AHSANULLAH UNIVAERSITY OF SCIENCE AND TECHNOLOGY

Department of Computer Science and Engineering

Course Title: **Artificial Intelligence Lab**

Course No: **CSE 4108**

Group No: **B208**

Project Report

On

Spam Mail Prediction

Submitted To:

Mr. Md. Siam Ansary

Lecturer, Department of CSE, AUST

Ms. Tamanna Tabassum

Lecturer, Department of CSE, AUST

Submitted By:

Name: Tahmid Jawad Annoor
ID : 170204053

Name: Ifti Sam Ibn Rahman
ID : 170204117

Description of the Problem:

Spam mail, or junk mail, is a type of email that is sent to a massive number of users at one time, frequently containing cryptic messages, scams, or most dangerously, phishing content. Clicking on a spam email can be dangerous, exposing our computer and personal information to different types of malware. Therefore, it’s important to implement additional safety measures to protect the device, especially when it handles sensitive information like user data. In this project, we tried to implement some machine learning algorithms to detect whether a mail is spam.

Brief Description of the Dataset:

Our dataset contains 300 sample data. The features of our dataset are Category, Message, Senders Anonymity, Legitimate Contact Info and Appealing Subject. This dataset is created with the help of Kaggle. All the values of these features are in text. So when used in model these values were encoded into numerical values.

Description of Used ML Models:

Logistic Regression:

Logistic regression is a statistical analysis method to predict a binary outcome, such as yes or no, based on prior observations of a data set. A logistic regression model predicts a dependent data variable by analyzing the relationship between one or more existing independent variables. Though it is a regression model algorithm, it is widely used in binary classification problem.

Naïve Bayes:

It is a classification technique based on Bayes’ Theorem with an assumption of independence among predictors. Naïve Bayes is one of the fast and easy ML algorithms to predict a class of datasets. It can be used for Binary as well as Multi-class Classifications. It performs well in Multi-class predictions as compared to the other Algorithms. It is the most popular choice for text classification problems.

Random Forest Classifier Algorithm:

Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset. Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting. It takes less training time compared to other algorithms. On large dataset, it works efficiently.

- **K-Nearest Neighbors:**
  K-nearest neighbors (KNN) is a supervised learning algorithm used for both regression and classification. KNN tries to predict the correct class for the test data by calculating the distance between the test data and all the training points. Then select the K number of points which is closest to the test data. It is easy to implement and understand, but it has a major drawback as it becomes significantly slow as the size of the data in use grows.

- **Gaussian Naïve Bayes:**
  Gaussian naïve Bayes is a supervised machine learning model which is used for classification problem. It is used when the features of the dataset are continuous in nature. This algorithm is frequently used in natural language processing (NLP).

Performance Scores of the Models:

Logistic Regression:

Metrics	Values
Accuracy Score	0.8333333333333334
R^2 Score	-19.999999999999996
Explained Variance Score	-2.220446049250313e-14
Mean Absolute Percentage Error	7.505999378950826e+16

Multinomial Naïve Bayes:

Metrics	Values
Accuracy Score	0.8666666666666667
Precision	0.8620689655172413
F1 Score	0.9259259259259259
Recall Score	1.0

Random Forest:

Metrics	Values
Accuracy Score	0.9166666666666666
Precision	0.9090909090909091
F1 Score	0.9523809523809523
Recall Score	1.0

KNN:

Metrics	Values
Accuracy Score	0.9333333333333333
Precision	0.9259259259259259
F1 Score	0.9615384615384615
Recall Score	1.0

Gaussian Naïve Bayes:

Metrics	Values
Accuracy Score	0.8666666666666667
Precision	0.9565217391304348
F1 Score	0.9166666666666666
Recall Score	0.88

Conclusion:

Machine Learning is a branch of artificial intelligence that provides systems the ability to improve from experience. Here we applied 5 different models and gain some impressive result. We apply multinomial Naïve Bayes, Random forest, logistic regression, KNN model and Gaussian Naïve Bayes Model and receive 86%, 92%, 83%, 93% and 86% accuracy rate respectively. So, we can see KNN is the best fit model for our dataset.

Percentage of Contribution:

Tahmid Jawad Annoor – 50%

Ifti Sam Ibn Rahman - 50%