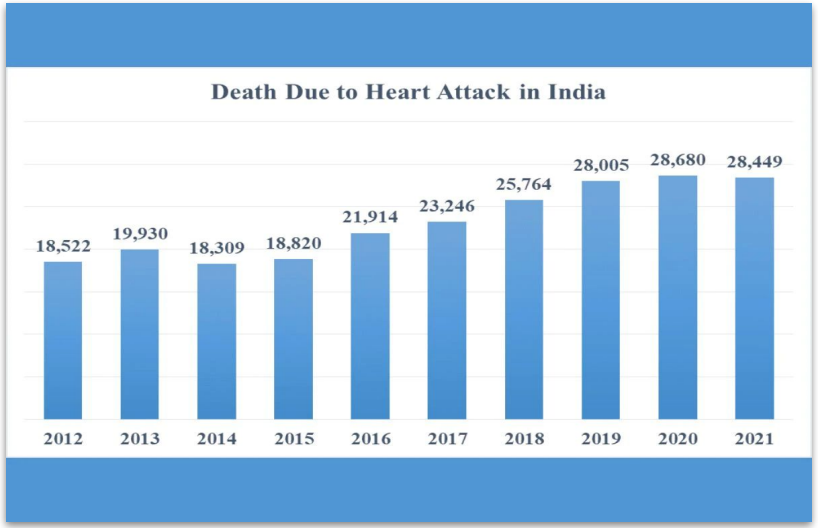
Cardiac Arrest Analyser

Piyush

Motivation

- Cardiac Arrest is one of the major diseases during this era after covid period (i .e in 2021 almost 28,449 people died due to cardiac arrest)
- Cardiac Arrest is become most dangerous for Younger generation the research show that most of died person belong to 25 to 50 years old.
- After covid the Cardiac Arrest probability has been increase.
- This motivate us to choose this topic and apply all we know about machine learning algorithm to ge a prevent solution for people.



Literature Survey

- **Cardiovascular disease detection using Artificial Immune System and other models** : In this study Ishan Gupta and his group presented a solution for detection of cardiovascular diseases by using clonal selection algorithm, which is an AIS with an average accuracy of 78%. 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, Artificial Neural Networks. They combined the k Nearest neighbour with Clonal Selection Algorithm which came out to be the best consistent and suited algorithm.
- **Machine Learning approaches to predicting the risk of in-ward cardiac arrest of cardiac patients** by Lahiru T. W. Rajapaksha shows that recurrent neural network (RNN) outperformed with 96% accuracy with 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.

Literature Survey

- **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 SVM classifier.

Dataset Description

- We have the dataset from Kaggle. Repository available at: https://www.kaggle.com/datasets/alexteboul/heart-disease-health-indicators-dataset
- The Data is consist of 253,680 data samples and there no null values in dataset
- There are 22 features in the Dataset including the Target Label.
- Target /Dependent variable is discrete and categorical in nature.
- We have to do binary classification in order to predict weather a person can have heart attack or not i.e "Heart/Cardiac Attack" score is between 0 or 1;where 1 having a heart/ Cardiac attack and 0 not having a heart disease/attack.

Preprocessing the dataset

- Mean value is less than median value of each column represented by 50%(50th percentile) in index column. Notably large differnece in 75th %tile and max values of predictors "Stroke","Diabetes","MentHlth", "PhysHlth", In "BMI","MentHlth", "PhysHlth" you can see the maxmimum value is completely/significantly outlier to the distribution of the column.

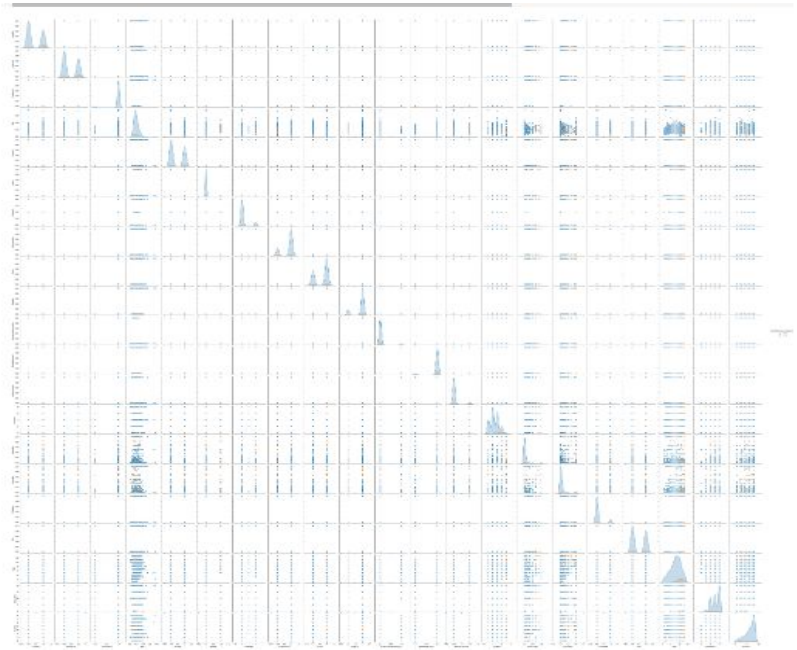
- To counter this we can apply normalization to "BMI","MentHlth", "PhysHlth" like min-max , standard-scalar. In this we are applying min-max because min-max difference is significantly large.

Preprocessing the dataset

- We noticed that the dataset doesn’t contain any null or missing values.
- The values of all the fields are small floating point integers.
- In case of attributes, like sex, diet, etc we see that the values are appropriately mapped to floating point numbers.
- Moreover, in case of fields like education, income etc. we notice that the values are still less than 10. This implies that all these attributes have been scaled appropriately to prevent any feature scaling issue.
- Therefore, there is no need for preprocessing the dataset to map various values, counter any missing values or for scaling the features for some features.

Visualizing the dataset

We have also plot Heat-map,Bar graphs, Box Plots etc to visualize and to get some insights of the data like Co-relation between the features by heat Map, information about the outliers by box plots etc. in the upcoming slides.



(Pair plots mentioned in above slide)

Visualizing the dataset

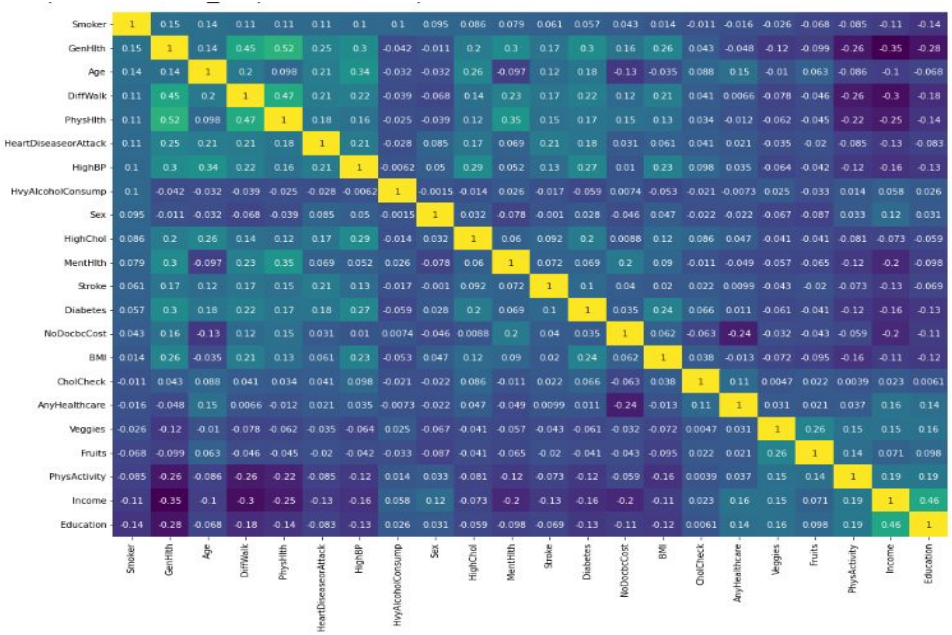
- Graphs at: https://colab.research.google.com/drive/1eEkrYYQVA8S8Dx3I4utQnOc0u167HWT4?usp=sharing
- Pair-plots are also a great way to immediately see the correlations between all variables.
- From the pair plots we can observe that among all the factors leading to a heart attack, the following features play a more effective role:
  - Diabetes
  - Physical Activity
  - Mental Health
  - Age
- Therefore, elderly people with diabetes less physical activity and poor mental health are more prone to heart diseases than others.

Visualizing the dataset

- Light shades represents positive correlation while Darker shades represents negative correlation.

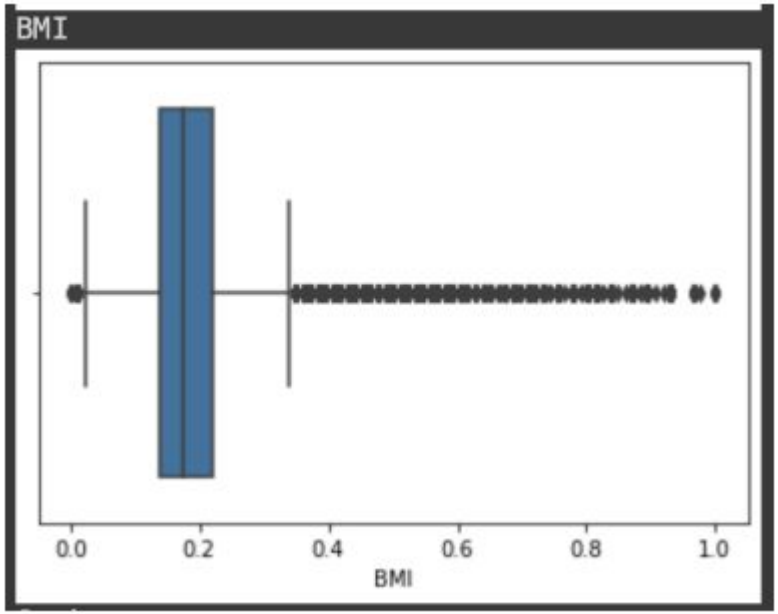
- Here we can infer that "PhysHlth" has strong positive correlation with "GenHlth" whereas it has strong negative correlation with "income".

- "AnyHlthCare" and "NoDocBcCost" has almost no correlation with "Heart/ Cardiac Attack" Since correlation is zero we can infer there is no linear relationship between these two predictors.

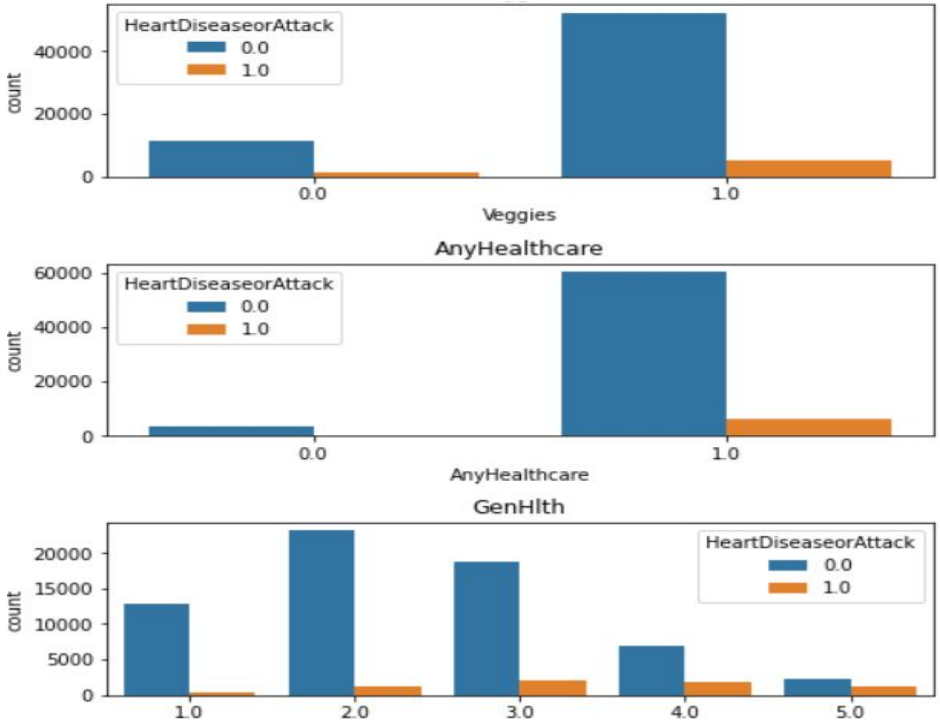


However it is safe to drop these features in case you're applying Logistic Regression model to the dataset.

BOX PLOT



BAR PLOTS

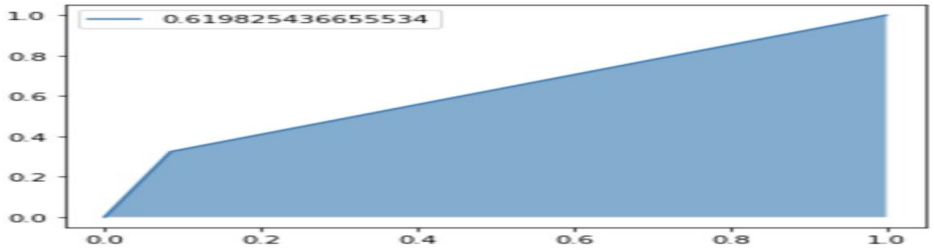


Methodology

- 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

❖ There are following models that we implemented for training the data

- ➢ **Logistic Regression**
- ➢ **Naives Bayes**
- ➢ **Decision tree**
- ➢ **Random forest**
- ➢ **K-Nearest Neighbours**
- ➢ **Ada-Boosting**
- ➢ **Support Vector Machine**
- ➢ **Multilayer Perceptron**
- ➢ **Probabilistic Neural Network**



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.

(Ex of ROC curve for binary classification for Decision Tree)

Methodology

- The Sequential Search algorithm similar to the PCA that features 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.
- Finally, we created a cumulative model from the models providing the best accuracies. The combined model that is Gaussian Bayes, MLP, KNN on the basis of Ensembling techniques that is give the output on the basis the maximum number Output from all the these models.

Result Analysis

- We used the following performance metrics to test our models

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
K Neighbor Classifier	0.8946	0.914	0.8945	0.938

Result Analysis

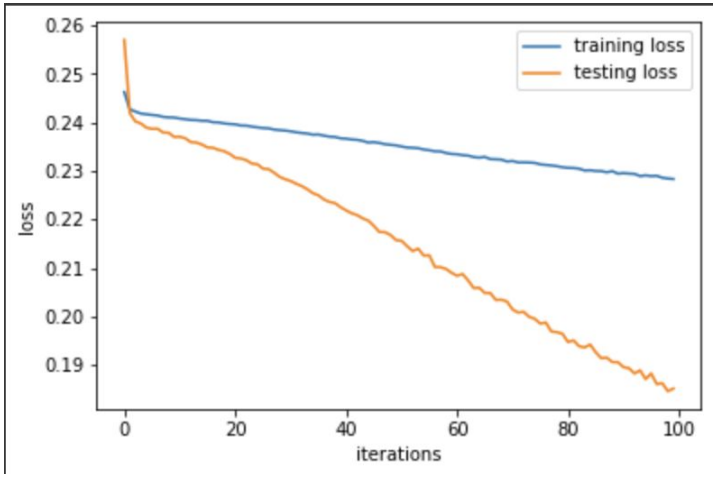
- The models trained after performing a sequential forward feature selection Algorithm give the following results:

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

Results And Analysis

- 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.



Results And Analysis

- After Analysing the data by EDA , from finding the correlation between the features , 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,Random Forests
- Multi Layer Perceptron, Adaboost etc.
- For feature selection i have used 8,10,12,14 output dimension in pca (principal
- Component Analysis). In the case of 8 accuracy came out to be largest by Gaussian Naive Bayes. With this my preprocessing of the data is complete.
- Now just have to tune the parameters for the model.
- Like which model is best for pca transformed data, by training diff models.

Conclusion And Future-Improvement

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

Conclusion And Future-Improvement

As existing Dataset have some imbalance among target labels, we can have Dataset which have equally or appro. proportion of target variable/Label .

We can have a subset of questions that can be used for preventive health screening for diseases like heart disease.

The results can be improved by obtaining a larger dataset with diverse sociodemographic characteristics.

We can introduce new features , which have more Co-relation with the Label than the existing redundant ones.

Contributions

Piyush - 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.

Timeline

In terms of getting the work done, we adhered to the schedule and completed all of the tasks. Here is the work that we have completed:

1 Pre-processing data,Data visualization
2 Feature Analysis and Selection, Plotting Maps,Dimensionality reduction , Logistic regression
3 Naive Bayes,Decision Trees
4 Random Forests , K-Nearest Neighbours
5 Analysis of Model Performance , HyperParameter Tuning
6 Lr,NB,DT,KNN
7 BOOSTING
8 SVM,MLP
9 Analysis of Model Performance , Hyper-parameter Tuning
10 Report Writing and Presentation Making

References

1. https://towardsdatascience.com/feature-extraction-techniques-d619b56e31be
2. https://jainendra.in/2021/01/07/best-projects-machine-learning-course-cse343-ece343/
3. https://www.javatpoint.com/data-preprocessing-machine-learning
4. https://towardsdatascience.com/how-to-perform-exploratory-data-analysis-with-seaborn-97e3413e841d
5. https://www.kaggle.com/datasets/alexteboul/heart-disease-health-indicators-dataset

THANK YOU