CS5812 Predictive Data Analysis (A 2023/4)

Student Id: 2349128

Data Description & Research Question

The CDC (Centres for Disease Control and Prevention) runs Behavioural Risk Factor Surveillance System which performs annual surveys. Our dataset is derived from a telephonic survey for heart disease and is called “Indicators of heart disease”, it contains more than 400K rows and 40 columns. If you wanted to , you can access the dataset from the appendix. The dataset contains majority of categorical variables and few numerical variables. The description of each of those variables is stated in the data dictionary below:

Variable Name	**Description**	**Variable Type**
Race	The racial or ethnic background of the individual surveyed.	Multi-class (Categorical)
State	The geographic state where the individual resides in	Multi-class (Categorical)
SleepHours	The average number of hours of sleep per night reported by the individual.	Numerical
TetanusLast10Tdap	Whether the individual received a tetanus vaccination or Tdap vaccine within the last 10 years.	Multi-class (Categorical)
HIVTesting	Whether the individual has undergone testing for HIV.	Binary(Categorical)
ECigaretteUsage	Whether the individual uses electronic cigarettes (e-cigarettes).	Multi-class (Categorical)
HighRiskLastYear	Whether the individual engaged in behaviors considered high risk for health in the past year.	Binary(Categorical)
HeightInMeters	The height of the individual in meters.	Numerical
HadDepressiveDisorder	Whether the individual has been diagnosed with a depressive disorder.	Binary(Categorical)
HadAsthma	Whether the individual has been diagnosed with asthma.	Binary(Categorical)
CovidPos	Whether the individual has tested positive for COVID-19.	Binary(Categorical)
MentalHealthDays	The number of days in the past month the individual experienced poor mental health.	Numerical

BMI	Body Mass Index, a measure of body fat based on height and weight.	Numerical
WeightInKilograms	The weight of the individual in kilograms.	Numerical
FluVaxLast12	Whether the individual received a flu vaccine within the last 12 months.	Binary(Categorical)
HadSkinCancer	Whether the individual has been diagnosed with skin cancer.	Binary(Categorical)
DifficultyConcentrating	Whether the individual experiences difficulty concentrating.	Binary(Categorical)
LastCheckupTime	Time elapsed since the individual's last medical checkup.	Multi-class (Categorical)
BlindOrVisionDifficulty	Whether the individual experiences blindness or vision difficulties.	Binary(Categorical)
Sex	The gender of the individual.	Binary(Categorical)
AlcoholDrinkers	Whether the individual consumes alcohol.	Binary(Categorical)
SmokerStatus	The smoking status of the individual.	Multi-class (Categorical)
DifficultyDressingBathing	Whether the individual experiences difficulty dressing or bathing.	Binary(Categorical)
PhysicalActivities	The frequency of physical activity performed by the individual.	Binary(Categorical)
DifficultyErrands	Whether the individual experiences difficulty running errands.	Binary(Categorical)
DeafOrHardOfHearing	Whether the individual is deaf or hard of hearing.	Binary(Categorical)
HadKidneyDisease	Whether the individual has been diagnosed with kidney disease.	Binary(Categorical)
HadArthritis	Whether the individual has been diagnosed with arthritis.	Binary(Categorical)
PneumoVaxEver	Whether the individual has ever received a pneumococcal vaccine.	Binary(Categorical)
HadCOPD	Whether the individual has been diagnosed with Chronic Obstructive Pulmonary Disease (COPD).	Binary(Categorical)
PhysicalHealthDays	The number of days in the past month the individual experienced poor physical health.	Numerical

HadDiabetes	Whether the individual has been diagnosed with diabetes.	Multi-class (Categorical)
DifficultyWalking	Whether the individual experiences difficulty walking.	Binary(Categorical)
RemovedTeeth	Whether the individual has had teeth removed.	Multi-class (Categorical)
ChestScan	Whether the individual has undergone a chest scan.	Binary(Categorical)
AgeCategory	The age bracket the individual fits in.	Multi-class (Categorical)
HadStroke	Whether the individual has had a stroke.	Binary(Categorical)
GeneralHealth	The general health status of the individual.	Multi-class (Categorical)
HadAngina	Whether the individual has experienced angina (chest pain).	Binary(Categorical)
HadHeartAttack	Whether the individual has experienced a heart attack.	Binary(Categorical)

Research Question:

After looking at the columns in the dataset, we decided to make ‘HadHeartAttack’ as our target binary variable.

My Research Question: “How accurately can we predict the occurrence of heart attacks based on health indicators, lifestyle factors, and other characteristics in the dataset?”

By examining how different predictors affect the likelihood of a heart attack, this research will contribute valuable insights into the early identification of individuals at high risk. Additionally, we intend to get a comprehensive understanding of the relationships between many factors that lead to occurrences of cardiovascular events. We calculated the number of missing (null) values in each column to determine the extent and the distribution of the missing data.

Data Preparation & Cleaning

During the early phase of our project, we performed a data cleaning procedure to ensure the validity and dependability of the dataset.

We began by examining basic information about the dataset to understand the structure, types of data included, and the overall completeness of the dataset. This helped us identify missing values in some of the variables.

Before doing any data manipulation, we checked for the proportion of rows which had 'HadHeartAttack' as 'Yes' vs the rows which had 'HadHeartAttack' as 'No'. This led us to the fact that the dataset was highly imbalanced and contain up to 95% of no heart attack values and only 5% of the rows were marked as yes for 'HadHeartAttack'. As we already had abundant no of rows, we decided to drop all the null values and check the proportion again. We found that the proportion of Nos and Yeses in 'HadHeartAttack' remained similar(96:4). Even after dropping all the Nans we were left with around 246K rows which was still plenty for our needs.

At this point, we knew that the majority of the variables are binary (categorical) so we decided to convert the Yeses to 1 and Nos to 0, basically converting them into proper binary variables. This would make it much easier for further processing and will be extremely useful in model building for various machine learning and deep learning algorithms. Some of the variables like 'HadDiabetes', 'TetanusLast10Tdap' and 'CovidPos' were multi class but the values weren't that impactful and would server much better as binary values.

Then we took all the ordinal categorical variables and mapped them with integers. The variables GeneralHealth & AgeCategory were mapped as 1,2,3.. and so o. SmokerStatus & ECigaretteUsage were mapped as 0(representing non-smokers),1,2,3. LastCheckupTime was mapped as 0,1,2 and so on where the higher the number the recent was their last checkup. RemovedTeeth was mapped as 0,1,2 and so on where the higher the number the higher the number of teeth were removed. We also encoded the nominal variables like State and Race to integers. This makes our data highly compatible with numerous models. You can access the code for the steps mentioned here in the appendix.

Exploratory Data Analysis

Exploratory Data Analysis (EDA) is an essential stage in the data analysis process, and it involves examining and understanding the structure, patterns, and characteristics of a dataset.

As our data is to big for some of these visualizations, we sometimes sample 10% of the data, so that we could get the results/visualizations in decent amount of time.

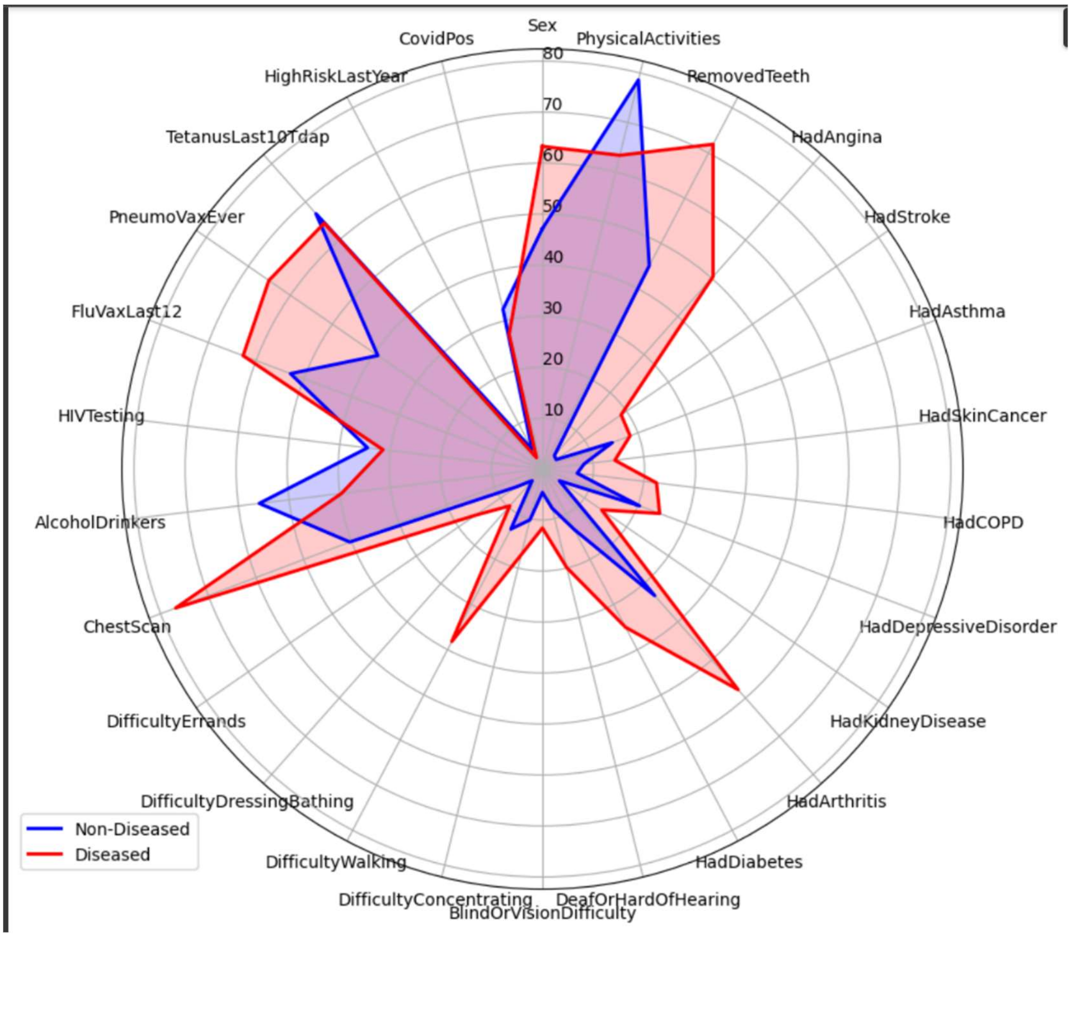
Using unsupervised learning such as clustering would be great way to find some patterns in relationships between some features.So, we decided to use K-means clustering on various features. First,We used oversampler to oversample the data make sure we have a dataframe with a balanced class then we use only the numerical variables & 'HadHeartAttack' to generate various scatterplots. With the visualizations, we find some patterns like BMI and WeightInKilograms have a linear relationship. But we don't find much interesting patterns except for some clusters forming in BMI vs HeightInMeters.

We also build a K-means clustering for combination of every numerical variable and we find that the scatter is similar but the clusters formed are different. The higher values in the BMI, HeightInMeters and WeightInKilograms form a different cluster. Though not all clusters seem meaningful.

So, we tried to discover the relation between BMI & AgeCategory. We used DBScan, as we had no clue how many clusters we needed and this method works well for such cases. The DBScan Clustering found 2 clusters and

We built few different bar charts to understand the distribution of class values for the categorical variables that we thought were interesting such as Sex.We see that there are higher number of men who have had heart attack than women. Also the people whose physical activities were on the lower end also have had more heart attacks then people who have had good level of physical activity in their life. You can look at the visualizations mentioned in the appendix

As we are aware of the presence of high numbers of binary variables, we think that the radar graph makes a good visualization to check their frequencies depending on the target variable 'HadHeartAttack'. So, first we do some mapping and change the categorical to binary as well just for the graph. Then we generated this radar graph :



Machine Learning

For the predictive modeling part, I have to focus on the target variable hadHeartAttack as my research question focuses on heart attacks. The target variable is binary in nature. I chose the Random Forest algorithm as the machine learning approach. Random Forest offers several advantages that align well with the dataset and my research question. As Random Forest is known for its robustness against overfitting, as instead of using a single decision tree, it combines multiple.

I went on and split the data into Training & Testing data (80:20 split). So, that we can check if the models work as good when its new data (data the model hasn't seen before). Then I applied a simple random forest model and I experimented with the amount of trees and found that n_estimators as 50 worked the best and took around 10 minutes for training. Increasing the amount didn't do much for the scores.

Now after running the model, we do its performance evaluation. Now my research question just focuses on accuracy of the model, but because of the high imbalance in the dataset, most of the training is done on values with Target variable 'HadHeartAttack' as 0. This is the reason why looking at accuracy might not be optimal. So we can also look at the f-1 score we get, as it already is a combination of precision and recall. We get 0.28 as our F1-score which signifies that our model isn't performing that well especially on the 1s.

AUC: 0.587

Classification Report:

```
      precision   recall  f1-score   support

   0     0.96     0.99     0.97    46573

   1     0.57     0.18     0.28     2632
```

As we are aware of the high imbalance in the dataset, I decided to try K-fold cross validation. This method ensures that each class is represented in both the training and validation sets across multiple iterations. This helps prevent bias towards the majority class. But even after using the k-fold cross validation and averaging out the precision, recall and f1-scores over the 5 splits(I used k=5). I found out that the results were the same as my simple base random forest model.

Accuracy: 0.9479, Precision: 0.5730, Recall: 0.1807, F1-score: 0.2747,AUC Score: 0.5865

Even the AUC we get for both of them is the same.

Deep Learning

For Deep Learning, I decided to simply use Deep Neural Networks as they are simple to implement and the good thing about them is they learn the patterns on their own. The only downside is the higher computational resource requirement which is doable in this day & age.

I do the train-test split again(80-20). I make a Neural Network with 3 hidden layers, all 3 layer are RELU (Rectified Linear Unit). The first layer has 64 units , the second one has 32 units and the third one has 16 units and then we end with the final layer with 1 unit using Sigmoid activation function. Using sigmoid activation, we can convert the values we get into either 0 or 1.

For optimizers we use 'adam' optimizer, for loss functions we use binary_crossentropy as we are performing binary classification. With some experimentation, I find 20 epochs and batch size 64 to be in balance with the time the neural network takes for learning. Then we look at the classification report for our performance evaluation

Accuracy: 0.946

ROC AUC Score: 0.6146666699733471

Classification Report:

```
      precision   recall  f1-score   support

  0      0.96      0.99      0.97     46573

  1      0.50      0.24      0.33      2632
```

Here we see the F1-score is still not that great but 0.33 is still an improvement from the random forest model. The Neural Network didn't take much time to get through the epochs. The runtime was around 5 mins. This was because of the low no of epochs and small batch size. But we can notice that our AUC score is higher than the ML model, which is an improvement worth noting. You can access the deep learning model mentioned from the appendix.

Performance Evaluation & Comparison of Methods

Firstly, we look at members of the group and their selected models.

For Machine Learning,

[R.M.] Rahil (Me) – Random Forest,

[D.P.] Dilan– Logistic Regression,

[H.H] Hamza – Decision Tree,

[O.K] Omar – Support Vector Machines (SVM),

[F.A.] Faisal – K-Nearest Neighbour (KNN)

Model	Performance
Rahil’s Random Forest	Acc: 0.95 F1: 0.28 AUC: 0.58
Dilan’s Logistic Regression	Acc: 0.95 F1: 0.34 AUC: 0.62
Hamza’s Decision Tree	Acc: 0.91 F1: 0.27 AUC:0.624
Omar’s SVM	Acc: 0.83 F1: 0.35
Faisal’s KNN	Acc: 0.85 F1: 0.35 AUC: 0.88

Model	Performance
Rahil’s Deep Neural Network	Acc: 0.95 F1: 0.36 AUC: 0.627
Dilan’s Convolutional Neural Network	Acc: 0.95 F1: 0.32 AUC: 0.60
Hamza’s Convolutional Neural Network	Acc: 0.95 F1: 0.20 AUC:0.55
Omar’s Deep Neural Network	Acc: 0.82 F1: 0.31 AUC: 0.91
Faisal’s Transfer Learning + LeNet Convolutional Neural Network	Acc: 0.95 F1: 0.24 AUC: 0.88

After all the model building, we all share the results our respective models performed and compare it. The Metrics everybody commonly used was the classification report as it included precision, accuracy, f1-score and the accuracy. This covered various aspects for us to evaluate the model.

I chose and implemented the random forest model for my machine learning approach and with or without K-Fold Cross validation I got the same results.

When compared to others’ model, we notice that the scores for the class 0 is always high 0.96 and above except for Omar’s SVM model which had 0.84 recall. However, for Class 1 everybody’s model is not performing well, this is all due to the aforementioned data imbalance. This also means that accuracy won’t be that good of a metric to look at. Even then, the lowest accuracy we see Is in Faisal’s KNN and Omar’s SVM model (84%).

Even then, Faisal’s KNN model and Omar’s SVM model had the highest F1 Score of 0.35.

If we focus on some other metric like AUC score(shows how good the classifier is in differentiating between positive and negative classes) then we can see that my random forest model is performing the poorest with 0.58 AUC score, meanwhile Dilan’s Logistic Regression model performs better at 0.62 along with Hamza’s Decision Tree at 0.624. This is ironic as random forests is supposed to be using multiple decision trees. While Faisal’s KNN model performs the best in AUC score at 0.88.

Now we look at the Deep Learning models,

Everybody’s models performing at the same accuracy of 95% except for Omar’s DNN which is at 82%. When we look at the F1 score, my Deep Neural Network has the highest F1 score of 0.36.However, if you look at the AUC, Omar’s Deep Neural Network is performing much better at 0.91and even Faisal’s CNN performing pretty well at 0.88 AUC score.

Discussion of Findings

One of the most the crucial thing that I learnt from this project is that data imbalances affects process very heavily. It had an effect on every model. It was a bottleneck for our models.

The Random Forest classifier demonstrated strong performance in predicting instances of 'no heart disease' (class 0), achieving high precision, recall, and F1-score. However, its performance on 'heart disease' instances (class 1) was comparatively weaker, highlighting the challenge posed by class imbalance. I learnt that the class imbalance was also the reason why some methods didn't work better even with K-fold cross validation.

Similar to the Random Forest, the Deep Neural Network model was also much better at predicting class 0 and was weaker for class 1. We found that increasing the number of epochs and batch size doesn't always leads to better results. Sometimes, simpler models with lesser run time perform similarly bad.

Visualizations, including scatterplots, radar graphs, and bar charts, provided insights into the relationships between variables and potential predictors of heart disease. They even uncovered some of the internal relations like BMI and Weight (which is expected).

To answer the research question, How accurately can we predict the occurrence of heart attacks based on health indicators, lifestyle factors, and other characteristics in the dataset. Technically, we get 95% accuracy but mostly its for No heart attack. So our models are actually pretty weak in predicting who could actually get a heart attack. It works well for telling if a person won't have an heart attack.

Additionally, not all scores are important. Usually, I would be looking at accuracy of a model and be impressed by such high numbers but we realize that It means nothing of value in our case. We need to actually figure out the correct metric to look at or combination of metrics.

DMP & ACS

O.K. worked on Data cleaning

D.P. implemented and applied Unsupervised Clustering for EDA

H.H.& F.A. did Data collection/Data preparation

R.M.(Me) & F.A. performed the Exploratory data analysis

The Data Management Plan (DMP) has been attached in the appendix.