Machine learning project

SPAM SMS DETECTION

Dataset Link :

https://www.kaggle.com/datasets/uciml/sms-spam-collection-dataset

https://raw.githubusercontent.com/AbhishekYadav-01/Encryptix/main/SPAM_SMS_DETECTION/spam.csv

Link of the google collab file :

https://colab.research.google.com/drive/1u-6Bz6Ts6s9LV8Jx7Rwgb8H1LbKYSbwk?usp=sharing

Github Repository Link :

https://github.com/AbhishekYadav-01/Encryptix/tree/main/SPAM_SMS_DETECTION

Aim :

Build an AI model that can classify SMS messages as spam or legitimate. Use technique TF-IDF with classifiers Naive Bayes, Logistic Regression, or Support Vector Machines to identify spam messages.

Dataset Loading :

The dataset used for this project is sourced from Kaggle. It consists of SMS messages labeled as 'spam' or 'ham' (legitimate).Number of samples: 5576

This is how the dataset look like :

```
ham     Go until jurong point, crazy.. Available only in bugis n great world la e buffet... Cine there got amore wat...
ham     Ok lar... Joking wif u oni...
spam    Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receive entry question(std txt rate)T&C's apply 08452810075over18's
ham     U dun say so early hor... U c already then say...
ham     Nah I don't think he goes to usf, he lives around here though
spam    FreeMsg Hey there darling it's been 3 week's now and no word back! I'd like some fun you up for it still? Tb ok! XxX std chgs to send, å£1.50 to rcv
```

Data Preprocessing :

After loading the dataset, the following preprocessing steps were undertaken:

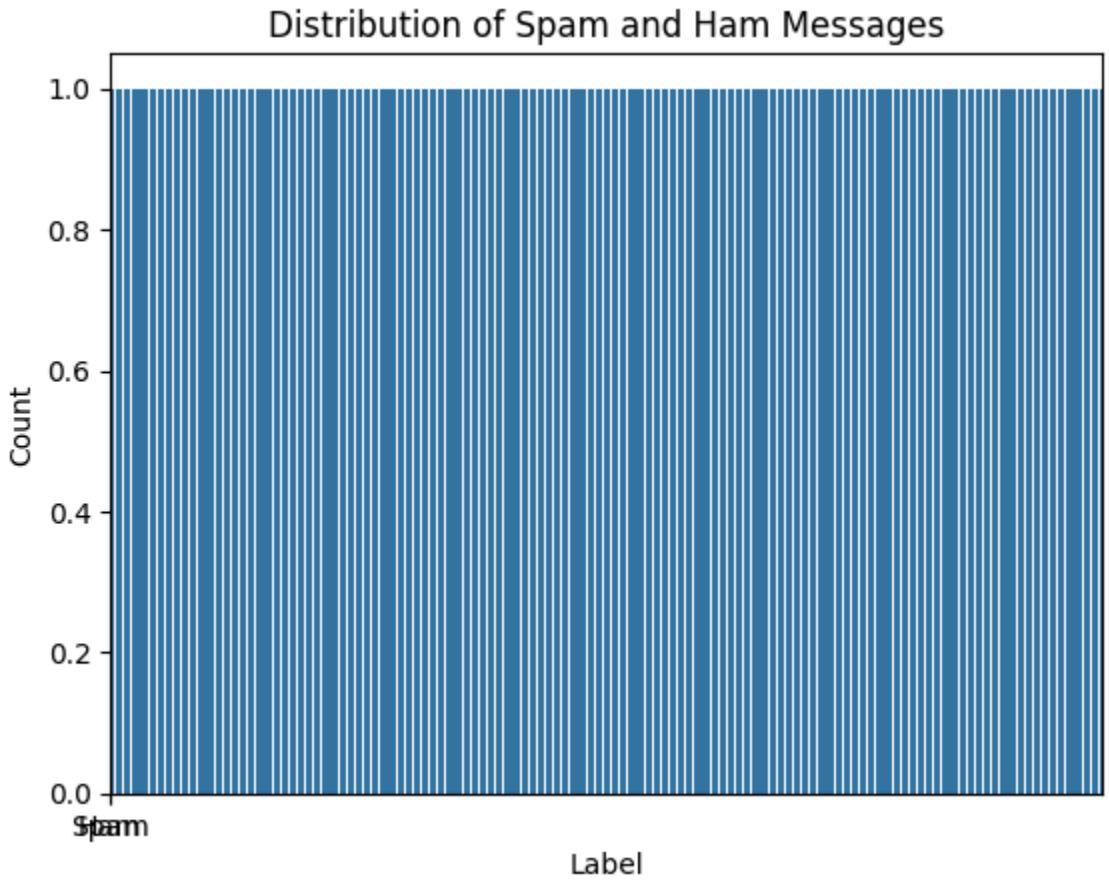
- Checking for missing values if any missing values then Dropping rows with missing values but we don't got any missing row.
- Mapping labels to binary values: 'ham' -> 0, 'spam' -> 1.
- Checking for any empty messages if their are any empty messages then Remove it but we don't got any empty messages from the dataset.

Distribution of Messages :

After these steps we got :

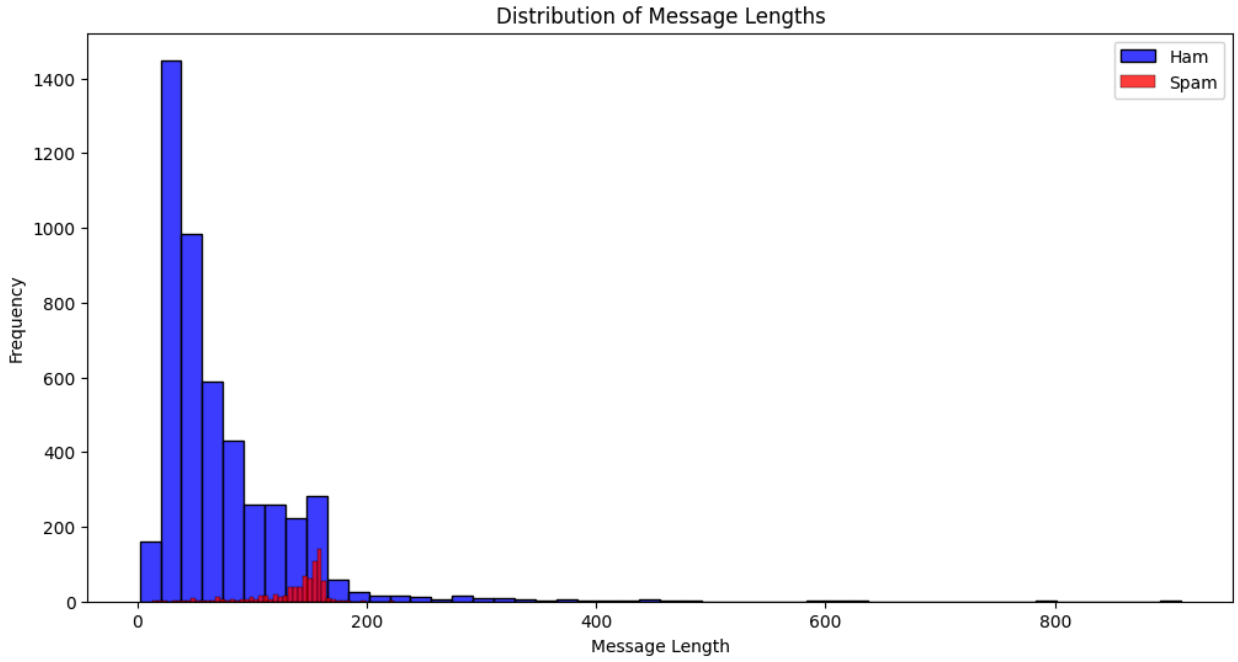
Label	Count
0	4829
1	747

This is the distribution of the Spam and Ham messages :



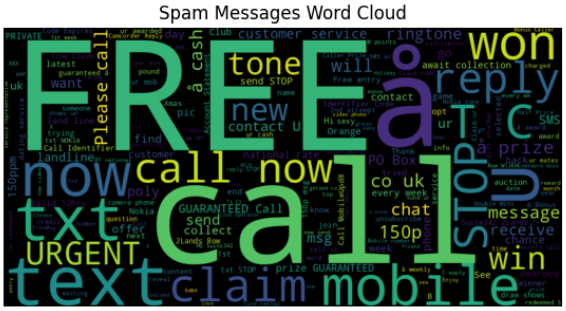
Message Length Analysis :

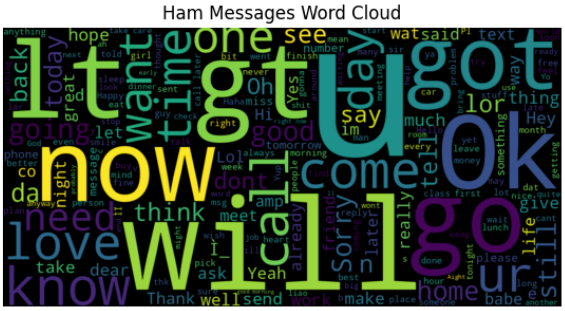
Here , we calculated message lengths and plotted their distributions for both spam and ham messages.



Word Clouds :

This is the Generated word clouds to visualize frequently occurring words in spam and ham messages.





TF-IDF Vectorization :

We have done this step to Utilize TF-IDF Vectorizer to transform text data into numerical features.

Model Selection and Evaluation :

We have trained these models -

- Naive Bayes
- Logistic Regression
- SVM (Linear Kernel)
- SVM (Polynomial Kernel)
- SVM (RBF Kernel)
- Logistic Regression Model with hyparameter tuning

And after training with 80% data we tested with 20% data and got these result :

Naive Bayes Model :
Accuracy: 0.98
Precision: 0.99
Recall: 0.86
F1 Score: 0.92

Logistic Regression Model :
Accuracy: 0.95
Precision: 0.98
Recall: 0.68
F1 Score: 0.80

SVM Linear Kernel Model :
Accuracy: 0.98
Precision: 0.96
Recall: 0.86
F1 Score: 0.91

SVM Polynomial Kernel Model :
Accuracy: 0.94
Precision: 1.00
Recall: 0.58
F1 Score: 0.73

SVM RBF Kernel Model :
Accuracy: 0.98
Precision: 1.00
Recall: 0.86
F1 Score: 0.92

Logistic Regression Model with hyparameter tuning :
Accuracy: 0.98
Precision: 0.99
Recall: 0.90
F1 Score: 0.94

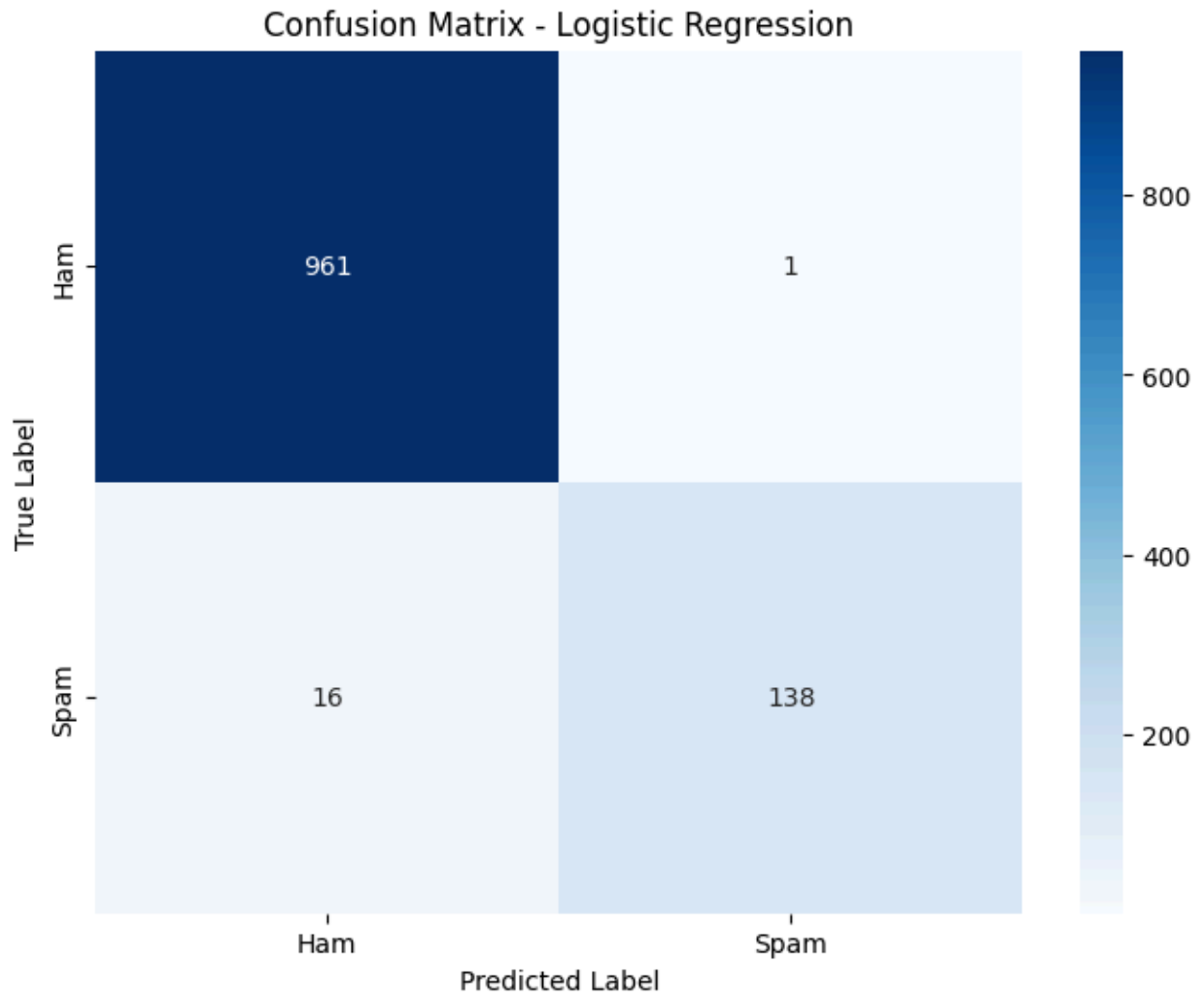
Here is the some actual vs predicted :

```
Model: Naive Bayes
Actual vs Predicted:
      Actual  Predicted
3690       0          0
3527       0          0
724        0          0
3370       0          0
468        0          0
...      ...        ...
4864       0          1
3227       0          0
3796       0          0
2879       0          0
1350       0          0

[1116 rows x 2 columns]
```

```
Model: Logistic Regression
Actual vs Predicted:
      Actual  Predicted
3690       0          0
3527       0          0
724        0          0
3370       0          0
468        0          0
...
1350       0          0

[1116 rows x 2 columns]
```

confusion matrix :

This is the confusion matrix for the Logistic Regression Model after applying hyparameter :



Classification Report :

Detailed classification report for Logistic Regression Model with hyparameter tuning , showing precision, recall, and F1 score for each class (Ham and Spam) :

	precision	recall	f1-score	support
Ham	0.98	1.00	0.99	962
Spam	0.99	0.90	0.94	154
accuracy			0.98	1116
macro avg	0.99	0.95	0.97	1116
weighted avg	0.98	0.98	0.98	1116

Cross-Validation :

- Conducted 10-fold cross-validation on the Logistic Regression model with hyparameter tuning.
- Obtained mean accuracy of 0.98 with a standard deviation of 0.00.

Cross-Validation Accuracy Scores:

- 0.99462366
- 0.98028674
- 0.98387097
- 0.98566308
- 0.98028674
- 0.98566308
- 0.98384201
- 0.97845601
- 0.98743268
- 0.98922801

Mean Cross-Validation Accuracy: 0.98
Standard Deviation of Cross-Validation Accuracy: 0.00__

Models Comparison based on accuracy :

Model	Accuracy
Logistic Regression	0.95
Naive Bayes	0.98
SVM (linear)	0.98
SVM (poly)	0.94
SVM (rbf)	0.98
Logistic Regression(hyp. tuning)	0.98

Conclusion :

Finally the project successfully completed and we have developed an AI model for spam SMS detection using machine learning techniques. The best-performing model, Logistic Regression Model with hyparameter tuning and SVM linear model both achieved high accuracy 98% and robustness across various evaluation metrics. The results suggest that the TF-IDF approach coupled with Logistic Regression is effective for distinguishing between spam and legitimate SMS messages.We are getting excellent precision for almost all models especially in svm rbf we are getting precision 1.0 ,it tells us that almost all the predicted labels were actually correct.