Practical - 02.

- Aim: Classify the email using the binary classification method. Email spam detection has two states:
  a) Normal state: Not spam b) Abnormal state - spam,
  Use Knearest neighbours and support vector machines for classification. Analyse their performance.

- Theory:

1) Objective:
The aim of this project is to classify emails as either Spam, or Not Spam using machine learning techniques. Specifically two classification algorithms - K nearest KNN and Support Vector Machine (SVM). are applied and their performance compared using evaluation metrics such as accuracy, precision, recall, and F1 Score.

2) Theory -

1. Data preparation -
The dataset consists of email texts labeled as spam or not spam. Preprocessing is necessary to convert raw email text into structured numerical features suitable for machine learning

- Text cleaning - Removal of stop words, punctuation and special symbols.

9.

- Tokenization and Stemming lemmatization – Breaking text into tokens and lemmatization : reducing words to their base forms.

- Feature Extraction – Converting text into numerical representations using methods like Bag of words or TF IDF.

- Splitting the dataset – Data is divided into training and testing subsets for model building and evaluation.

2) K-nearest neighbours –

KNN is a lazy learning, instance based classification algorithm. It classifies a new email by comparing it with the closest training examples in feature space.

* Working –

1. Calculate the distance (commonly euclidean of cosine distance in text datasets.) between the new email vector and all training samples.

2. Identify the K-nearest neighbours (emails most similar to the new one).

3. Perform a majority vote among these neighbours to assign the class label (spam or not spam)

* Key characteristics -

- Non parametric - Makes no assumptions about the underlying data distribution.
- Simple yet powerful - Works well when decision boundaries are irregular.
- Hyperparameter K: Choosing the right value of K is critical; too small may overfit k, too large may oversmooth.

* Strengths -

- Easy to implement and understand.
- Naturally handles multi class problems.

* Limitations -

- Computationally expensive for large datasets (requires storing and comparing with all training samples)
- Sensitive to irrelevant features and high dimensionality. (common in text data).

3] Support Vector Machines (SVM)

- SVM is a supervised learning algorithm widely used for binary classification tasks like spam detection. It finds the optimal hyperplane that separates the classes with maximum margin.

* How it works –

1. Each email is represented as a feature vector (TF-IDF weighted terms).
2. SVM identifies a decision boundary (hyperplane) that best separates spam and not spam classes.
3. The margin is maximized and only support vectors (data points closest to the hyperplane) determine the boundary.

* Key characteristics –

- Maximizes generalization – Focuses on robust classification by maximizing margins.
- Effective in high dimensional spaces – works well with text data where feature space is large.

* Strengths –

- High accuracy for binary classification.
- Effective even with sparse data (common in text classification).
- Less prone to overfitting in high dimensional space.

* Limitations –

- Computationally intensive for very large datasets.

- Choice of Kernel and tuning hyperparameters (C, gamma) is crucial for performance.

4] Model Evaluation -

The classification performance is assessed using:
- Accuracy: Percentage of correctly classified emails.
- Precision: Fraction of predicted spam emails that were actually spam.
- Recall: Fraction of actual spam emails correctly identified.
- F1 score: Harmonic mean of precision and recall (balances both).

* Expected Outcome:

- KNN provides decent performance but struggles with high dimensional text data, making predictions slower and less scalable.

- SVM generally outperforms KNN, offering higher precision and recall and proving more robust for spam detection tasks.

* CONCLUSION:

This practical demonstrated the use of KNN and SVM for binary email spam classification. KNN, while simple and intuitive was computationally heavy and less accurate in handling high dimensional text features. On the other hand, SVM effectively modeled the decision boundary between spam and not spam emails, acheiving higher precision, recall, and F1-scores

Thus SVM emerges as the superior model for email spam detection, highlighting its effectiveness in binary classification problems involving high dimensional feature spaces such as natural language text.