Spam Detection

Comparison of Machine Learning Techniques

Maria Al Jassim 222455334
Walaa Al Asmakh 221436687

Introduction

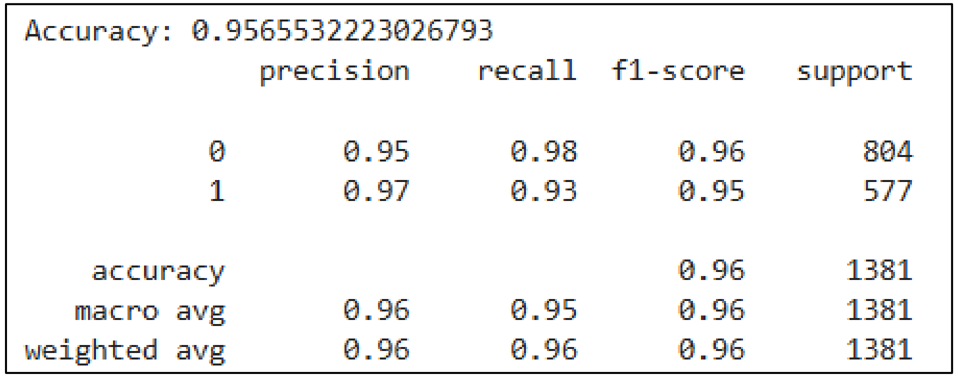
In this project, we focus on building a model to classify emails as spam or non-spam (ham) using machine learning techniques. The primary objective is to accurately identify spam emails, which are a significant source of unwanted information and often contain malicious content.For this project, we selected a publicly available dataset from Kaggle called the Spambase Dataset. This dataset contains various features derived from email content, such as word frequencies and special characters, which are used to predict whether an email is spam or not. To solve the problem, we applied two powerful machine learning algorithms:

- Decision Tree (DT): A tree-based model that splits the dataset into smaller, more manageable groups to make predictions based on features.
- Random Forest (RF): An ensemble method that combines multiple decision trees to improve prediction accuracy and reduce overfitting.

We evaluate the performance of both models and compare their results to identify the best solution for spam classification.

Results

we'll take a look at the results from both models, focusing on key metrics like Accuracy, Precision, Recall, and F1-Score for both Decision Tree (DT) and Random Forest (RF). These metrics will give us a clear picture of how well each model performed.

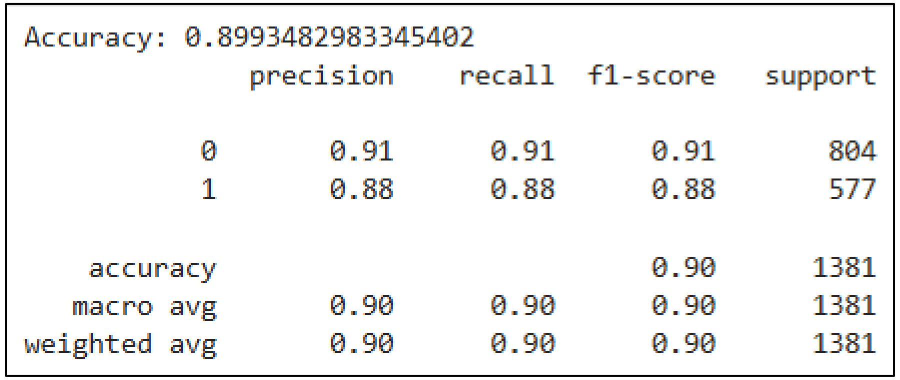


```
Accuracy: 0.9565532223026793
              precision    recall  f1-score   support

           0       0.95      0.98      0.96       804
           1       0.97      0.93      0.95       577

    accuracy                           0.96      1381
   macro avg       0.96      0.95      0.96      1381
weighted avg       0.96      0.96      0.96      1381
```

RF



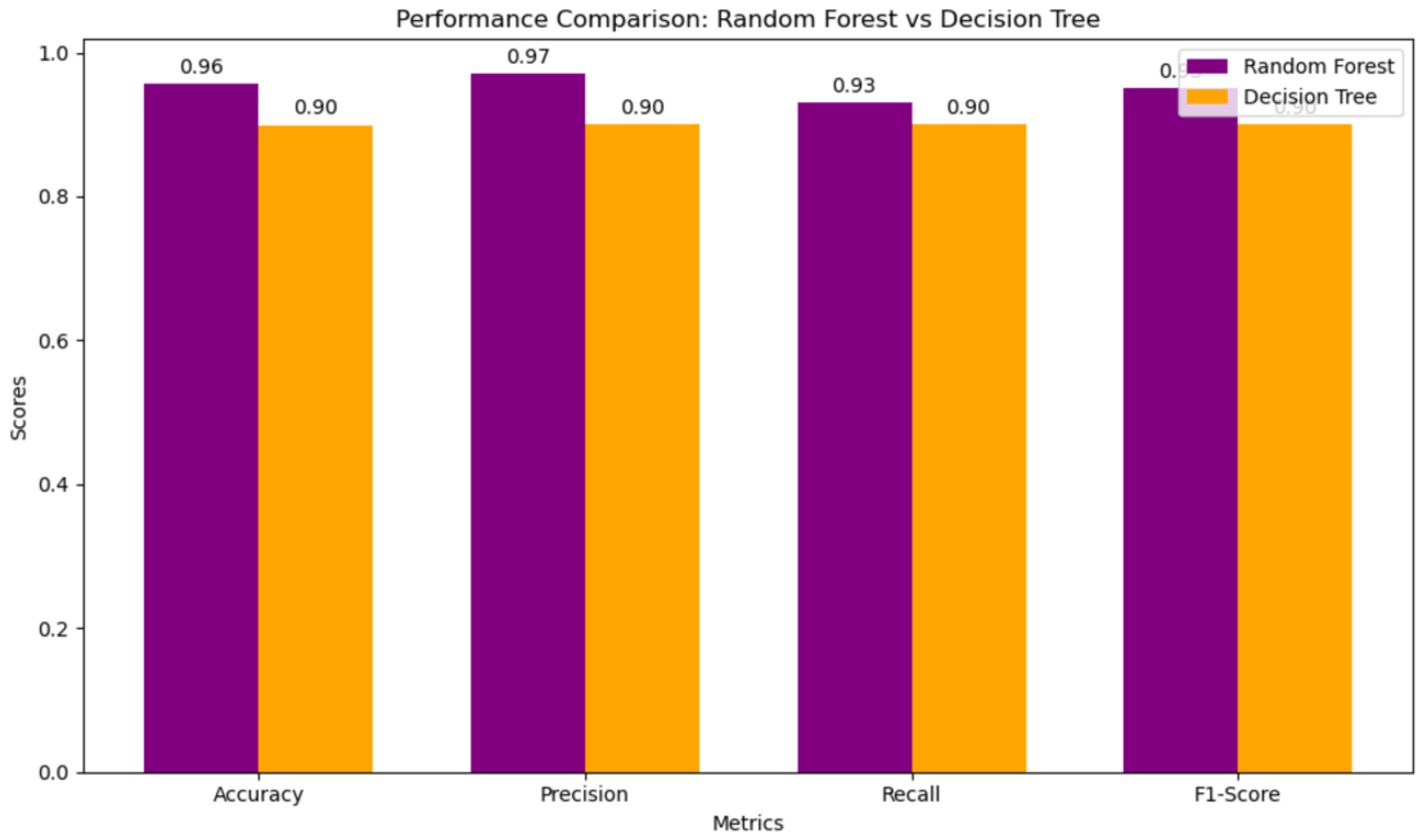
```
Accuracy: 0.8993482983345402
              precision    recall  f1-score   support

           0       0.91      0.91      0.91       804
           1       0.88      0.88      0.88       577

    accuracy                           0.90      1381
   macro avg       0.90      0.90      0.90      1381
weighted avg       0.90      0.90      0.90      1381
```

DT

Bar Chart


Performance Comparison: Random Forest vs Decision Tree
Random Forest
Decision Tree
Scores
1.0
0.8
0.6
0.4
0.2
0.0
0.96
0.90
0.97
0.90
0.93
0.90
0.95
0.90
Accuracy
Precision
Recall
F1-Score
Metrics

Analysis of Bar Chart

The bar chart displays a comparison between Random Forest (RF) and Decision Tree (DT) for four key performance metrics: Accuracy, Precision, Recall, and F1-Score.

As shown, Random Forest (RF) outperforms Decision Tree (DT) across all metrics, including accuracy, precision, recall, and F1-score.

Random Forest provides higher precision and recall, especially in identifying spam emails, and also maintains better overall accuracy.

Why RF is Better Than DT

- Random Forest is an ensemble method, meaning it uses multiple decision trees to make predictions. This reduces the chances of overfitting, which is a common issue with a single Decision Tree. The Decision Tree (DT) while is simple and easy to understand, can easily overfit the data, as it tries to make perfect splits in the data. This may lead to a lower recall, meaning it misses some actual spam emails. Random Forest, being an ensemble method, combines the results of multiple trees, which allows it to make more stable and reliable predictions. This explains its superior performance in comparison to Decision Tree (DT).

In conclusion, Random Forest is the better model for spam detection in this case, offering a more reliable, accurate, and stable solution compared to Decision Tree.