Capstone Project Submission

Instructions:

i) Please fill in all the required information.

ii) Avoid grammatical errors.

Team Member's Name, Email and Contribution:
1. **Poonam Dattu Shevkar** **Email-id :-** dattupoonam@gmail.com ❖ **Contribution-** ● Checking for NAN values using missingno matrix. ● Visualize the smoking ratio of male & female . ● Correlation analysis to drop the highly correlated features. ● One-hot encoding to create dummies in the dataset. ● Scaling the dataset and training the model into a train and test data set. ● Logistic regression & KNN Algorithm models were performed. ● ROC curve for implemented models of machine learning algorithms. 2. **Sanjay Ramkishan Verma** **Email-id :-** shankyverma1998@gmail.com ❖ **Contribution-** ● Visualization of dependent & independent features using histogram ● Outliers detection to remove noise from the dataset. ● Value counts visualization for features like age, diabetes patients , hypertensive patients,etc. ● SMOTE method to balance the dataset. ● Decision Tree Classifier & Support Vector Machine Algorithm models were performed. ● Hyperparameter Tuning applied on KNN algorithm.
Please paste the GitHub Repo link-

Github Link:-

https://github.com/Sara19598/Cardiovascular-Risk-Prediction-ML-Classification

Please write a short summary of your Capstone project and its components. Describe the problem statement, your approaches and your conclusions. (200-400 words)

Cardiovascular sickness is a general class for a scope of infections that are influencing heart and veins. The early strategies for estimating the cardiovascular sicknesses helped in settling on choices about the progressions to have happened in high-chance patients which brought about the decrease of their dangers.The dataset is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. The classification goal is to predict whether the patient has a 10-year risk of future coronary heart disease (CHD). The dataset provides the patients' information. It includes over 4,000 records and 15 attributes.

In this project , we have used Machine Learning Classification algorithms.Machine Learning examines how systems can learn or improve their exhibition in a view of Cardiovascular information.Our aim is to develop the machine learning model which has the highest accuracy and precision in predicting if a patient is at a risk of having Coronary heart diseases in future .

To train our model we used the dataset of 'Coronary Heart Diseases'.We load the necessary libraries and preprocess the data.For better understanding and analysis , EDA was conducted.

Further, data transformation was taken into account by validating outliers and correlation matrix to remove the noise from the dataset.Also, one-hot encoding is used to create the dummies in the dataset for future appearance of proper data. As data is imbalanced , SMOTE method was applied to overcome this part.We performed standard scalar transformation , to train the model by splitting up 70% data as supervised or train data set and 30% as test data set.

We load the machine learning models with the dataset and train the model. Further, we test the machine learning model namely Logistic Regression, K-nearest neighbors,Decision Tree Classifier and Support Vector Machine and check its accuracy,classification report,Confusion matrix & ROC curve for interpreting better results. We compare the accuracies of all the machine learning models and we learn that the KNN algorithm has the highest accuracy after hyperparameter tuning hence it is the best algorithm.

The machine learning algorithms under study were able to predict cardiovascular disease in patients with accuracy between 77.06% to 100%. It was shown that KNN algorithm has better Accuracy (Train Accuracy = **85.30** & Test Accuracy = **84.07**) when compared to different Machine-learning Algorithms based on hyperparameter tuning.

Drive Link:-
https://drive.google.com/drive/folders/1CmiyfyicbsQ-QUOStwgWYLw5En5JSuaj