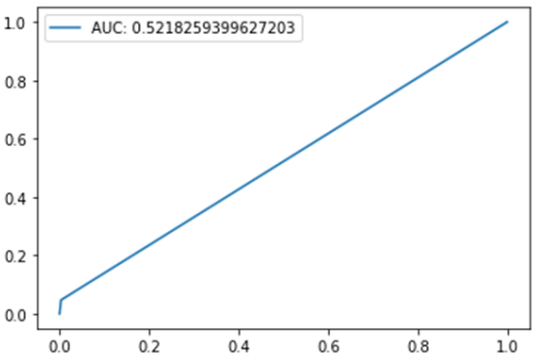
Score: 0.9086882523017614




For XGBoost:
{'learning_rate': 0.1, 'max_depth': 3, 'n_estimators': 100}

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0.0 | 0.91 | 0.99 | 0.95 | 44164 |
| 1.0 | 0.57 | 0.08 | 0.14 | 4603 |
| accuracy | | | 0.91 | 48767 |
| macro avg | 0.74 | 0.54 | 0.55 | 48767 |
| weighted avg | 0.88 | 0.91 | 0.87 | 48767 |

Score: 0.9072528554145222




We can see that out of all the algorithms, Logistic Regression performs better but it's minute difference in performance with all the other models with an accuracy of 0.9094. According to the AUC scores, we can see that the XGBoost Algorithm leads when compared with the other algorithms.

## 5. Summary and conclusions

In the United States alone, a person suffers from heart attack every 40 seconds and about 805,000 people across the year in the US, and most of these cases are first time users having a heart attack. Getting proper resources and developments in this field will help the patients immensely.

**Learnings from this project:**

This project helped us immensely to utilize all the methods and techniques we have learned from the CDS GA 1001 class, as we have iterated from the data collection, data pre-processing, inference, machine learning stage to find out the patterns and to solve this kind of use case, where we used techniques like cross validation, Hypothesis testing, Dimensionality reduction and then learning about pros and cons about all these tools in general.

**Improvements and Conclusions:**

- Getting the data of these data points (patients) historically will help the model to improve and generalize well over time, as this is considered as one of the important steps if we have access to this restricted data. Let's say, if the patient's cholesterol levels are less in this dataset, and periodically, if that patient acquires more cholesterol levels than the minimum, then that person might be treated urgently. So, getting this periodic data fed into this model will help in better generalization of the future data.
- Getting classified data in the "Heart Attack" will give us more insights about the data and could even more improve the model performance and can also reduce Type 2 Errors to some more extent.
- We would like to use deep learning models to improve the generalization of the data i.e., the performance of the model and new data predictions, as there has been outstanding developments on deep learning models and its research.

**EXTRA CREDIT**:

- Good thing about the dataset that we collected is that it contains all the real time data of all the patients, so that we will be able to generalize the model in a much better manner as we will be able to relate it to real world scenarios than generated datasets.
- Interesting point was the data set has 0 Null values, so we didn't spend any time on removing the null values, but we did handle out the other outliers and processing functions.

## 6. Author contributions

• Sri Vikas Prathanapu – Hypothesis Testing, Regression

• Kaushik Tummalapalli – Classification

• Prathyusha Kadali – Regression and Classification

## 7. References

**[1]** **https://www.kaggle.com/datasets/alexteboul/heart-disease-health-indicators-dataset**

**[2]** **https://www.kaggle.com/datasets/cdc/behavioral-risk-factor-surveillance-system**

**[3]** **https://www.cdc.gov/brfss/annual_data/annual_data.htm**

**[4]** **https://www.cdc.gov/brfss/annual_data/2015/pdf/codebook15_llcp.pdf**