logo.png

Assignment Report
Spam Classification Evaluation

MTech Data Science

Submitted by: Ranjan Kumar
Roll No: 2422118
Supervised by: Prof. Ramanujam

November 28, 2024

Contents

1 Dataset Used

Dataset Details:

- Total samples: 2,000
- Class Distribution: 1,237 negative (ham), 763 positive (spam)
- Features: 8 features extracted from text messages.
- Remarks: The dataset is imbalanced.

1.1 Dataset Sample

The first five samples of the final dataset are shown below:

`dataset_sample.png`

Figure 1: Sample Dataset

1.2 Spam Distribution

- **Ham (Negative Messages):** 1,237 samples

- **Spam (Positive Messages):** 763 samples

2 Features Extracted

The following features were extracted from the messages:

- **message_length:** The length of the message.
- **word_count:** Total number of words in the message.
- **has_url:** Whether the message contains a URL.
- **has_phone:** Whether the message contains a phone number.
- **has_money:** Whether the message mentions monetary values.
- **special_char_count:** Number of special characters in the message.
- **spam_word_count:** Number of spam-related words present.
- **tfidf_score:** The Term Frequency-Inverse Document Frequency score of the message.

Figure 2: Feature Extraction Process

3 Classification and Evaluation

For classification, the Lazy Predict library was used to evaluate various machine learning models. The code and results are as follows:

3.1 Code Snippet

code_snippet.png

Figure 3: Code Snippet for Lazy Predict Evaluation

3.2 Results

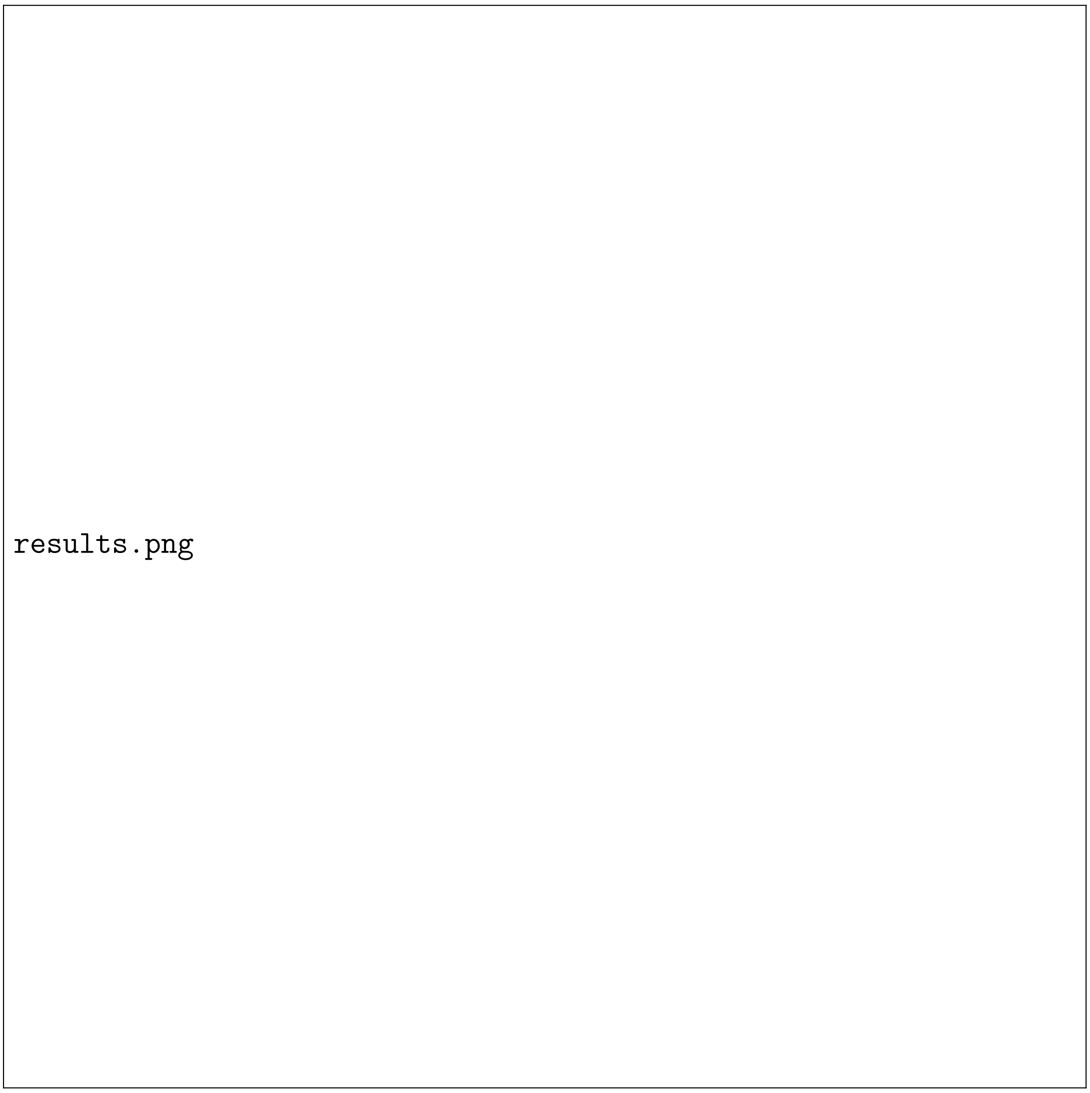

Figure 4: Lazy Predict Results

3.3 Observations

- The **ExtraTreesClassifier** achieved the highest accuracy (96%) and balanced accuracy (95%).
- The **RandomForestClassifier** achieved an accuracy of 92%.

4 Conclusion

The Lazy Predict evaluation provides a quick benchmarking of models. Based on the results:

- The **ExtraTreesClassifier** is recommended for its high accuracy and balanced accuracy.

- Future work could include addressing class imbalance and testing ensemble methods.