Harshil Mukesh Bhavsar (C0863170), Ritu Vinodbhai Kumbhani (C0865677), Shivam Arvindbhai Machhi (C0871186)

2214 2 [B125] Big Data Fundamentals - Data Storage Networking

Parisa Naraei

***Data Set Information:***

* Processed Cleveland data contains 14 classes such as age','sex','cp','trestbps', 'chol', 'fbs', 'restecg', 'thalach', 'exang', 'oldpeak', 'slope', 'ca', 'thal', 'num'.
* As the num column is the one which defines if the patient has the heart disease or not we have 13 independent variables and 1 dependent variable.
* We aim to train the classification model which has the highest accuracy for predicting the value of the num.

**Importing modules and data:**

A picture containing text

Description automatically generated

**Data preprocessing:**

* First we will change the value of num to 1 if it is greater than 0 as it means that the person has the heart disease.

Graphical user interface, application, table

Description automatically generated

Table

Description automatically generated

* Looking at the information of the dataset we found that 2 of the attributes(“ca”,”thal’) has the datatype as object so we decided to look into them to check if there is any value in string or other format

Text

Description automatically generated

* looking at the unique values of these 2 attributes we found that they contain “?” so we will replace that with NA and then we’ll fill those null values with the mode of the respected column values.

Table

Description automatically generated with medium confidence

* After the preprocessing we will split the data into training and testing set and use standard scaler from sklearn library to keep the data into the same scale.

Graphical user interface, text, application, email

Description automatically generated

* ***Algorithm used :***
  + We imported the classifier models such as logistic regression, Gaussian naive bayes, SVC, Decision Tree classifier, random forest classifier, gradient boosting from sklearn library for the training and comparing accuracy.
  + We have used gridsearchcv for the hyperparameter tuning of all the algorithms while training.
  + Lastly, we have stored the model results and model classifier into their respective folders.

1. ***Gaussian naive byes –***

* We have used gaussian naïve byes as it is the classifier which greatly works for the classification machine learning tasks by using bayes theorem in this implementation.
* We achieved 86% accuracy for training set and 80% accuracy for the testing set.

Graphical user interface, text, application

Description automatically generated

1. ***Logistic regression –***

* As logistic regression is best for doing binary classification we considered training our data with this model.
* We used L1 and L2 loss penalties with various ‘C’ values (Inverse of regularization strength lambda) with gridsearchcv to find the best parameters.
* We achieved the accuracy of 86% with the training data and 78% testing data

Graphical user interface, text, application, email

Description automatically generated

1. ***Decision tree –***

* We have used the decision tree classifier here with tuning of max\_leaf\_nodes, gini or entropy, max\_depth and min\_samples\_split to find the best parameters.
* We achieved the accuracy of the 86% with training data and 78% with the testing.

Text

Description automatically generated

1. ***Supprt vector classifier–***

* Support vector machine uses the svc to create the hyperplane after placing the data into high-dimensional space to separate the data into different classes. It works well with the classification tasks.
* We used various kernals such as linear, rbf and sigmoid with various regularization parameter values.
* The accuracy which we achieved on training set was 92% and with the testing data 80%.

Graphical user interface, text, application, email

Description automatically generated

1. ***Random forest –***

* In order to increase the dataset's predictive accuracy, a classifier called Random Forest uses many decision trees on different subsets of the input data. It is based on the idea of ensemble learning, which is the practise of integrating various classifiers to solve a challenging problem and enhance the model's performance. We did the hyperparameter tuning here as well just as we did with the decision tree classifier.
* We achieved accuracy of 100% in the training and 75% in the testing. We interpreted that the model is overtrained here as it gives perfect accuracy on training dataset but performs poorly on the testing data.

Text

Description automatically generated

1. ***Gradient boost –***

* In especially with large and complex data, gradient boosting is a technique that is gaining popularity for its prediction speed and accuracy. With only 303 observations, other classifiers are prone to overtraining, thus we employed our model to avoid this problem.
* We achieved the highest accuracy of 81% in the testing data and the accuracy of the 81% on the training data.

Text

Description automatically generated

***Graphs –***

* We have plotted learning curves, scalability and performance for the models which we have trained as below.

***Chart, line chart, scatter chart

Description automatically generated***

***Chart, line chart, scatter chart

Description automatically generated***

***Chart, line chart

Description automatically generated***

***Chart, line chart, scatter chart

Description automatically generated***

***Chart, line chart

Description automatically generated***

Chart, line chart

Description automatically generated

***Conclusion:***

From the overall training of the 6 classifier models and testing them on the testing set we conclude that we achieved the highest accuracy of **81**% with the gradient boosting classifier. The model seems to be performing well and is also not overtrained.