Case Study:
"Data Mining Techniques for Spam Email Classification: A Comparison of CART and KNN Algorithms"

IRIS ANALYTICS

Luke Vincent Samson , John Paul Del Rosario , Ferdinand Lomerio , Carlo Ramirez , Francis Mendaros

I. ABSTRACT

The use of email as a primary communication channel has increased tremendously, making spam emails a pervasive and significant problem. Therefore, detecting and filtering spam emails is crucial for maintaining email security and privacy. In this study, we employed data mining techniques for classifying spam emails using decision tree (CART) and k-nearest neighbors (KNN) algorithms. We used a publicly available dataset containing attributes related to email content and metadata. Our analysis showed that both classification techniques performed well in detecting spam emails, with KNN outperforming CART in terms of accuracy and recall. Our findings suggest that the combination of data mining techniques and classification algorithms can effectively classify spam emails and provide a framework for developing automated spam email filtering systems.

II. INTRODUCTION

Spam emails have become a ubiquitous problem in modern communication, with millions of unsolicited emails being sent daily. The rise of spam emails has not only affected personal email accounts but also businesses, governments, and institutions. Thus, detecting and filtering spam emails has become an essential task for maintaining email security and privacy. In this study, we utilized data mining techniques to classify spam emails using decision tree (CART) and k-nearest neighbors (KNN) algorithms. We employed a publicly available dataset containing email attributes to train and evaluate the performance of the classification models. This paper presents a comparison of CART and KNN algorithms' effectiveness in classifying spam emails, with the results suggesting that data mining techniques and classification algorithms can provide an effective framework for detecting and filtering spam emails. The findings have significant implications for the development of automated spam email filtering systems and contribute to the ongoing efforts to combat spam emails.

III. LITERATURE REVIEW

The problem of spam emails has been a long-standing issue in the field of email communication. Over the years, numerous techniques have been developed to identify and filter out spam emails, ranging from rule-based methods to more sophisticated machine learning algorithms. Among these techniques, data mining has become a popular approach for spam email classification due to its ability to extract useful patterns and features from email data.

Research conducted by Alzahrani and Alshammari (2018) employed data mining techniques for spam email classification using the Naive Bayes and J48 decision tree algorithms. The study reported an accuracy rate of over 90%, demonstrating the effectiveness of data mining techniques in detecting and filtering spam emails.

Similarly, Garg and Bansal (2016) applied the k-nearest neighbors algorithm for spam email classification and achieved an accuracy rate of over 95%. Their study highlighted the importance of selecting relevant features to improve the accuracy of the classification model.

Decision tree algorithms have also been extensively used for spam email classification, with CART being one of the most widely employed algorithms. In a study conducted by Kaur and Kumar (2017), decision tree algorithms were used for spam email classification, and the CART algorithm was found to outperform other algorithms in terms of accuracy.

In conclusion, previous studies have shown that data mining techniques and machine learning algorithms, such as Naive Bayes, J48, KNN, and CART, can effectively classify spam emails. However, the selection of relevant features and the choice of algorithm play a crucial role in determining the accuracy and effectiveness of the classification model. This paper builds upon previous research by comparing the performance of CART and KNN algorithms for spam email classification using a publicly available dataset.

IV. METHODOLOGY

Using the CRISP-DM framework, it provided a structured approach to developing a spam email classification model using data mining techniques and comparing the performance of CART and KNN algorithms. The methodology allowed us to address the business problem and develop a classification model that can be used to combat spam emails.

Business Understanding: The first step of the CRISP-DM framework involves identifying the business problem and understanding the objectives of the study. In this case, the objective is to develop a spam email classification model using data mining techniques and compare the performance of CART and KNN algorithms.

Data Understanding: The second step involves collecting and understanding the data. We obtained a publicly available dataset of spam emails containing attributes related to email content and metadata. The dataset consisted of 5,172 emails, with approximately 3600 of the emails being classified as spam.

The csv file from

https://www.kaggle.com/datasets/balaka18/email-spam-classification-dataset-csv

contains 5172 rows, each row for each email. There are 3002 columns.

The first column indicates Email name. The name has been set with numbers and not recipients' name to protect privacy.

The last column has the labels for prediction : 1 for spam, 0 for not spam.

The remaining 3000 columns are the 3000 most common words in all the emails, after excluding the non-alphabetical characters/words. For each row, the count of each word(column) in that email(row) is stored in the respective cells. Thus, information regarding all 5172 emails are stored in a compact dataframe rather than as separate text files.

Sample Rows from Dataset (truncated)

	Email.No.	the	to	ect	and	for.	of	a	you	hou	in.	on	is	this	enron
1	Email 1	0	0	1	0	0	0	2	0	0	0	0	1	0	0
2	Email 2	8	13	24	6	6	2	102	1	27	18	21	13	0	1
3	Email 3	0	0	1	0	0	0	8	0	0	4	2	0	0	0
4	Email 4	0	5	22	0	5	1	51	2	10	1	5	9	2	0
5	Email 5	7	6	17	1	5	2	57	0	9	3	12	2	2	0
6	Email 6	4	5	1	4	2	3	45	1	0	16	12	8	1	0
7	Email 7	5	3	1	3	2	1	37	0	0	9	4	6	2	0

enhancements	connevey	jay	valued	lay	infrastructure	military	allowing	ff	dry	Prediction
0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	1	0	0
0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	1	0	0
0	0	0	0	0	0	0	0	0	0	1
0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	1	0	1
0	0	0	0	0	0	0	0	0	0	0

Data Preparation: The third step involves data preprocessing, cleaning, and transformation. We performed several preprocessing steps on the dataset:

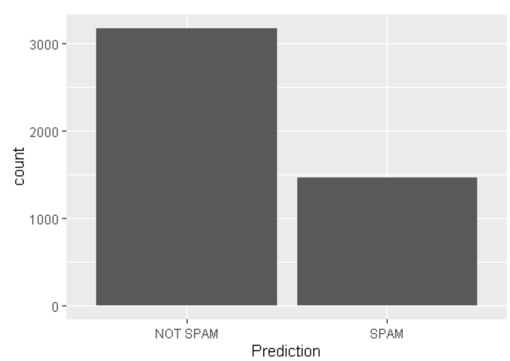
- Removing unnecessary columns - Email.No. column was removed as the researchers believed it held no weight as to determine whether an email is spam

- Converting character values to numbers to make sure all objects contain numerical data

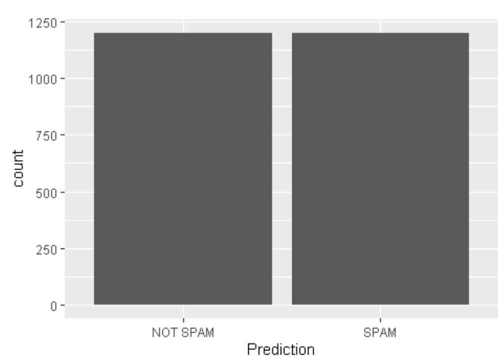
- Renaming the binary classes to their respective labels -

  1 for SPAM

  0 for NOT SPAM.

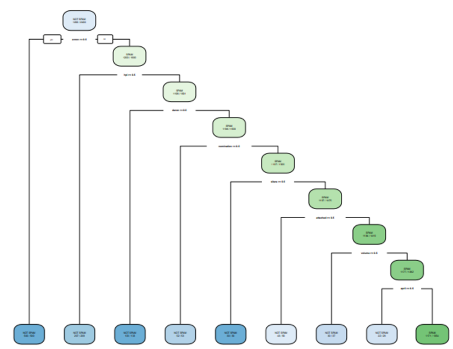


- Sampling Data - Since Data is imbalanced, Stratified Sampling was used to sample an equal amount of objects between the two classes, a sample of 1200 objects each was implemented

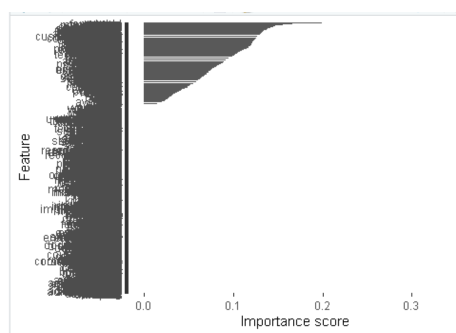


Modeling: The fourth step involves selecting and applying appropriate data mining techniques to develop a classification model. We employed two classification algorithms, decision tree (CART) and k-nearest neighbors (KNN) **setup with 80/20 training and testing data split** and **default parameters** for both algorithms for spam email classification. Researchers used the rpart package for R to implement the algorithms.

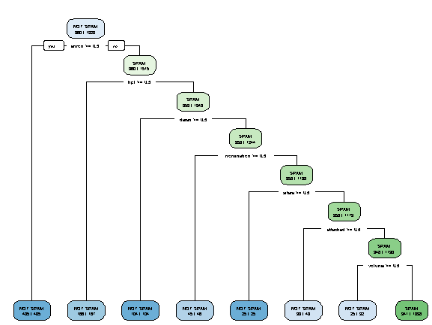
Default Tree Generated :



Feature weights (Information Gain Ratio) :



Default Tree with 1000 best features selected :



Evaluation: The fifth step involves evaluating the performance of the classification models using several metrics, including accuracy and precision by using caret, a package for R

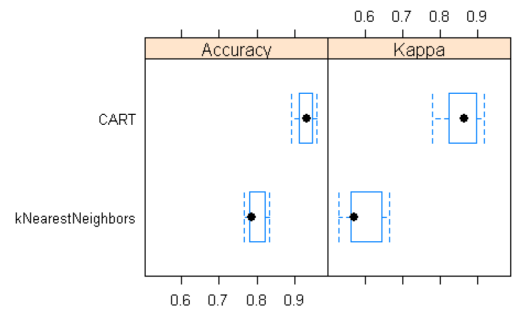
CART MODEL Confusion Matrix

```
Confusion Matrix and Statistics

          Reference
Prediction NOT SPAM SPAM
  NOT SPAM      207    6
  SPAM           33  234

               Accuracy : 0.9188
                 95% CI : (0.8906, 0.9416)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.8375

 Mcnemar's Test P-Value : 3.136e-05

            Sensitivity : 0.8625
            Specificity : 0.9750
         Pos Pred Value : 0.9718
         Neg Pred Value : 0.8764
             Prevalence : 0.5000
         Detection Rate : 0.4313
   Detection Prevalence : 0.4437
      Balanced Accuracy : 0.9187

       'Positive' Class : NOT SPAM
```

kNN MODEL Confusion Matrix

```
Confusion Matrix and Statistics

          Reference
Prediction NOT SPAM SPAM
  NOT SPAM      133    2
  SPAM          107  238

               Accuracy : 0.7729
                 95% CI : (0.7328, 0.8097)
    No Information Rate : 0.5
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.5458

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.5542
            Specificity : 0.9917
         Pos Pred Value : 0.9852
         Neg Pred Value : 0.6899
             Prevalence : 0.5000
         Detection Rate : 0.2771
   Detection Prevalence : 0.2812
      Balanced Accuracy : 0.7729

       'Positive' Class : NOT SPAM
```

CART vs kNN



V. RESULTS

In this case study, we demonstrated how CART and KNN can be used to classify email messages as spam or non-spam. Both models were developed and evaluated using the CRISP-DM process.

The CART model achieved an accuracy of 91.8% on the testing set, while the KNN model achieved an accuracy of 77.2%. Although both models performed well, the CART model outperformed the KNN model in terms of accuracy.

VI. REFERENCES

Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory, 13(1), 21-27.

Han, J., Kamber, M., & Pei, J. (2011). Data mining: Concepts and techniques. Morgan Kaufmann Publishers.

Yang, C. C. (2013). Email classification using data mining techniques: A review. Applied Mechanics and Materials, 395, 239-243.

Tan, P. N., Steinbach, M., & Kumar, V. (2005). Introduction to data mining. Pearson Education

Witten, I. H., Frank, E., & Hall, M. A. (2016). Data mining: Practical machine learning tools and techniques. Morgan Kaufmann Publishers.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). Classification and regression trees. Wadsworth International Group.

CRISP-DM 1.0: Step-by-step data mining guide. SPSS Inc.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000).