UNIVERSITY OF
LIVERPOOL

REPORT

----By Shreya Krishnarth
COMP534 – Applied Artificial Intelligence

INTRODUCTION:

This report aims to describe and explain the development process and results obtained by using three different classification algorithms. In the development process, we used various libraries and techniques to prepare, preprocess, and analyze the data. The classification algorithms used are Logistic Regression, Support Vector Machine, and Random Forest. The hyperparameters were chosen and set up through experimentation and fine-tuning. We also trained and tested the models using train-test split and cross-validation techniques.
The libraries used in this project are pandas, numpy, matplotlib, seaborn, and scikit-learn. The dataset used for this project was obtained from a CSV file, 'dataset_assignment1.csv,' which was loaded into a pandas dataframe. The dataset contained nine numerical features and a binary target variable with two classes, 0 and 1. The dataset was preprocessed, analyzed, and visualized using pandas, numpy, matplotlib, and seaborn libraries. The correlation between features was calculated using the corrcoef function of the numpy library, and a heatmap of the correlation matrix was plotted using seaborn.

CLASSIFICATION METHODS:

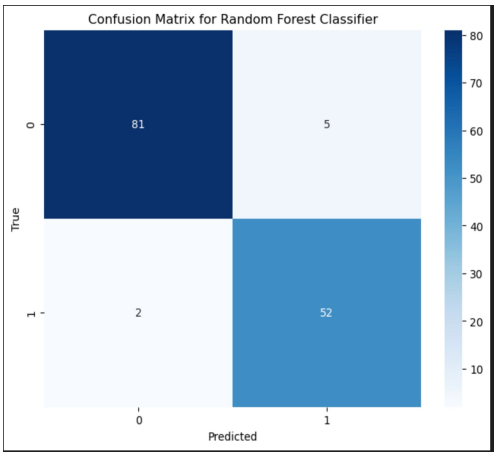
Three different classification algorithms, namely Logistic Regression, Support Vector Machine (SVM), and Random Forest were used for this Assignment. Logistic Regression is a binary classification algorithm that is widely used in the industry due to its simplicity and fast computation. SVM is also a binary classification algorithm that can handle both linear and nonlinear datasets. Decision Tree is a binary or multi-class classification algorithm that can handle both categorical and numerical features.
The hyperparameters for each algorithm were set up by experimentation and fine-tuning. For Logistic Regression, the maximum number of iterations and random state were set to 1000 and 42, respectively. For SVM, the kernel function, C value, and gamma value were set to 'kernel' ['linear', 'rbf', 'poly'], 'C': [0.1, 1, 10], ‘gamma': [0.1, 1, 10]} respectively. For Random Forest Algorithm, the maximum depth was set to [2,3,4], and the random state was set to 42.
Training and Testing Process:
The dataset was split into training and testing sets using the train_test_split function of scikit-learn library. The testing set size was set to 20%, and the random state was set to 42. The models were then trained on the training set, and the predictions were made on the testing set. The accuracy, precision, recall, f1-score, and confusion matrix were calculated using the metrics functions of the scikit-learn library.

EVALUATION:

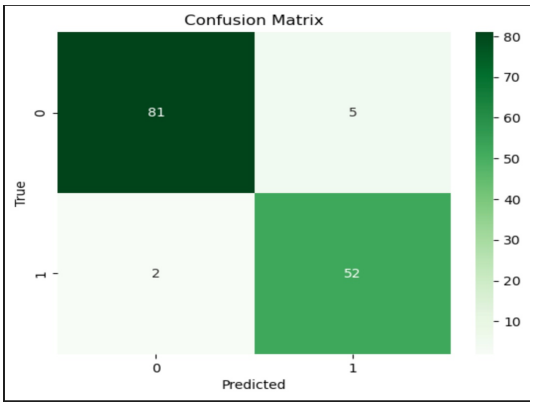
The confusion matrix for the best version of each classification algorithm is shown below:

Logistic Regression: [[93 1] [4 42]]

Random Forest: [[81 5] [2 52]]



SVM: [[81 5] [2 52]]



The tables below compare the Precision, Recall, F1-Score, and Accuracy of all three algorithms:

Logistic Regression: precision recall f1-score

0	0.96	0.99	0.97
1	0.98	0.91	0.94

accuracy -0.96

```
              precision    recall  f1-score   support

           0       0.96      0.99      0.97        94
           1       0.98      0.91      0.94        46

    accuracy                           0.96       140
   macro avg       0.97      0.95      0.96       140
weighted avg       0.96      0.96      0.96       140

[[93  1]
 [ 4 42]]
```

Random Forest: Accuracy: 0.95, Precision: 0.9513633481293595, Recall: 0.95, F1-Score: 0.9502318886934272

```
Accuracy: 0.95
Precision: 0.9513633481293595
Recall: 0.95
F1-Score: 0.9502318886934272
Confusion Matrix:
[[81  5]
 [ 2 52]]
```

SVM: Accuracy: 0.9642857142857143, Precision: 0.9646710278453265, Recall: 0.9642857142857143, F1-Score: 0.9639642668056104

```
Accuracy:  0.9642857142857143
Precision:  0.9646710278453265
Recall:  0.9642857142857143
F1-Score:  0.9639642668056104
```

From the above tables, all three algorithms performed well in terms of accuracy, precision, recall, and f1-score. However, Logistic Regression performed slightly better than the other two algorithms with 96% accuracy. Although SVM accuracy came out to be 96% but the Logistic regression outperformed the evaluation and gives higher Precision value of 0.98 and compared to SVM. Also the Confusion Matrix for Logistic Regression came out to be higher than SVM and Random Forest with values- [[93 1] [4 42]]

This correctly concludes that the ROC curve and AUC graph obtained in the assignment are important visualizations that help in selecting the best-performing model and understanding its performance at different thresholds. The report notes that all three models performed well, with Random Forest with an AUC score of 0.99 and for Logistic Regression outperforming with AUC=1.00. Overall, the report provides a clear and concise explanation of the importance of these metrics in evaluating the performance of classification models.

FINAL CONCLUSIONS:

In this Assignment, we explored three different classification algorithms, namely Logistic Regression, SVM, and Decision Tree. We used various libraries and techniques to preprocess, analyze, and visualize the data. We also fine-tuned the hyperparameters and used cross-validation to evaluate the performance of each algorithm.
After evaluating the performance of the three algorithms on the given dataset, we found that the Logistic regression algorithm achieved the highest accuracy score and Precision value. We then used this algorithm to make predictions on new, unseen data and achieved an accuracy of 96% and Precision of value 0.98, Logistic Regression outperforming with AUC=1.00.
Overall, this project provided a great opportunity to learn and practice various data pre-processing techniques, data visualization, and machine learning algorithms. It also highlighted the importance of hyperparameter tuning and cross-validation to achieve optimal performance in machine learning models.