Drishya Uniyal - MT21119

SML ASSIGNMENT 3

METHODOLOGIES

- ★ Importing all the necessary libraries.
- ★ Load the given datasets, training, and testing
- ★ Check the shape of the dataset.
- ★ Data Preprocessing
    - ○ Check for Null values
    - ○ Check for missing values
    - ○ Check for NaN values
    - ○ Remove outliers.
    - ○ Remove Duplicates
    - ○ Standard Scalar
- ★ Load the classifiers
- ★ Report the training and testing accuracies
- ★ Predict on the test dataset

OBSERVATIONS

Classifier	**Training Accuracy**	**Testing Accuracy**
Logistic Regression(LR)	86.3	81.63
Naive Bayes	83.42	73.47
K Nearest Neighbours(KNN)	70.47	69.39
Decision Tree(ID3)	100	73.47
Random Forest(RF)	100	83.67
ADABoost	96.37	75.51
ADABoost + LR	96.37	73.47
ADABoost + RF	96.37	77.55

Drishya Uniyal - MT21119

EXPERIMENTS

- ★ The dataset was first made fit for the experiment by removing the duplicates, outliers, checking the missing values, removing irrelevant data and other pre-processing techniques.
- ★ The first classifier used is Logistic Regression, which models the probability of the target variable belonging to a particular class.
- ★ The second classifier used is KNN which works by finding the K nearest neighbors to a new data point based on a similarity measure, such as Euclidean distance or cosine similarity.
- ★ The third classifier used is DT, which recursively splits the data into smaller subsets based on input variable values, to make predictions about the target variable.
- ★ The fourth classifier used is RF, which uses a technique called bagging (bootstrap aggregating) to create these subsets, which helps to reduce overfitting and improve the model's generalization performance. The accuracy achieved by Random Forest is better than by Decision trees.
- ★ The fifth classifier used is ADABoost ((Adaptive Boosting), which works by iteratively training a series of weak classifiers (e.g., decision trees or simple linear classifiers) on weighted versions of the training data.
- ★ The sixth classifier used is ADABoost _LR
- ★ The seventh classifier used is AdaBoost + RF

RESULTS

- ★ ADABoost + LR gives the best Test accuracy on Kaggle ***(93.3%)***
- ★ Logistic Regression gave the best amongst the other classifiers and so was used as a base_estimator for Adaboost.

ANALYSIS

Why logistic regression performed better than other classifiers?

Logistic regression may perform better than other classifiers for heart attack prediction because it is a simple yet powerful algorithm well-suited for binary classification problems, can handle both categorical and continuous predictor variables, and is less prone to overfitting than some other classifiers. However, the performance of any classifier depends heavily on the quality and relevance of the predictor variables used in the model, and the choice of performance metrics used to evaluate the classifier. Therefore, it is always important to carefully assess and compare multiple classifiers using appropriate metrics and statistical tests before drawing any conclusions about their relative performance.

Drishya Uniyal - MT21119

Logistic regression with Adaboost gives the best accuracy. Logistic regression combined with AdaBoost may perform better than logistic regression alone for heart attack prediction because AdaBoost can improve the classification accuracy of weak learners, such as logistic regression. When combined with AdaBoost, logistic regression may benefit from the ability of AdaBoost to improve the accuracy of weak learners. Logistic regression can be considered a weak learner because it models a linear decision boundary and may struggle to capture more complex relationships between the predictor variables and the outcome. However, by combining logistic regression with AdaBoost, the model can learn from the misclassified examples and gradually improve its accuracy over time.

INFERENCES

1. Preprocessing the data can help improve the data quality and performance of models.
2. Logistic regression with Adaboost outperformed other tested models, indicating that this combination of algorithms may be the most effective.
3. The performance of any classifier depends on the quality and relevance of the predictor variables used in the model and the choice of performance metrics used to evaluate the classifier.
4. Further analysis may be needed to identify which specific predictor variables are most strongly associated with the risk of heart attacks, which could help inform prevention strategies.

Overall, the results of my analysis suggest that logistic regression with AdaBoost is a promising approach for predicting heart attack risk in this dataset, but further research may be needed to validate these findings and identify the most critical risk factors for heart attacks.