SPAM EMAIL CLASSIFIER

OUTLINE

1. Problem Statement
2. Preprocessing
3. Models
4. Performance
5. Conclusion

1. PROBLEM STATEMENT

Email spam, often referred to as junk email, consists of unsolicited messages sent in bulk by email. These messages can contain advertising, scams, or malicious content intended to harm or deceive the recipient.The goal of this project is to create a machine learning model capable of distinguishing between spam and non-spam (ham) emails.

DATASET

The dataset used for this project is sourced from Kaggle, containing labeled emails as spam or ham. The dataset includes various features such as the email text, subject lines, and other metadata. It will serve as the foundation for training and testing our model.

kaggle

EXAMPLE

Ham:
“Did you catch the bus ? Are you frying an egg ? Did you make a tea? Are you eating your mom's left over dinner ? Do you feel my Love ? “

Spam:
“Thanks for your subscription to Ringtone UK your mobile will be charged £5/month Please confirm by replying YES or NO. If you reply NO you will not be charged. “

2. DATA PREPROCESSING

We clean and tokenize data by :

Remove stopwords, markups, punctuation marks

Remove all strings that contain a non-letter

Convert to lower

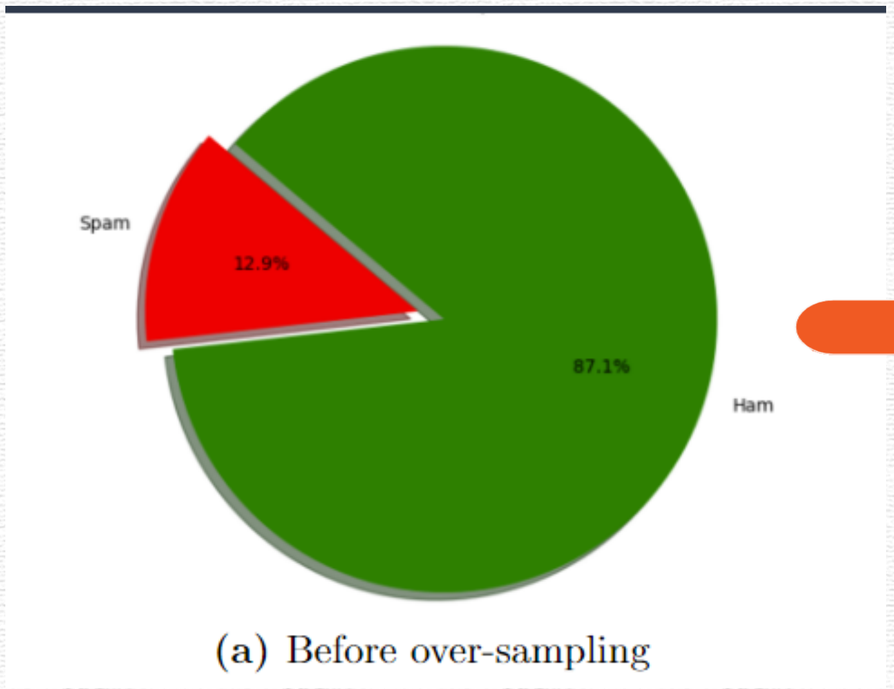
Reduce words to their root form

Remove empty emails

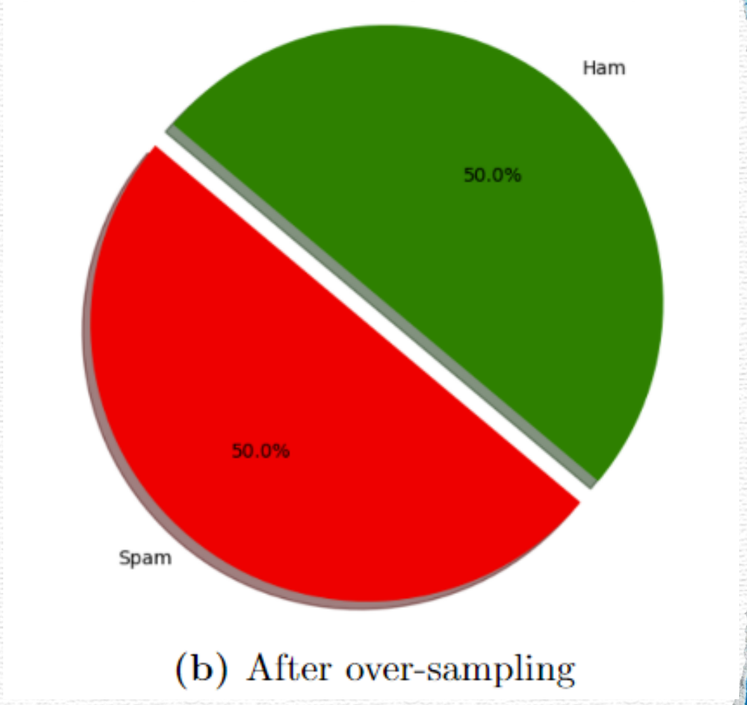
Term frequency - Inverse document frequency: is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus

2. DATA PREPROCESSING

Balance spam emails and ham emails by Over-sampling



(a) Before over-sampling



(b) After over-sampling

2. DATA PREPROCESSING

TF-IDF (term frequency-inverse document frequency)

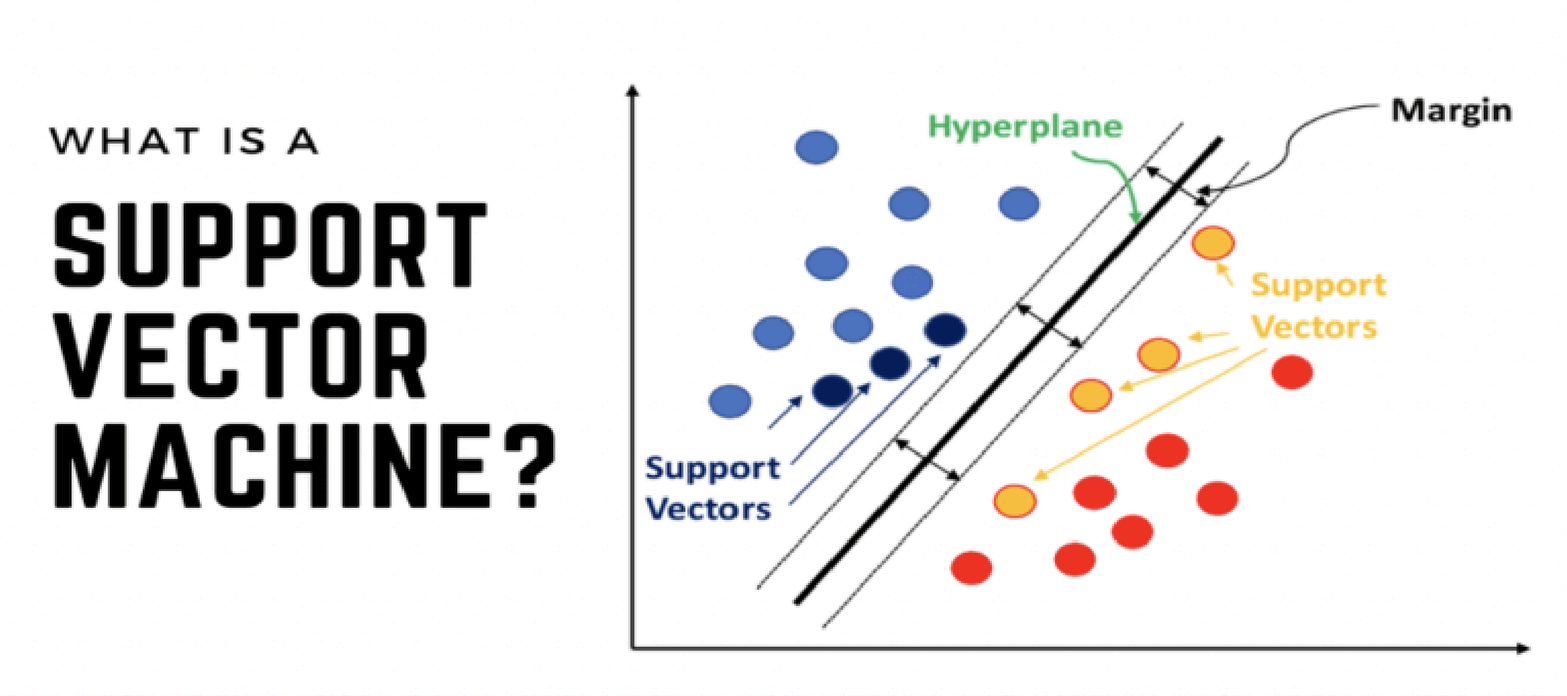
Evaluating how relevant a word is to a document in a collection of documents

$$tfidf(t, d, D) = tf(t, d) \times idf(t, D)$$

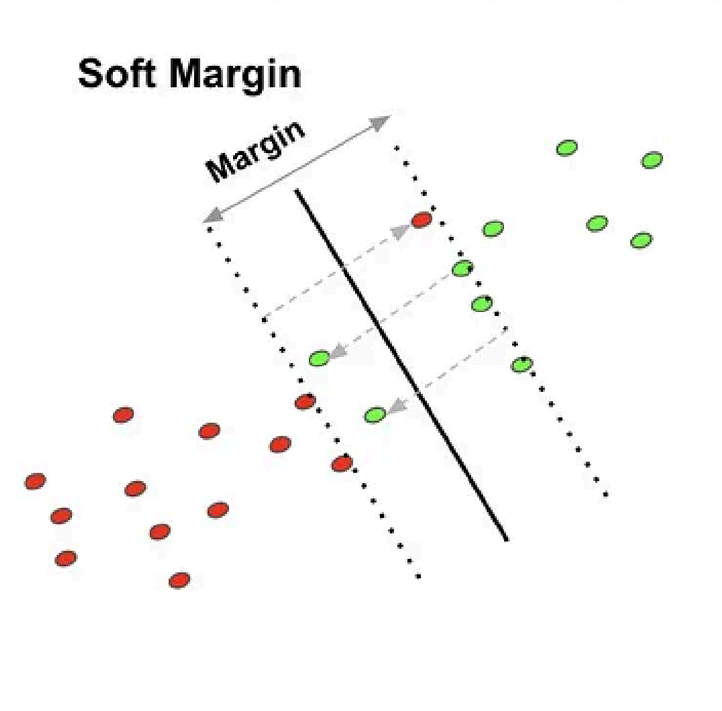
3. MODELS

1. SVM
2. XGBoost
3. Random Forest
4. Logistic Regression

3.1. SVM



The main objective of the SVM algorithm is to find the optimal hyperplane in an N-dimensional space that can separate the data points in different classes in the feature space



When data is not perfectly separable or contains outliers, SVM employs a soft margin technique by introducing slack variables. This approach softens the strict margin requirement, permitting some misclassifications or margin violations. It strikes a balance between maximizing the margin and minimizing classification errors.

$$\underset{\mathbf{w},\, b,\, \zeta}{\text{minimize}} \quad \|\mathbf{w}\|_2^2 + C \sum_{i=1}^{n} \zeta_i$$

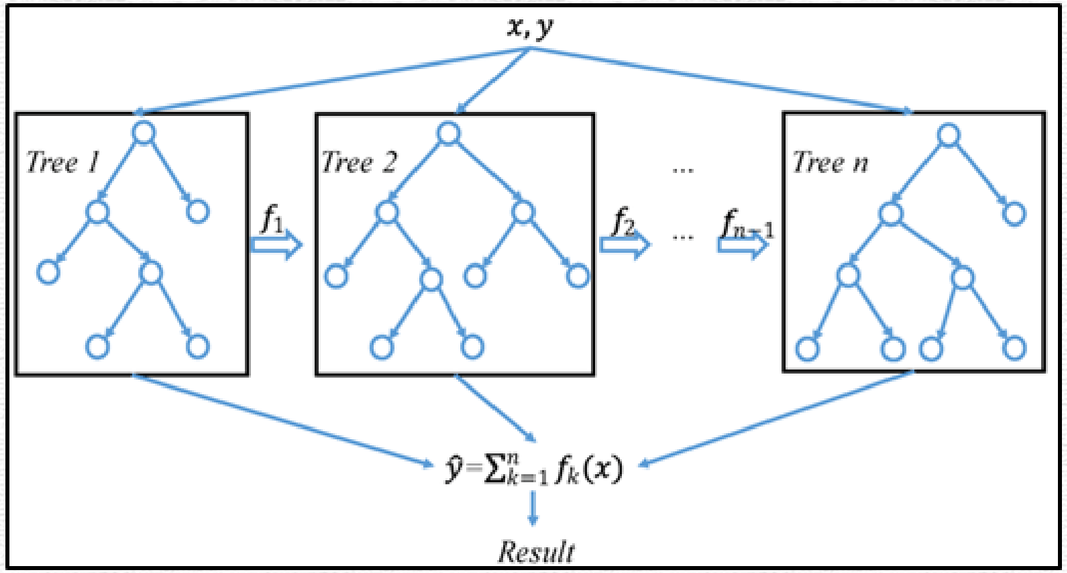
$$\text{subject to} \quad y_i(\mathbf{w}^\top \mathbf{x}_i - b) \geq 1 - \zeta_i, \quad \zeta_i \geq 0 \quad \forall i \in \{1, \ldots, n\}$$

3.1. SVM

PROS	CONS
○ Having excellent accuracy ○ Effective in high dimensions ○ Robust to overfitting ○ Handles non-linear data with kernel tricks	○ Sensitive to parameter tuning ○ Memory intensive due to support vectors ○ Computationally expensive for large datasets

3.2. XGBOOST

- XGBoost is an ensemble learning method
- Boosting techniques is a method that tries to combine multiple weak learners sequentially, with each one correcting its predecessor



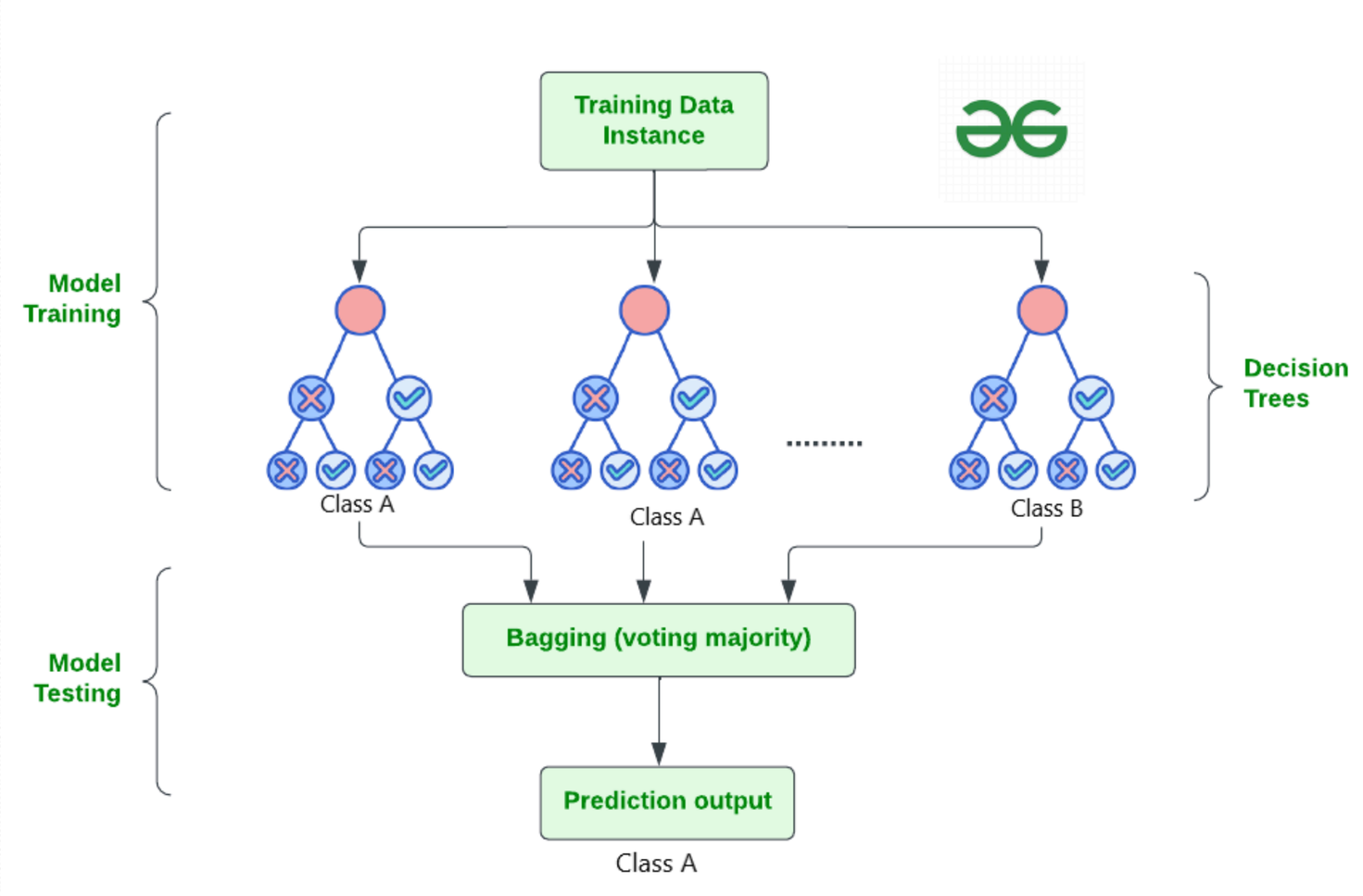
3.2. XGBOOST

PROS	CONS
○ Handles missing values automatically ○ Optimized for parallel proc	○ Time-consuming parameter tuning ○ Significant memory usage

3.3. RANDOM FOREST

- An ensemble learning method
- Bagging Techniques
  - Involves combining multiple weak learners in parallel.
  - Reduces overfitting and improves accuracy.
- Decision Trees
  - Constructs numerous decision trees during training.
  - Each tree contributes to the final prediction:
    - Regression Tasks: Averaging the results.
    - Classification Tasks: Majority vote.

3.3. RANDOM FOREST



3.3. RANDOM FOREST

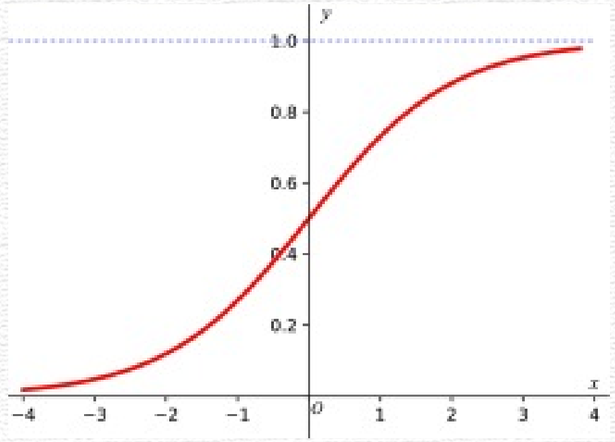
PROS	CONS
◦ High accuracy ◦ Reduces overfitting ◦ Versatile for classification and regression ◦ Tolerant of missing data	◦ More complex ◦ Longer training time ◦ Memory intensive ◦ Slower for real-time predictions

3.4. LOGISTIC REGRESSION

- Logistic Regression is a statistical method for analyzing datasets in which there are one or more independent variables that determine an outcome.

Sigmoid function

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$



3.4. LOGISTIC REGRESSION

The loss function in logistic regression with L2 regularization

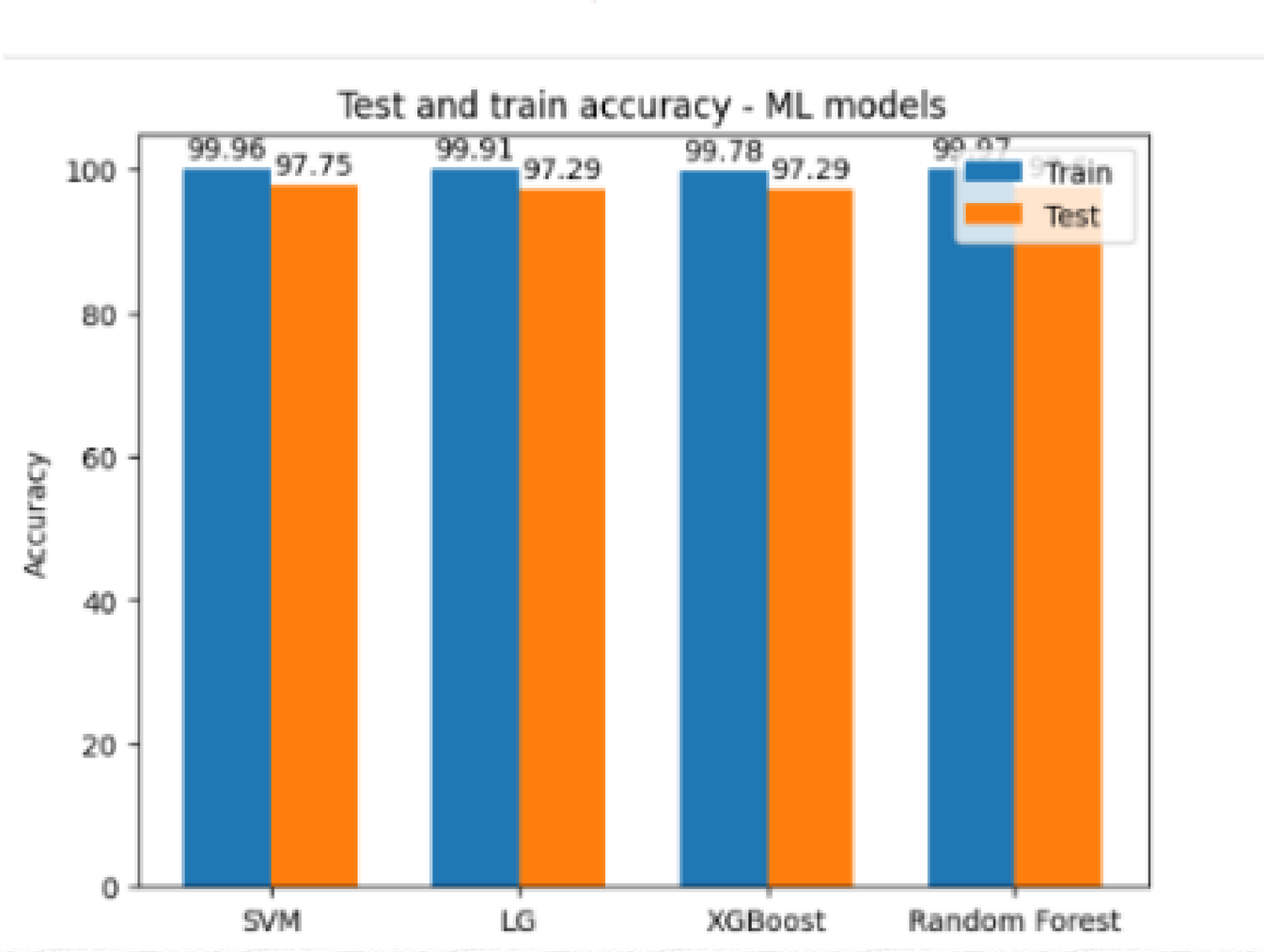
Similar to linear regression, we can handle overfitting by adding a regularization term to the error function:

$$J(w,b) = \frac{1}{N}\sum_{i=1}^{N}[-y_i \log(h_w(x_i)) - (1-y_i)\log(1-h_w(x_i))] + \frac{\lambda}{2}\|w\|^2$$

3.4. LOGISTIC REGRESSION

PROS	CONS
◦ Fast and efficient training. ◦ Requires few assumptions about the data. ◦ Provides useful probability predictions.	◦ Sensitive to linearly inseparable features. ◦ Prone to overfitting with many features. ◦ Ineffective with datasets containing many missing values.

4.MODEL VALUATION



CONCLUSION

- Our project on email spam classification using machine learning has successfully demonstrated the effectiveness of advanced algorithms in identifying and filtering out unwanted emails.
- By leveraging techniques such as natural language processing and supervised learning, we have developed a robust model that can distinguish between legitimate emails and spam with high accuracy.

THANK YOU

for listening