



BASICS OF MACHINE LEARNING


PROJECT

SPAM EMAIL DETECTION





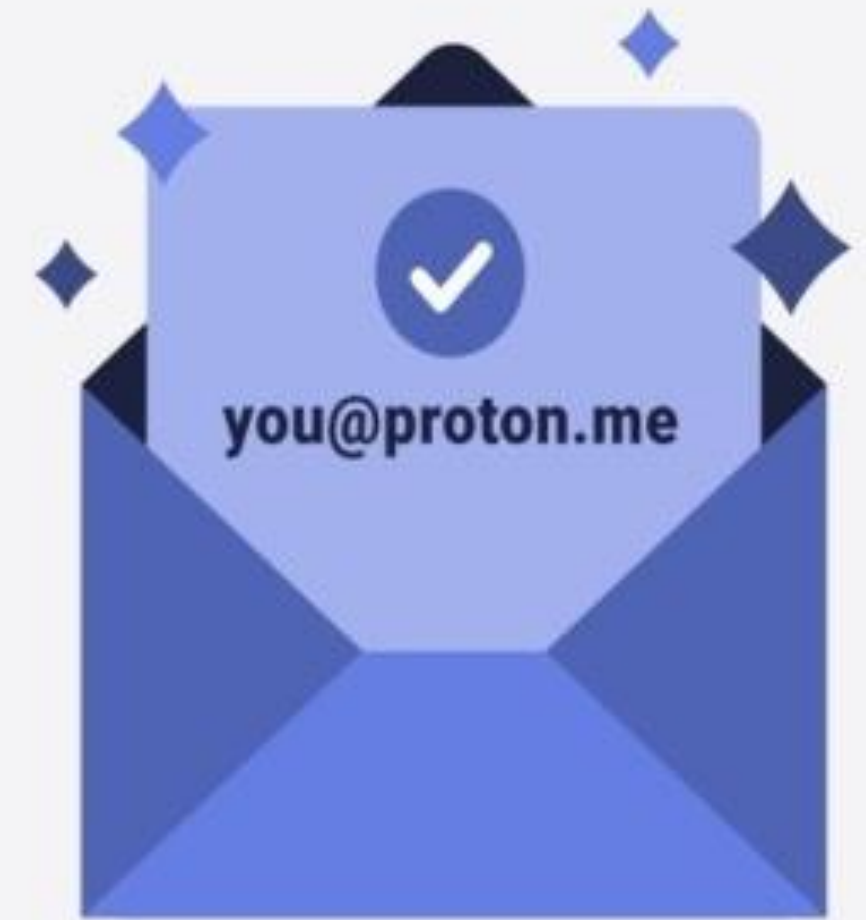
CONTENT

- 
- | | |
|----|------------------------------|
| 01 | INTRO |
| 02 | DATASET |
| 03 | FEATURE ENGINEERING |
| 04 | MODEL SELECTION AND TRAINING |
| 05 | STATISTICS |
| 06 | EVALUATION METRICS |
| 07 | FUTURE WORK |
| 08 | CONCLUSION |

INTRODUCTION

Spam email is unsolicited and unwanted junk email sent out in bulk to an indiscriminate recipient list. Typically, spam is sent for commercial purposes. It can be sent in massive volume by botnets, networks of infected computers.

Email spam also poses a security threat because messages can contain malicious links or malware that can allow a cyber-criminal access to a user's device or ability to find sensitive data/account information.



DATA SET

Explaining the use of Data set for this project and how the collected data set “email.csv” has been used to train our model.

We have taken roughly about 5000 units of data to train our model. This data has been classified into ham and spam for the model to detect and train itself using the data set.

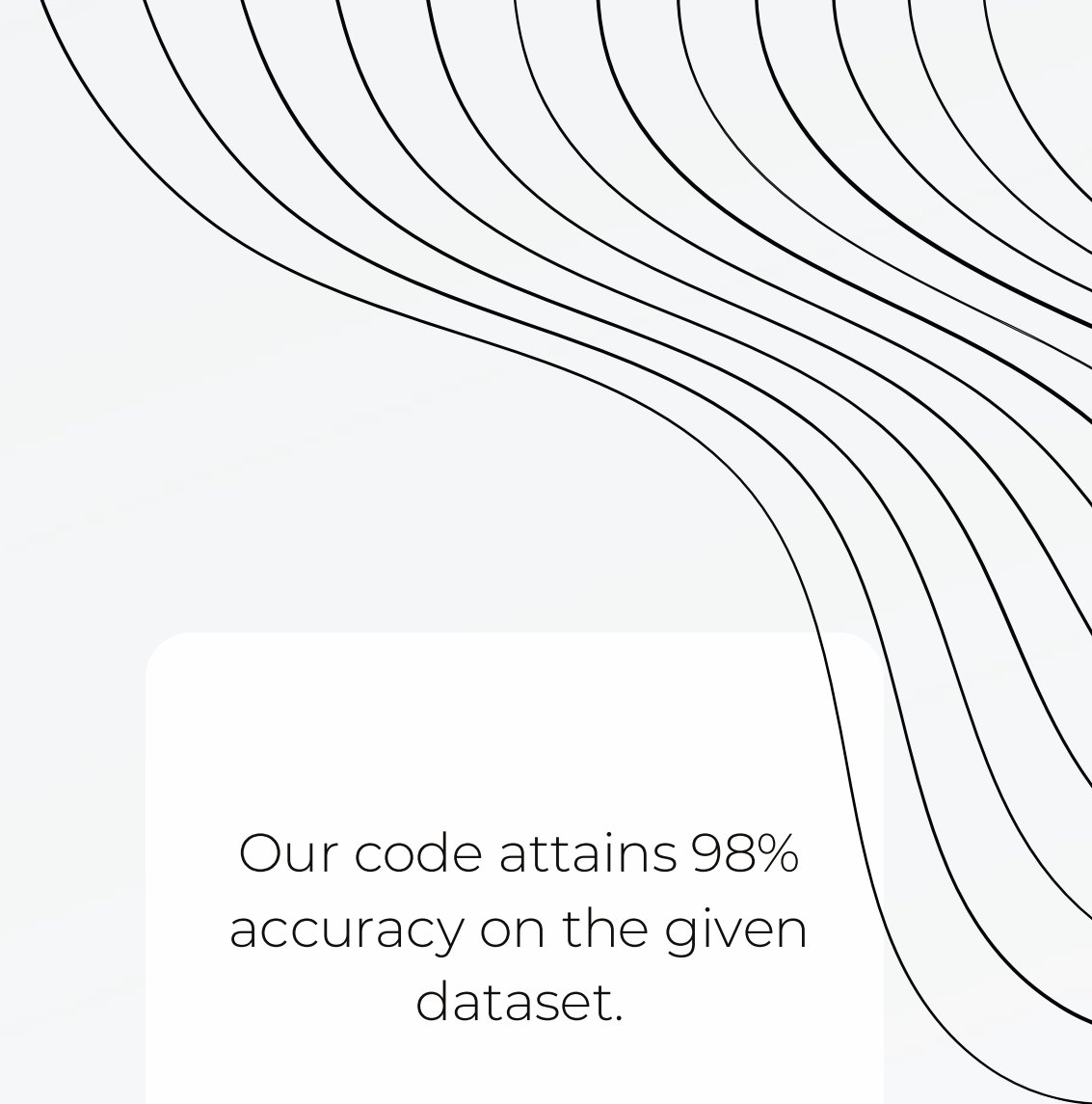
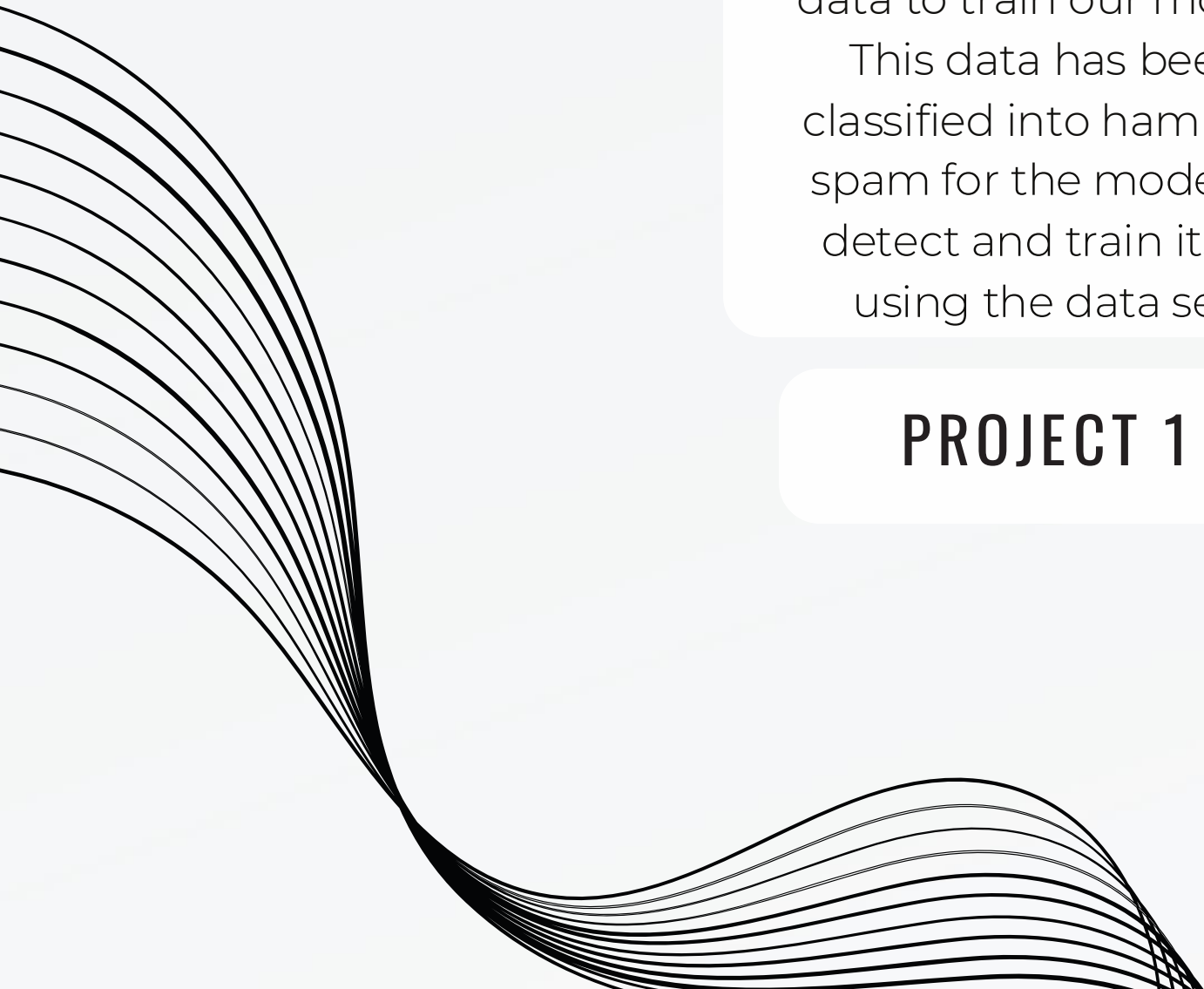
PROJECT 1

Here, we will divide the data into two sets the training set and the test set and based on the training set that is divided into the 1/4th ratio where the 25% data set will be used as the test set and the rest as the training set. Now based on the trained dataset the test set is evaluated.

PROJECT 2

Our code attains 98% accuracy on the given dataset.

PROJECT 3



FEATURE ENGINEERING

Mission



- Tokenization: Breaking down the email text into individual words or tokens.
- Word Frequency: Counting the occurrence of each word in the email corpus.
- N-grams: Extracting sequences of words to capture contextual information.

- Text Normalization: Converting words to their base forms and removing punctuation.
- Structural Features: Including metadata like email length and number of attachments.
- Feature Selection: Identifying the most relevant features for model training.

Vision





MODEL SELECTION AND TRAINING

- **Data Preparation:** The first step is to prepare your data. You'll need a labeled dataset where emails are marked as either "spam" or "not spam" (ham).
- **Feature Extraction:** Convert the text data into a format that can be used by machine learning algorithms. Common methods include Bag-of-Words, TF-IDF, and word embeddings like Word2Vec or GloVe.
- **Model Selection:** Choose a machine learning model to train on the data. Commonly used models for text classification tasks like spam email detection include:
 - **Naive Bayes Classifier**
 - **Support Vector Machines (SVM)**
- **Traning:** Split your dataset into training and testing sets. Train the chosen model on the training set and evaluate its performance on the testing set using metrics like accuracy, precision, recall, and F1-score.
- **Hyperparameter Tuning:** Optimize the model by tuning its hyperparameters to improve performance.
- **Validation:** After tuning, validate the model using cross-validation or a separate validation set to ensure its generalization capabilities.

STATISTICS

Accuracy of the model = 0.9829457364341085

Precision Score = 0.9675324675324676

Recall Score = 0.8975903614457831

f1 score = 0.9312499999999999

98%



EVALUATION METRICS

- False Positive Rate (FPR): Measures the proportion of legitimate emails incorrectly classified as spam.
- True Negative Rate (TNR): Indicates the proportion of legitimate emails correctly identified as non-spam.
- Area Under the Precision-Recall Curve (AUC-PR): Summarizes model performance, especially useful for imbalanced datasets.
- Time-based Metrics: Detects model degradation or drift over time.



FUTURE WORK

ENHANCING MODEL ROBUSTNESS AND GENERALIZATION

Explore methods to improve model robustness against evasion techniques and ensure effectiveness across diverse email platforms.

INTEGRATING US FEEDBACK AND ACTIVE LEARNING

Develop mechanisms for user feedback integration and active learning to enhance model performance and user engagement.

SCALING FOR LARGE SCALE DEPLOYMENT

Investigate scaling techniques for real-time processing and optimize system performance for high-throughput email environments.

CONCLUSION

Our project demonstrates the efficacy of machine learning in combatting spam emails, showcasing a robust detection system that accurately identifies and filters out unwanted messages. With ongoing advancements in algorithmic sophistication, user engagement, and scalability, our solution represents a crucial step forward in ensuring email security and enhancing user experience in an increasingly digital landscape.

