Assignment -2

Spam detection by using the frequency of word or character.

Name : Hari teja Narayanabhatla

Panther ID : 002566679

Group : MSA

**Business Question:** How to identify a given mail is spam or not spam based on the given words in the mail?

**Dataset Overview:**

The spam base dataset consists of frequencies of certain words that are found in mails. It

includes 4601 observations corresponding to email messages, out of which 1813 belong to

the spam category. From the original email messages, 58 different attributes were

computed. There are no categorical variables in the dataset, all the 57 attributes are

continuous variables. The data had no null values, every attribute had a value. A spam

column is a binary value with 0 as notspam and 1 as spam.

**Data Visualization:**

First tried to check if the data is balanced properly or not. Hence using pandas and

matplotlib tried to visualize the data

Bar chart

Description automatically generated with medium confidence

Fig 1: Shows the distribution of the classes in the given dataset.

There exists no null value or unknown value for any column in the given dataset. Hence, it’s

a clear dataset and we went ahead with further processing.

**Data Correlation**

Since we had so many columns, the computation for training the model or making a

machine to learn the algorithm is time-consuming. Hence, we tried to find if there is any

correlation between any of the columns and if yes, we can remove them the so that model

can give good results.

Chart

Description automatically generated

Fig 2: It gives us information that there are some columns which are correlated to each other.

Hence, I had considered only the threshold to be 0.5 and above. As a result, it would give us

the columns that have to be dropped to train the model with good accuracy.

Graphical user interface, text, application

Description automatically generated

­­­

**Optimized Feature selection or Dimensionality Reduction**

As we have applied correlation to the dataset to find any similarities. Another method called

PCA helps in understanding which features are important from a long set of features. PCA is

one of the famous techniques for dimensionality reduction. In this problem, dimensions

mean the features i.e. word frequencies of words/characters.

After applying the PCA over the spam dataset, we obtained a graph for the components

based on explained\_variance\_Ratio.

Chart, histogram, scatter chart

Description automatically generated

**Fig 3:**  It shows the components covering how much % of the variance in data.

**Conclusion**

Finally, we could consider holding onto principal components whose eigenvalues are greater

than or equivalent to 1. These eigenvalues are the sum of squares of distances between the

projected data points and the origin along an eigenvector associated with a principal

component. These are stored in the explained variance attribute. Hence, we would consider

those values which had their cumulative greater .99 value. This would be enough to give

our model a good accuracy which is 56 components.