Ministry of Science and Higher Education of the Republic of Kazakhstan
L.N. Gumilyov Eurasian National University

Faculty of Information Technology
Department of Information Systems

COURSEWORK

ON THE SUBJECT
"Mathematical Foundations of Intelligent Systems"
For third-year students of the specialty 6B06103 – Information Systems

Topic: Detecting spam and phishing emails using machine learning models

Completed by:
Student of group IS-35, Iskakov Y.K.
Coursework Supervisor: Prof, Zhukabayeva T. K.

TABLE OF CONTENTS

INTRODUCTION

Goal

Develop a machine learning system with a client application and PostgreSQL database for accurate spam detection.

Objectives

- Apply advanced techniques, models from existing literature.
- Build a practical, real-world application.
- Ensure high accuracy and documentation.

Tasks

- Make literature review.
- Preprocess labeled email datasets.
- Implement and optimize machine learning models.
- Deploy a client application for real-time detection.
- Evaluate performance and document findings.

INTRODUCTION

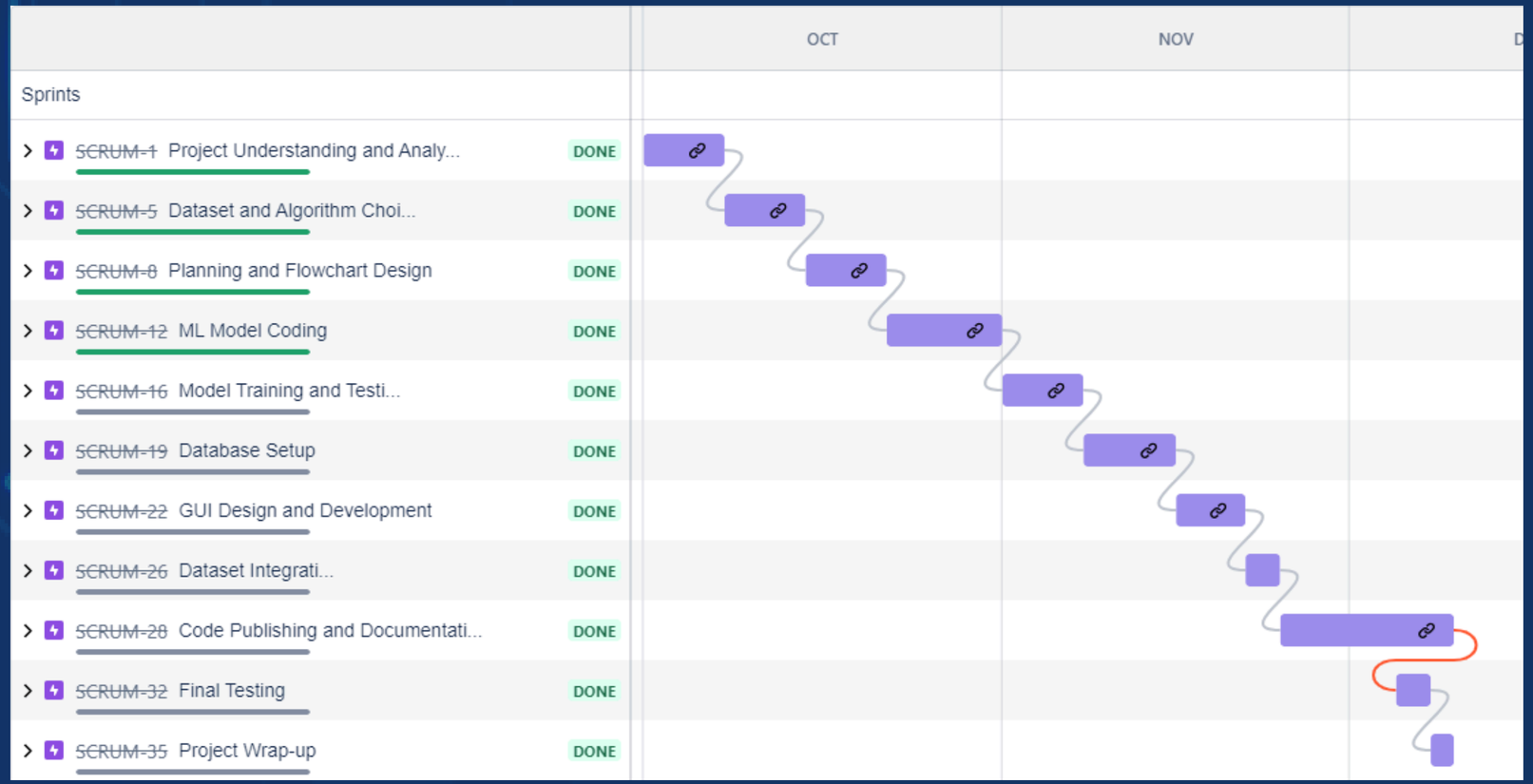


Figure 2.3. Project Timeline

https://iskakovk2016.atlassian.net/jira/software/projects/SCRUM/boards/1/timeline

LITERATURE REVIEW

Algorithms	Key Findings	Studies
Naive Bayes	- High accuracy but struggles with class-conditional independence assumptions. - Effective when enhanced with ensemble methods. - Performs well with text mining (e.g., TF-IDF).	Kumar et al. (2020), Sabiq et al. (2024), Singh et al. (2019), Shahzad and Ali (2018), Vijay and Verma (2023)
Random Forest	- Handles complex and imbalanced datasets effectively. - Often outperforms Naive Bayes and SVM in certain scenarios. - High precision, recall, and F1-score.	Cota and Zinca (2022), Saini et al. (2023), Negi et al. (2023), Gattani et al. (2023)
Support Vector Machine	- Excels in high-dimensional spaces. - Benefits significantly from ensemble methods like stacking. - Paired with TF-IDF for enhanced accuracy.	Negi et al. (2023), Al-Shanableh et al. (2024)
Deep Learning (LSTM, RNN)	- Superior at learning complex patterns in large datasets. - LSTM and RNN outperform traditional models. - Effective for non-linear relationships.	Siddique et al. (2021), Hossain et al. (2021)
Feature Engineering	- Improved Naive Bayes using term-weight aggregation and collaborative filtering. - Boosts classifier performance via feature selection and vectorization.	Song et al. (2009), Takci and Fatema (2023)
Ensemble Learning	- Combines strengths of multiple classifiers. - Significantly improves accuracy (e.g., +25% in Arabic spam detection). - Stacking is particularly effective	Saeed et al. (2022), Al-Shanableh et al. (2024)

METHODOLOGY

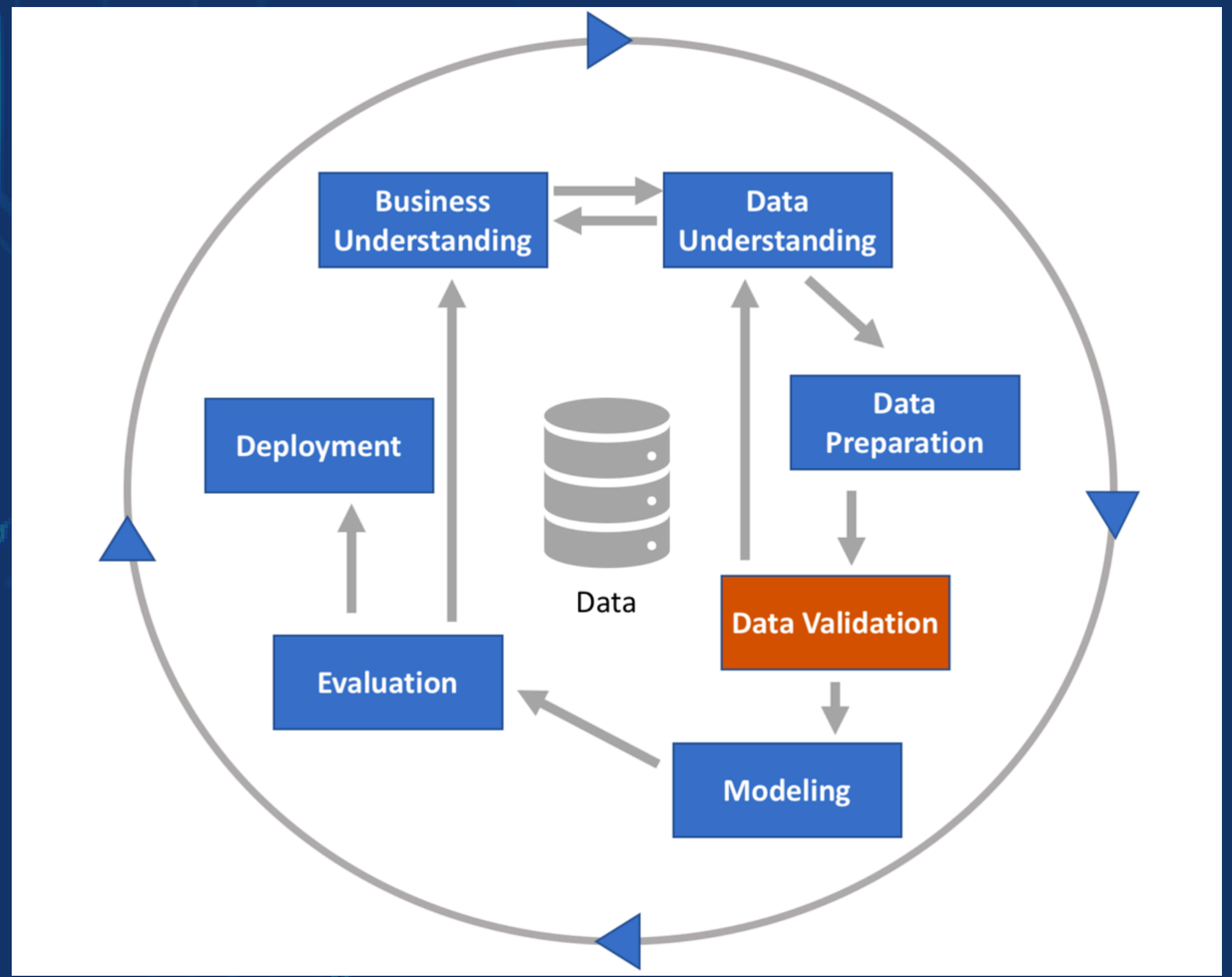


FIGURE 2.1 CRISP-DM LIFE CYCLE

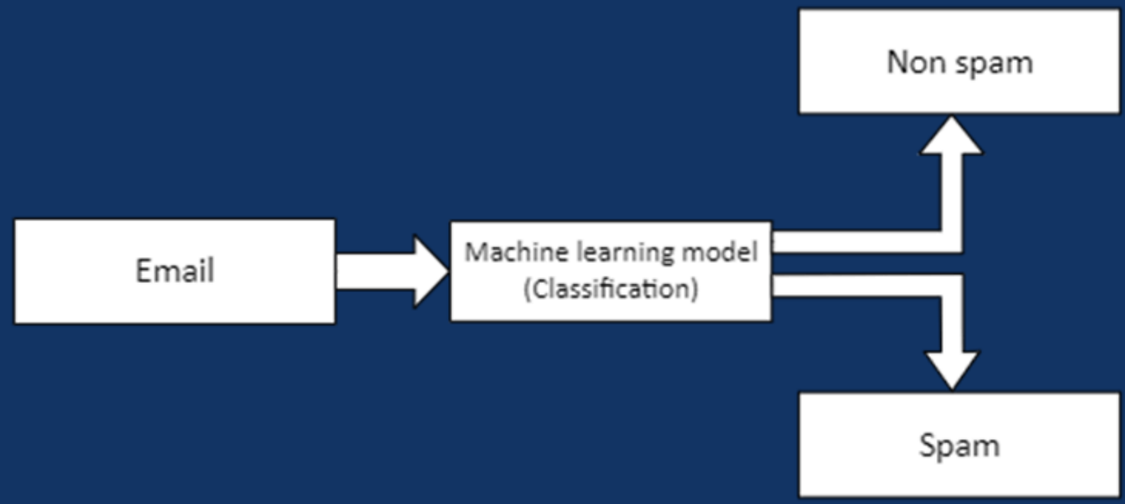


FIGURE 2.2. CLASSIFICATION DIAGRAM

DATASET

IN THIS STUDY, WE ASSESS THE MODELS' PERFORMANCE USING THE TWO DATASETS. THESE FILES ARE IN THE COMMONLY USED CSV FORMAT FOR DATA ANALYSIS PROCEDURES.

	label	message
0	0	Go until jurong point, crazy.. Available only ...
1	0	Ok lar... Joking wif u oni...
2	1	Free entry in 2 a wkly comp to win FA Cup fina...
3	0	U dun say so early hor... U c already then say...
4	0	Nah I don't th
...	...	
5567	1	This is the 2nd
5568	0	Will ì
5569	0	Pity, * was in
5570	0	The guy did s
5571	0	

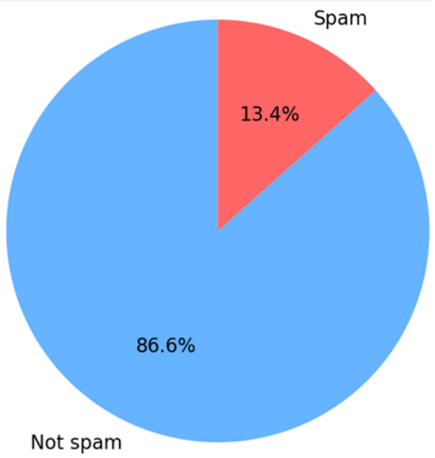


FIGURE 3.1. REPRESENTATION OF FIRST DATASET.

	message	label
0	Supply Quality China's EXCLUSIVE dimensions at...	1
1	over. SidLet me know. Thx.	0
2	Dear Friend,Greetings to you.I wish to accost ...	1
3	MR. CHEUNG PUIHANG SENG BANK LTD.DES VOEUX RD....	1
4	Not a surpr	
...		
4840	It has been pre	
4841		
4842	<html><head><style	
4843	Robinson talkin	
4844	PIR <preines(Wedn	

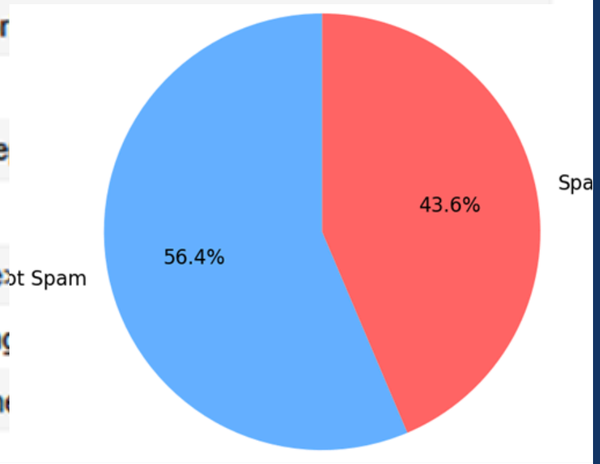


FIGURE 3.2. REPRESENTATION OF SECOND DATASET

DATA PROSSESING

FIGURE 3.4 FLOW CHART OF METHODOLOGY

Data preprocessing involves cleaning and preparing raw text data for analysis. This includes removing unnecessary characters (such as punctuation marks and stop words), converting text to lowercase, and normalizing it using stemming or lemmatization.

(TF-IDF) is a feature extraction method that estimates the importance of words in a document relative to the entire dataset.

DATA PROSSESING

FIGURE 3.5 DATA PREPROCESSING STEPS

Data preprocessing includes tokenization, stopword removal, stemming, ensuring the text is clean and transformed for effective analysis (see fig. 3.5).

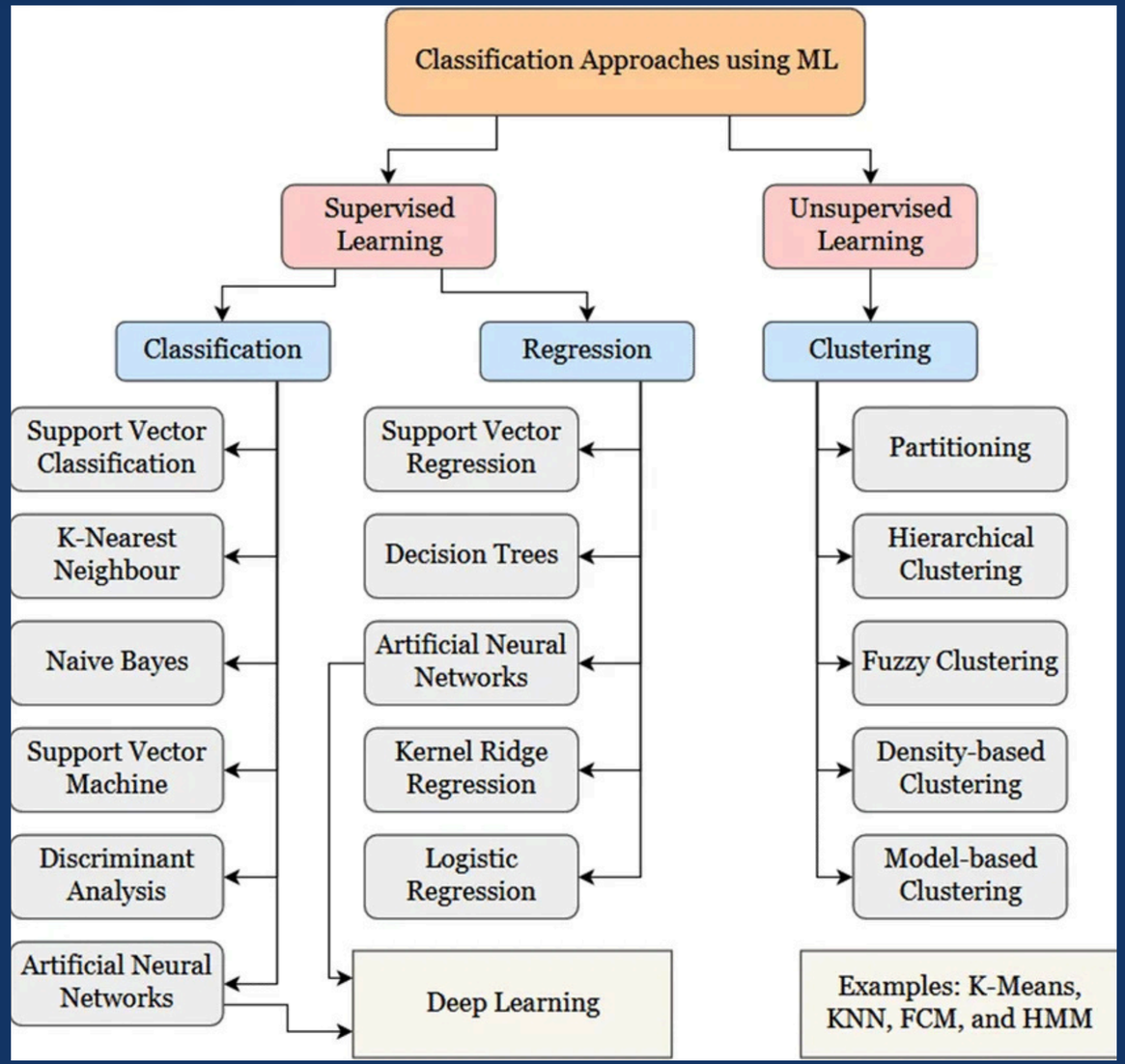
$$x_i' = S(T(L(x_i)) - S).$$

The goal of this step is to ensure that text is clean and standardized, which will allow for efficient feature extraction and improve the performance of machine learning models.

THEORETICAL FOUNDATION

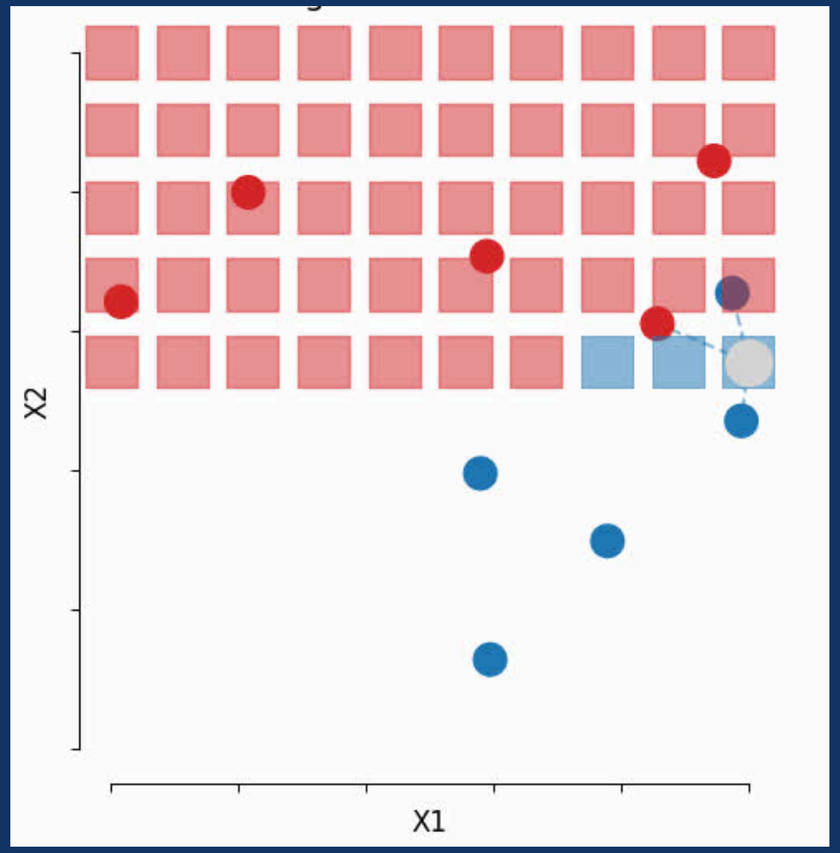
Machine learning makes use of computational techniques to enhance performance or forecast outcomes based on historical data, such as labeled datasets or interactions with the environment.

Classification is a supervised learning task where the model predicts discrete labels (categories) for each input. Examples include spam detection, sentiment analysis, or image categorization.



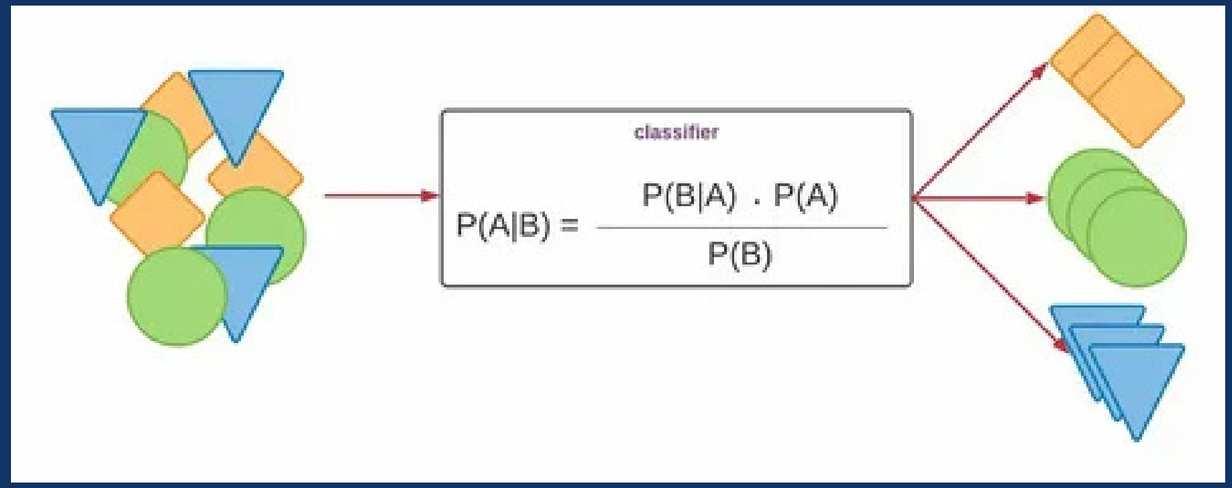
THEORETICAL FOUNDATION

A straightforward, non-parametric classification/regression algorithm called k-Nearest Neighbors bases its predictions on the k nearest data points in the training set. By using a majority vote of its neighbors, it assigns a new data point to the most prevalent class among its k closest neighbors. Ideal for datasets that are smaller and have distinct decision boundaries.



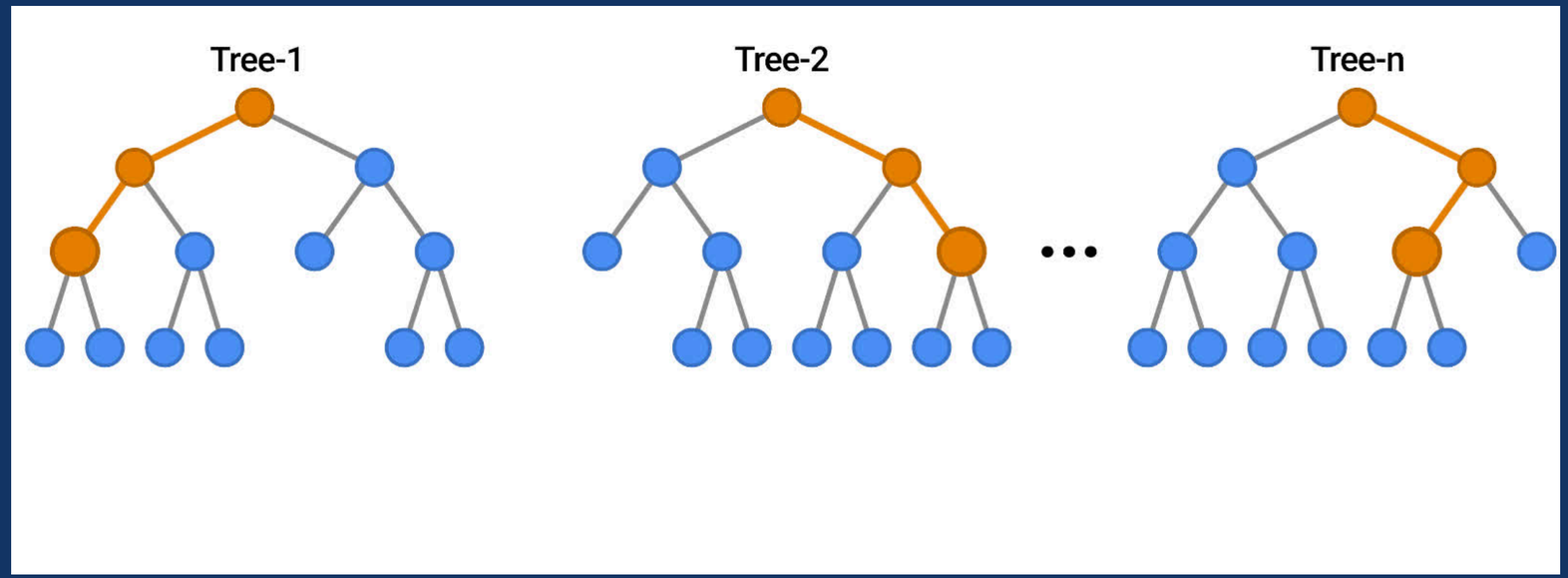
THEORETICAL FOUNDATION

Based on Bayes' theorem, Multinomial Naive Bayes (MNB) is a probabilistic classification algorithm. It performs especially well with categorical data, particularly text classification tasks, and makes the assumption that features are conditionally independent.



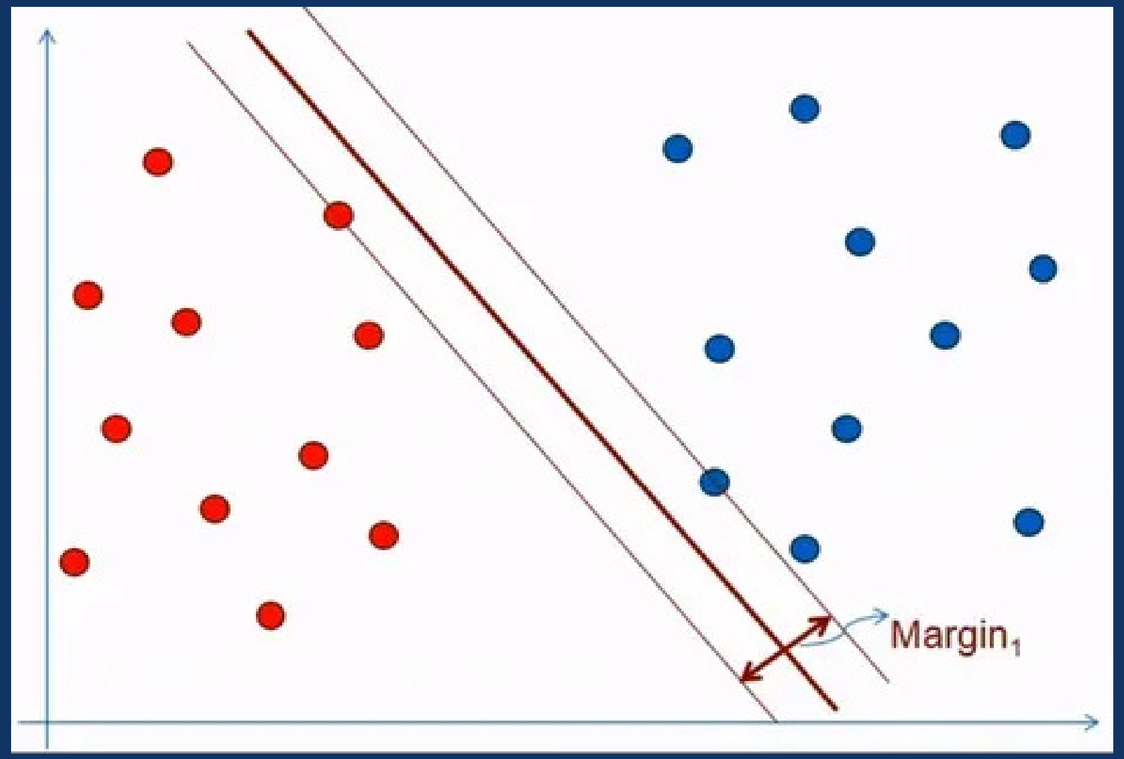
THEORETICAL FOUNDATION

Random Forest is an ensemble learning technique that builds several decision trees during training and produces a class that is the mean prediction (regression) or the mode of the classes (classification) of the individual trees.



THEORETICAL FOUNDATION

One supervised learning algorithm that determines the best hyperplane to divide classes with the greatest margin is called Support Vector Machine (SVM).



MACHINE LEARNING MODEL IMPLEMANTATION

Libraries we used

```
# SPLIT DATA INTO FEATURES AND LABELS
X = DF['MESSAGE']
Y = DF['LABEL']
TFIDF = TFIDFVECTORIZER(MAX_FEATURES=3000)
X = TFIDF.FIT_TRANSFORM(X)
```

```
X_TRAIN, X_TEST, Y_TRAIN, Y_TEST =
TRAIN_TEST_SPLIT(X, Y, TEST_SIZE=0.2,
RANDOM_STATE=42)
```

Transform the text data into numerical vectors in the feature extraction TF-IDF step

We divided the dataset into training and testing sets, using a fixed random seed (random_state=42) to ensure reproducibility and 80% of the data for training and 20% for testing.

MACHINE LEARNING MODEL PERFORMANCE

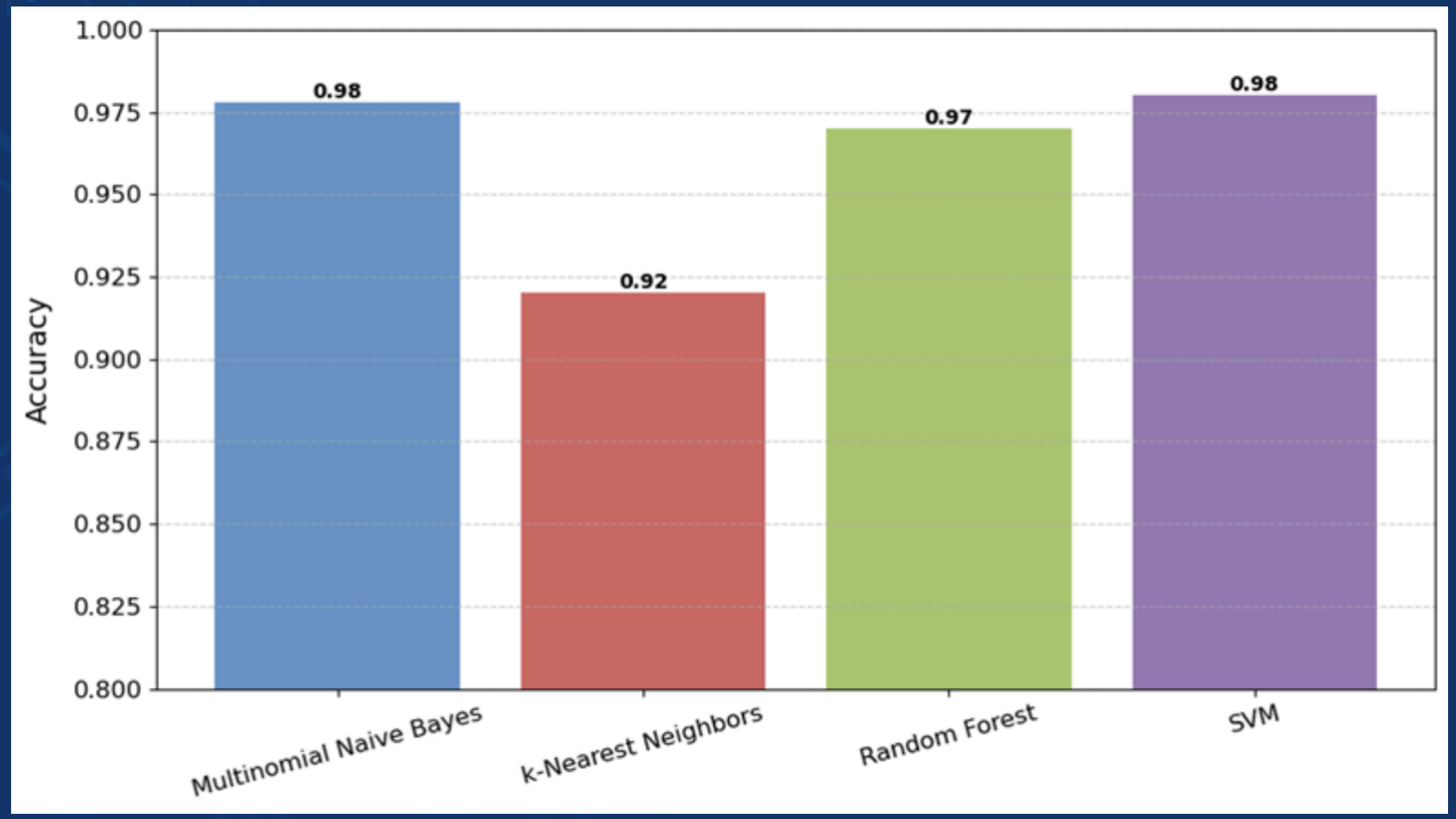


FIGURE 4.1. ACCURACY COMPARISON ACROSS DIFFERENT MODELS FOR FIRST DATASET.

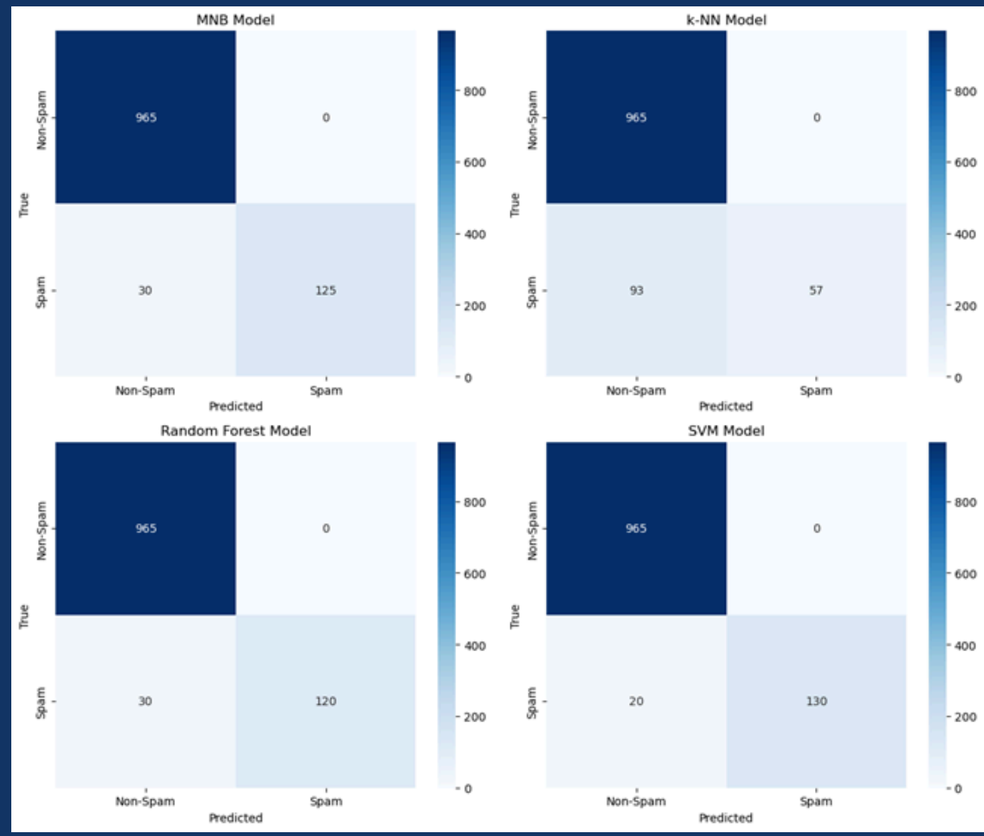


FIGURE 4.2. CONFUSION MATRIX FOR FIRST DATASET.

MACHINE LEARNING MODEL PERFORMANCE

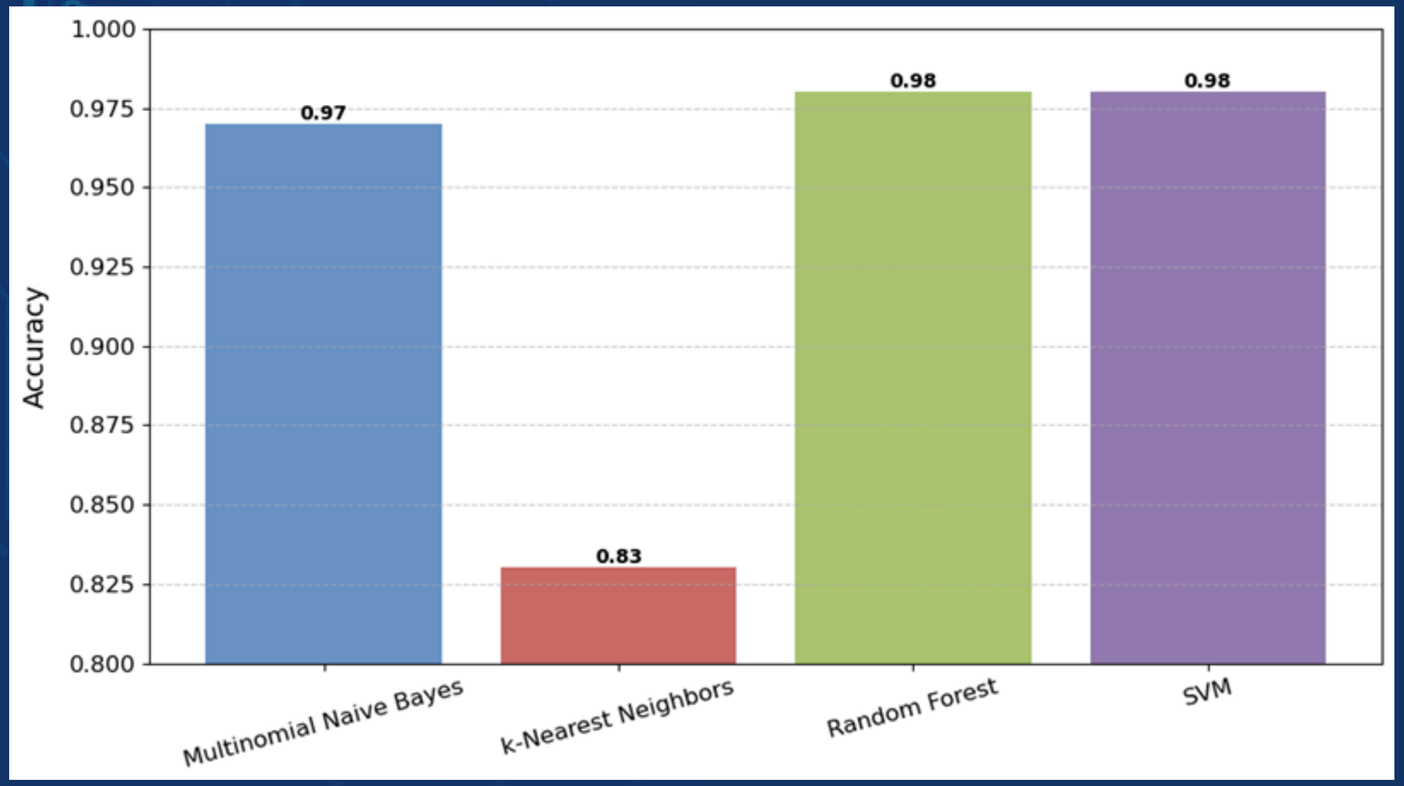


FIGURE 4.3. ACCURACY COMPARISON ACROSS DIFFERENT MODELS FOR SECOND DATASET.

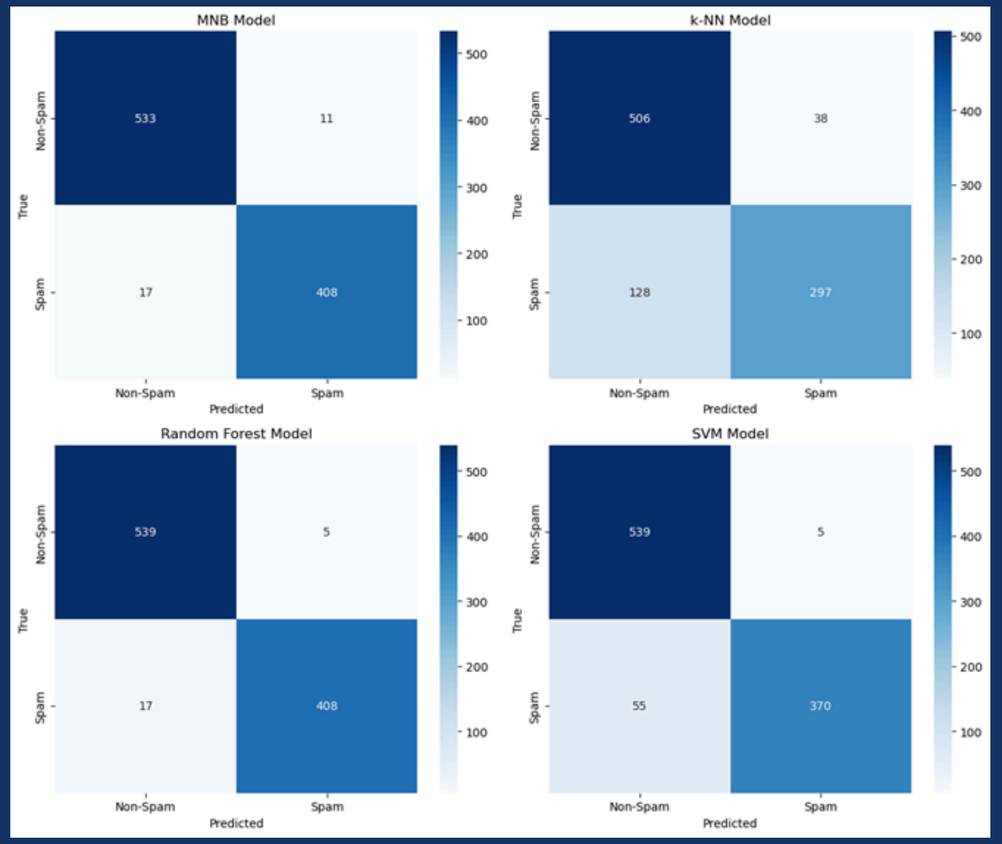
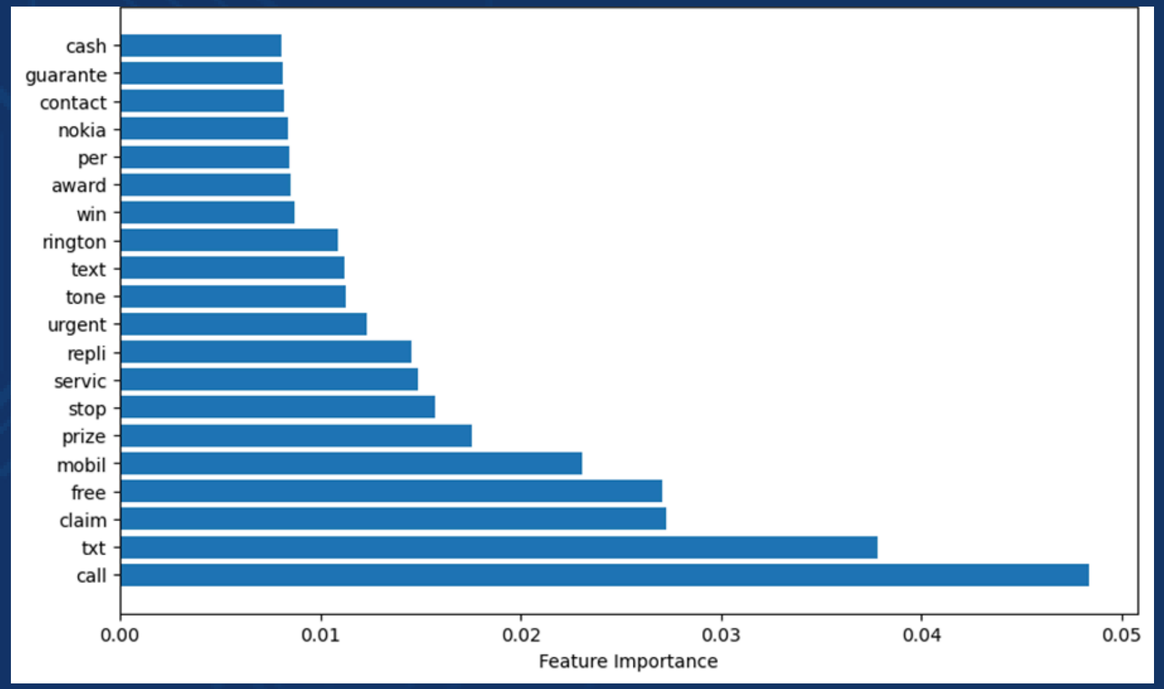
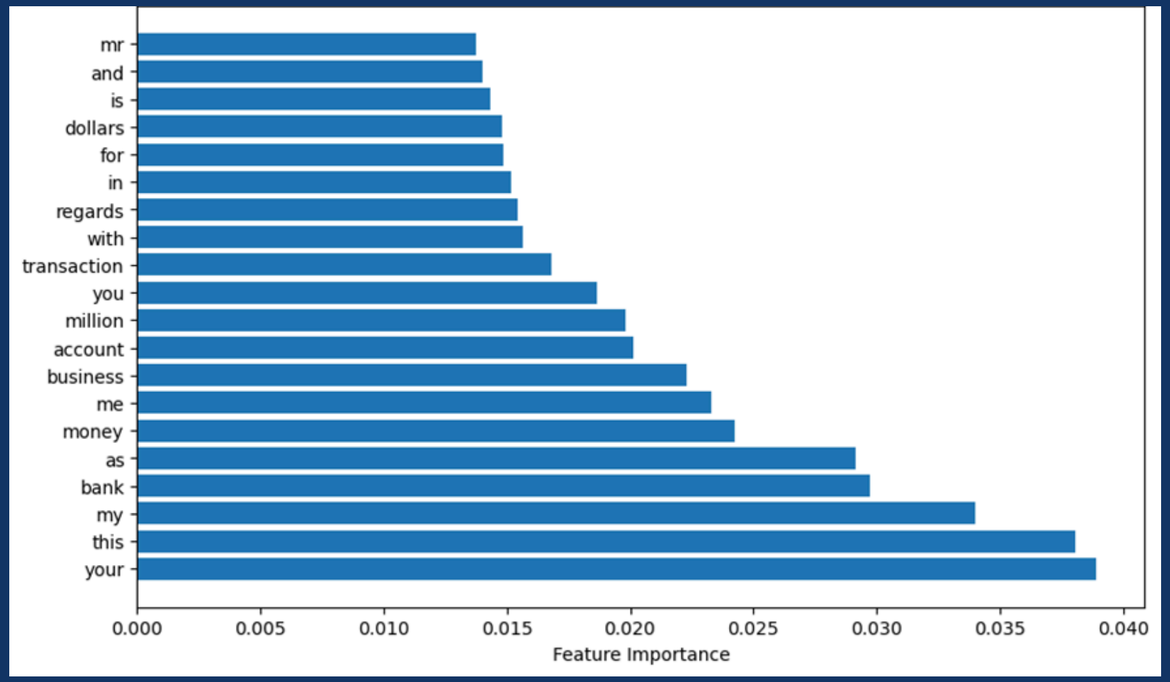


FIGURE 4.4. CONFUSION MATRIX FOR SECOND DATASET.

MACHINE LEARNING MODEL PERFORMANCE

IMPORTANT FEATURES IN RANDOM FOREST ACCORDING TO EACH DATASET





OVERALL, COMMUNICATION AND MONETARY-RELATED TERMS BEING PARTICULARLY SIGNIFICANT INDICATORS OF SPAM CONTENT.

CLIENT APPLICATION DEVELOPMENT

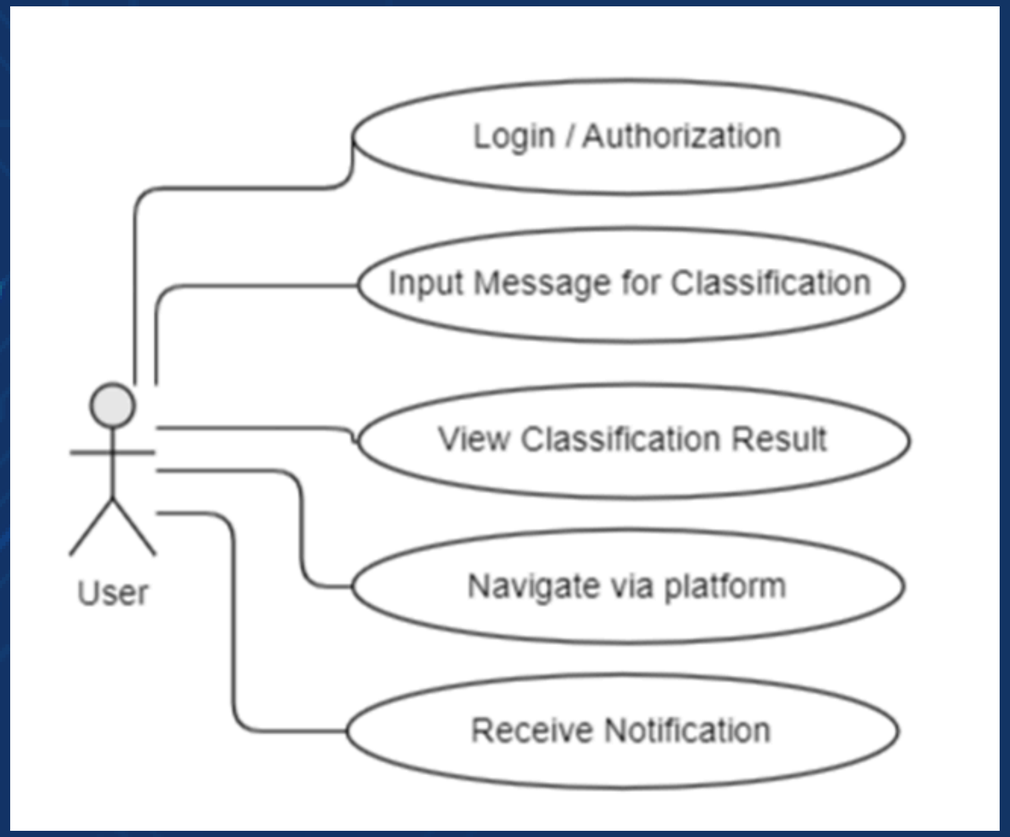


FIGURE 5.1 USE CASE DIAGRAM OF SYSTEM

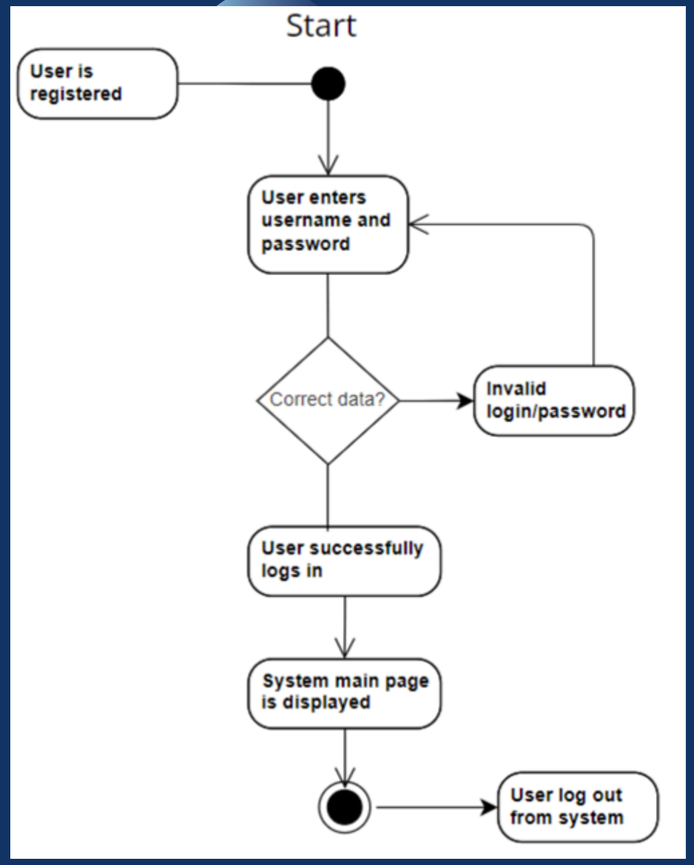
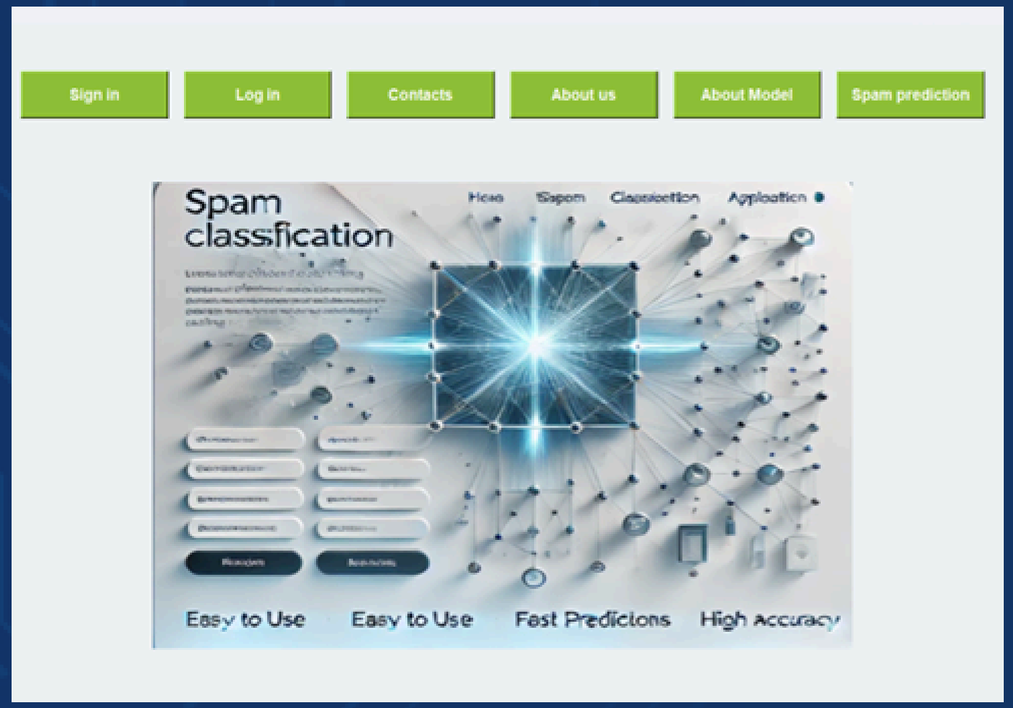
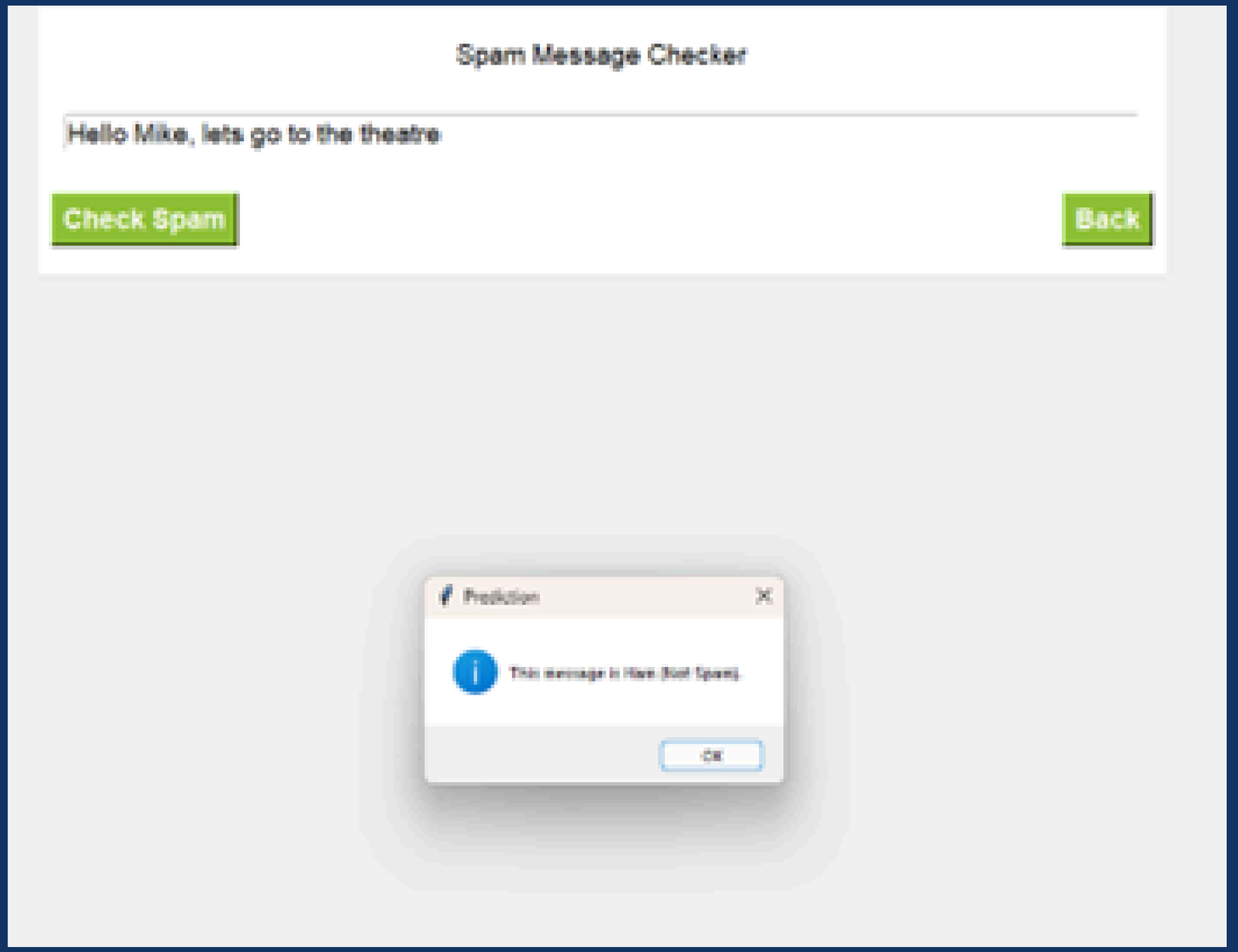


FIGURE 5.2. LOGIN ACTIVITY DIAGRAM

CLIENT APPLICATION CREATION





RESULTS AND DISCUSSION

Future research could examine deep learning techniques for sophisticated feature extraction, ensemble methods for increased accuracy, and expanding datasets to include a wider variety of spam types, even though the models demonstrated strong performance. Althnian et al. (2021) study found that classifier performance depends more on a dataset's representation of the original distribution than its size.

CONCLUSION

This project explores the role of machine learning in classifying phishing and spam threats through advanced techniques like TF-IDF, and classification algorithms. Models like SVM and Random Forest achieved high accuracies (97-98%), revealing that spam features like financial terms and urgency-inducing language. Practical implementations, including a client application and future research directions, highlight AI's potential in enhancing cybersecurity.

REFERENCES

Github: https://github.com/Yerassyl04/Spam-Prediction-System

Jira: https://iskakovk2016.atlassian.net/jira/software/projects/SCRUM/boards/1/timeline?shared=&atlOrigin=eyJpIjoiMGY3M2I0YThlOGRhNGI0NWE3Zjc3NGQ5ZDhkZmRhODQiLCJwIjoiaiJ9

THANK YOU!

thisyera@gmail.com