CSE/ECE 343: MACHINE LEARNING PROJECT REPORT

Title: Cardiac Arrest Analyser

Mehul Goel (mehul20311@iiitd.ac.in), **Sumit Sharma** (sumit20250@iiitd.ac.in)
Vishal Kumar (vishal20265@iiitd.ac.in), **Yashdeep** (yashdeep20160@iiitd.ac.in)

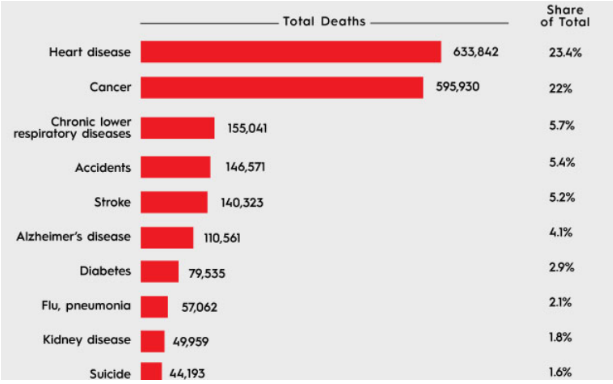
Abstract

Cardiac Arrest is the sudden cessation of cardiac activity so that the victim becomes unresponsive, with no normal breathing and no signs of circulation. If corrective measures are not taken rapidly, this condition progresses to sudden death. During this era after a covid period (i.e in 2021 almost 28,449 people died due to cardiac arrest). Cardiac Arrest is more dangerous for the younger generation. Research shows that most of the deceased belong from 25 to 50 years old. Our main aim is to develop a machine learning model which can predict such heart disease occurrence more accurately and precisely by analyzing multiple pre and post-health features. We have used a Heart Disease Health Indicator Dataset containing a total of 22 features. The target variable predicts the occurrence of heart maladies and the remaining 21 are normal health-indicative features with more than 2,50,000 data points from the Kaggle Machine Learning Repository. We have preprocessed the dataset by removing the null values and implemented Principal Component Analysis so that we can have a dataset robust to a variety of data values. According to our analysis, almost all the machine learning models performed well (accuracy of ~90% for both training and validation sets); however, in the case of PNN, the most efficient in classification with an accuracy of 84.85%. Furthermore, in order to improve the accuracy we used Sequential Feature Selection which led to increased accuracy of 90.65%. We hope that this would play a significant role in reducing cardiac arrest and such heart-prone deaths.

Keywords - Sequential Feature Selection, Principal Component Analysis,

1. Introduction

Cardiac arrest is a common and treatable cause of death and disability.It occurs when the heart stops beating suddenly. The lack of blood flow to the brain and other organs can cause a person to lose consciousness, become disabled, or die if not treated immediately.



Cardiovascular diseases are the leading cause of death globally, taking an estimated 17.9 million lives each year. Every 1 out of 5 people dies of cardiac arrest. It is said to be one of the most dangerous silent killers. Due to the spread of the coronavirus in recent years, this situation has worsened. Due to covid period, lots of people of all age groups have experienced depression, anxiety, and depreciation in their general quality of life. From a renowned survey, 12 lakh youngsters have died due to cardiac arrest.

In some people, the heart rate can vary from fast to slow but surveys said that shortness of breath, and chest pain after COVID-19 is a common complaints. Heart attack can be caused by increased stress on the heart, such as a fast heartbeat, low blood oxygen levels or anemia because the heart muscle isn't getting enough oxygen delivered in the blood in order to do this extra work. It is observed that in people with acute coronavirus disease, but it is less common in those who have survived the illness

We hope that this would play a significant role in reducing cardiac arrest and such heart-prone deaths and would also help raise awareness and help people in getting back on track.

Through the application of popular machine learning models, we aim to reduce the mortality rate due to various heart diseases. We have analyzed the dataset using various supervised machine-learning techniques in order to predict the probability of someone being prone to cardiac maladies. The study tries to increase the accuracy by utilizing various prognostic factors encountered in research papers in the field. Finally, we have compared and accumulated different models to get a better understanding of the field and give better outcomes, so that more lives can be saved.

2. Literature Survey

1. Cardiovascular disease detection using Artificial Immune System and other models: In this study, Ishan Gupta and his group presented a solution for the detection of cardiovascular diseases by using a clonal selection algorithm, which is an AIS with an average accuracy of 78%. The clonal Selection Algorithm is used for pattern recognition method problems and further, the result was compared with other models like Random Forest Classifier, Support Vector Machines, Decision tree Classifiers, and Artificial Neural Networks. They combined the k Nearest neighbor with Clonal Selection Algorithm which came out to be the best consistent and suited algorithm.

2. Machine Learning approaches to predicting the risk of in-ward cardiac arrest of cardiac patients by Lahiru T. W. Rajapaksha show that recurrent neural network (RNN) outperformed with 96% accuracy with a sensitivity of 95.83%. They developed a Deep learning model and used more than 15 medical details of all cardiac patients to develop this model. Their model demonstrated higher sensitivity and specificity than the earlier ones.

3. Machine Learning Approach for Sudden Cardiac Arrest Prediction Based on Optimal Heart Rate Variability Features by L.Murukesan and his team: In this study, they aimed to predict Sudden Cardiac Arrest(SCA) two minutes before its occurrence and used the proposed signal processing methodology for further results. They used a total of 34 features and then they applied a Sequential Feature Selection algorithm to select the optimal features. SVM gave a prediction rate of 96.36% and Probabilistic Neural Network (PNN) gave a prediction rate of 93.64%. Thus they used an SVM classifier.

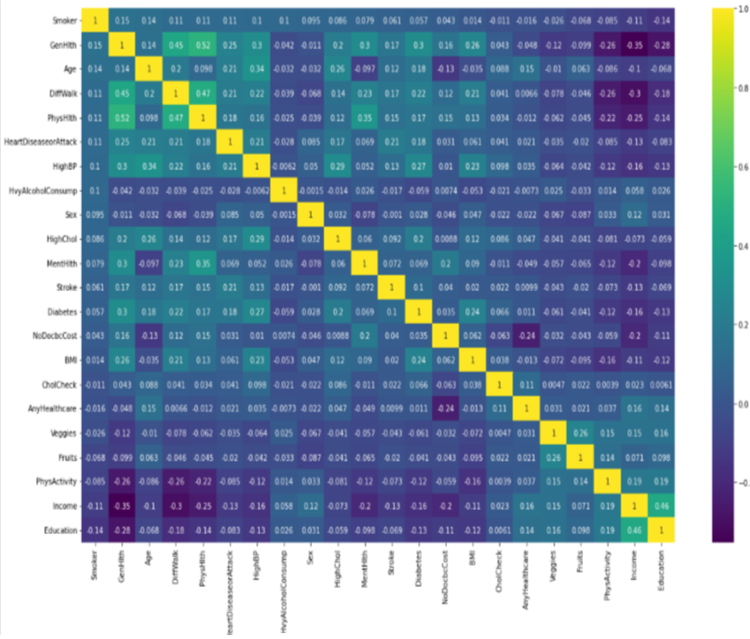
3. Data Preprocessing and Visualisation

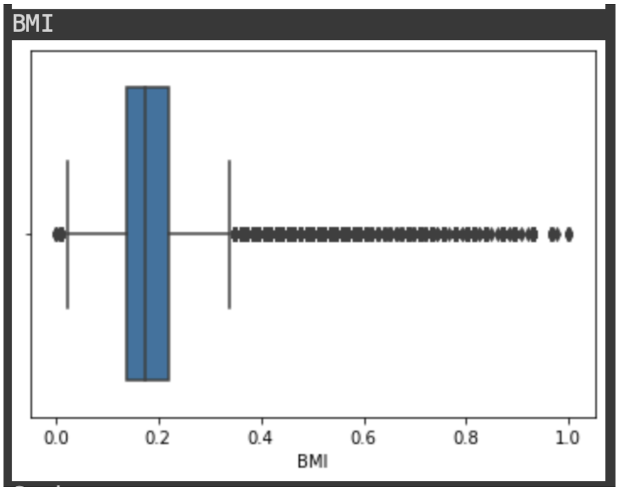
3.1. Data Description

The dataset contains 21 features and a target attribute. It has the following features: HighBP, High cholesterol, BMI, Smoker, History of Strokes, Physical Activity, Diet, Mental Health, Physical Health, Sex, Age, education, income, etc. The target variable in the dataset is "Heart Disease Attack" which states whether the person with the given attributes has had a heart disease/attack or not.

3.2. Data Visualization

We have plotted the correlation heat map of the features and selected the highly correlated ones. We have also shown the box plot of the features (fig 1.1) in order to visualize the distribution of a variable.





3.3 Data Preprocessing

Firstly we removed all the null values and removed all the duplicate values. After that, we went for the dimensionality reduction i.e. Principal Component Analysis Techniques (PCA) which ultimately eliminated the features and went further with 8 features for using the Gaussian Distribution to compute the accuracy. Secondly, after implementing all the models, in order to maximize the accuracy we used Sequential Feature Selection.

4. Methodology

We are splitting the dataset into a training and testing set using an 80:20 random split.After the division of our dataset, we chose supervised learning models to train and test on the dataset. We also performed hyperparameter tuning and chose the best model for training and testing.

Description Of Models:

1. *Logistic Regression:* It is a statistical model that, in its simplest version, models a binary dependent variable using a logistic function. By switching the loss function to a cross-entropy loss and the predicted probability distribution to a multinomial probability distribution, we were able to achieve multi-class classification using multinomial logistic regression.
2. *Naive Bayes:* It is used to categorize the data using probability theory. It uses Bayes Theorem. The characteristics are also considered independent.

3. *Decision Tree:* It is a non-parameterized supervised learning method with a pre-specified target variable that is frequently applied to classification issues.

4. *Random Forests*: It is an ensemble learning technique in which a large number of decision trees are built during the training phase. The class that

the majority of the trees chose is the result of the random forest.

5. *K-Nearest Neighbours:* It is based on supervised learning, where each data point's nearest k neighbors are determined. The projected class is the label that represents the majority of those data points.

6. *Ada-Boosting:* It is a boosting technique used as an Ensemble Method in Machine Learning. The weights are re-assigned to each instance, with higher weights assigned to incorrectly classified instances.

7. *Support Vector Machine:* It is a supervised machine learning algorithm used for both classification and regression.

8. *Multi-Layer Perceptron:* It is a feedforward artificial neural network that generates a set of outputs from a set of inputs. An MLP is characterized by several layers of input nodes connected as a directed graph between the input and output layers

9. *Sequential Frequency Selection:* The algorithm chooses many features from the collection of features and assesses them for the model, iterating between several sets while reducing and increasing the number of features, allowing the model to achieve the best performance and outcomes.
   These essentially form a portion of the wrapper methods that progressively add and remove features from the dataset. When this happens, the method is known as naïve sequential feature selection. It selects M features from N features based on individual scores after evaluating each feature independently. Because it doesn't take feature reliance into account, it works only on datasets with lesser feature correlation such as the chosen dataset.

We can use a variety of machine learning algorithms to classify the given problem. In this project, we can use the following models for classification like Logistic regression, Gaussian Naive Bayes, Decision Trees, Random Forests, Multi-Layer Perceptron, SVM, and ADA boost classifier.

5. Result And Analysis

After Analysing the data by EDA, from finding the correlation between the features and normalizing the features which have outliers in their dataset. As this a classification or logistic regression problem, we can use no of models for our problem like Gaussian Naive Bayes, and Random Forests. Multi-Layer Perceptron, Adaboost, etc. For feature selection, we have used 8,10,12,14 output dimensions in PCA (Principal Component Analysis). In the case of 8, accuracy came out to be largest by Gaussian Naive Bayes. With this our preprocessing of the data is complete. Now we just have to tune the parameters for the model. Like which model is best for PCA-transformed data, by training different models.

After training the models after feature selection using PCA, we get the following scores on the validation set.

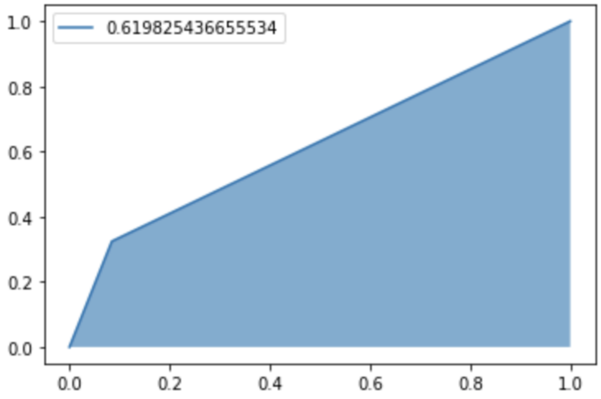
Model	Accuracy Score	F1 Score	Recall Score	Precision Score
Logistic Regression	0.9073	0.8777	0.9073	0.8788
Gaussian Naive Bayes	0.892	0.8815	0.8928	0.8739
Decision Tree Classifier (Entropy)	0.8562	0.8544	0.8558	0.8530
Decision Tree Classifier (Gini)	0.8558	0.8547	0.8558	0.8537
Random Forest Classifier	0.8921	0.9125	0.8921	0.9375
AdaBoost Classifier	0.8776	0.8877	0.8776	0.8992
PNN	0.8485	0.8470	0.8485	0.84564
K Neighbor Classifier	0.8946	0.914	0.8945	0.938

From the above table we can see that the model with the best overall accuracy score for the validation set is the Logistic Regression model, with an accuracy of 90.7%.

Other models like the K Neighbor Classifier, Random Forest Classifier, and Gaussian Naive Bayes give a decent performance with an accuracy score of 89%

Models Decision Tree Classifiers, Probabilistic Neural Networks(PNN), Adaboost Classifier however gives a worse performance with an accuracy of less than 88%

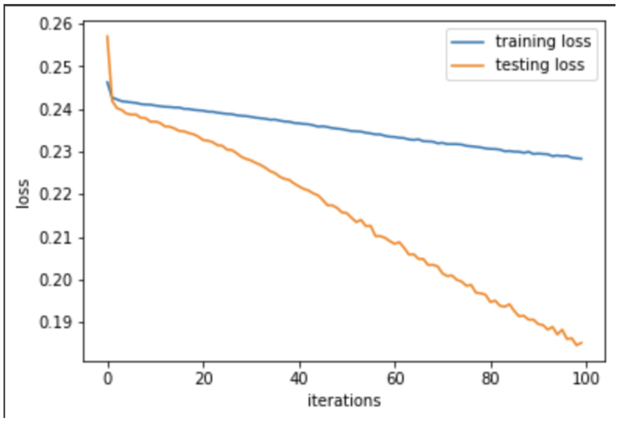
We get the following ROC-AUC curve using the decision tree classifiers.



As evident from the roc curve, the decision tree classifiers don't perform well. This is because the ratio of true positive and false positive rates is only marginally better than the 45-degree diagonal of the curve.

Apart from these models we also modeled the given dataset using sklearn's MLP Classifier model. In this model, the hyperparameters tuned were the max number of iterations, the activation function, the hidden layers, and the number of neurons present in them.

We get the following training and validation vs iteration plots when training MLP on the given dataset.



As we can observe from the graph, after 100 iterations the testing loss gradually decreases, faster than the training loss. However, taking the scale into account, we can notice that even after 100 iterations the training and validation losses are only about 0.05 units apart. Therefore, we can say that the variance in this model is minimal. Moreover, since the loss is so low, we can also say that the bias is also minimal. Therefore, in this model, the data is neither overfitted nor under fitted.

The models trained after performing a sequential forward feature selection Algorithm give the following results:

Model	Accuracy Score (Training)	Accuracy Score (Validation)	F1 Score	Recall Score	Precision Score
Gaussian Naive Bayes	0.9052	0.9052	0.9477	0.9052	0.9955
MLP Classifier	0.9057	0.90495	0.9384	0.90495	0.9794
KNN	0.9100	0.9027	0.9383	0.9027	0.98039

As we can observe from the given table, the models trained on the features selected from the forward sequential feature selector, perform better than those trained on the features selected from PCA. In all the above models we see that the training and validation set accuracy are similar. This indicates that the variance of the models is less. Moreover, the accuracy of the models is high. Thus the models are neither overfitted nor under fitted.

Among the models trained using the sequential feature selector the best validation set accuracy is obtained in the K Nearest Neighbors classifier of 91%. In this model, we chose the number of neighbors to be 8.

Finally, we created a cumulative model from the models providing the best accuracies. The combined model trained on features selected using the sequential feature selection gives a validation set accuracy score of: 0.9060

6. Conclusion

- After finishing with EDA and data preprocessing, we have built, computed, and tested our data on various models mentioned above.
- Training models on data with the

original set of features and data transformed by PCA and sequential feature selector, we got different accuracies on different models.

- Among Gaussian NB, logistic regression, Decision Tree Classifier, Random forest, KNN, etc we got the highest accuracy of 91%, which was given by KNN with the optimal number of neighbors being 8.
- We also tried to make an ensemble model by taking 3 models which were giving highest accuracies i.e Gaussian, MLP, KNN to enhance the result of our project.
- The ensemble model also gives a similar accuracy to KNN. At last, we ended up taking KNN as the best classifier among all the classifiers.

7. Contribution

- Yashdeep - Data Preprocessing and Visualization, Result Analysis, Training, Models and Hyper Parameter Tuning-[DT, KNN, LR, MLP, Boosting], Report Writing, Making Presentation.
- Mehul - Dimensionality Reduction(PCA), Sequential feature selection, Result Analysis, Training Models and Hyperparameter tuning-[RF, KNN, LR, MLP, Bagging], Report Writing, Making Presentation.
- Sumit - Plotting Maps, Model Selection, Training Models, and Hyper Parameter Tuning-[RF, NB, LR, SVM], Report Writing, Making Presentation.
- Vishal -Plotting Maps, Model selection, Training Models, and Hyper Parameter Tuning-[DT, NB, LR, Boosting, SVM], Report Writing, Making Presentation.

These were the assigned responsibilities for each team member. However, all the members equally contributed to all the work done.

8. References

- https://towardsdatascience.com/feature-extraction-techniques-d619b56e31be
- https://jainendra.in/2021/01/07/best-projects-machine-learning-course-cse343-ece343/
- https://www.javatpoint.com/data-preprocessing-machine-learning
- https://towardsdatascience.com/how-to-perform-exploratory-data-analysis-with-seaborn-97e3413e841d
- https://www.kaggle.com/datasets/alexteboul/heart-disease-health-indicators-dataset
- https://scikit-learn.org/stable/modules/feature_selection.html#sequential-feature-selection
- https://www.javatpoint.com/auc-roc-curve-in-machine-learning
- https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.fill_between.html#matplotlib.pyplot.fill_between
- https://www.statology.org/numpy-mode/