# 

Assignment#2: Machine Learning: Spam Classification.

Swapnali Dashrath

Oakland University

CSI6160 - Machine Learning

Prof. Mohammad Wardat.

October 12, 2024.

**SMS Spam Detection Modeling and Evaluation Report**

This report presents the implementation and evaluation SMS Spam Detection model using Iris simple datasets using an ML library and also builds their own model from scratch. on Iris Dataset using existing libraries (numpy, pandas, seaborn, matplotlib etc). Implementation is using Python (code file submitted along with report) and Evaluation is based on two evaluation metrics from Accuracy, Precision, Recall and F1-Score.Also, this report explore the building the model from scratch.

**Problem Statement:**

We are trying to use attributes of SMS Spam Dataset to predict the Spam and Ham messages, specifically, using Word of Bags Algorithm as well as building model from scratch.

**Approach:**

1. Analyzing Data – SMS Spam Dataset
2. Data Cleanup and formatting using Data Frames.
3. Visualize the data using Data Plots.
4. Defining Relationship of data features with target.
5. Exploratory Data Analysis (EDA) – Pairplots.
6. Training Data Split.
7. Data Processing for modeling
8. Vectorization using bag of words.
9. Evaluation based on Accuracy, precision and confusion Matrix.

**SMS Spam Dataset:**

The dataset consists of 5572 samples of messages, divided into two types: Spam and Ham with four features: text and new features such as number of character, number of words, number of sentences and transformed text.

**Classifiers Implemented**

1. Vectorization using bags of word

**Data Processing**

Data processing involves implementation of transformed text column performing Lower case, Tokenization, Removing special characters, Removing stop words and punctuation, Stemming

**Data Splitting**

The dataset was split into training (70%) and testing (30%) sets. This allows us to train the models on the training data and evaluate them on unseen test data.

**Modeling**

Our base line model detecting spam and ham messages based on Bags of word algorithm.

**Evaluation Metrics**

We used the following metrics to evaluate the models:

* Accuracy: Proportion of correctly classified instances.
* Precision: quality of accuracy.
* Confusion Matrix

**Results**

1. Based on the Approach:

**Gaussian Naïve Bays:**

Accuracy: 0.8694390715667312

Confusion Matrix:

[[788 108]

[ 27 111]]

Precision: 0.5068493150684932

**Multinomial Naïve Bays**

Accuracy:0.9709864603481625

Confusion Matrix:

[[896 0]

[ 30 108]]

Precision:1.0

**Bernoulli Naïve Bays**

Accuracy: 0.9835589941972921

Confusion Matrix:

[[895 1]

[ 16 122]]

Precision:0.991869918699187

**Interpretations:**

We are successfully able to near to 100% accuracy using

Bernoulli and Multinomial Classifier using Naïve Bays Model building using existing library and from scratch implementations.

**References**

Almeida, T. & Hidalgo, J. (2011). SMS Spam Collection [Dataset]. UCI Machine Learning Repository. <https://doi.org/10.24432/C5CC84>.

Normalized Neard. (2021),  [Naive Bayes Classifier in Python (from scratch!)](https://www.youtube.com/watch?v=IMrxB8Mq5KU)

[(youtube.com)](https://www.youtube.com/watch?v=IMrxB8Mq5KU) [(174) Naive Bayes Classifier in Python (from scratch!) - YouTube](https://www.youtube.com/watch?v=3I8oX3OUL6I)

CampusX (2021). [Email Spam Classifier | SMS Spam Classifier | End to End Project | Heroku Deployment (youtube.com)](https://www.youtube.com/watch?v=YncZ0WwxyzU) [Email Spam Classifier | SMS Spam Classifier | End to End Project | Heroku Deployment](https://www.youtube.com/watch?v=YncZ0WwxyzU)