SPAM CHECK PREDICTION SYSTEM REPORT

Introduction :-

Here I take a dataset that contains the message text out of this some message is a spam message and some message is ham means normal message. Now I try to make a prediction system that classify the message into spam and normal message basis on the text using some machine learning classification model and check accuracy , precision , recall and f1 score for each model . Here I use five ml classification model that are support vector machine , naïve bayes , k-nearest neighbours , random forest and decision tree.

Information about the dataset :-

Here we use the sms spam collection dataset . This is a set of sms tagged messages that been collected for sms spam research . It contains one set of sms messages in English of 5572 messages , tagged according being ham (legitimate) or spam.

- Size of dataset – this dataset contains 5572 english text messages.
- Collection of dataset - this dataset is collect from Kaggle.
- No of row and column – this dataset contains two column Category and message . In message column contains the text sms and the category column contains the lable for the text ie the text sms is ham or spam.
- No of spam messages – 13% of dataset ie 747 messages are spam.
- No of ham messages - 87% of dataset ie 4825 messages are ham (legitimate).

General procedure :-

- Import the dependencies.
- Data collections and pre-processing.
- Label encoding.
- Splitting the dataset into train data and test data.
- Feature extraction.
- Train the model.
- Evaluating the train model.

Different model evaluation on training data

Name of the model	Accuracy	Precision	Recall	F1 score
SVM	0.9952	0.9930	0.9712	0.9820
Naïve bayes	0.9807	1	0.8547	0.9216
KNN	0.9201	0.9957	0.4003	0.5710
Random Forest	0.9997	1	0.9983	0.9991
Decision Tree	1	1	1	1

Different model evaluation on test data

Name of the model	Accuracy	Precision	Recall	F1 score
SVM	0.9820	0.9927	0.8774	0.9315
Naïve bayes	0.9730	1	0.8064	0.8928
KNN	0.9094	1	0.3483	0.5167
Random Forest	0.9757	1	0.8258	0.9045
Decision Tree	0.9659	0.9680	0.7806	0.8642