Final Project
Spam Email Detection

1st YUJIE DONG
Ontario Tech University
Faculty of Information Technology and business
Oshawa, Canada
100741840
yujie.dong@ontariotechu.net

2th Niranjankumar Gangaraju
Ontario Tech University
Faculty of Information Technology and business
Oshawa, Canada
100767124
niranjiankumar.gangaraju@ontariotechu.net

***Abstract*—Spam emails are also including useless emails and spear phishing, which needs email filters to process the huge amount of emails[4]. Artificial intelligence(AI) can be utilized as the proper measures to filter emails by checking the routing information, originating source and destination domains, subject, text content and attachment[4].In the paper, support vector machine(SVM) is significant in filtering spam emails. The misclassification of emails reduce the weight value. We can judge the spam emails by using WSVM(weighted support vector machine), SVM with KPCM(Kernel based probabilistic c-means) and KFCM with WSVM[7]. The cause of the email spam is that this kind of email is easy and cheap. What is nominal and numerical attributes and what is the efficiency of computing can be handled by decision tree(DT). Hydrid combination Logistic Regression(LR) is utilized for filtering noisy data or instance before data will be processed by decision tree. Logistic Regression can generate correct prediction with the certain threshold by reducing noisy data.[10].**

I. Introduction

E-mail becomes a popular way to communicate over Internet, while the spam email has become a serious trouble to affect the efficiency of email.[8]. There are many different spams through email and phone or SMS[3]. The vicious spammers prefer sending more spam emails to propagate what they want to advocate[4]. It is profitable to spam some financial institution, for which the profit is 3.5 million USD[4]. From 2016 to 2018 in Australia, the losses of investment scams are over 20 million AUD, over 30 million AUD and almost 40 million AUD, which are much more than dating romance, false billing, advanced pay fee fraud. Spams have made the data allocation become difficult, which is a high financial loss for the whole company. Machine learning and non-machine learning can be two types of filtering techniques.[14].There are also other algorithms that have been proposed, including Jaro, Bi-gram, Longest Common Subsequence(LCS), Levenshtein Distance(LD) and TFIDF.There are also white list/black list. White list is to block emails by judging its email addresses and white list is the list for which senders to allow.[7]. Meanwhile, SVM is emphasized in some papers. SVM generally can give a binary decision for the final classification.[8]. Decision tree(DT) is another way to overcome the deficiency with missing attribute values.[10]. However, decision tree is sensitive to the training set, other unnecessary attribute, which can reduce the excellence of the DT. Therefore one of the solutions which is called hybrid combination Logistic Regression(LR) and DT. This can be effectively used for detecting email spam.[10]. Artificial neural network(ANN) are organized by the weighted units[4]. The combined particle swarm optimization, support vector machine and artificial neural network are combined to classify and separate the spam emails. [11]. Malicious spam detection cannot only depend on the rule-based filtering. SVM always tries to discover hyperplane, in which data can divide maximum margin for the hyperplane.[11]. The spam emails also include hacking attacks, malicious attack and phishing attacks by phishing URLs, which are targeting useful data. Bayesian classifier is utilized for evaluating the performance of Bayesian classifier by its accuracy, time, precision and error.[13]. The extreme learning machine(ELM) method is a single layer feed-forward neural network, which can randomly decide weights from hidden layers[13]. ELM can also be utilized as spam classification algorithm.[13]. Some paper also presents a new way to filter spam detection. This filter are combined with several attributes, which are also used previously and based on the usage of the words, and also uses the KNN algorithms to execute the classification of the emails.[15]. This method was used ten years ago. Meanwhile, there were other methods used ten years ago, which chose negative selection when detecting spam emails.[16]. Except these algorithms, ensemble classifiers with a voting mechanism, bagging and boosting(Adaboost) can also be chosen as the algorithms[2]. As for the data set, there are two different data sets chosen. The name of them are CSDMC2010 spam dataset and Enron corpus.[6]. The data sets in UCI database which is standard data sets have been utilized to evaluate how the proposed method is effective.[11]. The data set such as Kaggle website is chosen. Accuracy score, Recall, Percision and ROC are utilized to describe the final result[2]. Sometimes a model that utilized a new data set for selecting features improves classification. Training time can be improved by selecting feature and improving the accuracy of malicious spam detection.[11].

II. Detailed Description of the topic

A large number of unwanted email will consume the storage and cause security threat to users. A system knowing how to categorize the spam email is highly needed. Some malicious spam detection relies on filtering the emails by rules and the methods have to be updated from time to time because the spammers also updated their methods.[12]. Malicious spam emails targets some sensitive information while the emails will also consume the scarce resources such as network bandwidth and storage space.[12].There are also some physical approaches preventing users from being targeted by spam email, such as refusing replying to spam emails, clicking the URLs in Emails and posting your email ids on the unfamiliar web.[13].The first one is feature selection. After the subset generation, one method is filter-based and the other is based on wrapper. Meanwhile, the methods based on filtering assesses the features whether they are independent or not while the first method assesses the feature subsets based on the chosen classifier.[12]. Logistic Regression, KNN and SVM, Decision Tree, Naive Bayes are usually chosen as the ML model to classify spam email[1]. When we anatomize the spam email, the spam attack comprises email phishing, spear phishing and whaling. The Non-AI based anti-spam systems consists of domainkeys identified mail(DKIM) and sender policy framework(SPF).Non-AI based plans also include propositions that is based on architectural modification. Collaborative models include fuzzy hashing, cryptographic hashing, distributed checksumhouse, greylisting, DNS blacklisting and whitelisting and solutions which are social trust based [4]. Heuristic filtering models which are another filter systems utilizing regular expression(REGEX)[4]. Content based approaches checking the content of the email also include regular content filtering systems, context sensitive proposals, DNS lookup systems and fuzzy logic based systems. Except the content based filtering frameworks, these frameworks include filtering emails by IP and DNS. There are also other frameworks, filtering email based on countries or peer to perr. The combination of existing techniques includes spamassasin, zerospam. AI based detection and classification techniques include systems based on bio-inspired intelligence. Bio-technology often contains genetic algorithm, NSA and PSO or other related systems, such as biological immune system(BIS). As for the machine learning based systems, like other systems, it often includes supervised machine learning algorithms, semi-supervised machine learning algorithm, unsupervised machine learning algorithms and reinforcement learning. After processing by these algorithms, we can measure the performance. Other measures include feature selection and engineering, highlighting some key supervised and unsupervised algorithms. As for the supervised learning, it is always supervised by nature. If the algorithms are summarized, supervised algorithms comprise artificial neural networks, naive bayes, decison trees, logistic regression and support vector machine. Unsupervised algorithms contain K-Nearest Neighbour, K-Means clustering and expection.[4]. There is also an algorithm called particle Swarm Optimization(PSO), which has the stochastic distribution and can optimize globally the parameters of NB approach[5]. There are several aspects including training models, data, and features and classification algorithms.[6]. Detection using weighted support vector machine with KPCM or using improvised weighted SVM with KFCM are effective.[7]. The textual content is still a main way to filter the emails by their textual content.[11]. Bayesian method is one of the methods classifying texts and e-mail. [11]. There are also other approaches related such as analyzing traffic, method for intuition, hashing the content, signatures, black lists.[11]. Kennedy and Eberhart in 1995 introduced the idea Particle Swarm Optimization(PSO).[11]. PSO can be a global solution to answer one or more points in non-single dimensional space.[11]. Artificial neural network(ANN) can be applied to solve the complicated systems with higher projected output.[11]. Support vector machine(SVM) can make a super plane when making decision. Therefore, positive and negative samples can be the binary separation.[11]. It is the common knowledge that spam emails always have some connection to the recipient. Random Forest(RT) is also used for machine learning algorithm, optimizing parameters and selecting features.[12]. There are also many features that can be used such as payload(body) features, header features, subject features and attachment features. Detecting the malicious spam will use SVM, ANN, RF, Naive Bayes and Adaboost.[12]. Bayesian classifier use its classifier to calculating probability of being a spam email.[13]. Context- based filter also takes header of the mail like the suject about classifying text, HTML tage, tokenizing, word frequency calculation and stop-word removal into account.[13]. The neurons that are showed in the extreme learning machine(ELM) can be individually grown outside of the application.[14]. Ten years ago, detecting and filtering can be most feasible in filtering spam emails.[16]. Artificial immune systems is similar to a human body, mimicking the machinism of it. Negative selection is one branch of it.[16]. Finally, The evaluation criteria includes accuracy(TP), though error rate(classified instances wrongly) and error(TN,incorrectly classified instance).[13].

III. Related work in this area

The comparison among different algorithms shows that Decision Tree outperforms Logistic RegressionKNN and SVM and Naive Bayes[1]. Many researchers previously have done a lot of comparisons among different algorithms. Jerzy Stefanowaski and Joao Gama have studied the Ensemble learning for data stream analysis[2]. They utilize the concept drift to evolve patterns of a data stream, whose result can even perform better to distinguish the spam email data set in small data sets. Horvitz, Heckerman, Sahami has utilized the Bayesian approach to filtering unwanted email. In their research, both numeric and non-numeric features are dealt by Naive Bayes[2]. Dey S and Trivedi utilized the boosting algorithm and Naive Bayes while the selection of feature is done in Naive Bayes. A threshold is set and the distance will be checked whether it is crossed the threshold or not. If it did not happen, new records will be set. It is referred by Elsevier B.V. about the simplicity and effectiveness of Naive Bayes [2]. Drucker and Vapnik also verified that Boosted Decision tree and Support vector machine outperform than other algorithms but for SVM, it is time-consuming to test and not insured for small dataset[2]. SVM carries out only small changes in data points when generalizing. It will not damage the model a lot. It also resists overfitting. Meanwhile, its weakness is that tunning the hyper-parameters by SVM is complex. The training time of SVM can be greatly consuming on large data sets. Meanwhile, SVM is not optimal for multi-class classification.[4]. Support vector machine can process the imbalanced data set effectively.Cheng-Chi Lee and others also have adopted Decision Tree, in which only categorical attributes are supported and various accuracy matrix can filter the important attribute[2]. The advantages of Decision trees virtually do not have no hyperparameters to be tuned for optimization. Explaining and visualizing the decision trees are significantly easier while decision trees can also handle both categorical and numerical data. If it is performed in the large data sets, it can also perform appreciably. The weaknesses of decision trees can become inconsistent with minor variance in data. For decision trees, significant overfitting will be caused by noisy training data sets. The final results will be hard to interpreted in various forms or shapes, making it difficult to choose the best one[4]. There are also some related researches executed related with these algorithms. Artificial neural networks excel in performing with large data sets but it is expensive and sometimes it can only solve the scratch for problems.[4]. Naive bayes process well if the data is missing or noisy. Naive bayes can also operate well with any kind of data, whatever is continuous or discrete[4]. Meanwhile, it is fast, highly scalable and easy to implement[4]. However, it must be assured that all features are independent of each other and handling imbalanced data sets will be inefficient for Naive bayes[4]. Adaboost excels in classifying the classes into multi-class and can overcome the deficiency of overfitting. However, computation by adaboost is really expensive and sometimes it also makes the classification more complex rather substantially. In logistic regression, the interpretation can be convenient and intuitive. If the input function are not scaled, they can still are able to function well even. However, it does not suit non-linear problems. If the outcome is not continuous, it cannot be predicted. The data points utilized by logistic regression are also demanded to be independent of each other, which can not be realistic sometimes for every problem faced. Unsupervised algorithms comprise K-Nearest Neighbour, Expectation Maximization and K-Means Clustering. K-Nearest Neighbour can function well against noisy data set and be utilized for both classification and regression problems.It can also process multi-class problems effectively. However, if we determine the optimum value of K, it often becomes cumbersome. The cost for computing by K-Nearest Neighbour is higher than other algorithms. It is also extremely sensitive to outliers in the data points.[4]. Implementing K-Means clustering can be straightforward. The implementation of K-Means Clustering becomes easy and low cost in computation. However, it is highly sensitive to scaling though normalization or standardization has been executed. If we want to get a relatively optimal value for K, it requires multiple runs.[4]. Expectation maximization can process missing values efficiently and it overcomes clusters to be any specific geometry. The convergence time can be longer. Meanwhile, the computational cost is high enough to become a barrier to advocate this method widely.[4]. So all the algorithms have been listed above. There are also other related work done in paper.[5]. PSO is also utilized for detecting the spam email. This paper also illustrates the Table I, which includes many authors' results. Renuka and Visalakshi proposed that SVM-LSI performs well comparing with other researches. Harisingh aney et al reports huge time has been consumed by authors' data. Idris et al achieved better accuracy by NSA-PSO. Mohamad and Selamat focus on feature extraction. Tuteja and Bogiri utilizing K-means clustering and artificial neural network can get better results if it is with preprocessing steps when compared with the results without pre-processing. Kaur and Sharma has combined the concepts of PSO and decision tree algorithm, in which there is no information related with the usage of feature extraction approach. Feng et al mixed Support Vector Machine with Naive Bayes, which improves the results comparing with individual Support Vector Machine and Naive Bayes approaches. Kumaresan and Palanisamy stepsized support vector machine with cuckoo search resulting in superior algorithm and speeding up the classification. A paper also mentions a flowchart of spam email detection. The first step is to consider email in raw format. The second step is to pre-process the email. The third step is to apply the steps of tokenization, stemming and the removal of stop words. The fourth step is to select the feature using CFS. In this paper, the fifth step as for the probability distribution, we apply NB. The sixth step is for optimization, for which the author applies PSO. Then it is the step to judge whether there is the presence of spam keywords or not. The answer no will result in declaring the email as non-spam and the answer yes will result in classifying the email as spam.[5]. Renuka et al compared multilayer perceptron(MPL),Naive Bayes and J48, among which multiplayer performs. The filtered bayesian learning(FBL) can improve the Naive Bayes accuracy. Harisinghaney et al and Mohamad Selamat have judged whether it is the email spam detection by the image and textual data set.[5]. Most of the researchers focused on the text based filtering methods, which requires pre-processing the data initially[5]. Renuka et al used three machine learning algorithms, including Naive Bayes, J48, and Multiplayer Perceptron(MPL).[5]. The authors also used algorithm LCS algorithms in biological files, which found the accurate DNA sequence resemlance by checking the randomly generated sequence of DNA. Cheng et al. also published a method that could detect DNS packets in a real-time tunnel through Bi-gram algorithm.[6]. The used Levenshtein distance algorithm is applied for calculating phonetic distances, which can also be utilized to tell Scandianvian language from standard Danish. The 3D segmentation algorithms can be evaluated and compared by Jaro distance. The 3D segmentation comprise automatic segmentation and multiple ground truth segmentation, which can give a score.[6]. Pre-processing spam data set is the first step and the second step is to compare spam emails with ham emails. Similarity scores is set to check whether it cross the threshold or not and give a final result.[6]. The improved WSVM algorithm for preprocessing data have 7 steps. The first step is text-containing data set , for which the output can be a document term matrix. The second step is to select the highly repeated words from the frequently repeated words from each text-containing data. The third step is to make a vector space model from the text-repeated data. The fourth step is to extract these basic statistics about ham or spam message and then assign numeric value to spam and ham data column. Some basic statistics like mean and variance of spam and ham message are found. The sixth step is to extract text feature with several measures. The measures comprise deleting the white spaces, changing from lower case to upper case or vice versa and stripping stop words.[7]. After the text feature extraction, a document matrix has been formed.[7]. These steps are for preprocessing. The algorithm is for applying improved WSVM. The first step is to get the train document term matrix. The output from the input is to use many standards to measure the date, such as misclassification rate, precision and accuracy. In the second step, the kernel function played as radial is set and the DTM for performing SVM is trained. In the third step, the prediction rate of SVM algorithm is calculated. In the fourth step, the function is set while the DTM for processing WSVM is trained. In the fifth step, the prediction rate of WSVM algortihm is calculated. In the sixth step, KFCM algorithmic function is applied to obtain weights and the weights are assigned to a variable S. In the seventh step, the weighted variable S is applied in the WSVM instead of the other variables. In the eighth step, the result is generated and the rate for predicting is calculated.[7]. OCR is called optical character recognition, which is using technology and the classification for images, which are several studies and development of the approaches[8]. Here is to use the images to distinguish whether it is the spam or ham image. Uemura et al and Krasser et al use few image metadata for detecting image spam fast. SVM-based detected their own merit and demerit.[8]. SVM includes Gussian Kernel utilized for find non-linear decision boundaries between two variables. Step 1 is to processing, including lower casing, normalizing URLs, email URLs, number, dollar, stripping HTML, word stemming, removing of non-word, white space. The step2 is to list vocabulary. Step 3 is to extract feature. Step4 is training and analysis.[8]. However, in one paper, more features have been utilised resulting in more significant results.[9]. Tafazzoli et al raised a new way intending for the algorithms used for the weight selection. G.Poonkuzhali et al declared the tool TANAGRA utilized to observe the dataset along with various classifiers.[9]. R. Sanjiban sekhar et al discussed various methods for detecting spam detection. Alsmadi et al collects the assembled large individual data sets related with e-mails. Classifying the emails are based on their content. Perez-Diaz et al executes different schemes to apply rough set.[9]. Salehi et al focused on the shortcomings in simple Artificial Immune system(SAIS) and proposed an sample that SAIS and Particle Swarm Optimization(PSO) are involved. PSO mixed with mutation operator can select the feature which is wrapper-based. It can also be used to refine and get the features when dealing with the basic decision tree classifier and parameters.[10]. A novel model which can greatly raise the random generation of the dector when stochastic distribution modelling the data point was applied. The data point is using particle swarm optimization(PSO). The other PSO is combined with NSA-PSO, which is utilizing a local outlier factor(LOF).[10].Patil et al uses Naive Bayes and j48 classification to classify the correct and incorrect classification. Idris et al utilized negative selection algorithm instead of using detector.[9]. There are several dataset attributes used here. The first attribute is A1 to A48, which used for find ratio of words that are in the e-mails and judge whether they can match the word or not. A49 to A54 is also used for judging ratio of characters in the email which can march the characters.[9]. A55 understands the average length of uninterrupted sequence of upper cases.[9]. A57 is used to get the sum of length of uninterrupted sequence of upper cases. A58 is utilized to denote whether the email was considered spam or not.[9]. In the study, they found that Bayesian, thresholds and probability were sensitively cost to filter spam emails. Meanwhile, it can reduce the mistakes for classifying spam emails and performing better because of its cost-sensitive characteristics.[10]. Meanwhile, the integration of Negative Selection Algorithm(NSA) with the differential evolution(DE) has been put forward.[10]. Here, some authors also proposed a new algorithm, which is called LRFNT+DT and with FN Threshold for detecting email spam. Firstly, logistic regression is to process noisy data, decision tree is applied to reduce the noise of the data. [10]. FN which is false negative is to measure the accuracy of final results.[10]. Most of the steps may include data collecting, data preprocessing, obtaining features and data classification. Data collecting includes spam emails and ham emails. Obtaining features include feature extraction, feature normalization and feature selection.[12]. NICT(National Institue of information and communication) utilized SVM with feature selectionand outlier detection.[12]. Spam base is mapping reduce-based SVM and SPONTO(Spam ONTOlogy). SpamAssian utilized naive bayes, suppport vector machine and random forest. J. Nazario, phishing corpus and spam assassin can utilize Random Forest, bayes net and AdaBoost. Spambase can utilize random forest with selecting feature and optimizing parameter. HABUL and Botnet utilize Random Forest. Emails from 500 fortune companies utilize random forest. TREC and CEAS utilize random project and random boost. Spam assassin, spam track and TREC2006 utilized QUANT(QUAdratic neuron-base Neural Tree). Spambase utilizes ANN, GA and NSA(Negative selection algorithm). Enron and Bruce Guenter utilize symbiotic filtering. 100 emails utilize Fuzzy Classification. SpamAssassin utilizes

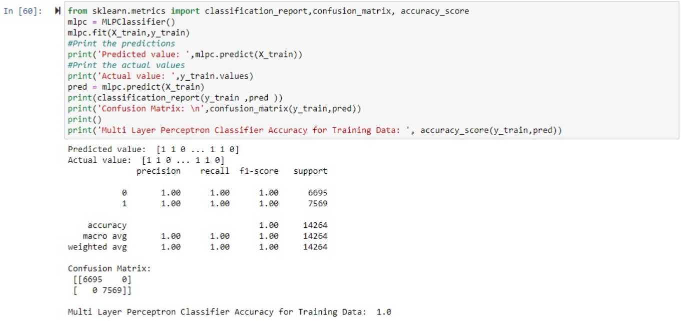
RoughSets.[12].

IV. Your thoughts remarks

Literature Review Report About naive bayes multinomial classifier The features are expected to be produced from a straightforward distribution of multinomial. The distribution of multinomial denotes the probability of identifying the numbers between various categories and so multinomial naive Bayes is being reserved for those features that mentions numbers or number values.The concept is to model the data distribution with a best suited multinomial distribution. The only place where this classifier is commonly used is in the message classification where those features are associated with word numbers or how frequent they repeat. About k-nearest neighbor classifier The knn classifier may be a machine learning classifier which will be implemented to solve both regression and classification difficulties. This classifier is not so difficult to deploy and analyze but encompasses a huge backlog of appearing possibly slows down because of the quantity of data in implementation increases. This classifier runs by identifying the spaces between a question and every one of the samples within the provided data, choosing the required quantity of samples k nearest to the problem and chooses for the foremost periodic label in the view of this classification or mean of those labels in the view of the regression. within the case of classification and forward, as we see by selecting the correct k with respect to provided data is finished by attempting various k's and then choosing the individual that performs better. About support vector machine classifier Svm may be a AI classifier which may be implemented for both categorization and forward challenges. Moreover, it is utilized in categorization obstacles. within the svm, we denote every data object to some extent in n-dimensional capacity where n is the quantity of features with the rate of every feature living the rate of a correlate. Post that we implement the classification by identifying the hyper-plane that differentiates the 2 objects. Svm's are the correlates of the introspection. This algorithm may be a border line that defines the 2 objects which are hyper plane and the line. Support Vector Machine is very strong classification algorithm. When utilized in coordination with the random forest and various AI tools which offers another space to ensemble the categorizations. Such classifiers are very difficult to imagine thanks to the difficulty in the preparation. About the classifier random forest The random forest could be an arrangement calculation comprising of the numerous choices' trees. It utilizes sacking and have haphazardness while building every individual tree to embrace to make an uncorrelated forest of trees whose forecast is more exact than that of some other tree. It's one among various premier exact learning classifiers accessible. It delivers a profoundly precise classifier for various data sets. It runs productively on enormous databases. It can deal with a large number of information factors without the cancellation of given factors. An estimation is given of what factors are significant inside the grouping. This is a proficient strategy for assessing missing information and keeps up precision when an outsized extent of the information is absent.

About multi-layer perceptron classifiers This classifier could be a neural network which creates a gathering of yields through an assortment of data sources. Mlp is recognized through different portions of information hubs related as a directed graph among input and the yield layers. MLP might be a profound study method. An mlp could be an operation relating huge portions inside a synchronized graph. Mlp could be a profound learning strategy as a different portion of neurons. Mlp is generally utilized to tackle issues that need regulation, adapting moreover as examinations through statistical neuroscience.

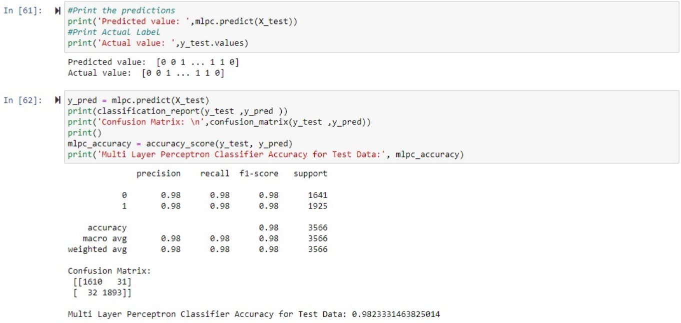
Implementation of classifiers on a given dataset In order to implement the classifiers to detect the accuracy of the spam emails, we have a dataset consisting of 19998 emails that includes both spam and legitimate emails with two columns. One column named 'text' consists of the email body with subject and the other column named 'spam' consists of spam as 1 or legitimate as 0 values. Email spam portrayed likewise as garbage email are those messages sent in mass through email which is called spamming.



```
In [60]:  from sklearn.metrics import classification_report,confusion_matrix, accuracy_score
          mlpc = MLPClassifier()
          mlpc.fit(X_train,y_train)
          #Print the predictions
          print('Predicted value: ',mlpc.predict(X_train))
          #Print the actual values
          print('Actual value: ',y_train.values)
          pred = mlpc.predict(X_train)
          print(classification_report(y_train ,pred ))
          print('Confusion Matrix: \n',confusion_matrix(y_train,pred))
          print()
          print('Multi Layer Perceptron Classifier Accuracy for Training Data: ', accuracy_score(y_train,pred))

          Predicted value:  [1 1 0 ... 1 1 0]
          Actual value:  [1 1 0 ... 1 1 0]
                        precision    recall  f1-score   support

                     0       1.00      1.00      1.00      6695
                     1       1.00      1.00      1.00      7569

              accuracy                           1.00     14264
             macro avg       1.00      1.00      1.00     14264
          weighted avg       1.00      1.00      1.00     14264

          Confusion Matrix:
           [[6695    0]
           [   0 7569]]

          Multi Layer Perceptron Classifier Accuracy for Training Data:  1.0
```

Figure 1

Multi-Layer Perceptron Classifier for Training Data We can perceive how great the model is performed by assessing the multi-layer perceptron classifier and demonstrating the report that incorporates the accuracy rate and matrix. It looks like the Multi-Layer Perceptron classifier used is 100 percentage accurate.



```
In [61]:  #Print the predictions
          print('Predicted value: ',mlpc.predict(X_test))
          #Print Actual Label
          print('Actual value: ',y_test.values)

          Predicted value:  [0 0 1 ... 1 1 0]
          Actual value:  [0 0 1 ... 1 1 0]

In [62]:  y_pred = mlpc.predict(X_test)
          print(classification_report(y_test ,y_pred ))
          print('Confusion Matrix: \n',confusion_matrix(y_test ,y_pred))
          print()
          mlpc_accuracy = accuracy_score(y_test, y_pred)
          print('Multi Layer Perceptron Classifier Accuracy for Test Data:', mlpc_accuracy)

                        precision    recall  f1-score   support

                     0       0.98      0.98      0.98      1641
                     1       0.98      0.98      0.98      1925

              accuracy                           0.98      3566
             macro avg       0.98      0.98      0.98      3566
          weighted avg       0.98      0.98      0.98      3566

          Confusion Matrix:
           [[1610   31]
           [  32 1893]]

          Multi Layer Perceptron Classifier Accuracy for Test Data: 0.9823331463825014
```

Figure 2

Multi-Layer Perceptron Classifier for Test Data How about we test the classifier MLP on the accessible test data collection (Xtest and ytest) by printing the anticipated value and the real value to check whether the model can be precisely classifying the email body. In the wake of assessing the model on the test data collection, the classifier MLP precisely recognized the email messages as spam or real with 98.23 percentage exactness on the test data. To conclude from the results of the accuracy rates that obtained from the classifiers (Multinomial Naive Bayes, K-Nearest Neighbor, Support Vector Machine, Random Forest and Multi-Layer Perceptron Classifiers) that were implemented, the classifier MLP shows the highest accuracy rates for both the data test (98.23 percentage accurate) data and the training (100 percentage accurate) data for the spam email detection.

References

[1] Nandhini, S., KS, J. M. (2020, February). Performance Evaluation of Machine Learning Algorithms for Email Spam Detection. In 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE) (pp. 1-4). IEEE.

[2] Suryawanshi, S., Goswami, A., Patil, P. (2019, December). Email Spam Detection: An Empirical Comparative Study of Different ML and Ensemble Classifiers. In 2019 IEEE 9th International Conference on Advanced Computing (IACC) (pp. 69-74). IEEE.

[3] Annareddy, S., Tammina, S. (2019, December). A Comparative Study of Deep Learning Methods for Spam Detection. In 2019 Third International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC) (pp. 66-72). IEEE.

[4] Karim, A., Azam, S., Shanmugam, B., Kannoorpatti, K., Alazab, M. (2019). A Comprehensive Survey for Intelligent Spam Email Detection. IEEE Access, 7, 168261-168295.

[5] Agarwal, K., Kumar, T. (2018, June). Email Spam Detection Using Integrated Approach of Naïve Bayes and Particle Swarm Optimization. In 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS) (pp. 685-690). IEEE.

[6] Varol, C., Abdulhadi, H. M. T. (2018, December). Comparision of String Matching Algorithms on Spam Email Detection. In 2018 International Congress on Big Data, Deep Learning and Fighting Cyber Terrorism (IBIGDELFT) (pp. 6-11). IEEE.

[7] Vishagini, V., Rajan, A. K. (2018, August). An improved spam detection method with weighted support vector machine. In 2018 International Conference on Data Science and Engineering (ICDSE) (pp. 1-5). IEEE.

[8] Kumar, P., Biswas, M. (2017, February). SVM with Gaussian kernel-based image spam detection on textual features. In 2017 3rd International Conference on Computational Intelligence Communication Technology (CICT) (pp. 1-6). IEEE.

[9] Kaur, H., Sharma, A. (2016, October). Improved email spam classification method using integrated particle swarm optimization and decision tree. In 2016 2nd International Conference on Next Generation Computing Technologies (NGCT) (pp. 516-521). IEEE.

[10] Wijaya, A., Bisri, A. (2016, October). Hybrid decision tree and logistic regression classifier for email spam detection. In 2016 8th International Conference on Information Technology and Electrical Engineering (ICITEE) (pp. 1-4). IEEE.

[11] Zavvar, M., Rezaei, M., Garavand, S. (2016). Email spam detection using combination of particle swarm optimization and artificial neural network and support vector machine. International Journal of Modern Education and Computer Science, 8(7), 68.

[12] Zaid, A., Alqatawna, J. F., Huneiti, A. (2016, August). A proposed model for malicious spam detection in email systems of educational institutes. In 2016 Cybersecurity and Cyberforensics Conference (CCC) (pp. 60-64). IEEE.

[13] Rathod, S. B., Pattewar, T. M. (2015, April). Content based spam detection in email using Bayesian classifier. In 2015 International Conference on Communications and Signal Processing (ICCSP) (pp. 1257-1261). IEEE.

[14] Roy, S. S., Viswanatham, V. M. (2016). Classifying spam emails using artificial intelligent techniques. In International Journal of Engineering Research in Africa (Vol. 22, pp. 152-161). Trans Tech Publications Ltd.

[15] Firte, L., Lemnaru, C., Potolea, R. (2010, August). Spam detection filter using KNN algorithm and resampling. In Proceedings of the 2010 IEEE 6th International Conference on Intelligent Computer Communication and Processing (pp. 27-33). IEEE.

[16] Ma, W., Tran, D., Sharma, D. (2009, November). A novel spam email detection system based on negative selection. In 2009 Fourth International Conference on Computer Sciences and Convergence Information Technology (pp. 987-992). IEEE.