Spam or Ham Text Classification using Logistic Regression

[1]G. Siva Nageswara Rao,[2]P. Madhuri,[3]D.Sudheer,[4]D.Meghana

[1]Professor,DepartmentofCSE,KoneruLakshmaiahEducationFoundation,AP,India. E-mail:sivanags@kluniversity.in

[2]Student,DepartmentofCSE,KoneruLakshmaiahEducationFoundation,AP,India.E-mail:madhuripinnela@gmail.com
[3]Student,DepartmentofCSE,KoneruLakshmaiahEducationFoundation,Vaddeswara,Guntur(dt), AP, India. E-mail:sudheerdama111@gmail.com
[4]Student,DepartmentofCSE,KoneruLakshmaiahEducationFoundation,Vaddeswara,Guntur(dt), AP, India. E-mail:meghanadondapati22@gmail.com

Article History: Received: 10 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 20 April 2021

Abstract

Individual and business clients like to utilize wellsprings of correspondence. The use and significance of messages ceaselessly develop regardless of the predominance of elective methods, like electronic messages, versatile applications, and informal organizations. As the volume of business-basic messages keeps on developing, the need to robotize the administration of messages increments for a few reasons, for example, spam email order, phishing email characterization, and multi-envelope classification, among others. Sending a gigantic number of undesirable sends makes security danger clients. These are sent along informing frameworks as substance. Mobile telephones in a real sense are being made un-operational through these sorts of digital assaults. This undertaking means to fabricate an AI model that figures out how to distinguish the assaulting through malware caused inside informing that contains various types of sources that incorporate content, video and sound. The model intends to get familiar with the assaulting caused through malware and afterwards trigger an activity that counters the assault dependent on its sort•

Keywords: spam detection, classification, cyber-attack, malware, logistic regression, messaging.

1 Introduction

AIhas various purposes used in the field of software engineering from settling an organization traffic issue to identifying a malware. Messages are utilized routinely by numerous individuals for correspondence and for mingling. Security breaks that bargains client information permits 'spammers' to parody an erode email address on sending ludicrous (spam) messages. This is likewise misused to acquire unapproved admittance to their gadget by fooling the client into tapping the spam interface inside the spam email, that comprises a phishing assault [1].

.

Spamming remains monetarily reasonable considering the way that promoting experts have no working expenses past the association of their mailing records, workers, foundations, IP reaches, and space names is dense to examine senders liable for their mass mailings. The expenses, for example, lost advantage and shakedown, are borne by people when everything is said in done and by Internet master networks, which have added additional ability to acclimate to the volume. Spamming the themefor beginning in different areas [2]. Most email spam messages are business in nature. On the off chance that business, many are bothering just as unsafe considering the way that they may contain joins that lead to phishing destinations or regions that are encouraging malware - or fuse malware as record connections[3].

Spammers assemble email addresses from visit rooms, locales, customer records, newsgroups, and diseases that harvest customers' area books. These accumulated email addresses are occasionally similarly offered to various spammers .

A segment of such spams in different media include-

- Email spam
- Instant informing spam
- Newsgroup and discussion
- Mobile telephone
- Social organizing spam
- Social spam
- Blog, wiki, and guestbook spam

- Spam focusing on video sharing destinations
- VoIP Spam
- Academic search spam
- Mobile applications spam

"Ham" is an email that isn't Spam. In that capacity, "non-spam", or "incredible mail". It should be seen as a more restricted, snappier identical for "non-spam". Its utilization is particularly typical among threatening to spam

programming fashioners, and not extensively known elsewhere; with everything better to use the articulation "non-spam", in gleam in all.

Spam messages contain: call, presently, guarantee, free, txt, ensured
Ham messages contain: great, later, home, sorry, need, come

Taking apart and basically examining academic exploration work on a wide range of spam email is itself a mammoth assignment and frequently unthinkable in a solitary overview endeavor. Remembering that, this paper fundamentally focusses on the wise and computerized arrangements contrived against vindictive spam messages. Especially on the accompanying:

1) Containing vindictive connections

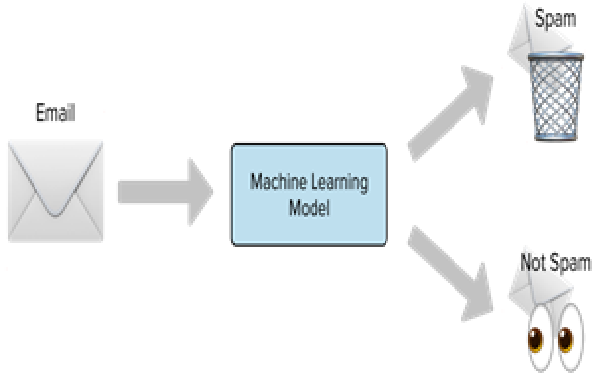
2) Containing malevolent connection

3) Phishing endeavors

4) Phishing and Spoofing efforts

In this paper, we proposed a fundamental AI model utilizing Logistic regression idea that orders text from messages and short messages either as spam or ham and assess the model precision.

Fig 1: Overview of Spam filtering



Related Information
In this specific segment, we will momentarily depict the exploration did around there. Here, we give a concise study of comparative methodologies, covering a wide range of techniques and its issues. Different writing articles pivot on instance models.

S. Nandhini et al., talked about the proficient strategies for utilizing a portion of the well an AI model which can group whether a mail is a spam or ham. UCI Data Set is utilized for the investigation. The exhibition of five significant AI grouping calculations viz. Strategic Regression, Decision Tree, Naive Bayes, KNN and SVM are assessed to prepare and fabricate a successful AI model

for email spam recognition. informational collection [4].

Salwa Adriana Saab, Nicholas Mitri, Mariette Awad& et al..[5] introduced a review of some normal separating calculations that depend on text characterization to decide whether. A correlation of these is made on Spam Base informational collection to decide the best order calculation regarding execution, figuring time, and accuracy/review speeds.

W.A. Awad , S.M. ELseuofi& et al… [6] looked into the absolute most well known AI techniques (Bayesian arrangement, k-NN, ANNs, SVMs, Artificial insusceptible framework and Rough sets) and their materialness to spam order issues. The calculation details are introduced and a correlation of their outcomes on the SpamAssassin spam corpus is introduced.

Amandeep Singh Rajput, Vijay Athavale, Sumit Mittal& et al… [7] proposed a local area arranged approach by using bundle figuring with the help of equivalent machines for speedy separation of SPAM and HAM. A gathering approach can construct the figuring power various folds with existing hardware and resources thusly by accelerating taking care of without achieving any extra cost. In this examination, we simply use header-based isolating methodology, thus by keeping the assurance of the customer faultless. The standard test set for HAM and SPAM from Spam Assassin is used. Two kinds of equivalent conditions are used in this assessment. First is where distinctive Anti Spam procedures are used in the equivalent environment against the test corpora and false certain and sham negative exactness recorded. The second equivalent environment is where standard test corpora are separated into parts and dealt with into equivalent machine environment with single adversary of spam method used at all machines and the proficient is recorded against autonomous machine being used.

Vangelis Metsis, Ion Androutsopoulos, Georgios Paliouras& et al… [8] discussed five remarkable versions of Credulous Bayes, and take a gander at them on six new, non-encoded datasets, that contain ham messages of explicit Enron customers and new spam messages. The new datasets, which we make straightforwardly open, are more commonsense than past commensurate benchmarks, since they keep up the transient solicitation of the messages in the two characterizations, and they mimic the changing degree of customers get as time goes on. We get a preliminary system that duplicates the continuous planning of altered spam channels, and we plot roc twists that license us to take a gander at the changed versions of NB over the entire compromise between clear positives and authentic negatives.
Aditya Gupta, Khatri Mrunal Mohan, SushilaShidnal& et al [9] clarified the assessment of the show of Naïve Bayesian AI count concerning antispam isolating is done here. The growing volume of unconstrained mass email (spam) has created a prerequisite for strong foe of spam channels.

PROPOSED Methodology

A. Technology Requirements

The machine learning athenaeumworn in the project are mentioned below-

1. NLTK Library

Common Language Processing is control or getting text or discourse by any product or machine. A similarity is that people connect, see each other perspectives, and react with the proper answer. In NLP, this association, understanding, the reaction is made by a PC rather than a human.

NLTK represents Natural Language Toolkit. This instrument stash is maybe the most noteworthy NLP libraries which contains packs to make machines understand human language and answer to it with a fitting response. Tokenization, Stemming, Lemmatization, Punctuation, Character tally, vocabletally are a portion of these bundles which will be examined in this enlightening exercise.

NLTK calculations, for example, tokenizing, grammatical form labeling, stemming, conclusion investigation, point division, and named substance acknowledgment. NLTK encourages the PC to examination, pre-measure, and comprehend the composed content.

NLTK is to work with human language information. Serene by the authors of NLTK, it directs the peruseracross the essentials of composing Python programs, working with corpora, sorting text, breaking down etymological construction, and that's just the beginning [10].

Language Datasets in Python 3, we use Natural Language Toolkit (NLTK). The procedure involves:
Step 1 — Importing NLTK.
Step 2 — Downloading NLTK's Data and Tagger.
Step 3 — Tokenizing Sentences.
Step 4 — Tagging Sentences.
Step 5 — Counting POS Tags.
Step 6 — Running the NLP Script

2. Scikit-learn library

Scikit-learn (once in the past scikits.learn and furthermore known as sklearn) is a free programming AI library for the

Python programming language. calculations including support vector machines, irregular backwoods, inclination boosting, and is planned to collaborate with the Python mathematical and logical athenaeum NumPy and SciPy.

Scikit-learn is generally written in Python, and utilizations numpy widely for superior direct variable based math and cluster tasks. Moreover, some center calculations are written in Cython to improve execution. Backing vector machines are executed by a Cython covering around LIBSVM; strategic relapse and straight help vector machines by a comparable covering around LIBLINEAR.

Scikit-learn is a library in Python that gives numerous unaided and administered learning calculations. It's endless supply of the innovation you may as of now be acquainted with, as NumPy, pandas, and Matplotlib.

For pip establishment, run the accompanying order in the terminal:

- pip introduce scikit-learn.
- conda introduce scikit-learn.
- import sklearn.

# Import scikit gain from sklearn import datasets # Load information iris= datasets.load_iris() # Print state of information to affirm information is stacked print(iris.data.shape).

The scikit-learn project started as scikits.learn, a Google Summer of Code project by David Cournapeau. Its name starts from the possibility that it is a "SciKit" (SciPy Toolkit), a freely made and passed on untouchable extension to SciPy. The first codebase was later altered by various specialists.

Scikit-learn is presumably the most helpful library for AI in Python. The sklearn library contains a great deal of effective apparatuses for AI and factual displaying including characterization, relapse, grouping and dimensionality decrease.

B. Software Requirements:

- The Operating System used: Linux 64-bit (Ubuntu)
- Python: Jupyter Notebook
- //Add other software requirements

C. Dataset Overview

This dataset incorporates the content of 5572 SMS messages alongside a name showing whether the message is undesirable. Garbage messages are marked spam, while authentic messages are named ham..

D. Implementation/Flow of the project

Step 1. Data pre-processing

It is advance in Machine Learning as the essence of information can be gotten from it straightforwardly influences the model; accordingly, it is critical that we pre-measure our information prior to taking care of it into our model.

Step 2. Wiping out Stop words

In trademark language dealing with, silly words (data), are implied as stop words. ... Stop Words: A stop word is a vocable, (for instance, "the", "a", "an", "in") that a web crawler has been modified to disregard, both when requesting segments for looking and keeping that delayed consequence of a question. NLTK upholds stop word evacuation, and you rundown of stop words in the corpus module. To eliminate prevent words from a sentence, you can separate your content into words and afterward eliminate in the rundown of stop words given by NLTK.[11].

For assignments like content order, where the content is to be characterized into variousstop vocable are taken out or barred from the given content can be given to those words which characterize the significance of the content.

Step 3. Tokenizing

It is the process of breaking down the text corpus into individual elements. These individual elements act as an input to machine learning algorithms.

For Example:

Every province has its own uniqueness

Step 4. Count Vectorization
Vectorization is a method by which you can cause your code to execute quick. It is an extremely fascinating and significant approach to upgrade calculations when you are executing it without any preparation. Presently, with the assistance of profoundly improved mathematical direct polynomial math libraries in C/C++, Octave/Matlab, Python etc.

The Count Vectorizer gives a basic method to both tokenize an assortment of text archives and assemble a jargon of known words, yet additionally to encode new reports utilizing that jargon. We can utilize CountVectorizer of the scikit-learn library. It of course eliminates accentuation and lower the records. It transforms every vector into the meagre lattice. It will ensure the word present in the jargon and if present it prints the quantity of events of the word in the jargon.

Step 5: Stemming
Stemming is significant in characteristic language understanding (NLU) and normal language handling (NLP). Stemming is likewise a piece of inquiries and Internet web crawlers.

Step 6: Separating training and test set
It is an example of information relevant to the representation. The real dataset that we use to prepare the representation (loads and inclinations on account of a Neural Network). The model sees and gains from this information.

6.1 Validation Dataset
It is used to give a fair assessment of a model fit on the assemble dataset while tuning model hyper boundaries. The estimation turns out to be extra one-sided as ability on the approval dataset is joined into the model design. The consent set is utilized to assess a given model, however this is for regular assessment. We, as AI to tweak the representation hyper boundaries. Consequently the model sometimes sees this information yet never does it "Learn" from this. We utilize the approval set outcomes, and update more significant level hyper boundaries. So the approval set influences a model, yet just by implication. The approval set is otherwise called the Dev set or the Development set. This bodes well since this dataset helps during the "improvement" phase of the model [12].

6.2 Test Dataset
i) The example of information used to give an impartial assessment of a last model fit on the preparation dataset. The Test dataset gives the best quality level used to assess the model. It is just utilized once a model is totally trained (using the train and approval sets (For instance on numerous Kaggle rivalries, the approval set is delivered at first alongside the preparation set and the genuine test set is possibly delivered when the resistence is going to close, and it is the consequence of the model on the Test set that chooses the champ). Commonly the approval set is utilized as the test set, yet it isn't acceptable practice. It contains painstakingly inspected information that traverses the different classes that the model would confront when utilized in reality.

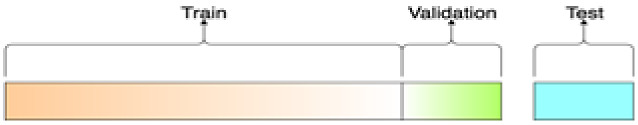
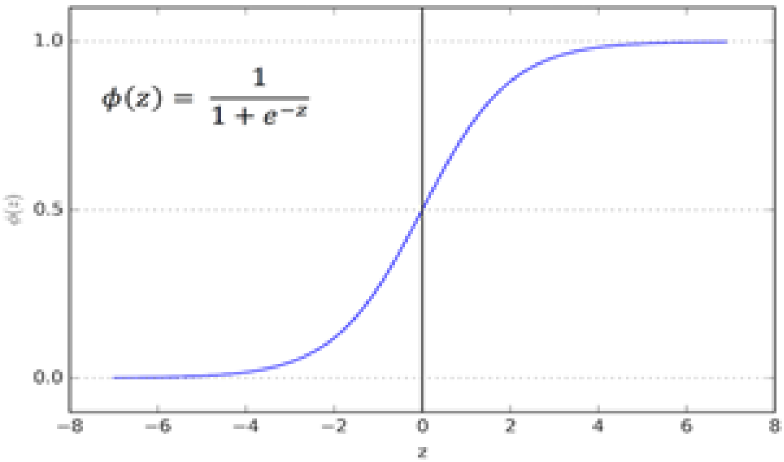


Fig 2: A visualization of the splits

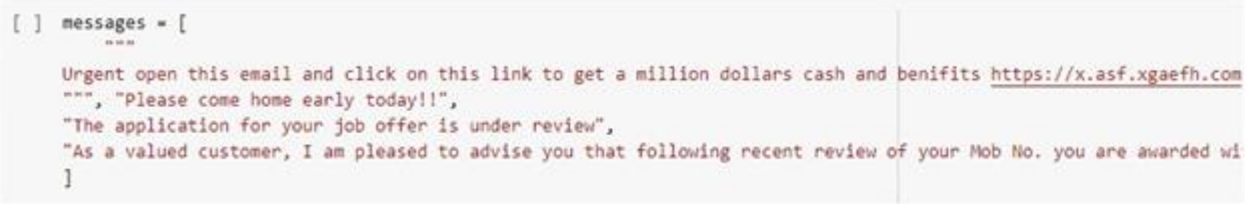
Step 7. Apply Logistic regression Algorithm
Logistic Regression measures the relationship joining categorical dependent variable and one or more independent variables by estimating probabilities using a logistic function**.**



RESULTS

This particular section provides experimental outcomes of the prefered method. Firstly, the dataset description is provided. Then, the presentation of logistic regression model is to detect if a text is spam or ham. The model is later evaluated with the help of confusion matrix, ROC AUC and Gini coefficient.

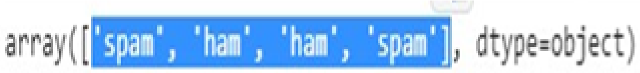
Fig 3: Input messages



```
[ ]  messages = [
         """
     Urgent open this email and click on this link to get a million dollars cash and benifits https://x.asf.xgaefh.com
     """, "Please come home early today!!",
     "The application for your job offer is under review",
     "As a valued customer, I am pleased to advise you that following recent review of your Mob No. you are awarded wi
     ]
```

The Fig.3 shows the list of input given (98%)

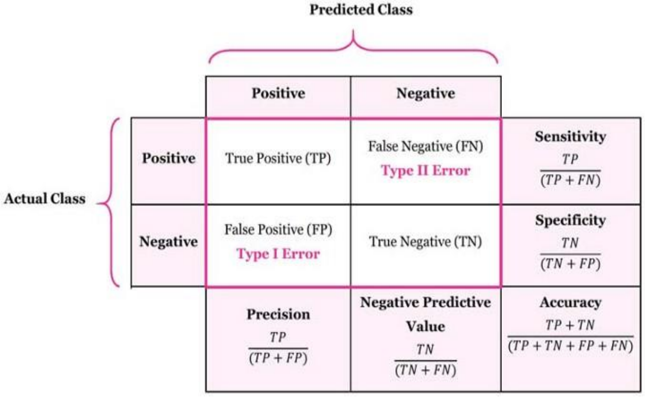
Output of the representation that detects text to be spam or ham looks like:



```
array(['spam', 'ham', 'ham', 'spam'], dtype=object)
```

1. Confusion matrix:

A confusion matrix is a table that helps envisions the presentation of a grouping model. It very well may be utilized to estimate Precision, Sensitivity (aka recall), Specificity and accuracy.



		Predicted Class		
		Positive	Negative	
Actual Class	Positive	True Positive (TP)	False Negative (FN) Type II Error	Sensitivity $\frac{TP}{(TP+FN)}$
	Negative	False Positive (FP) Type I Error	True Negative (TN)	Specificity $\frac{TN}{(TN+FP)}$
		Precision $\frac{TP}{(TP+FP)}$	Negative Predictive Value $\frac{TN}{(TN+FN)}$	Accuracy $\frac{TP+TN}{(TP+TN+FP+FN)}$

Fig 3: Confusion Matrix

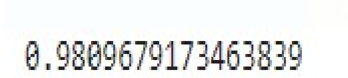
Definition of the Terms:

- True Positive (TP) : Observation is positive, and is predicted to be positive.
- False Negative (FN) : Observation is positive, but is predicted negative.
- True Negative (TN) : Observation is negative, and is predicted to be negative.
- False Positive (FP) : Observation is negative, but is predicted positive.

Proceeding with the vision of evaluating the proposed model, we calculate the performance using the given

formulae-
Accuracy is the capacity to decide the rightness or closeness of character arrangement. The equation for accuracy can be given by
Accuracy = (TP+TN)/(TP+TN+FP+FN)
The formula for precision can be given by
Precision = TP/(TP+FP)
The equation for recall can be given by
Sensitivity (recall)=TP/(TP+FN)

Specificity is defined as the proportion of negatives in a binary classification test which are correctly identified and can be given by
Specificity=TN/(TN+FP)



```
0.9809679173463839
```

Fig 4: Accuracy of the developed model

2. ROC AUC
AUC means Area Under Curve, which is calculated for the ROC curve.

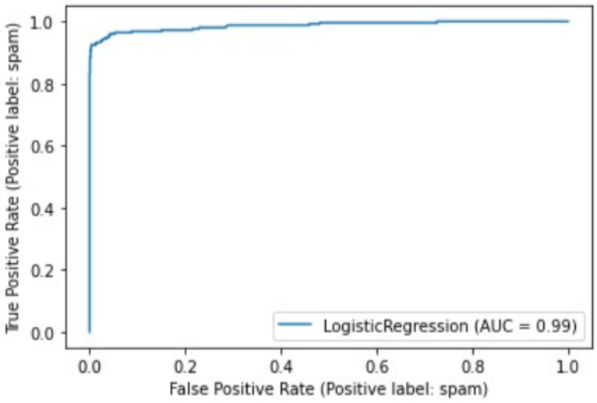


Fig 4: ROC AUC Curve
An ROC curve is a graph plotted between TPR and FPR. The nearest the value of AUC is to 1, the more the model is developed. It can be calculated using functions in both R and Python.

It is evident from the ROC-AUC curve that the AUC is 0.99, which is closer to 1. And hence, we can conclude that the model is fully developed to classify whether the text in any short message is ham or spam.

3. Gini Coefficient
Gini is most usually utilized for imbalanced datasets where the likelihood alone makes it hard to anticipate a result. Gini is estimated in qualities somewhere 0 and 1, where a score of 1 implies that the model is 100% exact in anticipating the result. A score of 1 just exists in principle. Practically speaking, the nearer the Gini is to 1, the better. Though, a Gini score equivalent to 0 methods the model is altogether wrong. To accomplish a score of 0, the model would need to credit irregular qualities to each forecast..

GINI = AUC*2–1
GINI (MNB)= 0.99*2-1= 0.98

0.98 is way closer to 1 it is developed in order to classify the spam or ham text messages with maximum possible efficiency [13].

CONCLUSION
In this work, we presented a basic machine learning model on the Spam Base dataset and developed the representation performance evaluation. Results are reliable with the hypothetical strength also, constraints of each approach. An intriguing expansion to this work is check the adequacy of individual classifiers in accurately collecting the text messages as spam or ham. Our model has generated an accuracy of 98%, 0.99 ROC AUC curve and with a 0.96 GINI coefficient value when developed using Logistic Regression.

References

1.W. Feng, J. Sun, L. Zhang, C. Cao, and Q. Yang, ''A support vector machine based Naive Bayes algorithm for spam filtering,'' in Proc. IEEE 35th Int. Perform. Comput. Commun. Conf. (IPCCC), Dec. 2016

2."Spam". Merriam-Webster Dictionary (definition & more). 2012-08-31. Retrieved 2013-07-05.

3."The Definition of Spam". The Spamhaus Project. Retrieved 2013-09-03.

4.S. Nandhini. Performance Evaluation of Machine Learning Algorithms for Email Spam Detection. International Conference on Emerging Trends in Information Technology and Engineering, 2020

5. Salwa Adriana Saab, Nicholas Mitri, Mariette Awad, "Ham or Spam? A comparative study for some Content-based Classification Algorithms for Email Filtering". 17th IEEE Mediterranean Electrotechnical Conference, Beirut, Lebanon, 13-16 April 2014.

6.Amandeep Singh Rajput, Vijay Athavale, Sumit Mittal, "Intelligent Model for Classification of SPAM and HAM". International Journal of Innovative Technology and Exploring Engineering (IJITEE), Volume-8 Issue-6S, April 2019.

7.Vangelis Metsis, Ion Androutsopoulos, Georgios Paliouras, "Spam Filtering with Naive Bayes – Which Naive Bayes?" . CEAS 2006 Third Conference on Email and AntiSpam, July 2006.

8.Aditya Gupta, Khatri Mrunal Mohan, Sushila Shidnal, "Spam Filter using Naïve Bayesian Technique". International Journal of Computational Engineering Research (IJCER), June 2018.

9..Michelle Morales(January 3, 2017) How To Work with Language Data in Python 3 using the Natural Language Toolkit (NLTK). Retrieved from https://www.digitalocean.com/community/tutorials/how-to-work-with-language-data-in-python-3-using-the-natural-language-toolkit-nltk

10. Tarang Shah(Dec 6, 2017) About Train, Validation and Test Sets in Machine Learning. Retrieved from https://towardsdatascience.com/train-validation-and-test-sets-72cb40cba9e7

11. Spam Detection with Logistic Regression. Natasha Sharma(May 2018). Retrieved from https://towardsdatascience.com/spam-detection-with-logistic-regression-23e3709e522

12. Tarang Shah(Dec 6, 2017) About Train, Validation and Test Sets in Machine Learning. Retrieved from https://towardsdatascience.com/train-validation-and-test-sets-72cb40cba9e7

13. Abhigyan(200, Mar 18) Calculating Accuracy of an ML Model. Retrieved from https://medium.com/analytics-vidhya/calculating-accuracy-of-an-ml-model-8ae7894802e