

NAME: ASHNA

REGISTRATION NUMBER: 25BSA10138

OVERVIEW OF THIS PROJECT:

This project gives a complete automated solution for text classification, which is built to differentiate between authorized email (Ham) and unauthorized and unwanted emails(Spam).The classifier is built on fundamental naïve bayes theorem, which also involves core machine learning and natural language processing.

FEATURES

1. Data preprocessing: preparation of raw email text data.
2. Transforming textual data into a numerical format good for machine learning models using the dictionary-of-words model , implemented through CountVectorizer.
3. Training the naives bayes model on a labeled dataset
4. Generating a report which includes efficiency, precision and confusion matrix.
5. Live prediction: allows for classification.

TECHNOLOGIES USED

- 1.Language :Python 3.x – used for data science
- 2.Data Handling: Pandas: necessary for loading, cleaning, and interpreting dataset
3. Numerical operations-Numpy- supports handling of numerical arrays by underlying ML model.
4. NLP: scikit-learn- for production-ready tools for modelling and text vectorization.
- 5.Live Prediction- Allows for the classification of user defined text messages.

HOW TO EXECUTE THE PROJECT

- 1.Python 3.6 to be installed for executing
2. installation- all important dependencies can be installed
pip install pandas scikit-learn numpy
3. running the classifier – ensure the python source file(e.g. spam_classifier.py)

4.Execute the script from the command line

Python spam_classifier.py