Spam Mail Detection Using Machine Learning

Project Link :- https://github.com/Nikhil-Dangeti/Spam-Mail-Detection

My contribution to this project involved enhancing an existing spam mail detection code in a Jupyter notebook. I fix the errors in the code . Initially, the code utilized Naive Bayes and Decision Tree algorithms. I improved it by:

Extending Algorithm Range: I incorporated two additional machine learning algorithms:

- Logistic Regression
- Support Vector Machine (SVM)

Implementing CNN: I also integrated a Convolutional Neural Network (CNN), which achieved an accuracy of 95%.

By updating and optimizing the code, I significantly enhanced the performance and accuracy of the spam detection model.

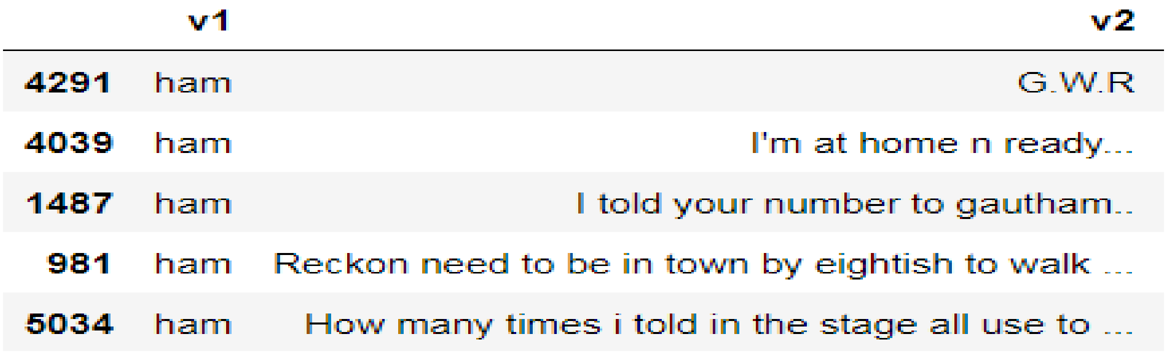
Introduction:-

Our project focuses on spam mail detection using Machine Learning (ML). Email is a form of communication that is regularly used to convey a wide range of information. Because of its convenience and capacity to convey messages, documents, pictures, videos, and links, it is a popular communication tool. However, spam mails are unwanted messages or junk messages sent in bulk. Spam mails can come in many forms, including advertisements for products or services, phishing scams, fraudulent offers, and malware-containing attachments.

To address this issue, we employed five machine learning classifiers. We trained these classifiers on a specific dataset and selected the best two based on their accuracy and precision.

Data set:

We have taken dataset(spam.csv) from Kaggle. These data set contain 2 features, one is for text and other is for target(spam or ham)



	v1	v2
4291	ham	G.W.R
4039	ham	I'm at home n ready...
1487	ham	I told your number to gautham..
981	ham	Reckon need to be in town by eightish to walk ...
5034	ham	How many times i told in the stage all use to ...

Process:

- **Dataset Overview**
  - Dataset used: "spam.csv."
- **Data Cleaning and Preprocessing**
  - Steps taken to clean the data:
  - Removed unnecessary columns .
  - Handled missing values .
  - Dropped duplicate entries.
- **Exploratory Data Analysis (EDA)**
  - Pie chart: Distribution of spam and ham messages.
  - Histograms: Visualizing character and word counts.
  - Heatmap: Correlation within the dataset
- **Data Preprocessing**
  - Overview of text processing tasks:
  - NLTK for tokenization and stemming.
  - Removal of special characters and stop words.
  - Lowercasing of text.
- **Model Selection and Training**
  - Overview of classification models used: SVM, Naive Bayes, Decision Tree, CNN, Logistic Regression .
  - Splitting data into training and testing sets.
  - Model training and evaluation metrics (accuracy, precision)
- **Model Evaluation**
  - Presentation of accuracy and precision scores for each mode

Observation:-

Algorithms	CNN	Naïve Bayes	Decision Tree	Logistic Regression	SVM
Accuracy	95	94	86	90	92
precession	96	96	93	95	97

Conclusion:-

- In this project we trained data set on Multiple classifiers .(Decision tree, Naive bayes , Logistic Regression, Support Vector Machine,CNN).
- Among those Naïve bayes and CNN are giving more accuracy
- CNN is Giving 95% accuracy and Naïve Bayes is giving 94% accuracy.