**UNIVERSITY OF LIVERPOOL**

**COMP 534 – APPLIED ARTIFICIAL INTELLIGENCE**

**DETAILED REPORT ON PERFORMANCE OF VARIOUS SUPERVISED LEARNING METHODS ON BINARY CLASSIFICATION PROBLEM**

**NAME: ASWIN MURALI**

**STUDENT ID: 201713299**

**INTRODUCTION:**

The objective of the project is to compare the performance of various supervised learning methods for binary classification. We were asked to consider any three classification models and predict their classes in the given sample based on their features. The entire project was carried out using the Python programming language. The 3 supervised learning methods are:

1. Decision Trees
2. Random Forests
3. Kernel SVM (Support Vector Machines)

**LIBRARIES USED:**

We have made use of the following libraries in this given project:

1. Numpy: Implemented for mathematical computing and provides support for arrays and matrices.
2. Pandas: Implemented as it offers high-performance data structures and helps in data analysis and manipulation.
3. Matplotlib: Implemented for plotting graphs such as bar, line, and scatter graphs based on the performance of models.
4. Sci-kit Learn: Implemented for machine learning algorithms in Python such as classification, evaluation metrics, and more.
5. Seaborn: Implemented for high-level functions of data visualisation.

**DATA EXPLORATION:**

The provided dataset consisted of samples of features ranging from 0 to 9 and had class variables indicating whether the samples belonged to 0 or 1. The dataset had 700 rows and 10 columns. The dataset was loaded into a Python file to understand its characteristics, and the number of samples for each class was printed. The dataset was visualised as a scatterplot to reveal any patterns or potential biases. Additionally, we computed statistical summaries for these features, which included metrics such as mean, median, standard deviation, minimum, and maximum values.

**TRAINING AND TESTING DATA:**

The given dataset is split into training and testing data at 80% and 20%, respectively. The training data was used for model training and hyperparameter tuning, and the testing data was used for the final performance evaluation.

We have also used KFold cross-validation to determine the best hyperparameter for the model in each classification method.

**MODEL EVALUATION:**

**Decision Trees**

A popular supervised learning model, primarily used for tasks such as classification and regression in machine learning.

After the evaluation of test results for decision trees, the best version of this classification method was seen in the Gini index parameter with an accuracy of 95%. The scores of precision, recall, and f1 were seen to be performing better in the Gini index when compared to other parameters. Therefore, the overall performance of the Gini index indicated the good performance of the model.

A diagram of a tree

Description automatically generated with medium confidence

*Figure 1: Decision trees confusion matrix*

The illustration, in figure 1 shows the confusion matrix, for decision trees. Upon examining the matrix, it is evident that there were 88 instances of positives (TP) where the model correctly predicted outcomes that were indeed true. Additionally, there were 44 negatives (TN) indicating predictions of false outcomes as false. The model also yielded 4 False positives (FP) and 4 false negatives (FN). False positives occur when the model predicts an outcome incorrectly while false negatives arise from negative predictions made by the model.

The precision score of the model was 94.28%; this is computed by [TP/(TP+FP)]. It can be defined as the proportion of true positive values among all the positive predictions made by the model. The recall score of the model was 92.28%; this is computed by [TP/(TP+FN)]. It can be defined by saying what portion of actual true positives is correctly classified. The F1 score of the model was 91.66%; this was computed by taking the harmonic mean of both precision and recall.

**Random Forests**

Random forests are known to be one of the popular machine learning algorithms that are built upon the concepts of decision trees. These are also widely used for classification and regression techniques.

A graph of a forest

Description automatically generated with medium confidence

*Figure 2: Random forests confusion matrix*

According to the confusion matrix obtained for random forests in Figure 2, we can see that the model has predicted 88 True positives (TP) and 4 False positives (FP). In the next column, there are 3 False negatives (FN) and 45 True negatives (TN).

After computing the various evaluation metrics for the random forest model, the precision score was found to be 91.83%, the recall score was 93.75%, the F1 score was 92.78%, and the overall accuracy of the model was 95%. The overall performance of the model for the given dataset was commendable and very good.

**Kernel SVM (Support Vector Machines)**

These are extensions of the general SVM model that are primarily used for tackling non-linear classification problems. These were the following hyperparameters used for this model: ‘kernel’ = [‘linear’, ‘poly’, ‘rbf','sigmoid’]. When these hyperparameters were passed through the GridSearchCV cross-validation method, it returned ‘rbf’ to be the best hyperparameter when compared to others.

A blue squares with white text

Description automatically generated

*Figure 3: Kernel SVM confusion matrix*

As per the confusion matrix obtained for Kernel SVM in Figure 3, there were 87 true positives (TP) and 5 false positives (FP); in the next column, there were only 1 false negative (FN) and 47 true negatives (TN).

We can conclude with the following results of evaluation metrics: the precision of the model was seen to be 90.38%, the recall score was great with 97.91%, the F1 score was 94%, and the overall accuracy of the model is 95.71%.

**Conclusion**

According to the overall performance of the models shown in Figure 4, we can see that all three models were used, i.e., decision trees (DT), random forests, and SVM. The overall scores of evaluation metrics for these models are similarly high.

The decision trees have expressed an excellent score of accuracy and precision when compared to other models. The recall scores and F1 score are also meant to have decent scores for the given dataset.

Random forests comparatively have better scores in all aspects except for the precision score when compared with other respective scores.

The kernel SVM linear classifier has the highest recall score amongst the three supervised models. This model showcases excellent balance between accuracy, recall, and F1 scores.

We can conclude this by stating that, with the overall choice of supervised classifier models along with the size, complexity, and characteristics of the given dataset, we have evaluated different classifier models and their performance along with the evaluation metrics. These are the overall analyses of their models and the dataset provided.

A graph of a bar chart

Description automatically generated

*Figure 4: Overall performance graph for the models*