VISVESVARAYA TECHNOLOGICAL UNIVERSITY
JNANASANGAMA, BELAGAVI – 590018

An Internship Report
on

SPAM MAIL PREDICTION USING MACHINE LEARNING

Submitted in partial fulfillment for the award of degree of

Bachelor of Engineering
In
Computer Science and Engineering

Submitted by
Rhea Benedicta D'souza
4SO18CS097

Internship Carried Out
at
Cognitive Solution
5th Floor, M.K. Shalimar Complex
Kankanady, Mangalore - 575002

Internal Guide
Ms Supriya Salian
Assistant Professor
St. Joseph Engineering College

External Guide
Mrs. Sibby Susan
Technical Trainer
Cognitive Solution

Department of Computer Science and Engineering
St Joseph Engineering College
Mangaluru – 575028

2021-22

VISVESVARAYA TECHNOLOGICAL UNIVERSITY
JNANASANGAMA, BELAGAVI – 590018

An Internship Report
on

SPAM MAIL PREDICTION USING MACHINE LEARNING

Submitted in partial fulfillment of the requirements for the degree

Bachelor of Engineering
in

Computer Science and Engineering

Submitted by

Rhea Benedicta D'souza **4SO18CS097**

Department of Computer Science and Engineering
St Joseph Engineering College
Mangaluru - 575028
2021-22

St Joseph Engineering College
Mangaluru – 575 028
Department of Computer Science and Engineering

CERTIFICATE

Certified that the Internship Work titled **"SPAM MAIL PREDICTION USING MACHINE LEARNING"** was carried out by **Ms. Rhea Benedicta D'souza**, bearing USN **4SO18CS097**, a bonafide student of final year B.E. in partial fulfillment for the award of Bachelor of Engineering in Computer Science and Engineering of the Visvesvaraya Technological University, Belagavi, during the year 2021-22. Further, it is certified that all corrections/suggestions indicated during Internal Evaluation have been incorporated in this report.

----------	----------	----------
Ms. Supriya Salian	**Dr Sridevi Saralaya**	**Dr. Rio D'Souza**
Internal Guide	**Head of the Department**	**Principal**

External Viva Voce Examination

Name of the Examiners	**Signature with Date**
1. ----------	----------
2. ----------	.----------

COGNITIVE SOLUTION

Date: 29/10/2021

CERTIFICATE OF INTERNSHIP

We the undersigned do hereby proudly present this Certificate of Internship for outstanding honorable effort of **Ms. RHEA BENEDICTA D'SOUZA**, Bonafide Student of **ST. JOSEPH ENGINEERING COLLEGE, MANGALORE** with registration number **4SO18CS097** for her successful completion in "Machine Learning using Python" internship at Cognitive Solution from 8th September to 8th October 2021.

We found **Ms. RHEA BENEDICTA D'SOUZA** is effective in discharging responsibilities assigned to her. We found her effective and her character and conduct were good.

We wish success in her future assignment.

Mr.Shameer Ahmed,
Managing Director,
Mangalore.

COGNITIVE SOLUTION
5th Floor, Shalimar Complex
Next To Mak Mall
Kankanady, Mangalore - 575 002

DECLARATION

I, **Rhea Benedicta D'souza,** bearing USN **4SO18CS097**, student of final year B.E. in Computer Science and Engineering, St Joseph Engineering College, Mangaluru, hereby declare that the Internship Work titled **"SPAM MAIL PREDICTION USING MACHINE LEARNING"** has been duly executed by me from 8th September – 8th October 2021, at Cognitive Solution, Kankanady, Mangalore. Further, the "Tasks Performed" section of this report represents the work done solely by me and does not contain any statements falsely claiming work done by others, as my own.

Date: 30/4/2022

Place: Mangalore **Rhea Benedicta Dsouza**

ACKNOWLEDGMENT

I dedicate this page to acknowledge and thank those responsible for shaping this project. Without their guidance and help, this experience would not have been so smooth and efficient.

I would like to extend my sincere gratitude to **Ms. Siby Sussan**, Technical Trainer at Cognitive Solution for giving me the opportunity to complete my internship, his guidance and encouragement helped me throughout the internship.

I sincerely thank **Ms Supriya Salian**, Assistant Professor, Department of Computer Science and Engineering for her guidance which helped us fulfill the requirements prescribed by the university and her valuable suggestions which brought this internship to fruition.

I am indebted to **Dr. Sridevi Saralaya**, Head of the Department of Computer Science and Engineering, whose kind consent and guidance helped us complete this internship successfully.

I am grateful to our Director, **Rev. Fr Wilfred Prakash D'Souza**, our Assistant Director, **Rev. Fr Alwyn Richard D'Souza** and our Principal, **Dr. Rio D'Souza** for their support and encouragement.

I am grateful to the staff of Computer Science and Engineering Department for their encouragement and support.

I would also like to thank my friends for their valuable suggestions and family members for their continued support during the period of my internship.

Executive Summary

I carried out my internship in Machine Learning using Python at Cognitive Solution, Kankanady, Mangalore from 8th September 2021 to 8th October 2021.

Cognitive Solutions is a service provider of Web-based Development & Web-based Software Development Solutions, Mobile Application Development, Graphic Design, and Windows Applications. Cognitive Solutions is headquartered in Mangalore, with the Business Development in UAE, Saudi Arabia, and Qatar. In a short span of 8+ years, our products, as well as services & solutions, have been widely accepted by the global market. Today, Cognitive Solution has the experience to undertake any IT development or deployment works on a single point responsibility basis. Their Products and Services are user-friendly with easy controls and are of superior specifications. They are always proactive to fulfill clients' needs and requirements to the best possible extent of their satisfaction. They manage interactive sessions with clients throughout the project development.

The Objective of the internship was to gain knowledge in machine learning and build a project for a better understanding of concepts and to enhance my skill. The language used was python.

This internship gave me an opportunity to enhance my professional skills and gain industry experience. It also encouraged me to learn new technologies and tools like Machine Learning. Further, it helped to develop my presentation skill and time management. I was able to complete the task in a short duration of time. This internship also provided me an opportunity to apply acquired knowledge to real work experiences. Overall, the internship helped me gain valuable work experience and enhanced my skills which is very essential to my career.

RHEA BENEDICTA D'SOUZA

CONTENTS

Figure Index

CHAPTER 1

ABOUT THE COMPANY

1.1 Brief History

Cognitive Solutions is a rapidly growing company in the field of computer application implementation, solutions, and services. Cognitive Solutions is a service provider of web-based Development & Web-based Software Development Solutions, mobile application Development, Graphic Design, and Windows Applications. Cognitive Solutions is headquartered in Mangalore, with the Business Development in UAE, Saudi Arabia, and Qatar. In a short span of 8+ years, our products, as well as services & solutions, have been widely accepted by the global market. Today, Cognitive Solution has the experience to undertake any IT development or deployment works on a single point responsibility basis. Their efficient and experienced team is a great resource. Intellect's infrastructure houses a team of young and competitive professionals having experience in Web Designing and Software Development who are dedicated to providing a high-end solution to their clients.

The team at Cognitive Solutions:

- Mr. Shameer Ahmed, Company Head
- Ms. Siby Sussan, Industry Guide

1.2 Services Offered by the Company

Cognitive Solutions develop software and web-based applications with the latest technologies. For web development projects, they also provide hosting and domain facility for customers, so they don't need to bother about that. Their Products and Services are user-friendly with easy controls and are of superior specifications. They are always proactive to fulfill clients' needs and requirements to the best possible extent of their satisfaction. They manage interactive sessions with clients throughout the project development.

1.3 Contact Details

Address	:	5th Floor, M.K. Shalimar Complex, Kankanady, Mangalore - 575002
Phone	:	0824-4283434
Email	:	contactuscognitive@gmail.com

CHAPTER 2
ABOUT THE DEPARTMENT

2.1 Introduction

The internship introduced training on machine learning using python. The training department of Cognitive Solution provides training for Mobile app development, Progressive Web app development, Angular JavaScript, Python, IoT, ML, AI, Data Science, Advanced Excel, and Digital Marketing. They are passionate about the subjects they teach and deliver enthusiasm through their webinars and courses. The professionals at Cognitive Solution are expertise in their subjective fields and are also certified by recognized industrial standards.

2.2 Roles and Responsibilities

During the start of the internship as we were beginners in Machine Learning, we were asked to practice what was taught and solve various tasks assigned by our external guide so that we understand and utilize how these methods could be applied in the real world while solving the daily tasks. After having a basic understanding in Machine Learning, I was assigned to work on a project assigned by the selecting any topic of my choice so that I am well versed with the concepts taught. The project selected by me was to detect spam and ham mail using Machine Learning. Here the data set which I chose was from Kaggle and I used it to train my model. Data collection and pre processing, label encoding, splitting dataset into training and testing etc.. were some of the steps implemented in my project.

CHAPTER 3
TASKS PERFORMED

3.1 Daily Work Schedule

Week 1 (08/09/2021 - 11/09/2021):

In week 1, in order to move ahead to machine learning with AI, basic python concepts were brushed up, like syntax, data types operators etc and a few exercises were given so that I was well versed with the basic concepts.

Week 2 (13/09/2021 - 16/09/2021):

In week 2, concepts such as loops, object-oriented programming, JSON, and Regular Expressions. I was given exercises on these concepts. The facilitator then introduced me to the basics of NumPy, Pandas and Matplotlib.

Week 3 (20/09/2021 - 23/09/2021):

In week 3, I was introduced to machine learning basics. Differences between supervised and unsupervised learning techniques was explained. Linear regression was taught by giving a demo exercise. I was then made to implement it in my system and analyze it.

Week 4 (27/09/2021 - 30/09/2021):

In week 4, classification algorithms were introduced. Other algorithms introduced were KNN (K Nearest Neighbor), SVM (Support Vector Machine), and Naive Bayes. Exercises on the same were taught and implemented.

Week 5 (01/10/2021 - 04/10/2021):

In week 5, Logistic regression, decision tree classification, and Kmeans clustering were taught. Problems on decision tree classification were given. Kmeans algorithm was taught and implemented. I was tasked with completing a project by selecting a topic of my choice on any of the algorithms I was aware of.

Week 6 (05/10/2021 - 08/10/2021):

In week 6, I I went through some of the topics taught and chose my topic of interest and started working on spam mail detection. This project used a logistic regression model to train the mail dataset. I created a kaggle account and chose one of the dataset which contained a good number of spam and ham mails. I altered the dataset by taking some mails from my gmail account and adding them onto the dataset. The next two days I spent on my project, by performing the following steps:

importing dependencies, data collection and pre-processing, label encoding, splitting data into importing dependencies, data collection and pre-processing, label encoding, splitting data into training and testing, feature extraction, training the model, evaluating the trained model and building a predictive system. I successfully completed my project on time by the 7th of October 2022.

3.2 Project Implementation

PROBLEM STATEMENT: In the new era of technical advancement, e-mails have gathered significant users for professional, commercial, and personal communications. Because of the high demand and huge user base, there is an upsurge in unwanted emails, also known as spam emails. Even in the current date, people lose a lot of money to frauds every day.

OBJECTIVE: The objective of identification of spam e-mails are:
a) To give knowledge to the user about the fake e-mails and relevant e-mails
b) To identify if the mail is spam or not.

WORKFLOW :

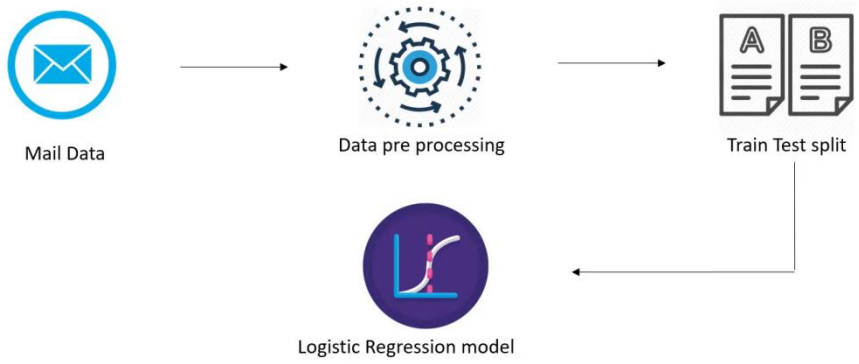

Fig 3.2a : Training Testing phase

Fig 3.2b : New email classification

3.3 PROJECT SNAPSHOTS:

3.3.1 Importing the dependencies

```
In [1]: import numpy as np
        import pandas as pd
        from sklearn.model_selection import train_test_split
        from sklearn.feature_extraction.text import TfidfVectorizer
        from sklearn.linear_model import LogisticRegression
        from sklearn.metrics import accuracy_score
```

Fig 3.3.1 : Importing the dependencies

Here, numpy is used to create an array,pandas is used to create data frame, train_test_split is used to split our data into training and testing, TfidfVectorizer is used to convert text data(mail data) into numerical value so our machine learning model can understand it, LogisticRegression is used to classify mails into spam mail or ham mail, and accuracy score is used to know how well our model is predicting.

3.3.2 Data Collection and Pre-Processing

```
In [29]: # loading the data from csv file to a pandas Dataframe
         raw_mail_data = pd.read_csv('mail_data.csv')

In [30]: print(raw_mail_data)

     Category                                            Message
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...
...       ...                                                ...
5567     spam  This is the 2nd time we have tried 2 contact u...
5568      ham                        wil you be going home?
5569      ham                    Please respond at the earliest
5570      ham         The due date for the assignment has passed
5571      ham                      Rofl. Its true to its name

[5572 rows x 2 columns]
```

Fig 3.3.2a: Display of raw mail data

```
In [31]: # replace the null values with a null string
         mail_data = raw_mail_data.where((pd.notnull(raw_mail_data)),'')

In [32]: # printing the first 5 rows of the dataframe
         mail_data.head()

Out[32]:
```

	Category	Message
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...

Fig 3.3.2b: Display first 5 rows

3.3.3 Label Encoding:

```
In [33]: # checking the number of rows and columns in the dataframe
         mail_data.shape

Out[33]: (5572, 2)

In [34]: # label spam mail as 0;  ham mail as 1;

         mail_data.loc[mail_data['Category'] == 'spam', 'Category',] = 0
         mail_data.loc[mail_data['Category'] == 'ham', 'Category',] = 1

In [35]: # separating the data as texts and label

         X = mail_data['Message']

         Y = mail_data['Category']

In [36]: print(X)

         0       Go until jurong point, crazy.. Available only ...
         1                           Ok lar... Joking wif u oni...
         2       Free entry in 2 a wkly comp to win FA Cup fina...
         3       U dun say so early hor... U c already then say...
         4       Nah I don't think he goes to usf, he lives aro...
                                                ...
         5567    This is the 2nd time we have tried 2 contact u...
         5568                                wil you be going home?
         5569                       Please respond at the earliest
         5570          The due date for the assignment has passed
         5571                             Rofl. Its true to its name
         Name: Message, Length: 5572, dtype: object
```

Fig 3.3.3a: Display of messages after separation

Here, the number of rows and columns consisting in the data frame are displayed. This is followed by label encoding which means labeling the spam mails as '0' and ham mails (non-spam mails) as '1'. The data is then separated as X and Y, where X contains the mails and Y contains the category of the mail which is nothing but 1 (ham) and 0 (spam).

```
In [11]: print(Y)

         0       1
         1       1
         2       0
         3       1
         4       1
                ..
         5567    0
         5568    1
         5569    1
         5570    1
         5571    1
         Name: Category, Length: 5572, dtype: object
```

Fig 3.3.3b: Displaying the category after separation

3.3.4 Splitting the data into training data and test data:

```
In [37]: X_train, X_test, Y_train, Y_test = train_test_split(X, Y, test_size=0.2, random_state=3)

In [38]: print(X.shape)
         print(X_train.shape)
         print(X_test.shape)

         (5572,)
         (4457,)
         (1115,)
```

Fig 3.3.4: Displaying the number of data split into training and test

Here, 80% of the data is going into training and 20% of the data is going into testing.

3.3.5 Feature Extraction:

```
In [39]: # transform the text data to feature vectors that can be used as input to the Logistic regression

feature_extraction = TfidfVectorizer(min_df = 1, stop_words='english', lowercase='True')

X_train_features = feature_extraction.fit_transform(X_train)
X_test_features = feature_extraction.transform(X_test)

# convert Y_train and Y_test values as integers

Y_train = Y_train.astype('int')
Y_test = Y_test.astype('int')
```

Fig 3.3.5a: Transforming text data into feature vectors

Here, we load TfidVectorizer into a variable called feature extraction. It is used to transform Text data into feature vectors. The two important steps taking place here are: fitting all our training data into vectorizer and following this it will transform all the x_train data into feature vectors which are nothing but numerical values.

```
In [16]: print(X_train_features)

  (0, 5413)     0.6198254967574347
  (0, 4456)     0.4168658090846482
  (0, 2224)     0.413103377943378
  (0, 3811)     0.34780165336891333
  (0, 2329)     0.38783870336935383
  (1, 4080)     0.18880584110891163
  (1, 3185)     0.29694482957694585
  (1, 3325)     0.31610586766078863
  (1, 2957)     0.3398297002864083
  (1, 2746)     0.3398297002864083
  (1, 918)      0.22871581159877646
  (1, 1839)     0.2784903590561455
  (1, 2758)     0.3226407885943799
  (1, 2956)     0.33036995955537024
  (1, 1991)     0.33036995955537024
  (1, 3046)     0.2503712792613518
  (1, 3811)     0.17419952275504033
  (2, 407)      0.509272536051008
  (2, 3156)     0.4107239318312698
  (2, 2404)     0.45287711070606745
  (2, 6601)     0.6056811524587518
  (3, 2870)     0.5864269879324768
  (3, 7414)     0.8100020912469564
  (4, 50)       0.23633754072626942
  (4, 5497)     0.15743785051118356
  :     :
  (4454, 4602)  0.2669765732445391
  (4454, 3142)  0.320144516777631 56
  (4455, 2247)  0.37052851863170466
  (4455, 2469)  0.35441545511837946
```

Fig 3.3.5b: Displaying x_train_features

3.3.6 Training the model

```
In [35]: model = LogisticRegression()

In [36]: # training the Logistic Regression model with the training data
         model.fit(X_train_features, Y_train)

Out[36]: LogisticRegression()
```

Fig 3.3.6: Training the model with the training data

3.3.7 Evaluating the trained model

```
In [37]: # prediction on training data

         prediction_on_training_data = model.predict(X_train_features)
         accuracy_on_training_data = accuracy_score(Y_train, prediction_on_training_data)

In [38]: print('Accuracy on training data : ', accuracy_on_training_data)

         Accuracy on training data :  0.9670181736594121
```

Fig 3.3.7a: Accuracy on training data

Here, we can observe that if we use this model to predict 100 different mails, it will predict them with 96% accuracy.

```
In [39]: # prediction on test data

         prediction_on_test_data = model.predict(X_test_features)
         accuracy_on_test_data = accuracy_score(Y_test, prediction_on_test_data)

In [40]: print('Accuracy on test data : ', accuracy_on_test_data)

         Accuracy on test data :  0.9659192825112107
```

Fig 3.3.7b: Accuracy on test data

3.3.8 Building a predictive system

```
In [64]: input_mail = ["You are awarded a SiPix Digital Camera! call 09061221061 from landline. Delivery within 28days. T Cs Box177. M221B

         # convert text to feature vectors
         input_data_features = feature_extraction.transform(input_mail)

         # making prediction

         prediction = model.predict(input_data_features)
         print(prediction)


         if (prediction[0]==1):
           print('Ham mail')

         else:
           print('Spam mail')

         [0]
         Spam mail
```

Fig 3.3.8a: Spam mail prediction

In [43]:
```
input_mail = ["Please respond at the earliest. Your assignment due date has passed"]

# convert text to feature vectors
input_data_features = feature_extraction.transform(input_mail)

# making prediction

prediction = model.predict(input_data_features)
print(prediction)


if (prediction[0]==1):
  print('Ham mail')

else:
  print('Spam mail')
```

```
[1]
Ham mail
```

Fig 3.3.8b: Ham mail prediction

CHAPTER 4

REFLECTION NOTES

4.1 Experience

My internship experience at Cognitive Solution, has taught me not only the technical concepts but also to work as a team by discussing the tasks assigned by our external with my peers assigned by our external guide. My External Guide, Ms. Siby Sussan is a continuous source of inspiration. Her guidance has helped me and my peers to learn new concepts and complete the project in this internship at ease. I gained valuable work experience during this period and the internship also made me realize how to work and discuss as a team by applying and following various work ethics and morals as well as executing the given tasks within the deadline.

4.2 Technical Outcomes

1. Learned new concepts and Technologies:

During the course of the internship, I had to constantly learn new concepts and technologies such as Data Pre-processing and prepare it for machine learning algorithms. After that, we were taught on Classification and Regression algorithms with the basics of Natural Language Processing.

2. Understood the importance of accuracy in Machine Learning:

Companies use machine learning models to make practical business decisions, and more accurate model outcomes result in better decisions. The cost of errors can be huge, but optimizing model accuracy mitigates that cost. There is, of course, a point of diminishing returns when the value of developing a more accurate model won't result in a corresponding profit increase, but often it is beneficial across the board. A false positive cancer diagnosis, for example, costs both the hospital and the patient. The benefits of improving model accuracy help avoid the considerable time, money, and undue stress.

3. Understood the importance of Machine Learning and its different types:

Machine learning is important because it gives enterprises a view of trends in customer behavior and business operational patterns, as well as supports the development of new products. Many of today's leading companies, such as Facebook, Google, and Uber, make machine learning a central part of their operations. Machine learning has become a significant competitive differentiator for many companies. There are four basic approaches: Supervised Learning, Unsupervised Learning, Semi-supervised Learning, Reinforcement Learning

4.3 Non-Technical Outcomes

1. Time Management: During the internship, I had to execute all the weekly tasks assigned by the External Guide within the given time frame. It helped me learn to manage my time better by maintaining a balance between my work and personal life.

2. Communication Skills: In this internship, after the project work was done, we had to present a seminar on the project which we worked on and we had to write report on it. Our verbal communication had to clear and concise in so that our questions and problems encountered in the project were clearly understood by others.

3. Work Ethics: The internship helped me develop work ethics i.e., submitting weekly assignments on time, attending all the sessions and completing all the project works within the deadline.

4. Adaptability Skills: During this internship, we had a lot of tasks which had to be completed within the week itself. So, managing both the assignments and session, I finally adapted to the work environment.

REFERENCES

1) Cognitive Solutions : http://www.dataqueuesystems.com/cognitive/
2) Kaggle: https://www.kaggle.com/datasets/venky73/spam-mails-dataset
3) Tutorial: https://www.javatpoint.com/machine-learning
4) IEEE research paper: https://ieeexplore.ieee.org/document/7863267