

School of Computing, Creative Technology and Engineering

|  |  |
| --- | --- |
| Student ID | 77356744 |
| Student Name | Hritick Jha |
| Module Name & CRN | Applied Machine Learning |
| Level | 5 |
| Assessment Name & Part No. | 1 |
| Project Title | Email Spam |
| Data of Submission |  |
| Course | Applied Machine learning |
| Academic Year | 2024 |

**Table of Contents PART 1**

1. Abstract……………………………………………………………………………………….….3

2. Introduction………………………………………………………………………………………4

3. Data overviews………………………………………………………………………………. ….4

3.1 Data setup of emails Spam test file……………………………………………………….….4

3.2 Total count of data in rows and columns of Data sets……………………………………………….………..4

4. Data preparation………………………………………………………………………………….5

4.1 Data cleaning……………………………………………………………………………….…5

4.1.1 Data cleaning of email spam…………………………………………………….….5

4.1.2 Drop last 3 columns……………………………………………………………….…5

4.1.3 Remaining the last columns……………………………………………………….…6

4.1.4 They will be changed target and to will be change text in dataset of email spam …..6

4.1.5 Missing value………………………………………………………………………..7

4.1.6 Check for duplicate values………………………..……………………………….…7

4.1.7 Remove duplicate values…………………………………………………………….7

5. Data Tranformation…………………………………………………………………………...…..7

5.1 Numeric attributes……………………………………………………………………..7

5.2 Categorical attributes…………………………………………………………………..8

5.3 High Correlated attributes……………………………………………………………..8

6. Exploratory Data Analysis………………………………………………………………………..8

6.1 Summary statistics…………………………………………………………………………...8

6.2 pie chart…………………………………………………………………………………...….8

6.3 Data is imbalanced……………………………………………………………………………9

6.3.1 Table of Data is imbalanced……………………………………………………….….9

6.3.2 Add number of words……………………………………………………………...…10

6.3.3 Add number of sentence……………………………………………………..……….11

6.3.4 check the describe the table…………………………………………………………..11

6.3.5 Ham message…………………………………………………………………………11

6.3.6 spam message…………………………………………………………………………11

6.4 Histogram plots………………………………………………………………………..……....11

6.4.1 code of Ham message of histogram……………………………………………..……11

6.4.2 Ham message of histogram………………………………………………………..….11

6.4.3 code of spam message of histogram………………………………………….…….…12

6.4.4 Spam message of histogram……………………………………………………..……12

6.5 Correlation and coefficienta of ham message and spam message…………………………...…13

6.5.1 code of correlation and coefficient of ham message and spam message…………...…13

6.5.2 Figure of correlation and coefficient of ham message and spam message………...….13

7. Data preprocessing……………………………………………………………………………...….13

7.1 Lower cases………………………………………………………………………………….....13

7.2 Tokenization………………………………………………………………………………...….13

7.3 Removing Special characters………………………………………………………….…….…14

7.4 Removing stop words and punctuation………………………………………………...………14

7.5 Stemming……………………………………………………………………………..………..15

8.Conclusion………………………………………………………………………………………….15

9.References…………………………………………………………………………………………..16

Abstract

There is a growing need for more dependable and potent antispam filters due to the increase of unwanted emails, or spam. It's now feasible to effectively recognize and filter spam emails thanks to recent developments in machine learning algorithms. We provide an in-depth examination of many popular machine learning-based email spam filtering methods. An overview of the main concepts, approaches, outcomes, and potential future directions in spam filtering research are given in our review. The application of machine learning techniques in email spam filtering systems by well-known ISPs such as Gmail, Yahoo, and Outlook is studied in the first half of the study backdrop.

The principles of email spam filtering were covered, along with the several methods that researchers are using machine learning approaches to fight spam. We juxtapose the limitations and benefits of current machine learning techniques with the unresolved issues in spam filtering research. We suggested deep learning and deep adversarial learning as viable solutions to the problem of spam emails.

**Introduction**

One of the easiest and least expensive methods of communication available today is email. Still, in recent years, spam emails have increased in frequency. Emails are classified as spam or non-spam using data mining classification techniques. In the future, they plan to add records that have incomplete data since they believe it will improve performance in the case that data collection expands significantly. Finally, they seek to predict survival time by discretizing it using data from specific cancers, such lung cancer, where survivorship is extremely low, and a year's worth of data.   
(May 5, 2011, Aman Kumar Sharma)

Researchers are trying to enhance bagging modeling techniques because the process yields models that are difficult to interpret. This will provide new insight into the issue at hand, aid in the understanding of the automatically generated information by doctors, and possibly even reveal connections and regularities that were previously unknown. In the future, they intend to incorporate records with incomplete data since they think it will enhance performance, particularly considering the anticipated expansion of the data collection. Finally, by using a year's worth of data on malignancies—like lung cancer, which has a very low survivability rate—they hope to estimate the duration of survival. (May 5, 2011, Aman Kumar Sharma)

**Data Overviews:**

Email spam is without a doubt one of the most potent cyberweapons. Not considered spam are newsletters and other high-quality bulk mailings. Unsolicited mass emails with malicious files, images that frequently contain Resource Locators (URLs), or text URLs that lead to malicious or phishing websites are all considered spam. Skillful social engineering tactics deceive email users into opening attachments or clicking on hyperlinks. (Emerging Trends in ICT Security, published in 2014)

**Data setup of email spam test file**

A screenshot of a computer

Description automatically generated

**Total count of data in rows and columns of Data sets**

A white background with black text

Description automatically generated

**Data preparation**

1. **Data Cleaning:** Email data cleansing, also known as email list cleaning, is the process of removing duplicate, invalid, or inactive email addresses from your email list. The goals are to improve email deliverability, decrease the likelihood that your emails will be marked as spam, and raise the caliber of your email list (Jessica Martinez Apr. 11, 2023). There are several approaches for cleaning up email data. They are presented in the following order:
2. Clean your email list Regularly.
3. Identify inactive and unengaged Subscribers.
4. Remove Duplicate and Invalid Email Addresses
5. Verify Email Addresses
6. Manage Bonce Rates
7. Use Double Opt-Ins
8. Comply with Data protection Regulations.

**Data cleaning of email spam**

A screenshot of a computer code

Description automatically generated

**Drop Last 3 Columns**

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**Remaining the Last columns**

A computer screen shot of text

Description automatically generated

**They will be changed target and to will be change text in dataset of email spam.**

A screenshot of a computer

Description automatically generated

A screenshot of a computer screen

Description automatically generated

**Missing Value**

**A screenshot of a computer screen

Description automatically generated**

**Check for duplicate values.**

**A screenshot of a computer code

Description automatically generated**

**Remove Duplicate values.**

**A computer screen shot of a code

Description automatically generated**

1. **Data Transformation:** Data transformation is the process of combining and standardizing raw data to scale to a particular value in a format or structure that facilitates its usage in an algorithm or model. (Mohammed A. M. Ali, 2022) There are many different types of data transformation, including these:
2. **Numeric attributes: -** Numerical qualities are quantifiable values given as real or integer numbers. Jian Pie (2012) states that they could be on the ratio or interval scale. Because interval-scaled attributes are measured on an equal-size unit scale with an order, it is possible to sort them and determine the difference between values. Jian Pian (2021). One can characterize a value as a multiple of another value due to the genuine zero-point of ratio-scaled features (Jian pie, 2021). This is.

A math equations on a white background

Description automatically generated

1. **Categorical attributes: -** Different names or symbols are used to represent the features of categorical data types. The color of the human iris is classified as a categorical data type since it can have hues like gray, green, blue, and black. Since the data values are not directly related to one another, no mathematical operator may be used other than the logical or "is equal" operator. (Data Science, Second Edition, 2019)
2. **High Correlated attributes: -** Qualities can be connected in a variety of ways. Correlated features are those that are related to each other in a way that makes it possible to determine the value of one feature by examining the values of the other features. For example, features can correlate positively or negatively. Composed on April 2, 2023, Christina Ellis

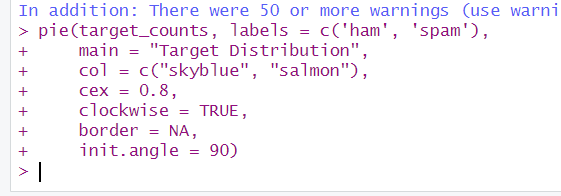
**Exploratory Data Analysis: -** Before moving on to more complex research methods, descriptive statistics are required to comprehend the fundamental properties of a dataset and obtain initial insights. Descriptive statistics provide insight into the distribution of key variables and enable us to identify patterns or anomalies in the data when doing an email spam investigation. An overview of the descriptive statistics that can be calculated for the dataset on email spam is provided below: -

1. **summary Statistics: -** For every feature in the dataset, compute the mean, median, mode, minimum, maximum, and standard deviation. This provides a quick summary of the variability and central trend of the data.

A number grid with numbers

Description automatically generated with medium confidence

1. **pie chart: -** For every feature in the dataset, find the mean, median, mode, minimum, maximum, and standard deviation. This provides a quick summary of the primary trend and variability in the data.



.

A diagram of a pie chart

Description automatically generated

1. **Data is imbalanced: -** In a dataset, an uneven distribution of occurrences across multiple classes is referred to as class imbalance. The distribution of emails that are and are not spam is of interest to us in this case.

**A screenshot of a computer

Description automatically generated**

**Table of Data is imbalanced.**

**A screenshot of a computer

Description automatically generated**

**Add Number of words**

**A screenshot of a computer

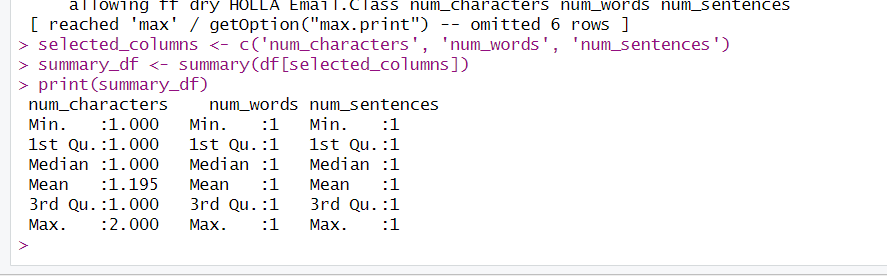
Description automatically generated**

**Add number of Sentences**

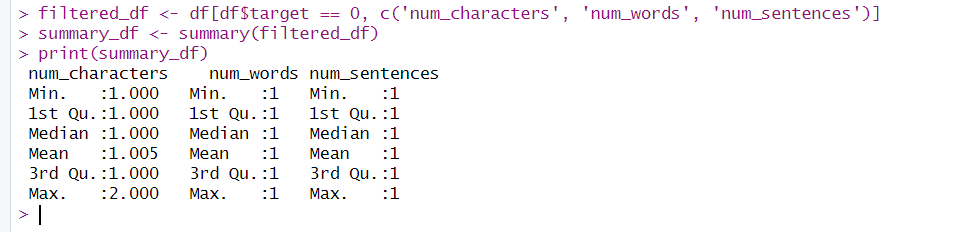
A screenshot of a computer screen

Description automatically generated

**Check the Describe the table.**

****

**Ham messages**

****

**Spam messages.**

A white background with black text

Description automatically generated

A lot of various kinds of spam use numeric characters, which stand for ham transmissions.

1. **Histogram plots:** - When conducting exploratory data analysis (EDA), histogram charts are a useful tool for comprehending the distribution of a continuous variable or feature inside a dataset. Histogram plots can be used to show the distribution of word frequencies or other numerical attributes associated with spam and non-spam emails in the context of classifying emails as spam.

**Code of Ham message of histogram**

**A close-up of a computer screen

Description automatically generated**

**Ham message of histogram**

**A graph with numbers and lines

Description automatically generated**

**Code of spam message of histogram**

A close-up of a computer code

Description automatically generated

**Spam message of histogram**

A graph with numbers and a bar

Description automatically generated

1. **corelation and coefficient of ham message and spam message: -** The correlation coefficient between "ham" and "spam" messages in your dataset can be found using R's 'cor()' function.

**Code of corelation and coefficient of ham message and spam message**

A screenshot of a computer program

Description automatically generated

**Figure of correlation and coefficient of ham message and spam message**

A screenshot of a computer screen

Description automatically generated

**Data preprocessing**

In the process of removing spam. It is essential and important to process the textual material beforehand. Eliminating any information that does not provide the document's class significant context is the main objective of text data preparation. In addition, we enjoy getting rid of irrelevant information. For textual retrieval tasks, removing stop words and stemming to reduce vocabulary are the two most widely used data cleaning methods. Beyond these two techniques, we also removed terms that were two characters or fewer.   
(IJSTE/Volume 1/Issue 11/007: Pre-processing of data detecting spam)

1. **Lower cases: -** A writing or graphic style known as "lowercase" uses entirely lowercase letters without any capitalization. Lowercase letters must be used in text analysis systems that require preprocessing and normalization of data, such as those that categorize spam emails.

A white screen with blue text

Description automatically generated

1. **Tokenization: -** The process of dividing a text document or sentence into smaller units called tokens is called tokenization. In natural language processing, or NLP, it is an essential stage. For many NLP applications, such as text categorization, emotion analysis, machine translation, and named entity recognition, tokenization is an essential preprocessing step.

A white screen with blue text

Description automatically generated

1. **Removing Special characters: -** Natural language processing (NLP) initiatives frequently begin with material that has special characters deleted. Text processing is sped up when special characters are removed. They are made up of punctuation, symbols, and non-alphanumeric characters that are used but don't have any particular meaning in the text.

**A white background with black text

Description automatically generated**

1. **Removing stop words and punctuation: -** Preprocessing in natural language processing (NLP) often involves removing punctuation and stop words. In addition, punctuation is usually eliminated from texts before processing because it is not as important to the meaning.

A white background with text

Description automatically generated

1. **Stemming: -** Stemming is a text normalization technique that breaks down words into their most basic or root form to aid with information retrieval and natural language processing (NLP). The stemming technique, which combines derived or inflected words into a single phrase, can be used to execute text analysis tasks more rapidly and accurately.

A screenshot of a computer

Description automatically generated

**Conclusion**

Ultimately, we have carefully examined the information supplied regarding email spam. In addition, a feature importance analysis was performed to determine the essential elements behind the spam email classification. This data can be utilized to refine the model via iterations and a better comprehension of the basic patterns observed in spam emails. To sum up, the information gained from this study is helpful in developing a practical model for identifying spam emails. Specifically, the random forest model exhibits potential for enhancing email security and protecting consumers from unwanted content in real-world settings. Spam email recognition models were created using a variety of machine learning techniques, including logistic regression decision trees, random forests, and support vector machines.

**References**

Ali, M. A. M., 2022. *DETECT AND ANALYSIS OF SPAM EMAIL BASED ON MACHINE LEARNING ALGORITHMS –COMPARATIVE APPROACH.* [Online]   
Available at: https://www.researchgate.net/publication/369022963\_DETECT\_AND\_ANALYSIS\_OF\_SPAM\_EMAIL\_BASED\_ON\_MACHINE\_LEARNING\_ALGORITHMS\_-COMPARATIVE\_APPROACH

Detection, D. P.-P. i. S., 2015. *Data Pre-Processing in Spam Detection.* [Online]   
Available at: https://www.ijste.org/articles/IJSTEV1I11008.pdf

Diogo A.B. Fernandes, .. P. R. I., 2014. *A Quick Perspective on the Current State in Cybersecurity.* [Online]   
Available at: https://www.sciencedirect.com/topics/computer-science/spam-email#:~:text=Email%20spam%20is%20undoubtedly%20one,servers%20is%20characterized%20as%20spam.

Ellis, C., April 22023. *Correlated features in regression models.* [Online]   
Available at: https://crunchingthedata.com/correlated-features-in-regression-models/

Martinez, J., 2023. *Email Data Cleansing: Best Practices & Choosing The Right Service.* [Online]   
Available at: https://blog.kickbox.com/email-data-cleansing-best-practices-choosing-the-right-service/#:~:text=What%20is%20Email%20Data%20Cleansing,of%20getting%20marked%20as%20spam.

Sharma, A., 2015. *Data Pre-Processing in Spam Detection,* Jaipur Campus: s.n.

Vijay Kotu, B. D., 2019. *Data Science (Second Edition).* [Online]   
Available at: https://www.sciencedirect.com/topics/computer-science/categorical-data#:~:text=Categorical%20data%20types%20are%20attributes,%2C%20blue%2C%20gray%2C%20etc.