Summer Training Report

On

Machine Learning

Submitted By:-

Saurabh Sharma

IT

158(2k18)

ACKNOWLEDGEMENT

It is my proud privilege and duty to acknowledge the kind of help and guidance received from several people in preparation of this report. It would not have been possible to prepare this report in this form without their valuable help, cooperation and guidance. First and foremost, I wish to record our sincere gratitude to 1Stop Coordinators for their constant support and encouragement in preparation of this report and for making available videos and interface facilities needed to prepare this report. The seminar on “Machine Learning” was very helpful to us in giving the necessary background information and inspiration in choosing this topic for the seminar. Their contributions and technical support in preparing this report are greatly acknowledged. Last but not the least, we wish to thank our parents for financing our studies in this college as well as for constantly encouraging us to learn engineering. Their personal sacrifice in providing this opportunity to learn engineering is gratefully acknowledged.

About the Organization

1stop is a technology company on a mission to equip students with relevant skills & practical exposure through internships and online trainings. Imagine a world full of freedom and possibilities. A world where you can discover your passion and turn it into your career. A world where your practical skills matter more than your university degree. A world where you do not have to wait till 21 to taste your first work experience (and get a rude shock that it is nothing like you had imagine it to be). A world where you graduate fully assured, fully confident, and fully prepared to stake claim on your place in the world.

Summer Training Certificate

IIT KANPUR

istop
Keep Growing

CERTIFICATE
OF PARTICIPATION

THIS CERTIFICATE IS PROUDLY PRESENTED TO

Saurabh Sharma

participated in "Machine Learning"
from 15th Jun, 2021 to 15th Aug, 2021
and successfully completed the program.

02-Sep-2021
DATE

PAUL MATHEW. I
OVERALL COORDINATOR

TIITK-2109000241

Introduction of Machine Learning

Machine learning (ML) is a type of artificial intelligence (AI) that allows software applications to become more accurate at predicting outcomes without being explicitly programmed to do so. Machine learning algorithms use historical data as input to predict new output values.

Type of Machine Learning:-

1.Supervised Learning
2.Unsupervised Learning

Supervised Learning:-

In this type of machine learning, data scientists supply algorithms with labelled training data and define the variables they want the algorithm to assess for correlations. Both the input and the output of the algorithm is specified.

Unsupervised Learning:-

This type of machine learning involves algorithms that train on unlabelled data. The algorithm scans through data sets looking for any meaningful connection. Both the data algorithms train on and the predictions or recommendations they output are predetermined.

Task best suited for Supervised Learning:-

- Binary classification. Dividing data into two categories.
- Multi-class classification. Choosing between more than two types of answers.
- Regression modelling. Predicting continuous values.
- Ensembling. Combining the predictions of multiple machine learning models to produce an accurate prediction.

Task best suited for Unsupervised Learning:-

- **Clustering.** Splitting the data set into groups based on similarity.
- **Anomaly detection.** Identifying unusual data points in a data set.
- **Association mining.** Identifying sets of items in a data set that frequently occur together.
- **Dimensionality Reduction.** Reducing the number of variables in a data set.

Uses of Machine Learning:-

- **Virtual Assistance**
  Smart assistants typically combine supervised and unsupervised machine learning models to interpret natural speech and supply context.

- **Self Driving Cars**
  Machine learning algorithms can even make it possible for a semi-autonomous car to recognize a partially visible object and alert the driver.

- **Business Intelligence**
  BI and analytics vendors use machine learning in their software to identify potentially important data points, patterns of data points and anomalies.
- **Human Resource Information System**
  HRIS systems can use machine learning models to filter through applications and identify the best candidates for an open position.
- **Recommandation Systems**
  Most of the companies, such as Amazon, Netflix, Google Play, etc., are using data science technology for making a better user experience with personalized recommendations. Such as, when you search for something on Amazon, and you start getting suggestions for similar products, so this is because of data science technology.

Training Content

- Introduction to Machine Learning
- Machine Learning Life Cycle
- Data Exploration and Manipulation
- Building Predictive model
- Understanding Different Type of Models
- Final Project

Final Project Problem Statement

Gmail Spam Detection:-

The objective of this project is to build a prediction model to predict whether a mail is spam or not by NLP (Natural Language Processing).

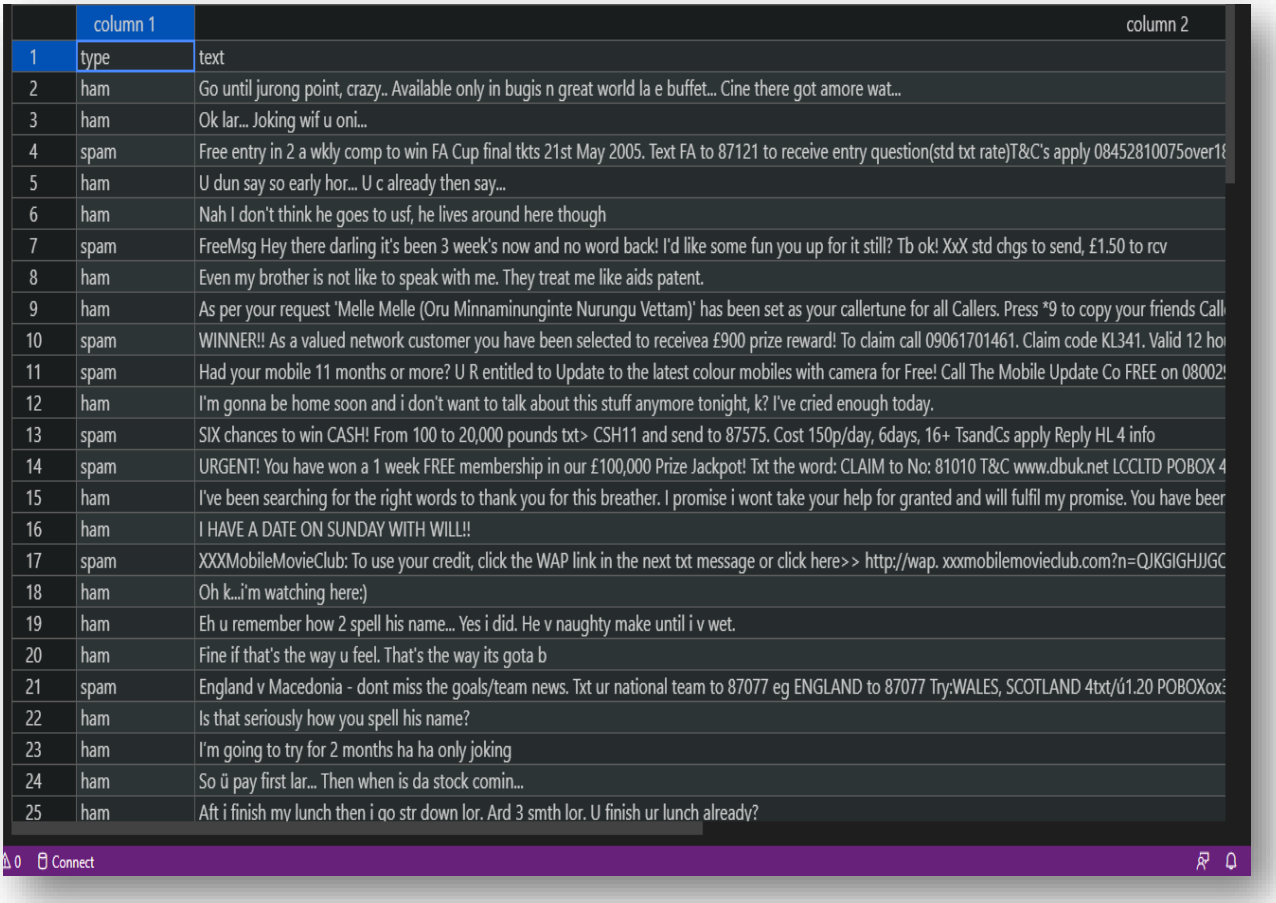
Step Involved:-

1. **Importing Dataset**
2. **Preprocessing Dataset**
3. **Vectorization**
4. **Training and Classification**
5. **Model Evaluation**

Data Dictionary:-

Data in dataset containing two column first is type of message (spam or not) and second column is text inside the mail.

Original Data Set



	column 1	column 2
1	type	text
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
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, £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 Call
10	spam	WINNER!! As a valued network customer you have been selected to receivea £900 prize reward! To claim call 09061701461. Claim code KL341. Valid 12 hou
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 08002
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 £100,000 Prize Jackpot! Txt the word: CLAIM to No: 81010 T&C www.dbuk.net LCCLTD POBOX 4
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 beer
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=QJKGIGHJJGC
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 that's the way u feel. That's 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/ú1.20 POBOXox3
22	ham	Is that seriously how you spell his name?
23	ham	I'm going to try for 2 months ha ha only joking
24	ham	So ü 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?

Mail Spam detection : Machine Learning

In [57]:
```
import numpy as np
import pandas as pd
import nltk
from nltk.corpus import stopwords
nltk.download('wordnet')
```

```
[nltk_data] Downloading package wordnet to
[nltk_data]     C:\Users\acer\AppData\Roaming\nltk_data...
[nltk_data]   Package wordnet is already up-to-date!
```

Out[57]: True

1) LOADING DATA

In [58]:
```
dt = pd.read_csv("spam.csv")
dt.head(10)
```

Out[58]:

	type	text
0	ham	Go until jurong point, crazy.. Available only ...
1	ham	Ok lar... Joking wif u oni...
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...
3	ham	U dun say so early hor... U c already then say...
4	ham	Nah I don't think he goes to usf, he lives aro...
5	spam	FreeMsg Hey there darling it's been 3 week's n...
6	ham	Even my brother is not like to speak with me. ...

6	ham	Even my brother is not like to speak with me. ...
7	ham	As per your request 'Melle Melle (Oru Minnamin...
8	spam	WINNER!! As a valued network customer you have...
9	spam	Had your mobile 11 months or more? U R entitle...

In [59]:
```
dt['spam'] = dt['type'].map( {'spam': 1, 'ham': 0} ).astype(int)
dt.head(5)
```

Out[59]:

	type	text	spam
0	ham	Go until jurong point, crazy.. Available only ...	0
1	ham	Ok lar... Joking wif u oni...	0
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...	1
3	ham	U dun say so early hor... U c already then say...	0
4	ham	Nah I don't think he goes to usf, he lives aro...	0

In [60]:
```
print("COLUMS IN THE GIVEN DATA:")
for col in dt.columns:
    print(col)
```

```
COLUMS IN THE GIVEN DATA:
type
text
spam
```

In [61]:
```
t=len(dt['type'])
print("NO OF ROWS IN REVIEW COLUMN:",t)
t=len(dt['text'])
print("NO OF ROWS IN liked COLUMN:",t)
```

```
NO OF ROWS IN REVIEW COLUMN: 116
NO OF ROWS IN liked COLUMN: 116
```

2)Tokenization

In [62]:
```
dt['text'][3]#before
```

Out[62]:
```
'U dun say so early hor... U c already then say...'
```

In [63]:
```
def tokenizer(text):
    return text.split()
```

In [64]:
```
dt['text']=dt['text'].apply(tokenizer)
```

In [65]:
```
dt['text'][3]#after
```

Out[65]:
```
['U',
 'dun',
 'say',
 'so',
 'early',
 'hor...',
 'U',
 'c',
 'already',
 'then',
 'say...']
```

3) STEMMING

In [66]:
```
dt['text'][3]#before
```

Out[66]:
```
['U',
 'dun',
 'say',
 'so',
```

```
 'early',
 'hor...',
 'U',
 'c',
 'already',
 'then',
 'say...']
```

In [67]:
```
from nltk.stem.snowball import SnowballStemmer
porter = SnowballStemmer("english", ignore_stopwords=False)
```

In [68]:
```
def stem_it(text):
    return [porter.stem(word) for word in text]
```

In [69]:
```
dt['text']=dt['text'].apply(stem_it)
```

In [70]:
```
dt['text'][] #after stemming
```

Out[70]:
```
['u',
 'dun',
 'say',
 'so',
 'earli',
 'hor...',
 'u',
 'c',
 'alreadi',
 'then',
 'say...']
```

4) LEMMITIZATION

In [71]:
```
dt['text'][80] #before
```

Out[71]:
```
['sorry,', "i'll", 'call', 'later']
```

In [72]:
```
from nltk.stem import WordNetLemmatizer
lemmatizer = WordNetLemmatizer()
```

In [73]:
```
def lemmit_it(text):
    return [lemmatizer.lemmatize(word, pos ="a") for word in text]
```

In [74]:
```
dt['text']=dt['text'].apply(lemmit_it)
```

In [75]:
```
dt['text'][80] #before
```

Out[75]:
```
['sorry,', "i'll", 'call', 'late']
```

5)STOPWORD REMOVAL

In [76]:
```
dt['text'][111] #before
```

Out[76]:
```
['what', 'is', 'the', 'plural', 'of', 'the', 'noun', 'research?']
```

In [77]:
```
#from nltk.corpus import stopwords
#print(stopwords.words('english'))
#set(['a','t','d','y'])
```

In [78]:
```
from nltk.corpus import stopwords
stop_words = stopwords.words('english')
```

In [79]:
```
def stop_it(text):
    review = [word for word in text if not word in stop_words ]
    return review
```

In [80]:
```
dt['text']=dt['text'].apply(stop_it)
```

In [81]:
```
dt['text'][111] #after
```

In [81]:
```
dt['text'][111] #after
```

Out[81]:
```
['plural', 'noun', 'research?']
```

In [82]:
```
dt.head(10)
```

Out[82]:

	type	text	spam
0	ham	[go, jurong, point,, crazy.., avail, onli, bug...	0
1	ham	[ok, lar..., joke, wif, u, oni...]	0
2	spam	[free, entri, 2, wkli, comp, win, fa, cup, fin...	1
3	ham	[u, dun, say, earli, hor..., u, c, alreadi, sa...	0
4	ham	[nah, think, goe, usf,, live, around, though]	0
5	spam	[freemsg, hey, darl, 3, week, word, back!, i'd...	1
6	ham	[even, brother, like, speak, me., treat, like,...	0
7	ham	[per, request, mell, mell, (oru, minnaminungin...	0
8	spam	[winner!!, valu, network, custom, select, rece...	1
9	spam	[mobil, 11, month, more?, u, r, entitl, updat,...	1

In [83]:
```
dt['text']=dt['text'].apply(' '.join)
```

In [84]:
```
dt.head()
```

Out[84]:

	type	text	spam
0	ham	go jurong point, crazy.. avail onli bugi n gre...	0
1	ham	ok lar... joke wif u oni...	0
2	spam	free entri 2 wkli comp win fa cup final tkts 2...	1
3	ham	u dun say earli hor... u c alreadi say...	0

3	ham	u dun say earli hor... u c alreadi say...	0
4	ham	nah think goe usf, live around though	0

6) Transform Text Data into TDF /TF-IDF Vectors

In [85]:
```
from sklearn.feature_extraction.text import TfidfVectorizer
tfidf=TfidfVectorizer()
y=dt.spam.values
x=tfidf.fit_transform(dt['text'])
```

In [86]:
```
from sklearn.model_selection import train_test_split
x_train,x_text,y_train,y_text=train_test_split(x,y,random_state=1,test_size=0.2,shuffle=False)
```

7) Classification

In [87]:
```
from sklearn.linear_model import LogisticRegression
clf=LogisticRegression()
clf.fit(x_train,y_train)
y_pred=clf.predict(x_text)
from sklearn.metrics import accuracy_score
acc_log = accuracy_score(y_pred, y_text)*100
print("accuracy:",acc_log )
```

```
accuracy: 87.5
```

8) Classification using LinearSVC Accuracy

In [88]:
```
from sklearn.svm import LinearSVC

linear_svc = LinearSVC(random_state=0)
```

7) Classification

In [87]:
```
from sklearn.linear_model import LogisticRegression
clf=LogisticRegression()
clf.fit(x_train,y_train)
y_pred=clf.predict(x_text)
from sklearn.metrics import accuracy_score
acc_log = accuracy_score(y_pred, y_text)*100
print("accuracy:",acc_log )
```

accuracy: 87.5

8) Classification using LinearSVC Accuracy

In [88]:
```
from sklearn.svm import LinearSVC

linear_svc = LinearSVC(random_state=0)
linear_svc.fit(x_train, y_train)
y_pred = linear_svc.predict(x_text)
acc_linear_svc =accuracy_score(y_pred, y_text) * 100
print("accuracy:",acc_linear_svc)
```

accuracy: 87.5

Saurabh Sharma