Machine Learning Supervised Classification

20PAIC53J

CT3 Mini Project

Inference:

Data Preprocessing:

- Number of rows and columns are 303 and 14 respectively.
- Obtained five-point summary of the attributes/features.
- Gathered information about the columns of the dataset.
- Checked for missing values and none were found.
- Checked the distribution of target variable and found that 165 people have defective heart(1) and 138 people have healthy heart(0)

Splitting the features and Target:

- Splitted the features and Target variable. Performed standard scaling.

Model Training and Evaluation:

- Built and Trained several ML algorithms and obtained achieved better results.
- Logistic Regression achieved training accuracy of 86.36% and testing accuracy of 88.52%
- Decision Tree Classifier with max depth of 3 achieved training of 97.42% but a lower testing accuracy of 81.97

- Random Forest Classifier performed achieved 100% training accuracy and testing accuracy of 85.25% which indicates overfitting.
- AdaBoost Classifier achieved training accuracy of 93.8% and testing accuracy of 88.33%
- Gradient Boosting Classifier achieved training accuracy of 100% and testing accuracy of 77.05% which shows overfitting.
- Voting Classifier combining multiple classifiers achieved training accuracy of 98.76% and testing accuracy of approximately 80.33%

Comparison and Visualization:

- The bar plot compares the training and testing accuracies of different classifiers. Random Forest and Logistic Regression seem to perform relatively well on both training and test data, while Decision Tree and Gradient Boosting show overfitting as they perform well on training data and poorly on testing data.

Conclusion:

- Overall, Logistic Regression and Random Forest seem to be better performing models based on the obtained results as they achieve high accuracies on both training and testing data.