SPAM HAM DETECTION

Detailed Project Report

Upendra Kumar
Data Science Intern at Ineuron.ai

INTRODUCTION

Nowadays, most people have access to the Internet, and they cannot survive without Smartphone and computers. They not only use the Internet for fun and entertainment, but they also use it for business, stock marketing, searching, sending e-mails, and so on. Hence, the usage of the Internet is growing rapidly.

One of the threats for such technology is a spam. Spam has a lot of definitions, as it is considered one of the complex problems in e-mail services. **Spam is a junk mail/message, or an unsolicited mail/message. Spam e-mails are also those unwanted, unsolicited e-mails that are not intended for a specific receiver.** It is basically an online communication sent to the user without permission. It takes on various forms like adult content, selling products or services, job offers, and so forth. The spam has increased tremendously in the last few years.

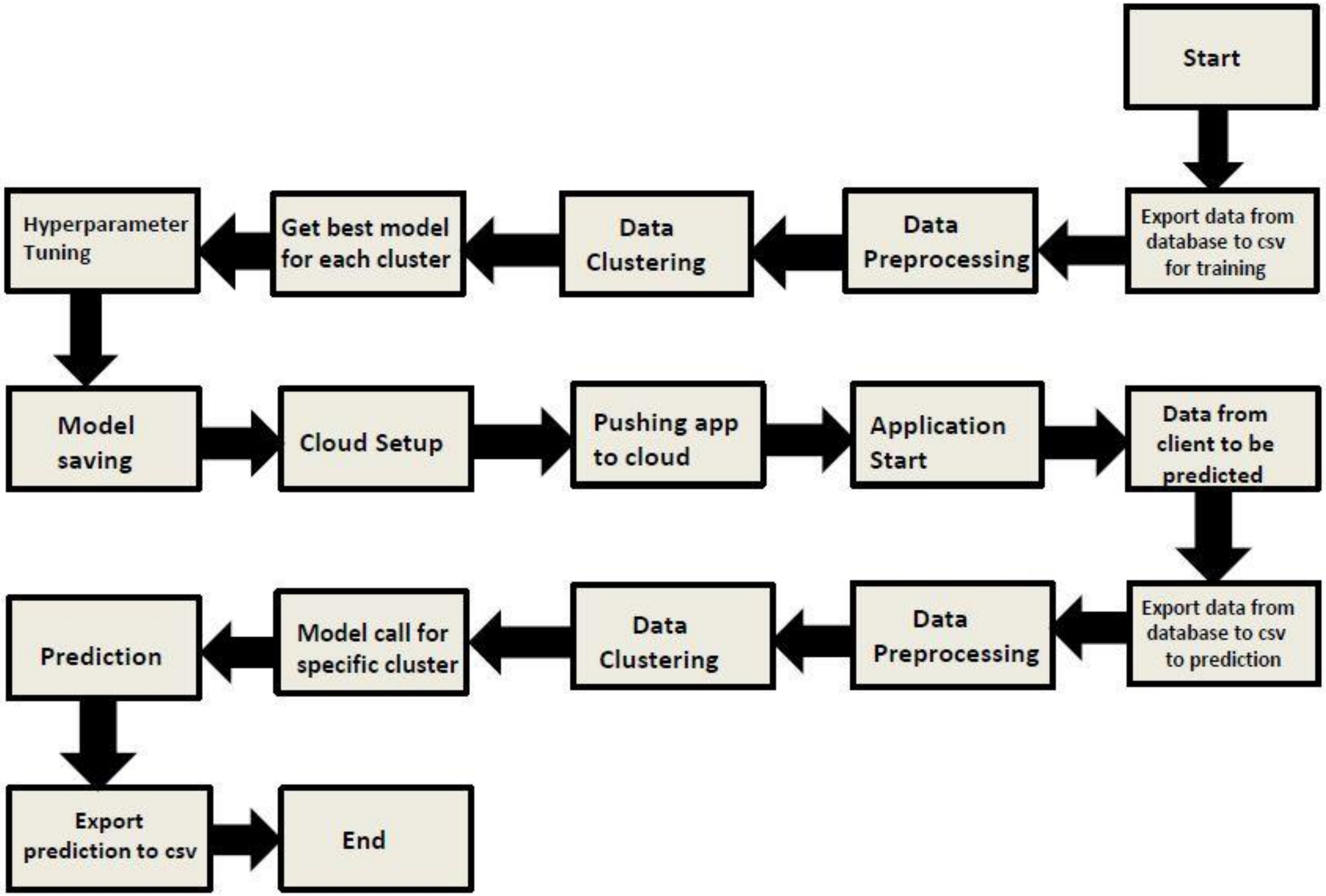
The good, perfect, and official mails are known as *ham*. It is also defined as an e-mail that is generally desired. Today, more than 85% of mails or messages received by users are spam. It costs the sender very little time to send, but most of the costs are paid by the recipient or the service providers rather than by the sender. The cost of spam can also be measured in lost human time, lost server time and loss of valuable mail/messages. The sent mail reserves a quota in the server, and the receiver may have a limited space, causing the server to reject another ham mail because it is out of space. Moreover, the reader may lose a lot of time in reading unuseful messages.

The **Machine Learning** field has a robust, ready-made and alternative way for solving this type of the problem. We can harness the power of Machine Learning to build a classifier type of thing, which help us to classify which message is/are spam or ham depending upon their context or content.

OBJECTIVE

The main goal of this project is to detect the class of text message that is Spam or Ham based on content of message. **Spam is a junk mail/message, or an unsolicited mail/message. Spam e-mails are also those unwanted, unsolicited e-mails that are not intended for a specific receiver. The good, perfect, and official mails are known as *ham*.** Our objective is to alert user from spam or junk mail/messages so that they can save themselves from being cheated from spammers.

ARCHITECTURE



Start

Export data from database to csv for training

Data Preprocessing

Data Clustering

Get best model for each cluster

Hyperparameter Tuning

Model saving

Cloud Setup

Pushing app to cloud

Application Start

Data from client to be predicted

Export data from database to csv to prediction

Data Preprocessing

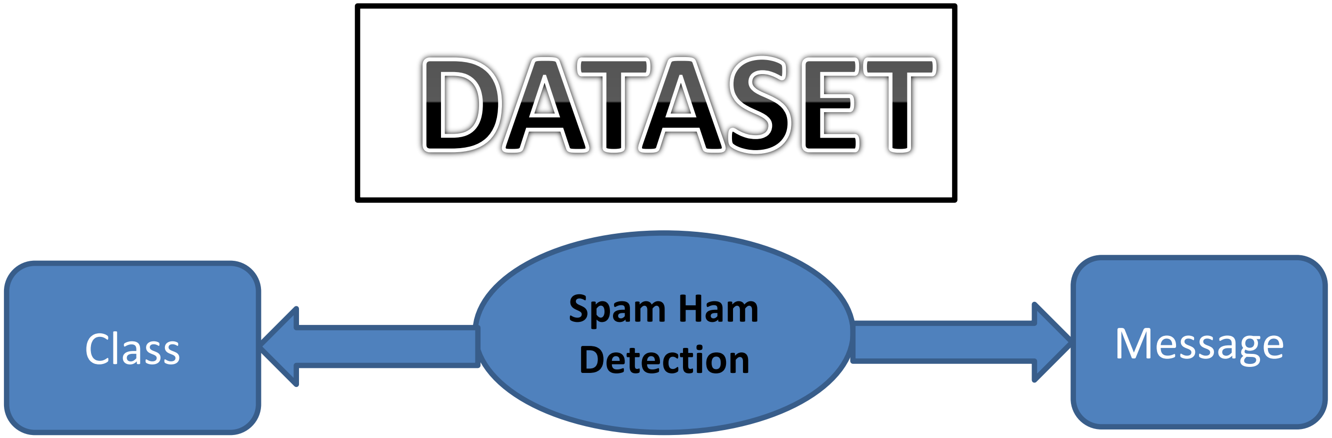
Data Clustering

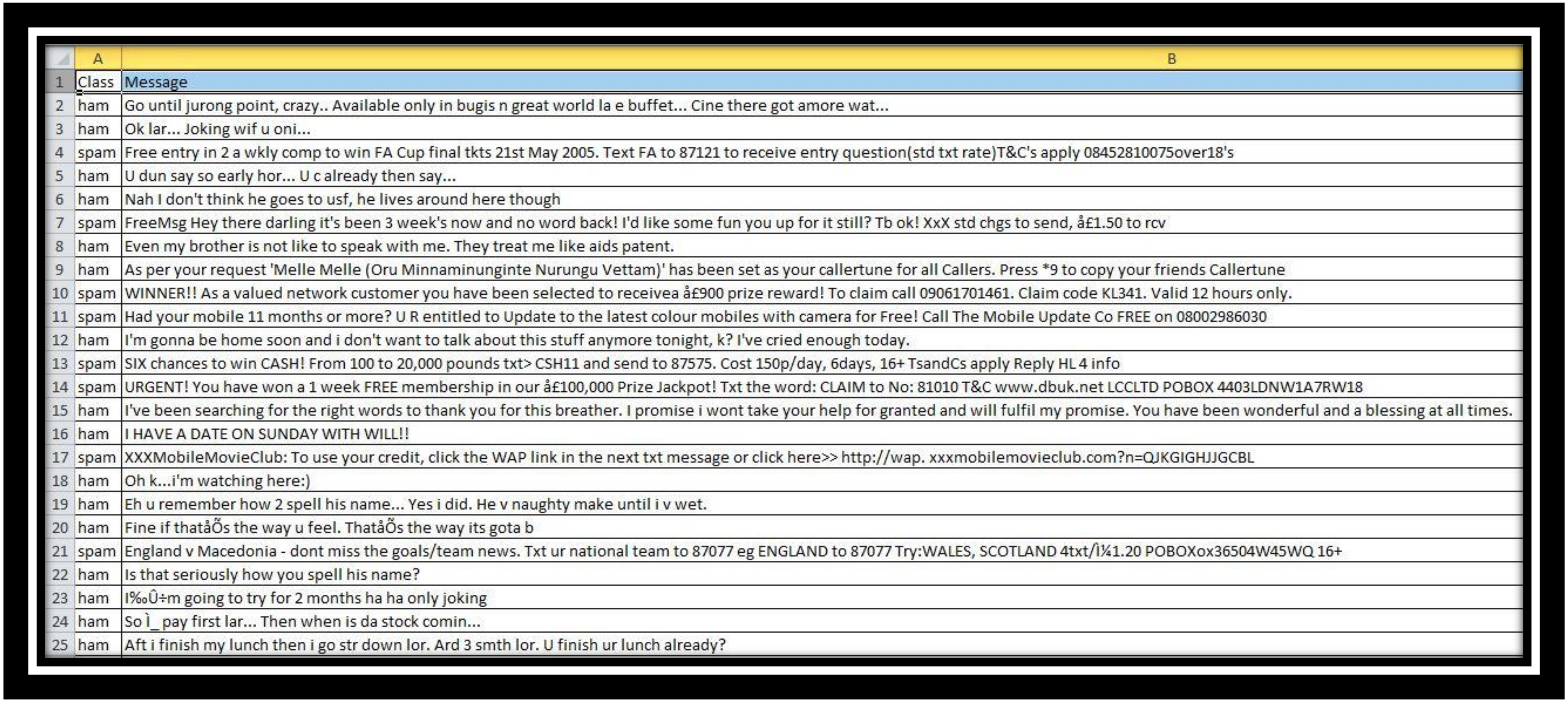
Model call for specific cluster

Prediction

Export prediction to csv

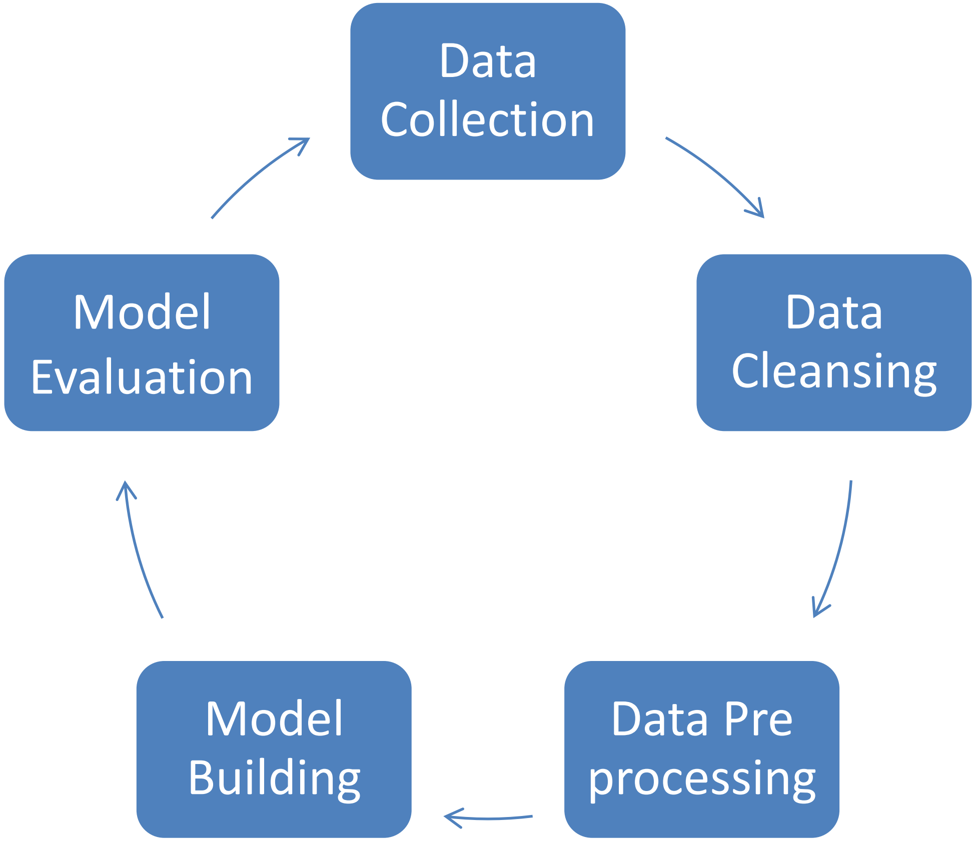
End





	A	B
1	Class	Message
2	ham	Go until jurong point, crazy.. Available only in bugis n great world la e buffet... Cine there got amore wat...
3	ham	Ok lar... Joking wif u oni...
4	spam	Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receive entry question(std txt rate)T&C's apply 08452810075over18's
5	ham	U dun say so early hor... U c already then say...
6	ham	Nah I don't think he goes to usf, he lives around here though
7	spam	FreeMsg Hey there darling it's been 3 week's now and no word back! I'd like some fun you up for it still? Tb ok! XxX std chgs to send, å£1.50 to rcv
8	ham	Even my brother is not like to speak with me. They treat me like aids patent.
9	ham	As per your request 'Melle Melle (Oru Minnaminunginte Nurungu Vettam)' has been set as your callertune for all Callers. Press *9 to copy your friends Callertune
10	spam	WINNER!! As a valued network customer you have been selected to receivea å£900 prize reward! To claim call 09061701461. Claim code KL341. Valid 12 hours only.
11	spam	Had your mobile 11 months or more? U R entitled to Update to the latest colour mobiles with camera for Free! Call The Mobile Update Co FREE on 08002986030
12	ham	I'm gonna be home soon and i don't want to talk about this stuff anymore tonight, k? I've cried enough today.
13	spam	SIX chances to win CASH! From 100 to 20,000 pounds txt> CSH11 and send to 87575. Cost 150p/day, 6days, 16+ TsandCs apply Reply HL 4 info
14	spam	URGENT! You have won a 1 week FREE membership in our å£100,000 Prize Jackpot! Txt the word: CLAIM to No: 81010 T&C www.dbuk.net LCCLTD POBOX 4403LDNW1A7RW18
15	ham	I've been searching for the right words to thank you for this breather. I promise i wont take your help for granted and will fulfil my promise. You have been wonderful and a blessing at all times.
16	ham	I HAVE A DATE ON SUNDAY WITH WILL!!
17	spam	XXXMobileMovieClub: To use your credit, click the WAP link in the next txt message or click here>> http://wap. xxxmobilemovieclub.com?n=QJKGIGHJJGCBL
18	ham	Oh k...i'm watching here:)
19	ham	Eh u remember how 2 spell his name... Yes i did. He v naughty make until i v wet.
20	ham	Fine if thatåÕs the way u feel. ThatåÕs the way its gota b
21	spam	England v Macedonia - dont miss the goals/team news. Txt ur national team to 87077 eg ENGLAND to 87077 Try:WALES, SCOTLAND 4txt/Ì¼1.20 POBOXox36504W45WQ 16+
22	ham	Is that seriously how you spell his name?
23	ham	I‰Û÷m going to try for 2 months ha ha only joking
24	ham	So ì_ pay first lar... Then when is da stock comin...
25	ham	Aft i finish my lunch then i go str down lor. Ard 3 smth lor. U finish ur lunch already?

Data Analysis Steps



Data Collection

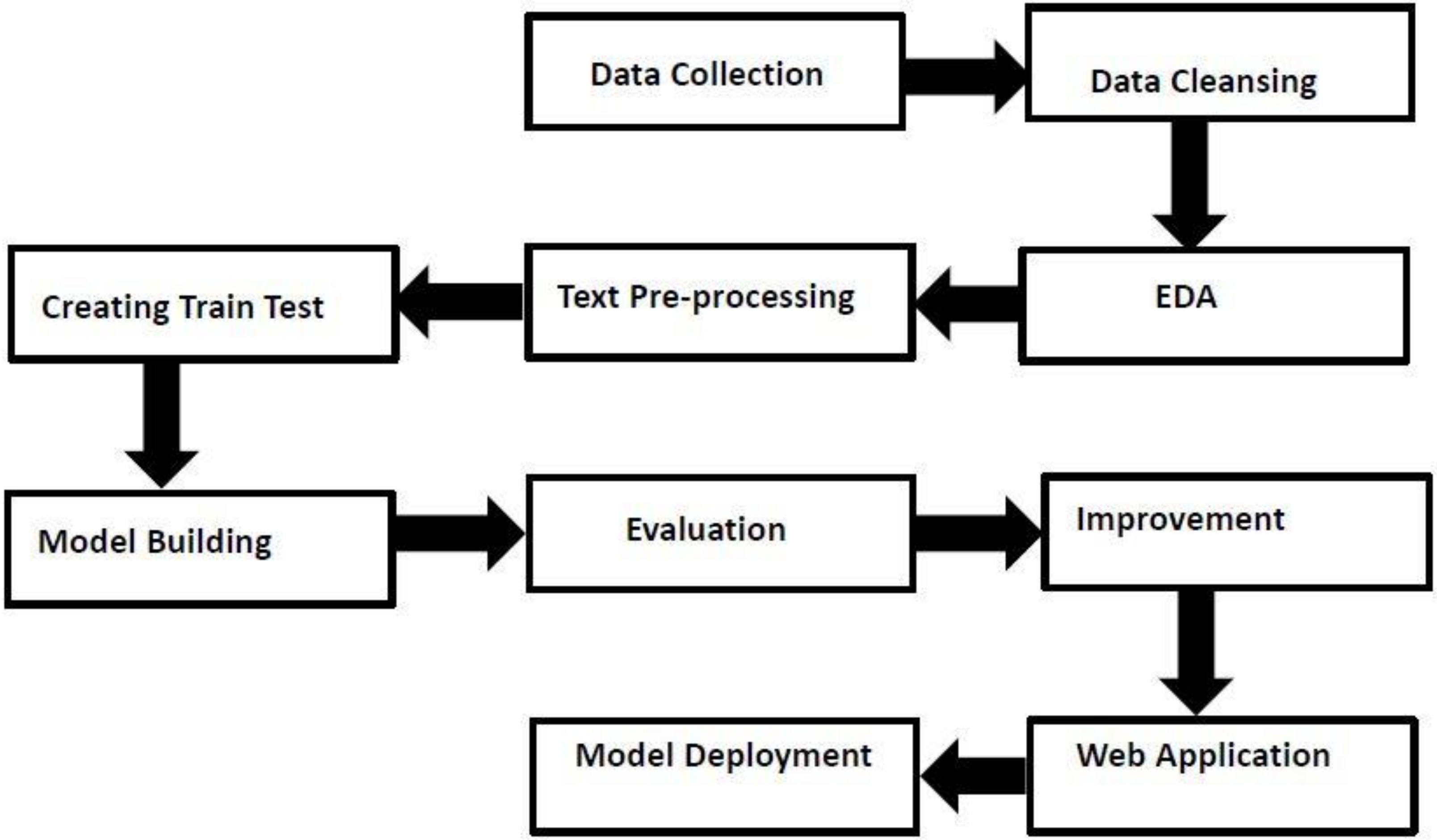
Data Cleansing

Data Pre processing

Model Building

Model Evaluation

MODEL TRAINING AND VALIDATION WORKFLOW



Data Collection

Data Cleansing

EDA

Text Pre-processing

Creating Train Test

Model Building

Evaluation

Improvement

Web Application

Model Deployment

MODEL TRAINING AND VALIDATION WORKFLOW

Data Collection

- **Spam Ham Message** Collection Set from UCI Machine Learning Repository

For Data Set: **https://archive.ics.uci.edu/ml/datasets/sms+spam+collection**

Data Cleansing

- Drop unnecessary columns.
- Drop duplicates rows from dataset.
- Rename required columns.
- Encode target column using Label Encoder.

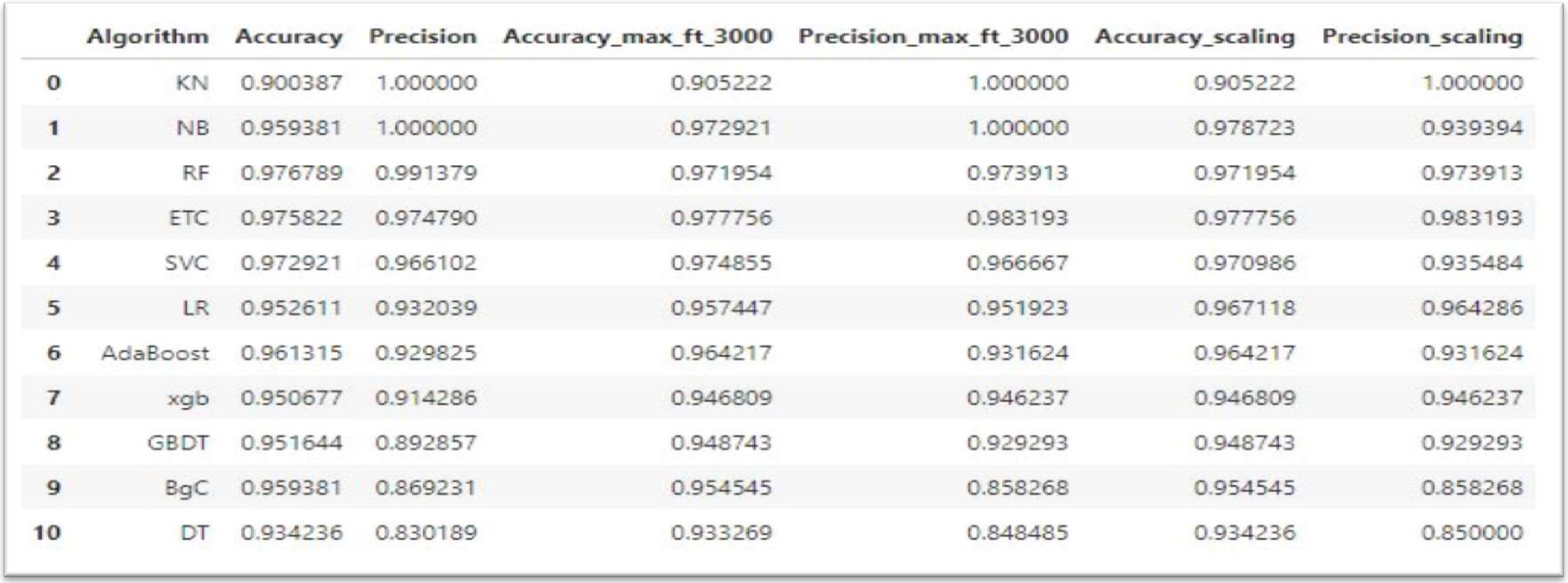
Data Pre-Processing

- Lower Case
- Tokenization
- Removing Special Characters
- Removing Stop Words and Punctuation
- Stemming

MODEL TRAINING AND VALIDATION WORKFLOW

Model Creation and Evaluation

- Various classification model created.
- Algorithms used are Naive Bayes, Logistic Regression, Decision Tree Classifier, Random Forest Classifier, Ada-Boost Classifier, Bagging Classifier etc.
- Multinomial Naïve Bayes classifier has given good accuracy and precision.
- Model performance evaluated based on accuracy, precision.
- In TFIDF Vectorizer and Count Vectorizer, We have selected TFIDF.



	Algorithm	Accuracy	Precision	Accuracy_max_ft_3000	Precision_max_ft_3000	Accuracy_scaling	Precision_scaling
0	KN	0.900387	1.000000	0.905222	1.000000	0.905222	1.000000
1	NB	0.959381	1.000000	0.972921	1.000000	0.978723	0.939394
2	RF	0.976789	0.991379	0.971954	0.973913	0.971954	0.973913
3	ETC	0.975822	0.974790	0.977756	0.983193	0.977756	0.983193
4	SVC	0.972921	0.966102	0.974855	0.966667	0.970986	0.935484
5	LR	0.952611	0.932039	0.957447	0.951923	0.967118	0.964286
6	AdaBoost	0.961315	0.929825	0.964217	0.931624	0.964217	0.931624
7	xgb	0.950677	0.914286	0.946809	0.946237	0.946809	0.946237
8	GBDT	0.951644	0.892857	0.948743	0.929293	0.948743	0.929293
9	BgC	0.959381	0.869231	0.954545	0.858268	0.954545	0.858268
10	DT	0.934236	0.830189	0.933269	0.848485	0.934236	0.850000

Multinomial Naive Bayes Algorithm

INTRODUCTION

Multinomial Naive Bayes algorithm is a probabilistic learning method that is mostly used in Natural Language Processing (NLP).

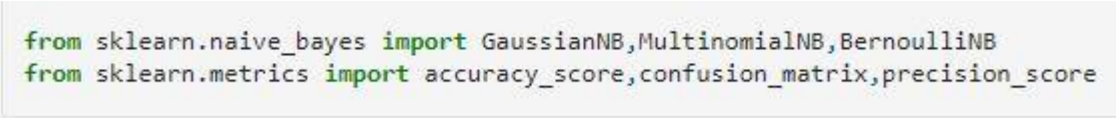
The algorithm is based on the Bayes theorem and predicts the tag of a text such as a piece of email or newspaper article. It calculates the probability of each tag for a given sample and then gives the tag with the highest probability as output.

Naive Bayes classifier is a collection of many algorithms where all the algorithms share one common principle, and that is each feature being classified is not related to any other feature. The presence or absence of a feature does not affect the presence or absence of the other feature.

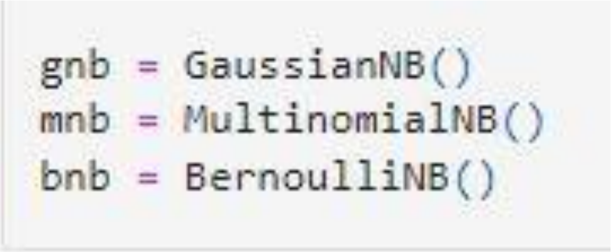
Advantages:

- It is easy to implement as you only have to calculate probability.
- You can use this algorithm on both continuous and discrete data.
- It is simple and can be used for predicting real-time applications.
- It is highly scalable and can easily handle large datasets.

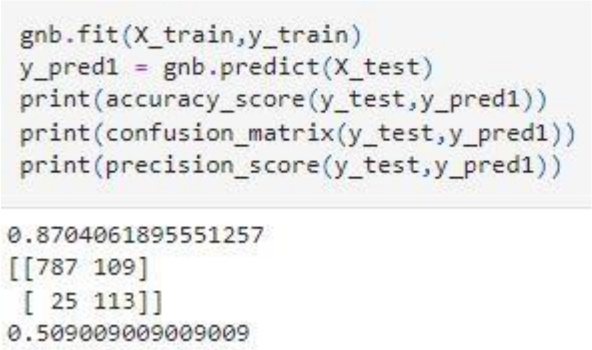
MODEL PREDICTION RESULTS ON TEST DATASET



```
from sklearn.naive_bayes import GaussianNB,MultinomialNB,BernoulliNB
from sklearn.metrics import accuracy_score,confusion_matrix,precision_score
```

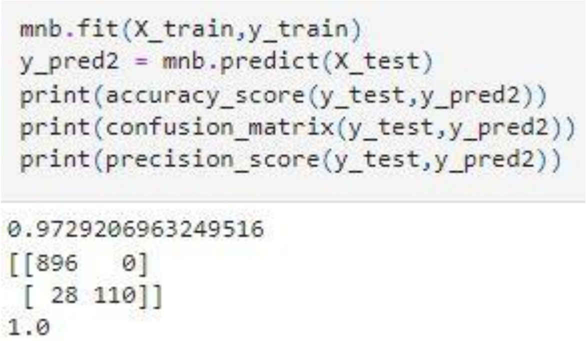


```
gnb = GaussianNB()
mnb = MultinomialNB()
bnb = BernoulliNB()
```



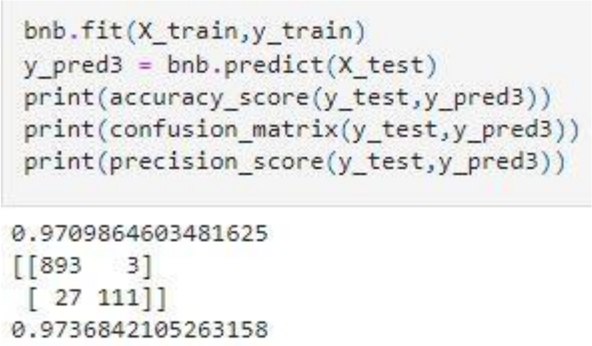
```
gnb.fit(X_train,y_train)
y_pred1 = gnb.predict(X_test)
print(accuracy_score(y_test,y_pred1))
print(confusion_matrix(y_test,y_pred1))
print(precision_score(y_test,y_pred1))
```

```
0.8704061895551257
[[787 109]
 [ 25 113]]
0.509009009009009
```



```
mnb.fit(X_train,y_train)
y_pred2 = mnb.predict(X_test)
print(accuracy_score(y_test,y_pred2))
print(confusion_matrix(y_test,y_pred2))
print(precision_score(y_test,y_pred2))
```

```
0.9729206963249516
[[896   0]
 [ 28 110]]
1.0
```



```
bnb.fit(X_train,y_train)
y_pred3 = bnb.predict(X_test)
print(accuracy_score(y_test,y_pred3))
print(confusion_matrix(y_test,y_pred3))
print(precision_score(y_test,y_pred3))
```

```
0.9709864603481625
[[893   3]
 [ 27 111]]
0.9736842105263158
```

FREQUENTLY ASKED QUESTIONS

Q1) What is the source of data?

The data for training is obtained from famous machine learning repository.
UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/sms+spam+collection

Q2) What was the type of data?

The data set was completely text data with two attributes, first class of text and second body of text.

Q3) What's the complete flow you followed in this Project?

Refer slide 7th, 8th and 9th for better understanding.

Q4) After the File validation what you do with incompatible file or files which didn't pass the validation?

Files like these are moved to the Achieve Folder and a list of these files has been shared with the client and we removed the bad data folder.

Q5) How logs are managed?

We are using different logs as per the steps that we follow in training and prediction like model training log and prediction log etc. And then sub log are inside those folder.

Q 6) What techniques were you using for data pre-processing?

- Lower Case
- Tokenization
- Removing Special Characters
- Removing Stop Words and Punctuation
- Stemming

Q 7) How training was done or what models were used?

- First Data validation done on raw data and then good data insertion happen in DB.
- Then Data cleansing and preprocessing done on final CSV file received from DB.
- We did clustering over the data to divide it on desired cluster based on elbow method.
- Various model such as Decision Tree, Random Forest , XGBoost and Naïve Bayes models are trained on all clusters and based on performance, for each cluster different model is saved.

Q 8) How Prediction was done?

- The testing files are shared by the client .We Perform the same life cycle till the data is clustered .
- Then on the basis of cluster number model is loaded and perform prediction. In the end we get the accumulated data of predictions.

Q 9) What are the different stages of deployment?

- After model training and finalizing all models. We created required files for deployment.
- Finally deployed our model over various cloud platforms such as Heroku and AWS.

Q 10) How is the User Interface present for this project?

- For this project I have made two types of UI.
- First is for bulk prediction.
- Second is for one user input prediction.
- Both UI are very user friendly and easy to use.

THANK YOU