Enhancement of Logistic Regression Algorithm Applied in Email Spam Detection

Vince Anthony S. Carlos[1], John Cedric C. Pancho[2], Vivien A. Agustin[3]

[1,2]Author, Pamantasan ng Lungsod ng Maynila
[3]Co-Author, Pamantasan ng Lungsod ng Maynila

ABSTRACT

Logistic regression is a popular binary classification approach, but like any machine learning algorithms, it has its limitations and possible concerns such as class imbalance, large datasets, and overfitting, which reduce its accuracy and efficiency. This study enhanced the Logistic Regression algorithm's performance for email spam detection by addressing these problems using the techniques of Term Frequency-Inverse Document Frequency for class imbalance, Recursive Feature Elimination for large datasets, and Principal Component Analysis for overfitting concerns. TF-IDF improves feature representation, highlighting key terms that differentiate spam from non-spam. RFE systematically eliminates irrelevant features, reducing computational complexity and enhancing efficiency, particularly for large datasets. PCA mitigates overfitting by reducing the dimensionality of feature spaces, ensuring the model generalizes effectively to unseen data. The enhanced Logistic Regression model demonstrated a significant improvement in spam detection accuracy, achieving up to 98% accuracy with TF-IDF. RFE reduced training time while maintaining robust performance on large datasets, and PCA improved model generalization, reducing overfitting risks. The proposed enhancements successfully address the key limitations of traditional Logistic Regression models in spam detection. This refined approach improves predictive accuracy, computational efficiency, and robustness, making it highly applicable to real-world email security systems.

CHAPTER ONE
INTRODUCTION
1.1 Background of the Study

Emails have recently become a crucial means of communication. Unsolicited commercial bulk emails, commonly known as spam, have become a major issue on the internet. Spam emails continue to be a significant problem for users. Spammers who send fraudulent emails collect email addresses from various sites and may include viruses and malicious codes in their messages. Spam interferes with internet users' ability to maximize their storage capacity and network bandwidth. With their increasing significance, emails have found applications in diverse industries such as finance, banking, and marketing. However, this growing reliance on email communication has also led to an escalation in cybersecurity threats, particularly in the form of email-based exploitation *(Deshpande et al., 2023)*. Cybercriminals exploit email channels to launch attacks that can cause significant harm to individuals and organizations. It has been estimated that up to 90% of cyber-attacks originate from emails (cyberattacks, 2020). On average, 46.59% of emails in 2023 were spam, which is 5.85 p.p. lower than in 2022. Q1 was the calmest quarter as it was, globally. *(Kaspersky, 2023).* Spam emails are extremely bothersome and harmful to users who have been deceived by fake internet mail and other dishonest practices that involve sending emails aimed at tricking unsuspecting individuals into revealing confidential information, such as usernames and passwords, as well as credit card numbers.

The Logistic Regression (LR) algorithm integrates various techniques that highlight the connections between independent variables and a methodology for assessing datasets, especially when one or more autonomous factors influence the outcomes. This algorithm employs a linear approach to predict probabilities and is effective in diminishing noise within data *(Arfah Anggraina et al 2019)*. Logistic regression is a highly suitable and commonly used algorithm for dataset classification. When classifying a dataset referred to as a "spam base," logistic regression serves as an exceptionally flexible decision-making tool for identifying spam emails within the dataset. This method conducts basic tests on the data distribution, which includes calculating statistical measures such as mean and standard deviation. Additionally, it performs operations like word and character counting, as well as maximum and minimum operations. *(Khanday & Pharvin 2021).*

Logistic regression is a widely used algorithm for binary classification, but like any machine learning algorithm, it has its limitations and potential issues. Logistic regression works by trying to get the best accuracy possible, but this method doesn't work well when the data is uneven. Most of the time, the models end up favoring the larger group, which can cause big problems in real-world applications *(Zhang L & Geisler T et al 2021)*. Selecting only statistically significant predictors can lead to severe distortions in regression coefficients. For example, the bias could be as high as 300% for certain odds ratios, indicating that selection bias can be more dominant than other types of bias in a model *(Steyerberg 2011)*. Main reason for this bias is the model's basic assumption that there is a straightforward, linear connection between the log odds of the result and the predictor variables. This assumption simplifies things too much and can incorrectly represent the more complex, non-linear relationships that exist in the real world, leading to biased results and decisions.

1.2 Statement of the Problem

As technology advances, the need for improving machine learning algorithms like Logistic Regression becomes critical to address various challenges. Logistic Regression, commonly used for binary classification, faces limitations that can affect its performance in more complex tasks. Here are the main issues that show why is there a need to make this algorithm enhanced:

1. Many statistics and machine learning methods, including logistic regression, which is known for being easy to understand, tend to be biased towards the larger group. This can lead to less accurate predictions for the smaller group, which might end up costing money and damaging reputations. *(Zhang L & Geisler T et al 2021)*. In logistic regression analysis, class imbalance is a significant issue, particularly evident when one class outnumbers the other, which complicates model specification and accuracy *(Suaad & Intesar 2023).*
2. Studies have shown that while standard machine learning models like Logistic Regression are straightforward to use for spam detection, they can still struggle to accurately classify spam, especially in larger datasets *(Bilge & Bariye 2019)*. As stated by the research study, the regression techniques are applicable only for minimal datasets having around some hundreds of records which is evident from the literature *(Dhamodharavadhani & Rathipriya Ramalingam 2019).*
3. Overfitting is an occurrence in a logistic regression that fits too closely to training data, but cannot make accurate predictions or conclusions from any data. While an overfitted model has a high accuracy for its training data, it will run poorly for the new or unseen data (Awan, 2023). As stated in a study

by Olaoye et al (2024), logistic regression is prone to overfitting if the model that is used is too complex relative to the data, or if it has too many irrelevant or noisy features.

1.3 Objective Of the Study

As data-driven threats and challenges become more complex, it is essential to enhance logistic regression to improve their predictive capabilities. This techniques aims to enhance the Logistic Regression Algorithm with the following key objectives:

1. Implementing ***TF-IDF (term frequency-inverse document frequency)*** can enhance logistic regression's performance on imbalanced datasets by improving the feature representation of textual content, thereby boosting the model's ability to distinguish important differences between classes. By emphasizing terms that are uniquely significant to the minority class and reducing the influence of common, less informative words, TF-IDF can help mitigate the inherent bias towards the majority class.
2. Utilizing ***Recursive Feature Elimination (RFE)*** it can greatly improve the efficiency of logistic regression, especially with large datasets. RFE systematically eliminates the less critical features, reducing data complexity and lessening the computational load on the logistic regression algorithm. Additionally, by concentrating on the most significant features, RFE maintains the robustness and efficiency of the logistic regression algorithm. This makes it more capable of handling and analyzing substantial data volumes, like those found in complex spam detection tasks.
3. Implementing ***Principal Component Analysis (PCA)*** can enhance the effectiveness of logistic regression models for detecting email spam by addressing overfitting concerns. PCA reduces the dimensions of feature spaces, which are typically large in email datasets due to diverse words and phrases, and concentrates on the most relevant data aspects. This reduction not only streamlines the logistic regression algorithm but also enhances its computational speed. By eliminating noisy and unnecessary features, PCA retains only the most crucial variables that aid in spam detection.

1.4 Significance of the study:

The study on enhancement of Logistic Regression algorithm applied in email spam detection is significant for several reasons and will benefit to:

Email Users. By improving the accuracy and efficiency of spam detection, the enhanced Logistic regression algorithm directly benefits all email users by reducing the volume of spam they encounter. This improvement not only protects users from potential scams and phishing attempts but also enhances their overall user experience by minimizing distractions and clutter in their inboxes.

Future Researchers. The advancements and insights derived from enhancing the Logistic Regression algorithm for spam detection provide a solid foundation for future research in the field. By addressing current limitations and exploring novel improvements, this research opens up new avenues for further studies. Future researchers can build upon the methodologies and findings to explore additional machine learning techniques, adapt the enhanced logistic regression to other forms of cybersecurity threats, This ongoing research will contribute to the continual development of more effective, adaptive, and resilient cybersecurity measures.

Organizations. Businesses and organizations of all sizes rely heavily on email for communication. Enhanced spam detection will safeguard these entities from cybersecurity threats carried via email, thus protecting their sensitive data and maintaining their integrity. Improved filtering mechanisms can prevent significant financial and reputational damage caused by email-based attacks.

Programmers/Developers. The advancements in logistic regression from this research offer valuable insights for developing more effective machine learning models. By understanding and applying these enhanced techniques, developers can create more sophisticated systems.

1.5 Scope and limitations:

The study focuses on improving the Logistic Regression algorithm specifically for email spam detection. The primary scope involves advancing the current Logistic Regression model by incorporating techniques such as Term Frequency-Inverse Document Frequency (TF-IDF), Recursive Feature Elimination (RFE), and Principal Component Analysis (PCA). These enhancements are aimed at addressing significant issues like class imbalance, managing large datasets, and preventing model overfitting. The study will utilize a spam email dataset, analyzing various attributes to categorize emails as spam or non-spam and will employ metrics such as accuracy, precision, and recall to measure the enhancements' impact on the Logistic Regression Algorithm.

However, the study encounters several limitations. Despite the use of machine learning techniques, it can still be extremely challenging and might still face some issues. Moreover, the enhancements are designed specifically for spam detection and might not be directly applicable to other types of classification tasks or datasets. Finally, the rapid evolution in spamming techniques and email systems could potentially outpace the model improvements, possibly rendering some aspects of the enhanced model less effective over time. Despite these limitations, the study aims to provide valuable insights into improving the Logistic Regression model for spam detection, which could guide further developments in the field.

1.6 Definition of Terms:

This section addresses the relevant terminologies and the concepts they represent, ensuring an accurate and precise interpretation of the central themes discussed throughout the study. The definitions are derived from reliable sources and reflect the context in which the terms are used within the study.

Bias. In statistics and machine learning, this refers to errors introduced in a model due to oversimplified assumptions in the machine learning process.

Class Imbalance. A situation in machine learning where the number of instances of one class is significantly higher than those of other classes, which can lead to a biased model favoring the majority class.

Cybersecurity. The practice of protecting systems, networks, and programs from digital attacks aimed at assessing, changing, or destroying sensitive information, extorting money from users, or interrupting normal business processes.

Dataset Classification. The process of organizing data into categories that make it more effective to retrieve, manage, and process.

Email-based Exploitation. The use of email as a medium to conduct harmful activities such as delivering malware, phishing for sensitive information, and executing fraudulent schemes.

Independent Variables. Variables in an analysis or mathematical model whose variation does not depend on that of another.

Logistic Regression (LR). A statistical method used for binary classification that models the probability of a certain class or event existing such as pass/fail, win/lose, alive/dead, or healthy/sick.

Overfitting. A modeling error in statistics and machine learning where a function is too closely fit to a limited set of data points, resulting in poor predictive performance on new data.

Principal Component Analysis (PCA). A statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components.

Recursive Feature Elimination (RFE). A feature selection method that fits a model and removes the weakest features (or feature) until the specified number of features is reached, helping to enhance model simplicity and performance.

Spam. Unsolicited commercial emails, often sent in bulk and containing advertisements, scams, or malicious links.

Statistical Measures. Quantitative measures that describe characteristics of data, such as mean (average), median (middle value), standard deviation (measure of variation from the mean), and others.

TF-IDF (Term Frequency-Inverse Document Frequency). A numerical statistic intended to reflect how important a word is to a document in a collection or corpus. It emphasizes words that are unique to specific documents and de-emphasizes words that are common across multiple documents.

CHAPTER TWO
REVIEW OF RELATED LITERATURE
2.1 Related Literature

In this research study by Acito 2023, titled "Logistic Regression". This research paper elaborates fundamental concepts underlying logistic regression, emphasizing its advantages over ordinary linear regression for modeling binary outcomes. It provides a detailed explanation of the logistic function and its role in transforming linear predictors into probabilities. Furthermore, the chapter explores performance evaluation metrics such as confusion matrices and receiver operating characteristic (ROC) curves, essential for assessing model accuracy and predictive power.

Based on the study of Kuiziniene and Krilavicius (2024), "Balancing Technique for Advanced Financial Distress Detection using Artificial Intelligence" it takes a thorough approach to tackle imbalanced datasets in predicting financial distress (FD). Researchers meticulously curated a Combined FD dataset, covering five distinct condition states. They rigorously assessed ten class balancing techniques, five dimensionality reduction methods, two feature selection strategies, eleven machine learning models, and twelve weighted majority algorithms (WMAs). Notably, their findings showcase superior performance, with the extreme gradient boosting machine (XGBoost) standing out for feature selection, alongside the experimental max number strategy. Furthermore, they highlight the efficacy of undersampling techniques for class balancing and emphasize the effectiveness of WMA 3.1 for weighted majority voting, as evidenced by high scores in the area under the receiver operating characteristic curve (AUC).

A research paper by Brightwood and Anthomy (2024) which has a title "Practical Considerations and Challenges in Applying Logistic Regression for Cyber Threat Detection", it highlights crucial factors necessary for effectively applying logistic regression in cyber threat detection. It stresses the importance of ensuring high-quality data and employing appropriate preprocessing techniques. This includes emphasizing the significance of feature selection, normalization, and addressing imbalances within datasets to improve the accuracy of the models. Additionally, the review underscores the critical role of feature engineering, which relies on domain expertise to identify and extract relevant features that encapsulate the intricacies of cyber threats. Furthermore, it emphasizes the interpretability of logistic regression models, allowing analysts to gain insights into the factors influencing threat detection decisions.

In a research by Brightwood (2024), titled "Regularization Techniques in Logistic Regression", it shows that Logistic Regression is a popular model for statistical technique which is used for binary classification. It is also used to predict the probability of a binary outcome. It is also stated in the paper that Logistic

Regression suffers from high-variance or overfitting especially when taking on high-dimensional datasets or when the number of features exceed the number of observations, which will result in poor generalization, decreased model performance, and unreliable predictions on new data. The researchers used a regularization technique called L1 or Lasso Regularization, which resulted in the shrinking of coefficients of irrelevant features to zero.

In a study entitled “Combining principal component analysis and logistic regression for multifactorial fall risk prediction among community-dwelling older adults” by Pan, et al (2024) which is aimed to create a robust fall risk prediction model tailored for older adults residing in the community. The research team gathered data on 45 variables associated with falls from a sample group comprising 1630 older adults in Taiwan. Employing a two-step approach, they utilized principal component analysis (PCA) as a preprocessing technique followed by logistic regression for modeling. Through this methodology, the final model achieved notable performance metrics, including an area under the receiver operating characteristic curve (AUC-ROC) of 0.78, with sensitivity, specificity, and accuracy rates of 74%, 70%, and 71%, respectively. The integration of PCA with logistic regression proved to be advantageous, offering a more dependable and practical tool for predicting fall risks compared to traditional methods.

According to a research paper by Wu, et al (2024), titled "Web Attack Detection Based on Honeypots and Logistic Regression Algorithm", similar to the paper, this research focuses on detecting Web attacks based on honeypots and Logistic Regression Algorithm. Logistic regression is employed in this study to train and test the classification of the text vectors, which generates a logistic regression model. It showed results that achieved rapid and accurate detection and recognition of web attack behaviors while ensuring performance efficacy.

According to the paper of Gifford and Bayrak (2023), entitled "A predictive analytics model for forecasting outcomes in the National Football League games using decision tree and logistic regression", it uses decision tree and Logistic Regression to construct a predictive analytics models to forecast the NFL games outcomes in a season using decision trees and Logistic Regression. The researchers compared the Decision Tree and Binary Logistic Regression models, the binary Logistic Regression displayed a lower misclassification rate, while Decision Tree model is higher than the misclassification rate of Logistic Regression.

Based on a study of Haq, et al (2023), it explores factors affecting Covid-19 recovery time, crucial for communities, medical practitioners, and governments. Using logistic regression and geographically weighted logistic regression (GWLR), the research aims to uncover nuanced insights into recovery dynamics. Analyzing 2021 Covid-19 data from West Sumatra's Regional Research and Development Agency, with 764 observations across 19 regencies/cities, the study finds no notable difference between logistic regression and GWLR models. Surprisingly, GWLR's inclusion of spatial factors doesn't provide additional insights, suggesting uniform recovery dynamics across West Sumatra's geographic regions.In a study to compare Logistic Regression to other models by Balboa, et al (2024), which has a title of “Logistic Regression vs Machine Learning to predict evacuation decisions in fire alarm situations”. The study compares logistic regression and various machine learning algorithms for predicting evacuation decisions during emergencies. Seven models were tested, including logistic regression, decision trees, Naïve Bayes, K-nearest neighbours, SVM, random forest, extreme gradient boosting, and artificial neural network. Logistic regression showed high predictive accuracy, while extreme gradient boosting excelled in overall accuracy, specificity, and precision. K-nearest neighbours demonstrated superior recall, and the artificial neural network achieved a high F1-score, indicating balanced precision and recall.

Based on a study by Wahyuningsih, et al (2024), entitled "Comparison of Effectiveness of Logistic Regression, Naive Bayes, and Random Forest Algorithms in Predicting Student Arguments", it focuses on comparing between three algorithms such as Logistic Regression, Naive Bayes, and Random Forest to predict student argumentation using essays from grades 6-12. It showed that Logistic Regression performs best with an accuracy rate of 94.34%, followed by Random Forest with 91.98% accuracy, and Naive Bayes with the lowest score of 88.93% accuracy.

2.2 Related Studies

CLASS IMBALANCED DATASETS IN MACHINE LEARNING ALGORITHM

The issue of class imbalance has gained significant attention in recent years, particularly in various practical domains of machine learning as stated by Guo et al. (2023) . In a class imbalance scenario, the majority of examples are labeled as one class, while only a small portion represents the minority class, which is often the class of greater importance. This imbalance causes standard machine learning algorithms to become biased toward the majority class, leading them to overlook the minority class, as they aim to optimize overall accuracy. Consequently, models may fail to detect critical instances of the minority class, reducing their effectiveness in applications where minority class predictions are essential. In a study by Rahman and Davis (2013), high-risk patients are often in the minority class, making their correct classification critical. Therefore, there is a need for robust sampling techniques tailored to medical datasets. It is also stated that Real-world data are often imbalanced, which is a significant factor contributing to the decrease in the generalization performance of machine learning algorithms. Conventional algorithms typically do not account for class imbalance, treating both the majority and minority classes with equal importance. However, when the imbalance is severe, it becomes challenging to build an effective classifier using these traditional approaches. In many class-imbalanced datasets, especially in medical applications, the cost of mispredicting the minority class is much higher than that of the majority class.

In this study by Buda et al. (2018), an investigation for the impact of class imbalance on classification performance of Convulutional neural networks and compare some of the methods that are frequently used to address the issue. The researchers stated that Class Imbalance is a common problem that has been comprehensively studied in classical machine learning, yet there are limitations in the context of deep learning. The several methods used in this study are the following: oversampling, undersampling, two-phase training, and thresholding. The study found out that the effect of class imbalance on classification performance is detrimental and others.

CHALLENGES IN HANDLING LARGE DATASET

According to a research study titled "Large language models overcome the challenges of unstructured text data in ecology" by Castro et al (2024), manual processing of large data can be labour-intensive, making it a significant challenge to process. The study assessed the application of three large language models which is GPT-3.5, GPT-4, and LLaMA-2-70B, to process information from unstructured textual sources. The assessment resulted with GPT-4 consistently outperforming other LLMs with the percentage of correct outputs often exceeding 90%.

Big Data presents both significant opportunities and challenges in modern data analysis. While large datasets allow for the discovery of complex patterns and population heterogeneities that small-scale data cannot reveal, they also pose unique issues such as scalability, storage constraints, noise accumulation, spurious correlations, and incidental endogeneity. These challenges necessitate new computational and statistical paradigms, as traditional methods often rely on exogenous assumptions that are difficult to validate with Big Data. Failure to address these issues can lead to inaccurate statistical inferences and misguided scientific conclusions (Fan, Han, & Liu, 2014).

In a research study by Cheng et al (2020), the researchers stated that large sizes of data may open up a lot of opportunities for scientific discoveries, it can also pose a lot of challenges when attempting to analyze these large data sets. The study on subdata selection methods under logistic regression models highlights the effectiveness of the information-based optimal subdata selection approach in handling Big Data challenges. This approach outperforms traditional random sampling strategies, such as uniform sampling, by ensuring that at least one eigenvalue of the information matrix grows as the full data size increases, even with a fixed subdata size. This finding underscores the method's capacity to balance computational complexity with statistical efficiency, making it a promising tool for Big Data analysis. The research suggests that information-based methods can be a superior alternative to traditional subdata selection techniques when dealing with large datasets, offering a practical solution to the trade-offs between accuracy and computational feasibility.

TECHNIQUES FOR OVERFITTING

Overfitting, which is characterized by high accuracy for a classifier when evaluated on the training set but low accuracy when evaluated on a separate test set, is a recognized issue in high-dimensional low-sample-size (HDLSS) settings. It has been emphasized that the training set accuracy is overly optimistic and thus unreliable, and accuracy should instead be evaluated on separate test sets or through resampling techniques where the model is redeveloped for each iteration. Despite overfitting concerns being relevant even in traditional settings where the number of predictors (p) is less than the number of observations (n), such practices are not as commonly applied (Subramanian and Simon, 2013).

According to a research entitled "Avoiding Overfitting: A Survey on Regularization Methods for Convolutional Neural Networks" by Santos et al (2022), CNN is machine learning used on computer vision-related tasks, such as image classification and object detection. Considering the information available, it may require more data variability than possible. This can cause the unavailability of new data which can cause overfitting, where it performs well on the training but now on new data. In this research, regularization methods are used to prevent the model from overfitting the training data. Regularization is an essential technique for improving the performance of CNNs, as it helps prevent overfitting to the training data. The study aims to highlight recent advancements in the field, providing a concise overview of how these methods function and summarizing their key outcomes.

One effective way to address this issue is through dimensionality reduction, which helps manage high-dimensional datasets by simplifying the feature space, making it easier to derive insights while maintaining the model's performance. Dimensionality reduction techniques also help reduce the number of variables in a dataset without sacrificing, and often enhancing, the model's efficiency. A research by Salam et al (2021), discusses nine dimensionality reduction techniques and their impact on mitigating overfitting: missing-values ratio, low variance filter, high correlation filter, random forest, principal component analysis (PCA), linear discriminant analysis (LDA), backward elimination, forward feature selection, and rough set theory. These techniques are essential in reducing overfitting while maintaining or even improving model performance.

2.3 Comparative Analysis

This section presents a comparative analysis, examining and contrasting the key elements relevant to the study. It aims to highlight differences, similarities, and significant trends, providing a thorough evaluation based on reliable sources and contextual relevance. The analysis is intended to offer a deeper understanding of the subject matter by drawing on established comparisons.

Figure 1 Comparison of Accuracy, Precision, Recall, F1 Score

Algorithm	Accuracy	Precision	Recall	F1 Score
Logistic Regression	94.34%	93.72%	94.96%	94.33%
Naive Bayes	88.93%	88.75%	89.12%	88.94%
Random Forest	91.98%	91.68%	92.31%	91.98%

Figure 1 shows the comparison of overall performance of the three algorithms which are Logistic Regression, Naive Bayes, and Random Forest. Findings have shown that Logistic Regression is the best performing model in all evaluation metrics. This table was derived from a study entitled “Comparison of Effectiveness of Logistic Regression, Naive Bayes, and Random Forest Algorithms in Predicting Student Arguments” by Wahyuningsih et al., (2024).

Algorithm	CV	Accuracy (Test Data)%	Accuracy (Training Data)%	F1 Score
Logistic Regression	10 folds	88.2	82.8	0.88
SVM	10 folds	87	82	0.87
Neural Networks	NA	83-85	83-85	

Figure 2 Comparison of Accuracy on Test Data and Training Data

Figure 2 shows the comparison of accuracy on test data and training data of the three algorithms which are Logistic Regression, Support Vector Machine, and Neural Networks. Findings have shown that Logistic Regression was the best performing model in all accuracy in terms of accuracy on test data and training data. This table was derived from a study entitled “Comparative Study of Classification Techniques (SVM, Logistic Regression and Neural Networks) to Predict the Prevalence of Heart Disease” by Khanna et al., (2015).

Model	AUC	Accuracy	Recall	Specificity	Precision	F1-score
LR	0.831	0.756	0.790	0.716	0.767	0.778
CART	0.716	0.772	0.762	0.783	0.805	0.783
NB	0.744	0.748	0.793	0.695	0.753	0.773

KNN	0.711	0.731	0.859	0.604	0.686	0.777
SVM	0.777	0.772	0.760	0.783	0.805	0.776
RF	0.699	0.754	0.829	0.770	0.791	0.757
XGBoost	0.761	0.780	0.793	0.810	0.820	0.776
ANN	0.770	0.773	0.732	0.783	0.805	0.785

Figure 3 Comparison of Performance of each model

Figure 3 shows the comparison of performance of the 8 algorithms which are Logistic Regression, Classification and Regression Tree, Naive Bayes, K-nearest neighbors, Support Vector Machine, Random Forest, XGBoost, and Artificial Neural Network. Findings have shown that Logistic Regression provided better prediction capability (AUC) among all models. This table was derived from a study entitled "Logistic regression vs. machine learning to predict evacuation decisions in fire alarm situations" by Balboa et al., (2024).

2.4 Synthesis

The reviewed literature highlights the widespread application of logistic regression and other machine learning techniques in various domains, including financial distress prediction, cyber threat detection, and health-related risk assessment. Logistic regression is particularly favored for its simplicity and interpretability, especially in binary classification tasks, and is frequently combined with dimensionality reduction methods like principal component analysis (PCA) to improve model performance. Many studies emphasize the critical role of class balancing, feature selection, and regularization techniques such as Lasso to address overfitting, particularly in high-dimensional or imbalanced datasets. Furthermore, recent advancements in ensemble methods and hybrid approaches, such as the integration of weighted majority algorithms (WMAs) with logistic regression models, demonstrate notable improvements in predictive accuracy across multiple fields.

A key similarity among the studies is their shared focus on addressing challenges inherent to specific data types, such as class imbalance and overfitting. For instance, both Kuiziniene and Krilavicius (2024) and Balboa et al. (2024) emphasize balancing techniques, including undersampling and the use of weighted majority algorithms, to improve model accuracy. Similarly, the works of Brightwood (2024) and Pan et al. (2024) highlight the use of dimensionality reduction techniques such as PCA and Lasso regularization to manage high-dimensional datasets. Overfitting is also a concern across these studies, with researchers employing strategies like regularization, feature selection, and resampling to mitigate its effects on model generalization and predictive performance.

Despite these similarities, the studies differ in their application contexts and specific modeling approaches. For instance, Kuiziniene and Krilavicius (2024) focus on predicting financial distress using a variety of machine learning models, with XGBoost emerging as the superior model, whereas Pan et al. (2024) rely on logistic regression combined with PCA for predicting fall risks among older adults. Meanwhile, Brightwood and Anthomy (2024) apply logistic regression in cyber threat detection, stressing feature engineering and data preprocessing, while Haq et al. (2023) compare logistic regression with geographically weighted logistic regression (GWLR) for Covid-19 recovery predictions, ultimately finding no added benefit from spatial factors. Each study, while tackling similar technical challenges, is tailored to the specific demands and nuances of its respective domain.

CHAPTER THREE
METHODOLOGY
3.1 Logistic Regression Algorithm

Logistic regression is a statistical method used to model conditional probabilities with discrete outcomes, often binary. The model, also referred to as the logit model, predicts the likelihood of a categorical result (such as "yes/no" or "pass/fail") based on one or more predictor variables, which are usually continuous *(Eldridge, 2023).*

Logistic regression helps calculate the chances of two possible outcomes and is useful for classification tasks. For example, it can predict if an email is spam, if a credit card transaction is fraudulent, if a tumor is benign or malignant in medicine, if a customer will buy a product in marketing, or if a student will finish their course on time in online education (Thanda, 2023).

3.2 Logistic Regression Framework

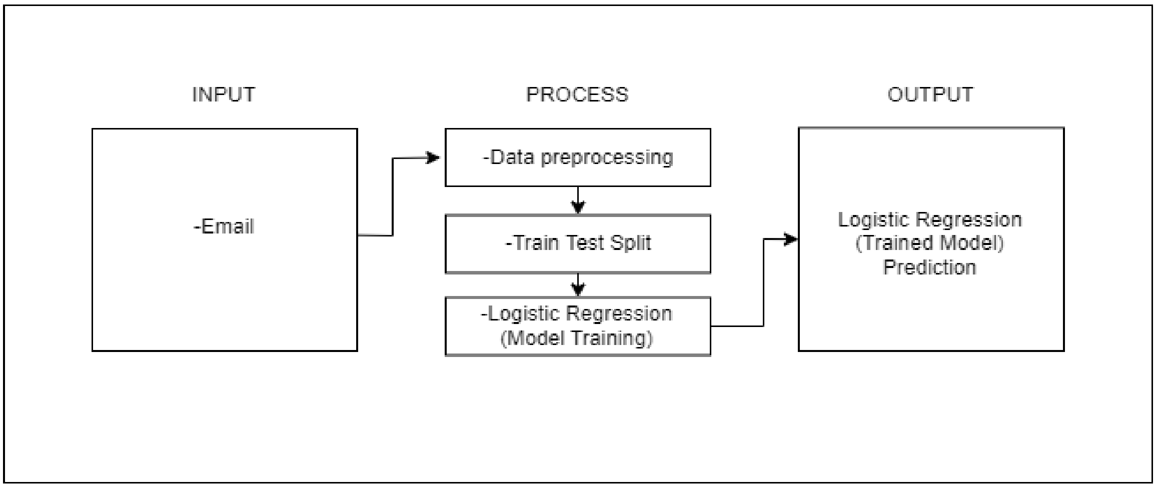

INPUT
PROCESS
OUTPUT
-Email
-Data preprocessing
-Train Test Split
-Logistic Regression (Model Training)
Logistic Regression (Trained Model) Prediction

Figure 4 Conceptual Framework of Logistic Regression

Figure 4 shows the conceptual framework of Logistic Regression Algorithm. The process begins with the input stage, where an email has been received. Each email is labeled as either spam or non-spam, which forms the foundation of the dataset.

During the process stage, numerous procedures are performed successively to convert the raw data into a format that the logistic regression algorithm can use. The data preprocessing phase consists of important tasks such as cleaning, which removes unnecessary symbols, punctuation, and whitespace, and tokenization, which turns the text into individual words or tokens. The dataset is then divided into training and testing sets through a train-test split. This split ensures that the model's performance is evaluated on unseen data, providing an unbiased measure of its generalization ability. The training phase involves the logistic regression model learning the optimal weights for each feature to minimize classification errors. Following that, the evaluation step assesses the model's performance using metrics such as accuracy, precision, recall, and the F1 score. In the output stage, the model outputs probabilities for each input message, indicating whether it is spam or not.

However, upon closer look it is clear that the original framework has limitations. While the feature extraction stage enables numerical representation, it does not automatically handle the selection of the most predictive features or address issues like high dimensionality. This lack of refinement may lead to suboptimal model performance, particularly with complex datasets. Additionally, the model does not incorporate techniques to reduce computational overhead or ensure robust feature selection, leaving room for enhancement

3.2.1 Proposed Logistic Regression Conceptual Framework

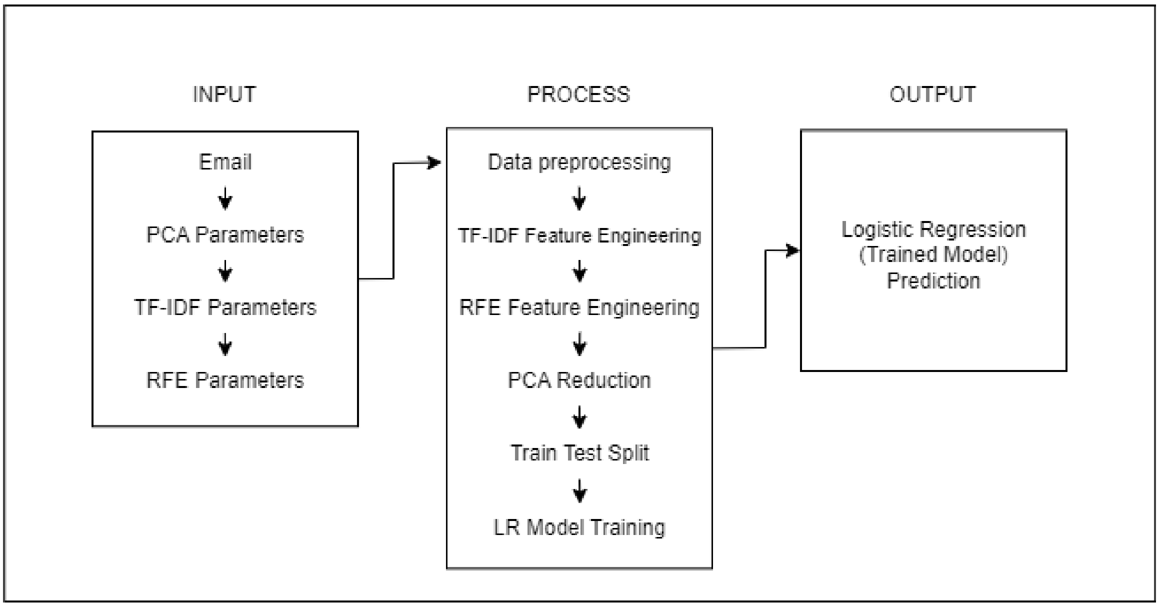


Figure 5 Conceptual Framework of the Proposed Algorithm

Figure 5 shows the proposed conceptual framework of the logistic regression algorithm, addressing the key limitations of its original framework. This proposed framework introduces a systematic approach to improve the effectiveness and efficiency of the logistic regression by incorporating advanced techniques at the stages of the process. These improvements target issues such as bias due to imbalance datasets, struggle handling large datasets, and overfitting issues, ensuring a more robust solution for email spam detection. The enhanced framework begins with the email labeled as spam or not spam, and is provided as the foundation for the dataset. This stage includes the configuration parameters for the enhancement techniques, such as settings for term frequency-inverse document frequency (TF-IDF), Recursive Feature Elimination (RFE), and Principal Component Analysis (PCA). These parameters allow flexibility in adapting the framework to different datasets and use cases.

In the process stage, involves the data preprocessing, where unwanted characters, punctuation, and stop words are removed, and the text is tokenized into individual words or phrases. Next, feature engineering is applied using TF-IDF to convert the cleaned text into numerical vectors, capturing the importance of words in the dataset. (*Wardana, N. S., Aditiawan, F. P., & Sari, A. P. 2024*). the Logistic Regression, integrating TF-IDF for feature extraction and FastText for feature expansion. Logistic Regression was selected due to its efficiency in binary sentiment classification and fast training capabilities. The approach is designed to address data imbalance and enhance classification performance. The results demonstrate that this method achieves optimal outcomes in terms of accuracy, recall, precision, and F-score, providing valuable insights for LinkedIn developers to improve service quality. RFE identifies and retains the most predictive features by eliminating less significant ones. *(Rasel Ahmed, Nafiz Fahad, et al. 2024).* The

Logistic Regression (LR) model, known for its simplicity and effectiveness in binary classification, is refined using Recursive Feature Elimination (RFE). This technique systematically removes less important features to enhance the model's performance. while PCA reduces the dimensionality of the selected features, ensuring computational efficiency and mitigating the risk of overfitting. *(Changsheng Zhu, Christian Uwa Idemudia, et al. 2019)* Through our experiment we have shown that an improved logistic regression model for predicting diabetes is possible through the integration of PCA.

Following feature preparation, the dataset is divided between training and testing subsets during the train-test split step. This ensures that the model is trained on one set of data and tested on another, resulting in an unbiased measure of its generalization capabilities. The model training step then optimizes the logistic regression model with the processed feature set, allowing it to understand the relationships between features and labels more efficiently. And, in the output stage the model produces predictions in the form of probabilities for each input message, representing the likelihood of being spam or not spam.

3.3 System Requirements

1. Local Machine

Processor: Intel Core i5-11400H (6 cores, 12 threads, up to 4.5 GHz).
Installed RAM: 8GB DDR4 RAM (expandable up to 64GB).
OS Edition: Windows 11
`**OS Version:** Version 21H2

2. Python

Python is essential in this study, supporting simulation, data visualization, and system development for email spam prediction. Python libraries include pandas for data handling, and scikit-learn modules such as LogisticRegression for classification and train_test_split for data partitioning. For feature extraction, we apply CountVectorizer and TfidfVectorizer to evaluate impacts on accuracy and overfitting. Additionally, RFE and PCA improve feature selection and reduce dimensionality, enhancing model performance. Visualization with matplotlib.pyplot and seaborn displays metrics like ROC curves and confusion matrices, and Gmail API integration enables real-time email fetching, supporting effective, interpretable predictions.

3. VSCode

Visual Studio Code (VSCode) has been used as the code editor for this study, Python data visualization, System development, and Gmail API integration within this research. Known for its flexibility and extensive library of extensions, VS Code facilitates efficient Python development through syntax highlighting, error checking, and debugging tools.

4. Kaggle

Kaggle has been an important tool in this study for gathering the data needed to train and test the spam prediction model. By using Kaggle's collection of datasets, we were able to find large sets of labeled emails, both spam and non-spam, that were essential for this study. These datasets, which are pre-processed and organized, helped us train the logistic regression model in a consistent way.

3.4 Methods and Tools

This section outlines the methods and tools used in which involves a comparative analysis of the existing algorithms with their limitations and the proposed algorithm to address the issues.

Problem 1 and Objective 1

The accuracy of logistic regression is affected by an imbalance in the data, favoring the more prevalent non-spam emails. This leads to a higher occurrence of mislabeled spam emails, compromising both security and the effectiveness of email systems.

The use of TF-IDF improves the situation by enhancing how features are represented in the model. It highlights key terms for determining spam while playing down less significant terms common to both spam and non-spam emails. This adjustment helps focus the model better and boosts its predictive precision. TF-IDF not only reduces the model's bias toward the majority class but also keeps up with evolving spam tactics, maintaining effective model performance. As a result, it greatly lowers the chances of spam emails being wrongly classified as non-spam, thus improving the security and efficiency of email filtering systems.

Problem 2 and Objective 2

The limitations of standard logistic regression models become apparent when they are applied to large datasets for tasks like spam detection. The main problem is the high computational demands of logistic regression, which cause inefficiencies and increased processing times as dataset sizes increase, rendering it less practical for analyzing large sets of data.

To improve logistic regression's scalability and efficiency with large datasets is through the implementation of Recursive Feature Elimination (RFE). RFE tackles these computational issues by progressively removing the least important features, thereby simplifying the data complexity. This reduction significantly decreases the computational burden on the logistic regression algorithm. Additionally, by concentrating on the most crucial features, RFE not only cuts down processing times but also maintains or even improves the model’s robustness and accuracy. This optimized method enables logistic regression to handle large-scale spam detection more efficiently and effectively, addressing its scalability and computational challenges.

Problem 3 and Objective 3

The challenge with using logistic regression is its vulnerability to overfitting. Overfitting is when a model performs well on training data but fails to generalize effectively to new or unseen data. In practice, an overfitted model performs well during training but poorly on real-world data, limiting its practical use.

To counteract overfitting in logistic regression models, implementing Principal Component Analysis (PCA) is an effective technique. PCA addresses overfitting by lowering the volume of the data, which in email datasets can be crowded with a vast array of words and phrases. This focus on the most relevant aspects of the data simplifies the logistic regression model, enhancing its ability to perform well with new data. This process of reducing dimensions gets rid of unneeded and distracting features and also makes the model work faster. Therefore, PCA helps the logistic regression algorithm keep only the most important variables, greatly improving its ability to detect spam by focusing on the truly important features.

CHAPTER FOUR
RESULTS AND DISCUSSIONS

This chapter presents the results of the techniques applied and discusses their implications in relation to the research objectives. The results are analyzed and interpreted, providing insights into the patterns, trends, and observations.

4.1 Term Frequency-Inverse Document Frequency (TF-IDF)

Implementing TF-IDF (term frequency-inverse document frequency) enhances the performance of logistic regression on imbalanced datasets by improving the feature representation of text. This allows the model to better identify important variations between classes. By emphasizing terms that are unique to the minority class and reducing the weight of common, less informative words, TF-IDF helps mitigate the bias toward the majority class.

Figure 6 Classification Report with TF-IDF Applied

	Precision	Recall	F1-Score	Support
Ham	0.99	0.98	0.99	958
Spam	0.90	0.95	0.92	157
Accuracy	-	-	0.98	1115
Macro Average	0.94	0.97	0.95	1115
Weighted Average	0.98	0.98	0.98	1115

The model's performance can be evaluated using several metrics. Precision indicates the accuracy of the model's predictions, with a precision of 0.99 for "ham" (99% of predicted "ham" emails were correctly identified) and 0.90 for "spam" (90% of predicted "spam" emails were correct). Recall reflects the model’s ability to identify all actual instances of each class, achieving a recall of 0.98 for "ham" (98% of actual "ham" emails were identified correctly) and 0.95 for "spam" (95% of actual "spam" emails were detected). The F1-score, a balance between precision and recall, is 0.99 for "ham" (indicating high accuracy for "ham") and 0.92 for "spam" (showing a slightly lower, but still strong, performance for "spam"). Support indicates the actual number of instances in the dataset, with 1825 emails labeled as "ham" and 223 as "spam." Overall, the model achieved an accuracy of 98%, demonstrating strong overall performance in classifying the emails.

True Negatives	False Positives	False Negatives	True Positives
1758	40	11	212

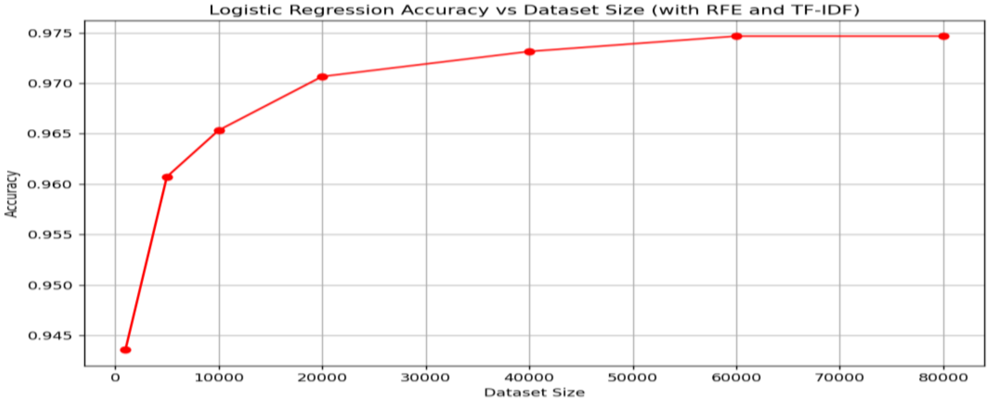
Figure 7 Confusion Matrix with TF-IDF Applied

The algorithm’s performance can also be analyzed through its confusion matrix components. It correctly identified 1785 emails as "ham" (not spam), which are the True Negatives (TN). However, there were 40 False Positives (FP), where emails were incorrectly classified as "spam" when they were actually "ham." Additionally, 11 False Negatives (FN) occurred, where "spam" emails were misclassified as "ham." On the positive side, the algorithm correctly identified 212 emails as "spam," representing the True Positives (TP).

4.2 Recursive Feature Elimination (RFE)

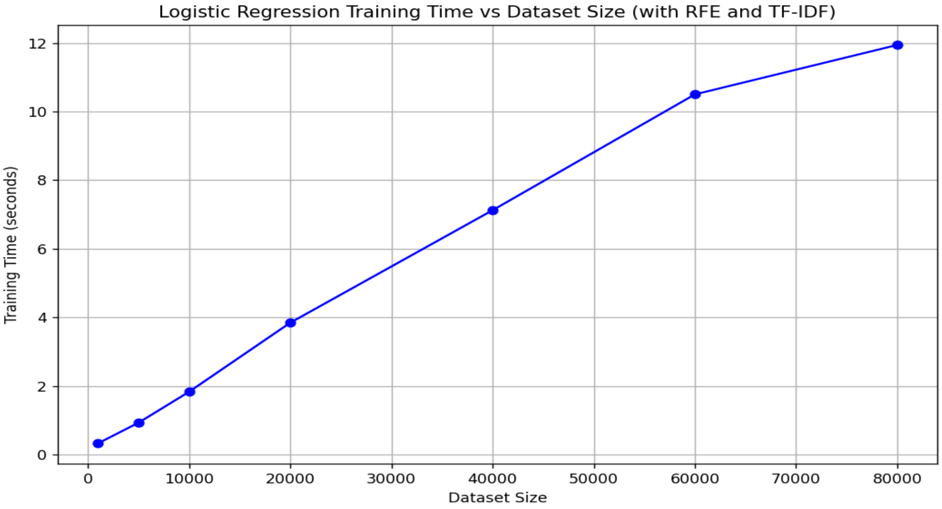
Using Recursive Feature Elimination (RFE) can greatly improve the efficiency of logistic regression, especially when working with large datasets. RFE works by progressively eliminating less relevant features, which reduces the complexity of the data and eases the computational burden on the logistic regression model. By retaining only the most significant features, RFE helps maintain the strength and effectiveness of the model, enhancing its capacity to process and analyze large datasets, such as those found in complex spam detection tasks.

Figure 8 Training Time vs. Dataset Size with RFE and TF-IDF



In this figure, as the dataset size increases, the training time of the algorithm also grows, although it remains relatively efficient. With a dataset of 1000 samples, the algorithm takes just 0.34 seconds to train. As the size expands to 5000 samples, the training time increases to 0.94 seconds, still under a second. With 10,000 samples, the training time rises to 2.18 seconds, and at 20,000 samples, it reaches 3.73 seconds. As the dataset size continues to grow, the training time increases more significantly—6.11 seconds for 40,000 samples and 8.97 seconds for 60,000 samples. Finally, with 80,000 samples, the training time peaks at 11.89 seconds. While the training time does grow with the dataset, the increases are gradual, with the largest jumps occurring between 40,000 and 60,000 samples, showing that the algorithm can still handle substantial datasets with relatively moderate training times.

Figure 9 Accuracy vs. Dataset Size with RFE and TF-IDF

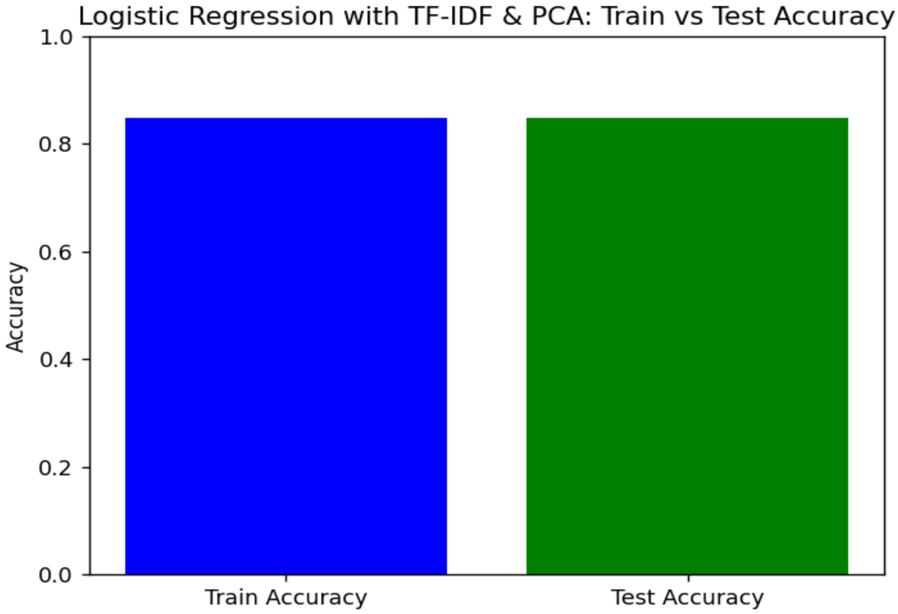


In this figure, as the dataset size increases, the algorithm's accuracy improves, though the gains become smaller over time. With a dataset of 1000 samples, the algorithm achieves an accuracy of 94.36%, which is good for a small dataset. As the dataset grows to 5000 samples, accuracy increases to 96.07%, and further improvement is seen with 10,000 samples, reaching 96.86%. With 20,000 samples, accuracy rises to 97.19%, but the improvements start to diminish. At 40,000 samples, accuracy reaches 97.47%, and this gain remains consistent even with 60,000 and 80,000 samples, indicating that the algorithm has likely reached its peak performance. Adding more data beyond this point does not result in significant accuracy improvements.

4.3 Principal Component Analysis (PCA)

Principal Component Analysis (PCA) can enhance the performance of logistic regression models in spam email detection by addressing overfitting problems. PCA reduces the dimensionality of the feature space, which is typically large in email datasets due to the variety of words and phrases, by emphasizing the most relevant data points. This reduction not only streamlines the logistic regression process but also boosts its computational efficiency. By eliminating unnecessary and noisy features, PCA ensures that only the most crucial variables for spam detection are retained.

Figure 10 Training Data Accuracy vs. Test Data Accuracy with TF-IDF and PCA



The algorithm demonstrates strong performance, with a train accuracy of 84.8%, indicating that it has learned the key patterns in the training data without overfitting. While not a perfect score, the model's ability to generalize to new data is evident, as shown by the test accuracy of 84.7%. The minimal drop in accuracy between the training and test sets suggests that the model generalizes well and avoids overfitting. This balance between training and test performance indicates that the model is robust and reliable, capable of effectively handling unseen, real-world data.

4.4 Overall Performance Metrics

Techniques	Accuracy (%)	Precision (Macro)	Recall (Macro)	F1-score (Macro)	Training Time(s)
TF-IDF	98.00	94.00	97.00	95.00	-
RFE with TF-IDF	97.47	-	-	-	11.89
PCA with TF-IDF	84.70	-	-	-	-

Figure 11 Overall Performance Metrics

This figure presents the results of techniques applied to enhance logistic regression for the efficient identification of spam messages. The TF-IDF approach demonstrates superior performance, achieving an accuracy of 98%, with macro-averaged precision of 94%, recall of 97%, and an F1-score of 95%. These metrics highlight its effectiveness in feature representation, particularly in addressing the challenges posed by imbalanced datasets with unequal distributions of spam and non-spam messages. However, it is noteworthy that the training time for the TF-IDF approach was not reported.

The RFE + TF-IDF method achieves a slightly lower accuracy of 97% while maintaining exceptional computational efficiency, with a training time of under 12 seconds for large datasets. Despite the lack of explicitly calculated precision and recall metrics for this method, it likely retains strong performance comparable to TF-IDF due to its focus on feature selection.

The PCA + TF-IDF method prioritizes dimensionality reduction, achieving a test accuracy of approximately 85%. While this suggests adaptability and computational efficiency, its effectiveness is notably lower than the other approaches. Metrics such as precision, recall, and F1-scores were not disclosed for PCA + TF-IDF, which may reflect its focus on balancing computational efficiency with classification performance.

CHAPTER FIVE

CONCLUSIONS AND RECOMMENDATIONS

Conclusions

The TF-IDF approach has demonstrated outstanding performance, achieving a high accuracy of 98%, along with strong macro-averaged precision (94%), recall (97%), and F1-score (95%). These results confirm that TF-IDF is highly effective for text feature representation, especially in handling imbalanced datasets. Its ability to emphasize class-distinctive terms while reducing noise makes it a reliable method for spam detection tasks.

The combination of RFE and TF-IDF achieved a slightly lower accuracy of 97.47% but demonstrated exceptional computational efficiency, with a training time of just under 12 seconds for large datasets. This indicates that RFE is effective in reducing the dataset’s complexity by retaining only the most significant features, making it suitable for scalable applications.

The PCA + TF-IDF approach achieved a lower test accuracy of 84.7%, reflecting its focus on reducing dimensionality rather than maximizing classification accuracy. By eliminating less significant features, PCA helps address overfitting and enhances computational efficiency, but it comes at the cost of reduced classification performance compared to TF-IDF and RFE + TF-IDF.

While TF-IDF excels inj accuracy and feature representation, RFE + TF-IDF offers a good balance of accuracy and computational efficiency. PCA + TF-IDF, while less accurate, is beneficial for applications requiring dimensionality reduction and reduced computational overhead.

Recommendations

Further research is advised to explore the operation complexities of the enhancements of Logistic Regression, focusing on specific modifications and its contributions to the overall performance, particularly handling imbalanced datasets, can provide deeper insights into their practical effectiveness.

It is also recommended to apply these methods to be tested on a wider range of datasets, including those from other classification methods. This would assess the versatility and scalability of the approaches, particularly the trade-offs between high-accuracy methods like TF-IDF and more computationally efficient techniques like RFE + TF-IDF.

Further research into the balance between computational efficiency and classification accuracy across various computational environments is crucial. For instance, examining how RFE + TF-IDF performs in resource-constrained environments compared to TF-IDF can help refine method selection based on application-specific needs.

Lastly, consider enhancing the interpretability of model predictions, methods like SHAP or LIME can be explored. This would ensure transparency in decision-making, especially for non-technical users, fostering trust and facilitating adoption in critical applications such as spam detection or other classification tasks.

REFERENCES

1. Zhang, L., Geisler, T., Ray, H., Xie, Y. (2021). Improving logistic regression on the balanced data by a novel penalized log-likelihood function. National Library of Medicine. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9542776/
2. Steyerberg, E., Schemper, M., Harrell, F., (2011). Logistic regression modeling and the number of events per variable: selection bias dominates. Journal of Clinical Epidemiology https://www.jclinepi.com/article/S0895-4356(11)00219-8/fulltext#articleInformation
3. Doshi, J., Parmar, K., Sanghavi, R., Shekokar, N. (2023). A comprehensive dual-layer architecture for phishing and spam email detection. ScienceDirect. https://www.sciencedirect.com/science/article/abs/pii/S0167404823002882
4. Khanday, S., Parveen, S. (2021). Logistic Regression Based Classification of Spam and Non-Spam Emails. ResearchGate. Logistic Regression Based Classification of Spam and Non-Spam Emails
5. Awan, A. (2023) What is Overfitting?. Datacamp. https://www.datacamp.com/blog/what-is-overfitting
6. Olaoye G., John, R., Luz, A. (2024) Application of logistic regression for cyber threat detection Application of logistic regression for cyber threat detection
7. Dhamodharavadhani, S., Ramalingam, S. (2019). Enhanced-Logistic-Regression-(ELR)-Model-for-Big-Data. ResearchGate. Enhanced-Logistic-Regression-(ELR)-Model-for-Big-Data - ResearchGate
8. Dedeturk, B., Akay, B. (2020). Spam filtering using a logistic regression model trained by an artificial bee colony algorithm. ScienceDirect. Spam filtering using a logistic regression model ... - ScienceDirect
9. Pan, P., Lee, C., Hsu, N., Sun, T. (2024). Combining principal component analysis and logistic regression for multifactorial fall risk prediction among community-dwelling older adults.

https://www.sciencedirect.com/science/article/abs/pii/S0197457224000843

10. Farag, S., El-saeiti, I. (2023). Effect and Influence of Class Imbalance and Multicollinearity in Binary Logistic Regression (A Comparative Simulation Study). Effect and Influence of Class Imbalance and Multicollinearity in Binary ...
11. Kuiziniene, D. and Krilavicius, T. (2024). Balancing Technique for Advanced Financial Distress Detection using Artificial Intelligence. ResearchGate. Balancing Techniques for Advanced Financial Distress Detection Using Artificial Intelligence
12. Brightwood, S. and Anthomy, L. (2024). Practical Considerations and Challenges in Applying Logistic Regression for Cyber Threat Detection. ResearchGate. Practical Considerations and Challenges in Applying Logistic Regression for Cyber Threat Detection
13. Brightwood, S. (2024). Regularization Techniques in Logistic Regression. ResearchGate. Regularization Techniques in Logistic Regression - ResearchGate
14. Acito, F. (2023). Logistic Regression. ResearchGate. https://www.researchgate.net/publication/376067939_Logistic_Regression
15. Balboa, A., et al (2024). Logistic Regression vs Machine Learning to predict evacuation decisions in fire alarm situations. ScienceDirect. https://www.sciencedirect.com/science/article/pii/S0925753524000754
16. Wahyuningsih, T., et al (2024). Comparison of Effectiveness of Logistic Regression, Naive Bayes, and Random Forest Algorithms in Predicting Student Arguments. ScienceDirect. https://www.sciencedirect.com/science/article/pii/S1877050924003715
17. Qu, W., et al (2024). Web Attack Detection Based on Honeypots and Logistic Regression Algorithm. ResearchGate. Web Attack Detection Based on Honeypots and Logistic Regression Algorithm
18. Eldridge, S. (2023, November 14). logistic regression. Encyclopedia Britannica. https://www.britannica.com/science/logistic-regression
19. Thanda, A. (2023, May 11). What is Logistic Regression? Beginners Guid https://careerfoundry.com/en/blog/data-analytics/what-is-logistic-regression/
20. Guo, X., et al. (2008). On the Class Imbalance Problem. IEEExplore. https://ieeexplore.ieee.org/abstract/document/4667275
21. Rahman, M., and Davis, D. (2013). Addressing the Class Imbalance Problem in Medical Datasets. International Journal of Machine Learning and Computing. Addressing the Class Imbalance Problem in Medical Datasets - ResearchGate
22. Buda, M., et al. (2018). A systematic study of the class imbalance problem in convolutional neural networks. ScienceDirect. https://www.sciencedirect.com/science/article/abs/pii/S0893608018302107
23. Castro, A. et al (2024). Large language models overcome the challenges of unstructured text data in ecology. ScienceDirect. https://www.sciencedirect.com/science/article/pii/S157495412400284X
24. Cheng, Q. et al (2020). Information-based Optimal Subdata Selection for Big Data Logistic Regression. ScienceDirect. https://www.sciencedirect.com/science/article/pii/S0378375820300331
25. Fan, J.(2014). Challenges of Big Data analysis. National Science Review. https://academic.oup.com/nsr/article/1/2/293/1397586
26. Santos, C., et al (2022). Avoiding Overfitting: A Survey on Regularization Methods for Convolutional Neural Networks. Association for Computing Machinery. https://dl.acm.org/doi/full/10.1145/3510413

27. Salam, M., et al (2021). The Effect of Different Dimensionality Random Techniques on Machine Learning Overfitting Problem. ResearchGate. The Effect of Different Dimensionality Reduction Techniques on Machine ...
28. Wahyuningsih, T. (2024). Comparison of Effectiveness of Logistic Regression, Naive Bayes, and Random Forest Algorithms in Predicting Student Arguments. ScienceDirect. https://www.sciencedirect.com/science/article/pii/S1877050924003715
29. Khanna, D. (2015). Comparative Study of Classification Techniques (SVM, Logistic Regression and Neural Networks) to Predict the Prevalence of Heart Disease. ResearchGate Comparative Study of Classification Techniques (SVM, Logistic Regression and Neural Networks) to Predict the Prevalence of Heart Disease
30. Wardana, N. S., Aditiawan, F. P., & Sari, A. P. (2024). Logistic Regression Classification with TF-IDF and FastText for Sentiment Analysis of LinkedIn Reviews. https://journal-laaroiba.com/ojs/index.php/visa/article/view/2835
31. Rasel Ahmed, Nafiz Fahad, et al (2024) A novel integrated logistic regression model enhanced with recursive feature elimination and explainable artificial intelligence for dementia prediction, Healthcare Analytics. https://www.sciencedirect.com/science/article/pii/S2772442524000649
32. Changsheng Zhu, Christian Uwa Idemudia, Wenfang Feng, (2019) Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques, Informatics in Medicine Unlocked. https://www.sciencedirect.com/science/article/pii/S2352914819300139