IDA & ML I
Spam Filter

Juan Danza
Potsdam University
Sep 18, 2025

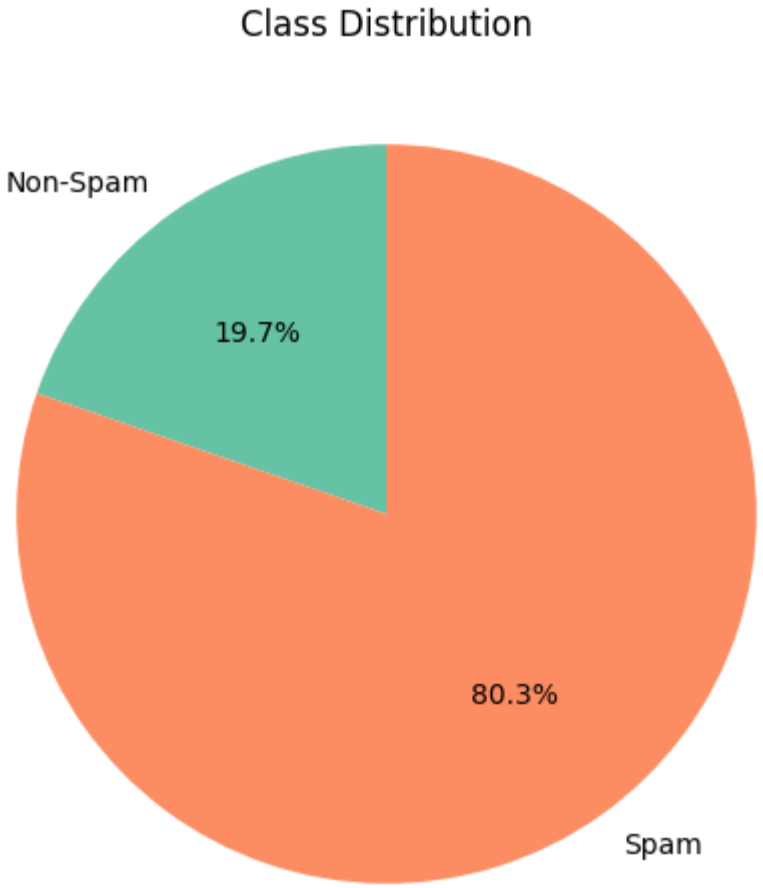
Problem Setting

Spam Detection: Binary Classification Problem (Supervised)

- **Input Attributes**: Numerical Count of Words
- **Target Variable**: Spam or not Spam (1/0)
- **Requirements**: Less than 0.2% False Positive Rate, Highest Recall as possible.

Data Overview

- 10000 examples.
- 57.173 features.
- Bag of words (independent).
- Class distribution?
  Are we under covariate shift?



Feature Extraction

- Already Bag of Words
- n-Grams and vectorization of words are not a choice.
- TF-IDF

$$TF(w) = \frac{Number\ of\ times\ word\ w\ appears\ in\ a\ document}{Total\ number\ of\ words\ in\ the\ document}$$

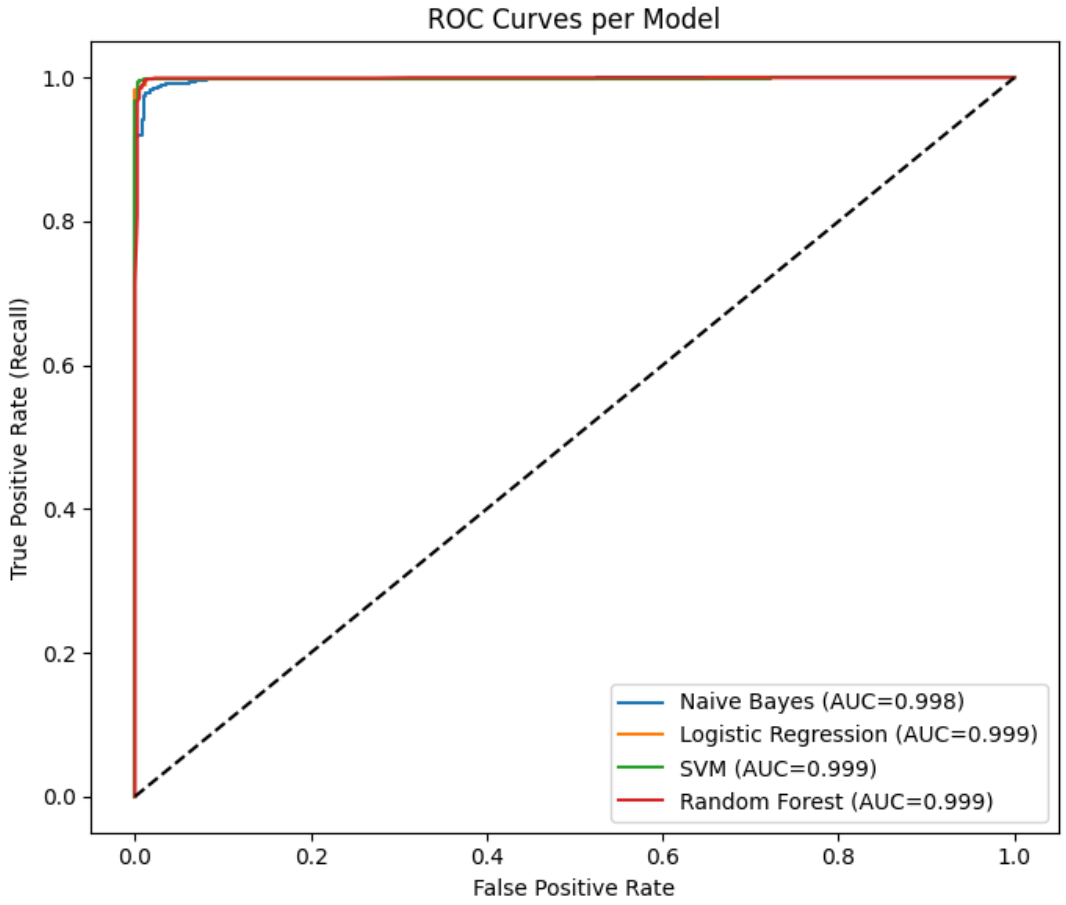
$$IDF(w) = \log\left(\frac{Total\ number\ of\ documents}{Number\ of\ documents\ that\ contain\ the\ word\ w}\right)$$

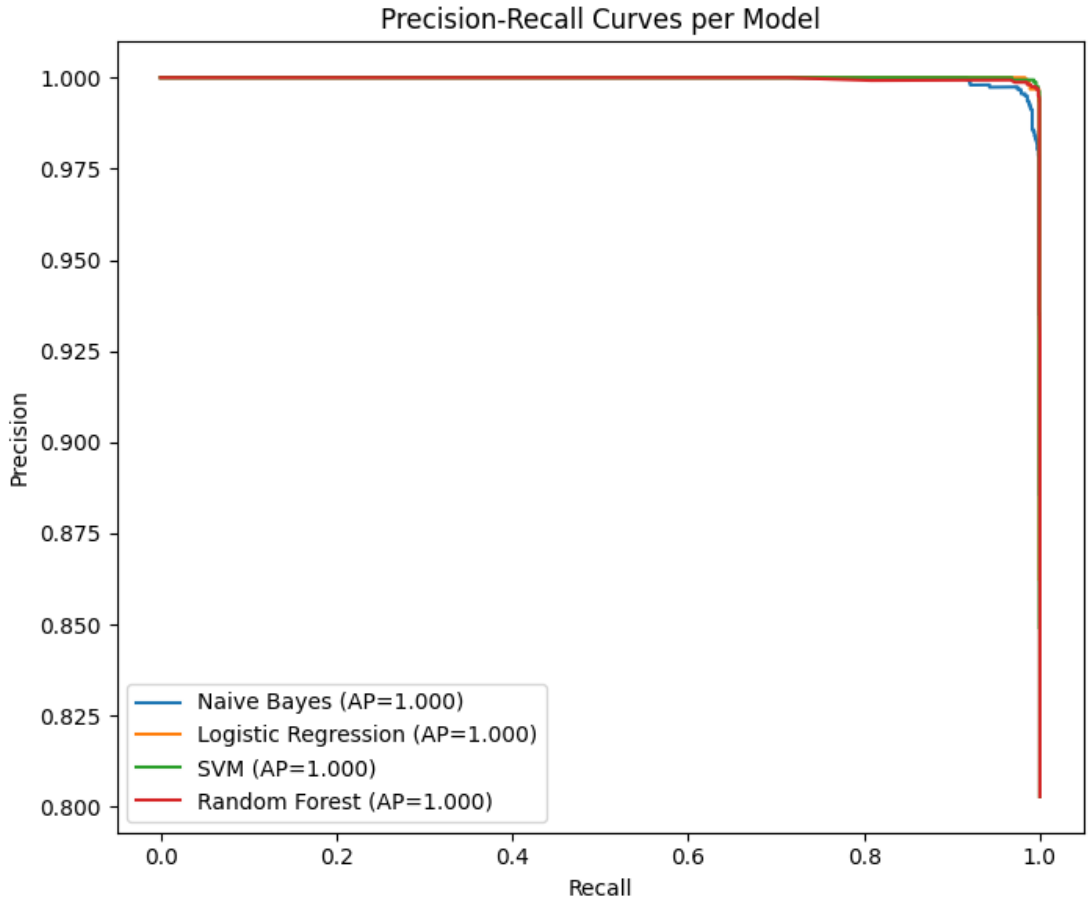
Model Selection

- Possible models according to the literature*:
  - Naïve Bayes
  - Random Forest
  - Logistic Regression
  - SVM

*Kaddoura, S., Chandrasekaran, G., Popescu, D.E., & Duraisamy, J.H. (2022). A systematic literature review on spam content detection and classification. *PeerJ Computer Science, 8*.,
Muath AlShaikh, Yasser Alrajeh, Sultan Alamri, Suhib Melhem & Ahmed Abu-Khadrah (2025) Supervised methods of machine learning for email classification: a literature survey, Systems Science & Control Engineering, 13:1, DOI: 10.1080/21642583.2025.2474450

Model Selection

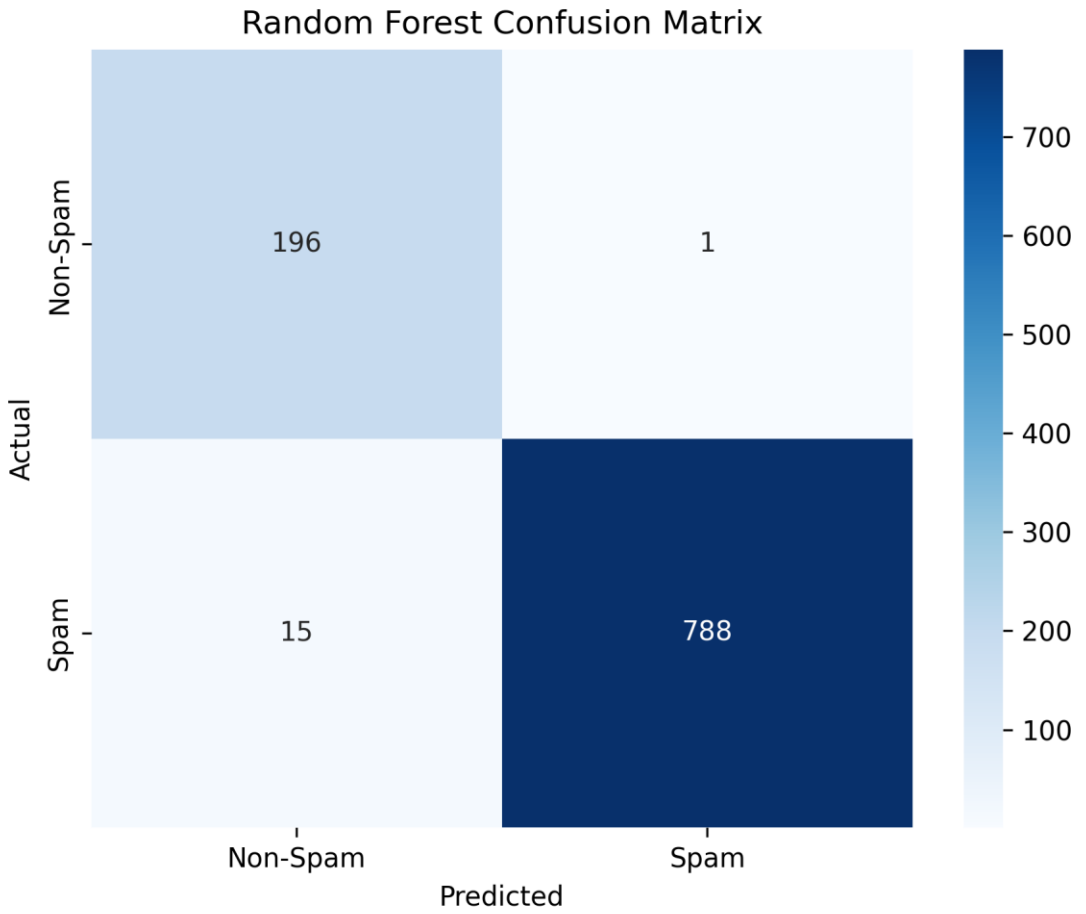


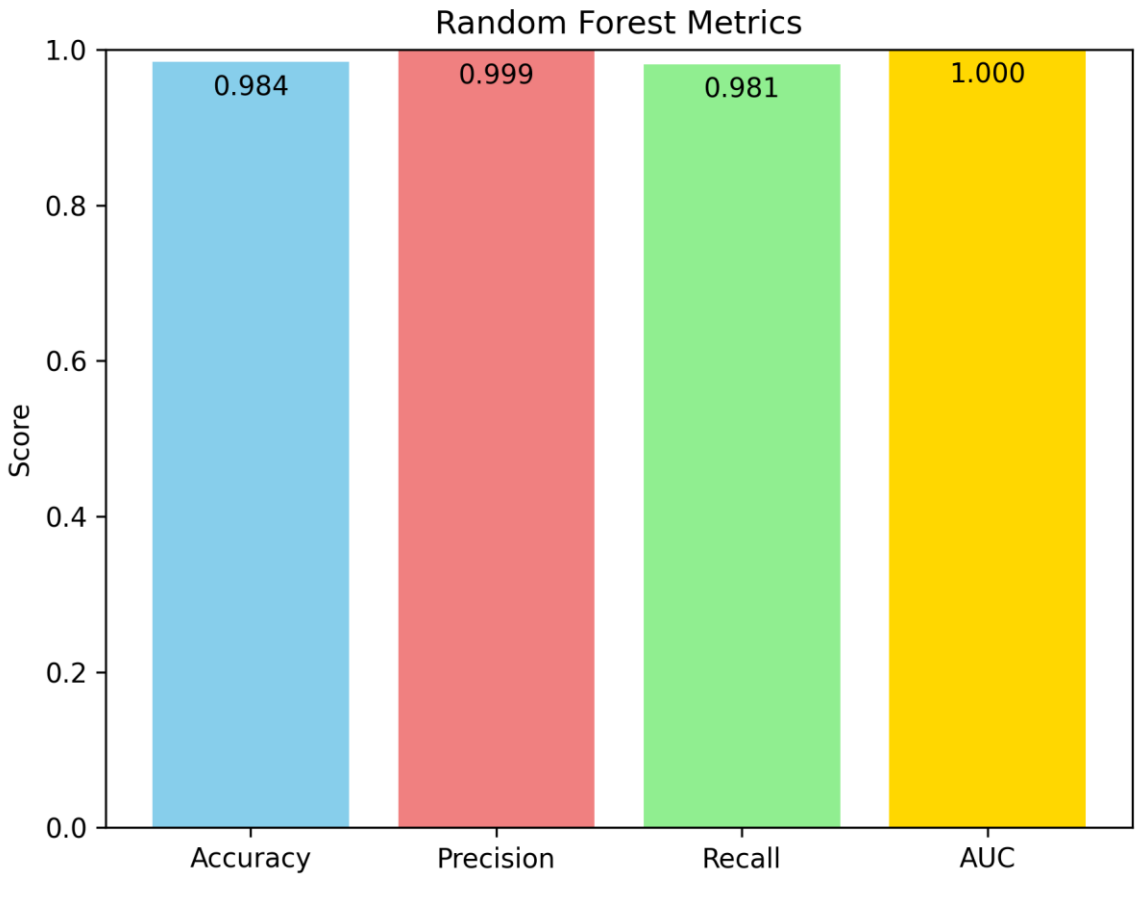


Model Selection

```
=== Average AUC across seeds ===
Naive Bayes: 0.9987 ± 0.0003
Logistic Regression: 0.9992 ± 0.0005
SVM: 0.9995 ± 0.0004
Random Forest: 0.9995 ± 0.0006
```

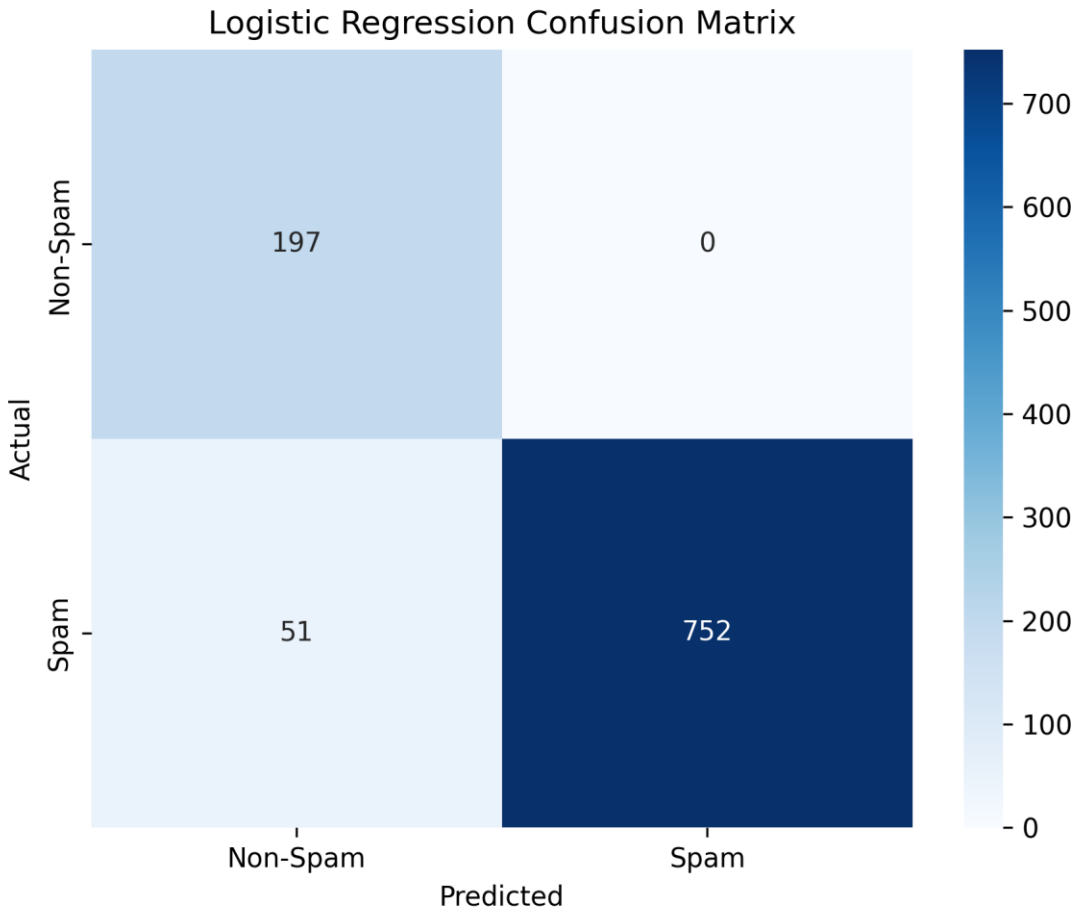
Results RF

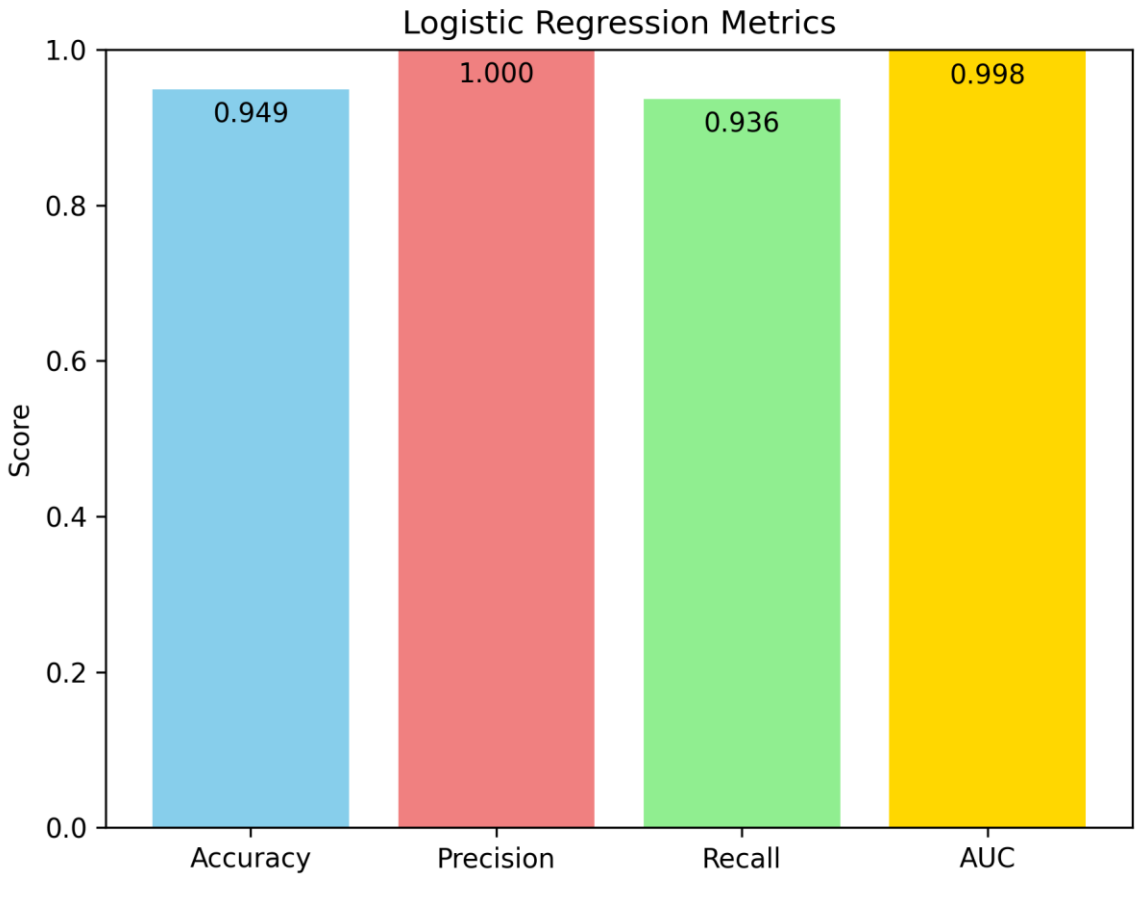




FP Rate: 0.51%

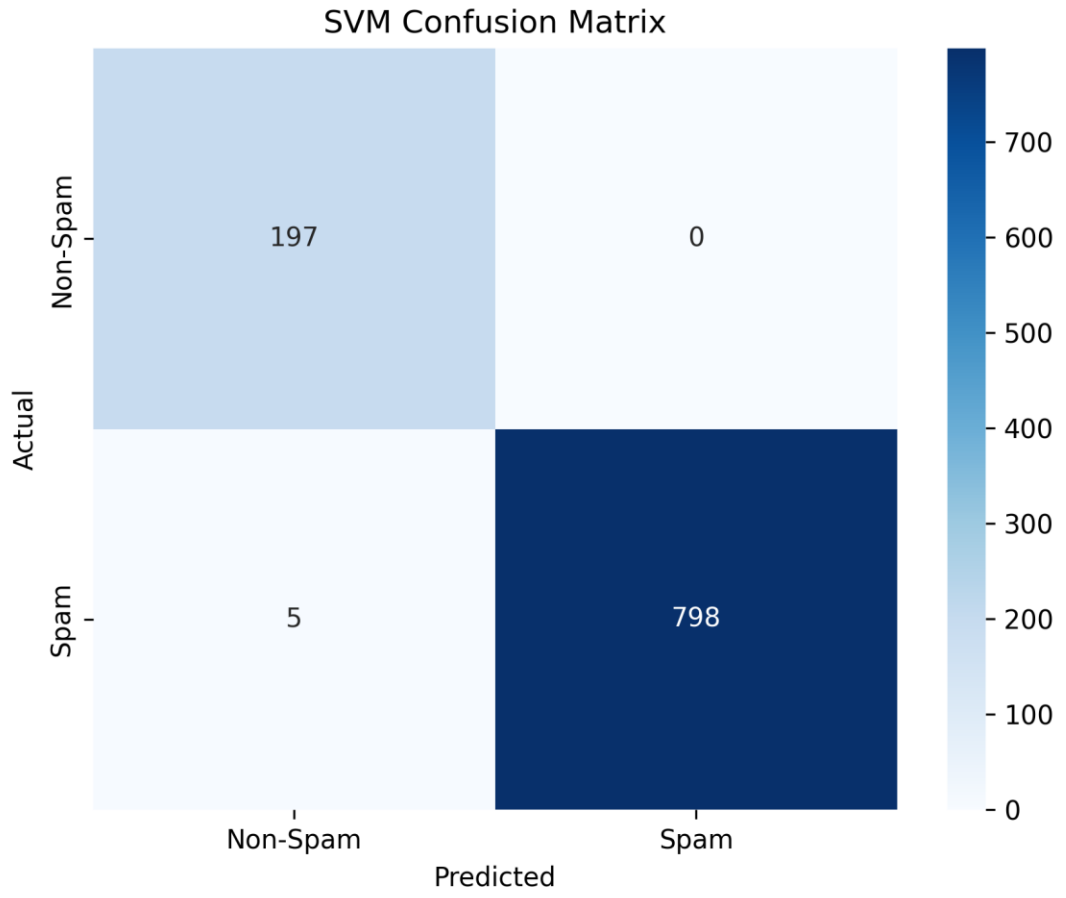
Results Logistic Regression

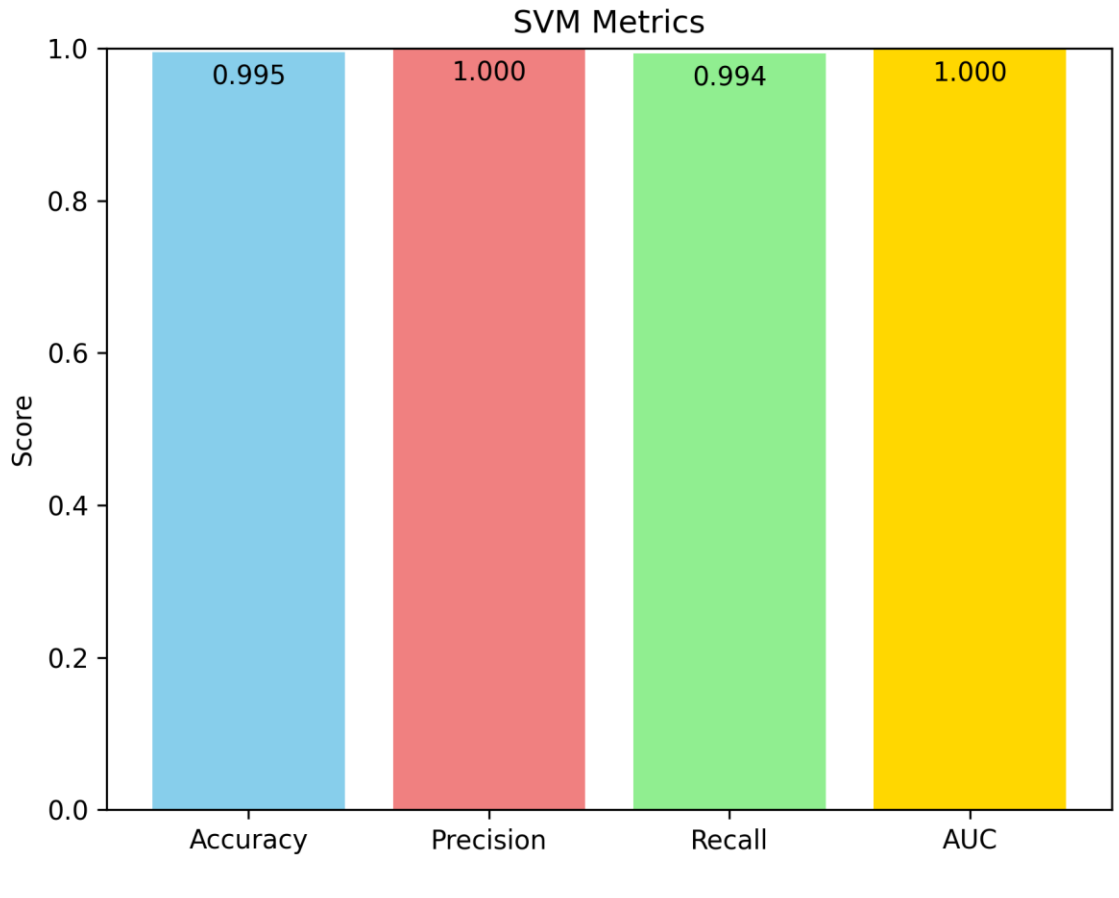




FP Rate: 0.0%

Results SVM





FP Rate: 0.0%