**Documentation**

**Description:**

Spam and Ham classification using Big dataset to detect spam messages in emails.

Using 3 model evaluating SVM, LR, KNN (cross validation) and 2 model for vectorizing TF\_IDF and Doc2VIC.

**logistic regression**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | **Accuracy** | **F1\_Score** | **Precision** | **Recall** |
| **TF\_IDF Modeling** | 0.973 | 0.972 | 0.974 | 0.973 |
| **Doc2Vec** | 0.968 | 0.967 | 0.967 | 0.968 |

**Support vector space**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | **Accuracy** | **F1\_Score** | **Precision** | **Recall** |
| **TF\_IDF Modeling** | 0.880 | 0.824 | 0.775 | 0.880 |
| **Doc2Vec** | 0.948 | 0.943 | 0.949 | 0.948 |

**TF\_IDF Manual -** KNN

|  |  |  |  |
| --- | --- | --- | --- |
| **Accuracy** | **F1\_Score** | **Precision** | **Recall** |
| 0.666 | 0.667 | 0.687 | 0.666 |

**Dataset Report**

Name: SMS\_Spam\_Collection

Size: 5574

Description: Have 2 column (Category,Message)

Every message show it is Category(Ham,Spam)