An Investigative Analysis of Two Distinct Machine Learning Techniques Across Three Dataset Using R .

ADEOLA DEBORAH ADENIJI
MASTERS IN DATA ANALYTICS
NATIONAL COLLEGE OF IRELAND
x23104201@student.ncirl.ie.

Abstract*— Machine learning techniques can demonstrate significant performance variation across different datasets as it is crucial for informed decisions making processes. This research investigates to compare the relationship between two distinct machine learning techniques, in the application of three datasets with the use of R programming language. To explore how the datasets influence the performance of disparities observed between the different implementations. In the cause of this study, which aims to address predictive analysis of classification problems, three datasets would be employed. This can be achieved by gaining deeper insights into their characteristics and complexities. With the use of performance metrics such as Accuracy, Recall, and F1- Score which would help determine their disparities and performance to know the best fit model. Machine learning techniques used during the investigation involve Random Forest using "randomForest and ranger libraries" and Support Vector Machine using "e1017 and Kernlab libraries". While the dataset of choice includes Adult Income focusing on predicting income exceeding $50,000 per year-based on census data; the Web Page Phishing dataset which aims at detecting phishing websites through URL characteristics; and the Statlog Shuttle dataset utilized for prediction operational states of space shuttle engines based on sensor readings. Using rigorous experimentation and analysis, this research highlights the strengths and weaknesses of both techniques in various classification tasks.

Keywords—Machine Learning, R Programming Language, Predictive analysis, Classification, Random Forest, Support Vector Machine (SVM), randomForest Library, Ranger Library, e1017 Library, Kernlab Library, Adult Income Dataset, Web Page Phishing Dataset, Statlog Shuttle dataset, Accuracy, Recall, F1- Score.

I. INTRODUCTION.

In the field of machine learning, the ability to learn and improve data is very important [1]. Unlike for humans who rely on their experiences, machines rely on vest amount of datasets .It is excellent at identifying patterns, enabling it to make predictions to unforeseen possibilities. This study explores two distinct machine learning techniques namely Random Forest and Support Vector Machine. This techniques analyzes the impact performance across three datasets with the use of R programming Language. The objective of this paper through predictive analysis in solving classification problem is to, assess and compare the strength and weaknesses of the above mentioned machine learning techniques, and explore how datasets characteristic influences the performance disparities between different implementation of machine learning techniques with the use of R Language.

II. LITERATURE REVIEW

A past study that aligns with the goal of this project wrote about how datasets characteristics can affect model performances. In the [2] research, it explores the effectiveness of decision trees and random forest which are machine learning techniques used for classification tasks. It investigated how said algorithm performed on a chosen type of dataset with the use of R language. The study compares the strengths and weaknesses of decision trees and random

forests when dealing with data types and classification challenges. The research also tries to depict whether Decision tree excels with dataset or if random forests were more resilient to data. Considering its limitations, the study focused on a set of datasets potentially limiting how broadly its findings can be applied. Moreover, the evaluation metric chosen such as only accuracy, did not fully capture all aspects of the model's performance that are relevant to the investigation.

Lots of research has shown that several studies have explored the use of Support Vector Machines (SVMs) for various classification analysis in R. Another paper [3] provides a proper guide on the implementation of support vector machine using the "e1071" package in R language. It had an edge in efficiently handling high-dimensional data and ability to perform only linear classifications using Kernel functions. However, it suggests that the use of the machine learning technique can take lots of computation time as well as being expensive to handle large datasets. The research also pointed out that support vector machines can be less interpretable compared to other classification methods, making it difficult to understand their reason behind their predictions.

A review by [4] that investigated Machine for Random Forest introduced a concept of Data Removal- Enabled (DaRE) Forest. This approach is a different method of removing data points from random forest models, as not easily achieved with traditional retraining methods. The method offers a significant advantage in terms of speed by selecting retraining subtrees, impacted by data removal, making it a suitable method for large datasets. However, it had some limitations. Retraining efficiency and predictions accuracy requires careful consideration. Another limitation was that the method might not be influenced by some specific characteristics of data and the chosen removal. While it promised feature research, it also needs to explore other methods of modelling across different domains.

Another report paper by [5] focuses on applying Knowledge Discovery in Database process on two machine learning techniques. The study compared Random Forest and Support Vector Machine in predicting radiation protection function and toxicity for radioprotection. According to the investigation, the support vector machine demonstrated better performance in predicting the radiation protection function, while random forest outperformed SVM in predicting toxicity. However, its limitations involved the difficulty in finding suitable datasets for machine learning on raw data for refining features selection to get a satisfactory outcome.

III. About Datasets

A. Adult Income

The Adult Income dataset is also known as the "Census Income". It is a collection of census information used for predicting income. Based on the information about the census data, it centers on whether an individual's income is less or exceeds $50,000 on a yearly basis. The datasets presents lots of opportunities as it includes features ranging from demographic information to employment details. It allows room to explore how machine learning models handles classification tasks, involving continuous and categorical features commonly found in the real-world. The aim of using adult income data is to predict weather an individual income exceeds $50,000 per year.

B. Web Page Phishing

The second dataset used during the investigation is the Web Page Phishing dataset. This dataset serves as valuable resources particularly in the detection of phishing websites in the realm of cybersecurity. It presents a unique challenge when solving classification problems. Phishing attacks poses a serious risk to individuals around the world [6]. This is a type of scam that is designed to trick users in sharing important information by downloading malicious software. Thus, the main aim in the use of this dataset is to train machine learning models that can detect and differentiate between phishing websites from the real ones based on their URL features.

C. Statlog (Shuttle)

Lastly the third dataset that would be analyzed is the Statlog (Shuttle) dataset. This is an essential dataset that provides real-world example based on sensor readings. It is an instrument in aerospace engineering that allows the predictions of operational states of space shuttle engines through those readings. The sensor readings provides

information which aids in ensuring the accuracy and safety missions by precisely analyzing engine conditions and spotting anomalies. This research aims to enhance its predictive capabilities , to enable a more proactive maintenance and risk mitigation strategies in the aerospace.

By exploring this three datasets, this research aims to utilize the power of Random forest and Support vector machine learning techniques to address critical challenges in a range of fields such as socioeconomic prediction, cybersecurity and aerospace engineering.

IV. About Techniques And Libaraies

A. Programming Language

R programming language is a software programming tool specifically designed for statistical computation and graphics. [7]. For this research R language would be considered because it excels in data exploration, visualization, and statistical modelling making it a good choice for this machine learning research.

B. Machine learning Techniques

Two machine learning techniques are used in this research "**Random Forest**" and "**Support Vector Machine"**. Random forest has proven to be a useful technique for handling complex datasets and to reduce bias during the investigation. It works by [8] combining multiple decision trees which made it flexible and resistant to overfitting. While the second technique, Support Vector Machine (SVM) is also another powerful technique been used. It proved useful in handling the classification tasks as a supervised learning algorithm, by finding an optimal separation line (hyperplane) [9] to classify data points.

C. R Libraries for Machine Learning Techniques.

To achieve the goal of this project, the use of four R language libraries were selected. The above-mentioned techniques made use of these libraries to complect each modelling task. First Random Forest technique used the "**randomForest**" and "**Ranger**" libraries, while Support Vector Machine made used of "**e1071**" and "**Kernlab**" libraries. The randomForest is a well-established R package for implementing random forest algorithms as it a user-friendly interface when building a model. [10] It provides the ability to specify parameters like number of trees and features selection methods as well as accommodating flexible approach to handling complex datasets when using Random Forest. Another R package used for random forest is the Ranger. It was excellent when modelling because of its speed and efficiency, especially while dealing with large datasets. Moving on to the third library which supports Support Vector Machine is the e1071. It provides different kernel [11] functions such as linear or radial for the SVM classification, paving way to explore the different ways to separate data points. The last library is Kernlab library. While the e1071 offers basic Support Vector Machine learning functionality, the Kernlab [12] provides a more extensive toolkit. All these libraries together helped in the comparison analysis of both techniques with the use of R language.

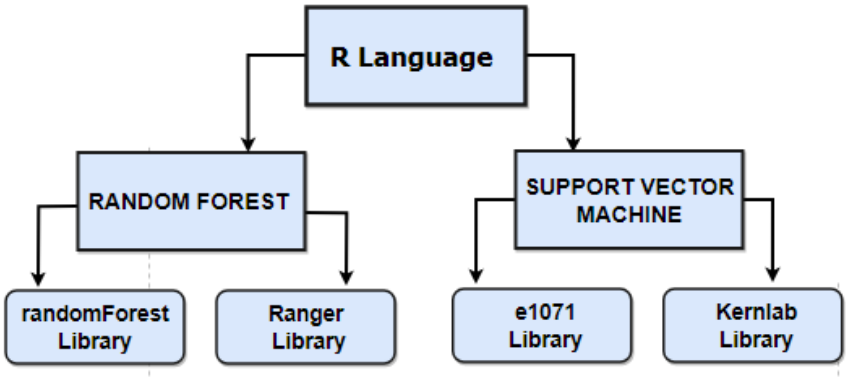

Figure 1: Overview of Programming Language and Machine Learning Techniques.

V. METHODOLOGY

This project will implement a Knowledge Discovery in Dataset (KDD) methodology as a structured framework to leverage the chosen datasets and compare how datasets characteristics can influence model implementation and performance. The KDD provides a systematic approach [13] that will be a provisional guide through various stages of the investigative analysis from data selection and preparation of each dataset to model building and final evaluation. By applying the KDD process, certain conditions must be accomplished. Below provides a general overview of what would be achieved in using this methodology.

A. Understanding dataset characteristics.

Data understanding is a fundamental aspect in this analysis as it involves the process of data selection. In this stage, a thorough investigation on Adult Income dataset, Web Page Phishing dataset, and Statlog Shuttle dataset are carried out to understand their features, data types and potential challenges.

B. Data Pre-Processing and Transfromation.

This stage is also a very important and essential stage as it involves preparing the chosen datasets to be ready and fit for analysis. Various tasks are involved as each datasets must undergo seven main characteristics, to ensure the data are suitable for carrying out machine learning algorithm.

C. Model Selection and Training.

This stage encompasses models' selection (Random Forest and Support Vector Machine), libraries selection (randomForest, Ranger, e1017, and Kernlab) as well as trained data gotten from the datasets for data mining.

D. Model Evaluation and Comparison

This stage is the final phase of the analysis as it involves the application of various evaluation metrics including **Accuracy**, **Recall** and **F1- Score** to assess the performance of the different models of each datasets. This stage helps to provide insights on how each dataset characteristics influences model effectiveness for classification task with the use of R language. It is also noteworthy to note how these performance metric would be calculated. Accuracy is calculated as the proportion of correctly classified instance as it makes use of the ***"accuracy ((TP + TN)/Total instances)"*** function in R. Recall, which is also known as the true positive because it calculates the proportion of actual positive cases that were predicted correctly, It uses a function **(*recall*(*TP / FN*))** in R**.** And the final evaluation metric is F1 - Score which is calculated based on the harmonic mean of Precision and Recall by balancing both measures. In R, it is calculated by ***(F1 (2 *(Precision*Recall))/ (Precision + Recall)).*** In using these metric R can compare the performance of different models on the three datasets to gain insights.

The Knowledge Discovery in Dataset presents a systematic approach as well as a comprehensive framework to extract knowledge for the chosen datasets. This knowledge would uncover valuable insights from the evaluation of the machine learning performance and how it influenced the datasets. At this stage, the goal is to ascertain the best performance model and it comparison disparities with the use of R.

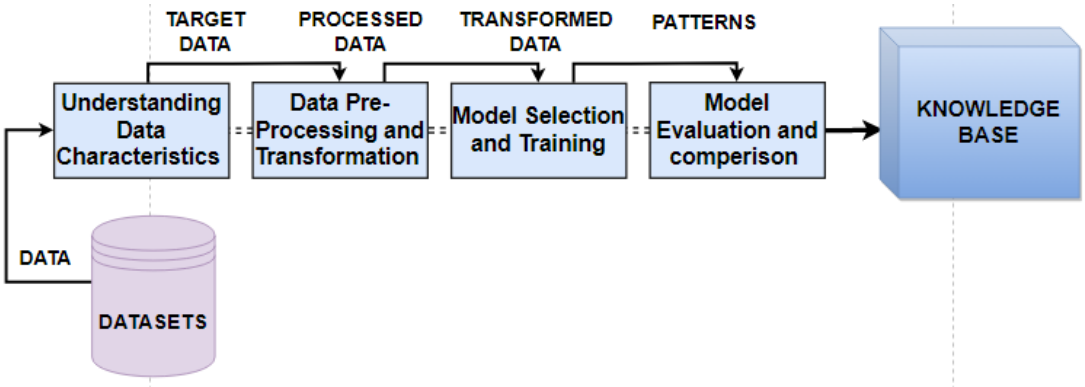

Figure 2: Overview Structure of the Knowledge Discovery in Dataset (KDD) Used in this Study.

In this section of the paper, a proper analysis is done on the three datasets using the above-mentioned Knowledge Discovery in Dataset methodology. Identification of dataset characteristics which may impact the performance are first analyzed, before the comparison of identical name techniques.

A. ADULT INCOME.

1. **Understanding data characteristics:** The adult income dataset was collected in two separate comma separated files (CSV). The first file of the data comprised 15 attributes and 32,562 instances. While the second file comprised 15 attributes and 16,282 instances. Data integration took place because both files are identical and have the same attributes. This was done for easier analysis to be carried out.

    i. **Number of independent and dependent variables:** The dataset consists of 14 independent variables, representing factors that can influence the outcome, with 1 dependent variable which indicates a single outcome or target variable (income) to be predicted.

    ii. **Number of records:** After integrating the two csv files, to one, the dataset then had a total of 48,844 records or instances to be observed.

    iii. **Data types of Combination:** In the datasets the observation combination includes:

Variable	Type	Description
Age	Integer (int)	Represents an individual's age.
Education_num	Integer (int)	Number of years of education completed
Workclass:	Character (chr)	Types of employment such as state govt. or private
Education	Character (chr)	Specifies education level like Bachelors or HS-grad
Marital_status	Character (chr)	Individual marital status such as "Never- married"
Occupation	Character (chr)	Job occupation category such as Adm-clerical
Relationship	Character (chr)	Relationship of household such as Not in family
Race	Character (chr)	Racial groups such as White or Black
Sex	Character (chr)	Genders of individuals such as Male or Female
Native_country	Character (chr)	Identifies country of origin such as United states
Fnlwgt	Numeri (int)	Weight given to each record for adjusting samples
Capital_gain	Numeri (int)	Cpital gain amount in previous year
Capital_loss	Numeri (int)	Capital loss amount in previous year
Hours_per_week	Numeri (int)	Specifies the number of hours worked per week
Income (Target)	Character (chr)	Indicates whether individuals' income exceeds $50,000 per week ("<= 50K" OR ">50K").

Table1: Overview of Data Type of Combinations in Adult Income dataset.

    iv. **Summary Statistics of Variables:** A descriptive statistic was conducted only on the numerical variables. This shows a summarized analysis of the adult income dataset. It exhibited a range of values with minimum values ranging from 1 to17, the maximum values ranging from 90 to 99,999. Age, having a median of about 38.64, fnlwgt having 178,145, 10 for education number, 0 for both gain and loss, 40.42 for hours per week. It also displayed NAs on each statistical summary, showing presence of missing values in each numerical data.

2. Data Pre-Processing and Transformation:

i. **Duplicate values:** An investigation was carried out to determine if the dataset consisted of duplicate values. The use of "*sum(duplicate(adult))"* function made this possible. A total of 59 duplicate instances were discovered, having a proportion of 0.1208% approximately. These duplicates were handled by removing them completely from the datasets. The removal process was used because the proportion of the duplicated are small, therefore removing them would not affect the analysis.

ii. **Missing Values:** This is a very important step when preparing the dataset for modelling. Once data is not properly cleaned, it would be inadequate for any type of analysis. A check for missing value using the "*sum(colSum(is.na(adult))>0)*" function was used. A total of 6 missing values with a proportion of 0.00082% was discovered in the data. Again, this was removed from the dataset as the proportions were too small to impact on the analysis. After much analysis more Na's valve was discovered having a proportion of over 9.78% (2,791 NA values) as was also removed.

iii. **Outliers** Through the examination depiction in Figure3.**,** a substantial proportion of outliers was identified. A total of 2,326 instances, approximately 7.96% of the dataset were observed to exhibit outlier behavior. The use of interquartile Range (IQR) was the method that was used to detect and eventually remove the anomalies from the dataset.

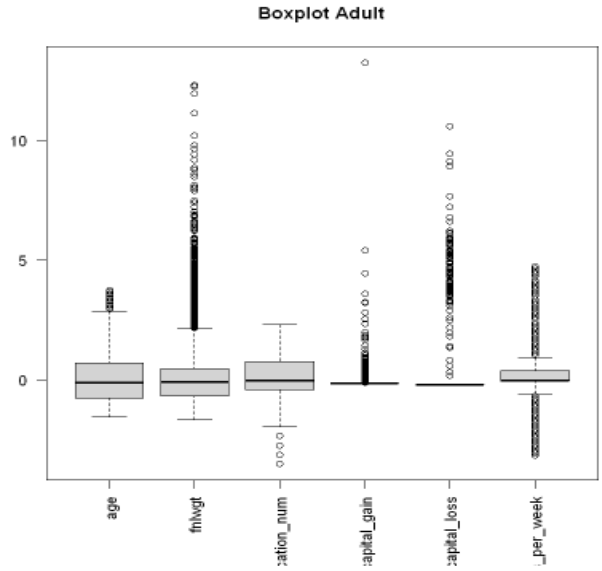

Figure 3: Outliers boxplot for adult income dataset.

iv. **Normalization and Label Encoding:** All the categorical variables were encoded to numerical values so that computers could understand them and prepare the data for modeling. For instance, the income variable was converted to 0's and 1's. Income which exceeded $50K was equated to 1 and income less than $50k was equated to 0. Furthermore, a standard scale was used to normalize all numerical variables. By ensuring that features are fairly compared, this normalization keeps variables with greater scales from controlling the model's training process.

After the preprocessing steps were over, the dataset now consists of 15 attributes and 26,899 instances. It is noteworthy that while categorical variables like workclass, education, marital_status, occupation, relationship, race, sex, and native_country have been converted into factor levels for

analysis, numerical variables like age, fnlwgt, education_num, capital_gain, capital_loss, and hours_per_week had also been normalized.

3. **Model Selection and Training:**

   i. **Data Balancing:** On this phase of the analysis, an investigation was carried out to know the number of records in the target variable that would be used for predictions. Upon inspection, the dataset was observed to be balanced with 14,033 records for class 0 (income <=$50K) and 12,866 records for class 1 (income >$50K).

   ii. **Data Splitting:** The data was then split into two sets; the Training and Testing. Having a ratio of 70% for the training and 30% for the testing. For the training set there was a total of 9,838 records in class 0, and 8,991 records for class 1, while for the testing set consisted of 4,195 records for class 0 and 3,875 records for class 1.

4. **Model Evaluation and Comparison**

Building upon the model selection and training, this next phase explores the performance of the two machine learning techniques using the four libraries in R. Below is a table that showcases the performance metrics used for the analysis. Based on the metrics, the Kernlab library for the support vector machine used in predicting adult income appears to be the best performing model among the four libraries. It achieved the highest level of accuracy, recall and F1- score. However, further analysis considering specific empirical designs and evaluation assessment may reveal additional insights into each performance implementation.

Library	Accuracy %	Recall %	F1- score %
randomForest	62.29	69.61	62.29
Ranger	63.11	72.54	62.51
e1071	62.12	65.67	63.01
Kernlab	64.06	76.90	62.56

Table 2: Overview of Model Evaluation in Adult income dataset.

B. WEB PAGE PHISHING DATASET.

1. **Understanding data characteristics:** The web page phishing dataset is the second dataset that would be explored in the section. A proper understating on how the data characteristics will influence the investigation is ascertained.

   i. **Number of independent and dependent variables:** This dataset consists of 19 independent variables, representing factors that can influence the data outcome, with 1 dependent variable (phishing) which indicates a single target for making predictions.

   ii. **Number of Records:** This dataset consisted of 20 attributes in total and 100077 instances representing the different characteristics of each data point in the phishing analysis.

   iii. **Data types of Combination:** Web phishing data is a unique data set as it only consists of numerical datatypes. The observation includes:

Variable	Type	Description
URL_length	Integer (int)	Length of the URL in characters
n_dots	Integer (int)	Count of dots (.) in the URL
n_hyphens	Integer (int)	Count of hyphens (-) in the URL

n_underline	Integer (int)	Count of underscores (_) in the URL
n_slash	Integer (int)	Count of slashes (/) in the URL
n_questionmark	Integer (int)	Count of question marks (?) in the URL
n_equal	Integer (int)	Count of equal signs (=) in the URL
n_at	Integer (int)	Count of at signs (@) in the URL
n_and	Integer (int)	Count of ampersands (&) in the URL
n_exclamation	Integer (int)	Count of exclamation marks (!) in the URL
n_space	Integer (int)	Count of spaces in the URL
n_tilde	Integer (int)	Count of tiles (~) in the URL
n_comma	Integer (int)	Count of commas (,) in the URL
n_plus	Integer (int)	Count of plus signs (+) in the URL
n_asterisk	Integer (int)	Count of asterisks (*) in the URL
n_hastag	Integer (int)	Count of hashtag symbols (#) in the URL
n_dollar	Integer (int)	Count of dollar signs ($) in the URL
n_percent	Integer (int)	Count of percent signs (%) in the URL
n_redirection	Integer (int)	Number of redirects in the URL before reaching the destination
phishing (Target)	Integer (int)	Binary indicator of phishing behavior: 0 for legitimate website, 1 for phishing website

Table 3: Overview of Data Type of Combinations in Web Page Phishing dataset.

iv. **Summary Statistics of variables:** Statistical summary was also carried out on this data. Its showed the different ranges of descriptive statistical analysis from minimum to the maximum value, then interquartile ranges, as well as the media and mode. It indicated that the highest mean value was approximately 39.18 character of URLs.

2. Data Pre-Processing and Transformation:

i. Series of investigations were made on the web phishing data such as checking for missing values, as well as duplicate values. It appeared that the dataset was devoid of such ensuring integrity and cleanliness for further analysis.

ii. **Outliers:** Through another examination depiction in Figure4., a sizable proportion of outliers was identified. A total of 10,1099 instances, approximately 5.05% of the dataset were observed to exhibit outlier behavior. The use of interquartile Range (IQR) was the method that was used to detect and eventually remove the anomalies from the phishing dataset.

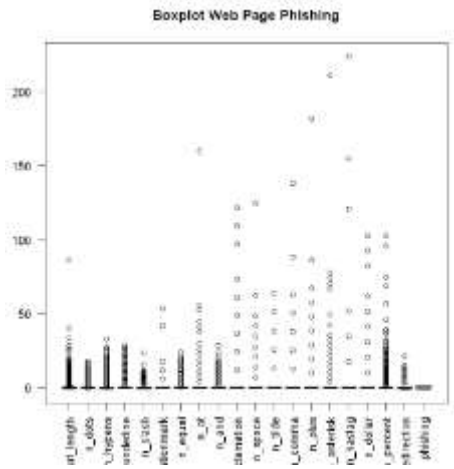

Figure 4: Boxplot for Web Page Phishing dataset.

iii. **Feature Engineering:** Feature engineering was conducted on the web page phishing dataset to remove irrelevant features that would not significantly impact the analysis. First, an investigation was made to find out the number and proportion of irrelevant variables that were present. A total of 9 attributes, approximately 0.45% of the data were identified by calculating the correlation of independent variables with the target variable (phishing). By setting a threshold of 0.05, features with higher correlations were retained. This approach ensured that only relevant features were considered for subsequent modeling, and predictive performance.

iv. **Label Encoding and Normalization:** This is a similar step that is applied across all three data sets in this project. As stated previously the benefits of label encoding and normalization, the target variable of this dataset was encoded, where the phishing class became levels of 0 and 1 formart. Additionally, all variables were then scaled to a normalized range for better analysis processes.

After the preprocessing steps were over, the dataset now consists of 12 attributes and 91,952 instances. It is noteworthy that while still retaining the numerical data type combinations, attributes such as 'url_length,"n_dots"n_hypens,"n_underline,"n_slash",n_questionmark,"n_equal",n_at,"n_and,"total_special_chars,"special_chars_ratio, and "phishing' were the only variables left for analysis.

3. Model Selection and Training:

i. **Data Balancing:** In the phase of the analysis, an investigation was carried out to know the number of records in the target variable that would be used for predictions. Upon inspection, the dataset was observed to be balanced with 62,803 labeled for class 0 (non-phishing) and 29149 instances labeled for class 1 (phishing).

ii. Data Splitting: Following the previous ratio used earlier, the dataset was split into training and testing sets, with 70% of the data allocated for training and 30% for testing. The training set consisted of 43,963 instances for class 0 and 29,402 instances for class 1

4. Model Evaluation and Comparison

Building upon the model selection and training, this is the last phase for the web page phishing analysis. With the use of the random forest and support vector machine libraries, performance evaluation had been made possible. Based on their performance metrics, model implemented using ranger library achieved the highest accuracy of 88.41%, closely followed by the kernel library with an accuracy of 88.08%, while randomForest had the highest recall of 92.24%. In considering the overall best performance, the ranger library consistently yielded higher values across all metrics, indicating its effectiveness in predicting phishing websites. Below is a tabular view that shows the performance and the evaluation outcomes of all models.

Library	Accuracy %	Recall %	F1- score %
randomForest	88.06	92.24	84.57
Ranger	88.41	91.93	85.07
e1071	87.69	90.59	84.35
Kernlab	88.08	90.62	84.80

Table 4: Overview of Model Evaluation in Web Page Phishing dataset.

C. STATLOG SHUTTLE DATASET.

1. Understanding data characteristics: The Statlog Shuttle dataset just like the adult data was collected in two separate comma separated files (CSV). The first file of the data comprised 10 attributes and 14,500 instances. While the second file comprised of 10 attributes as well and 43,500 instances. Again, data

integration took place because both files are identical and have the same attributes. This was done for easier analysis to be carried out.

i. **Number of independent and dependent variables:** The dataset consists of 10 independent variables, representing factors that can influence the outcome, with 1 dependent variable which indicates a single outcome or target variable (Class) to be predicted.

ii. **Number of records:** After the combination of the two csv files, to one, the dataset then had a total of 58,000 records or instances to be observed.

iii. **Data types of Combination: T**he datasets the observation combination includes:

Variable	Type	Description
Red_Flow	Integer (int)	Represent sensor reading relating to engine flow
Fpv_Close	Integer (int)	Represent sensor reading, possible valve position
Fpv_Open	Integer (int)	Represent sensor reading, possible valve position
High	Integer (int)	Sensor reading, potentially pressure of temperature
Bypass	Integer (int)	Sensor reading, related by bypass system state
Bpv_Close	Integer (int)	Represent sensor reading, possible valve position
Bpv_Open	Integer (int)	Represent sensor reading, possible valve position
Class	Integer (int)	Target variable representing the engines operational state
Unknown 1	Integer (int)	Represent unknown value
Unknown 2	Integer (int)	Represent unknown value

Table 5: Overview of Data Type of Combinations in Statlog shuttle dataset.

v. **Summary Statistics of variables:** A descriptive statistic was conducted on all variables in the Statlog data because they were numeric. It showed a summarized analysis that exhibited a range of values as each offers a quick overview of the data distribution of each feature.

2. Data Pre-Processing and Transformation:

i. **Irrelevant Variables:** Durning data preprocessing**,** two attributes were discovered to be irrelevant to the analysis which had a proportion of about 2% of the overall data. As seen in table 5, "Unknown 1 and 2", represented void information, hence they were removed from the data frame.

ii. **Duplicate values:** An investigation was carried out to determine if the dataset consisted of repetitive values. The use of "*sum(duplicate(clean_stalog))"* function made this possible. A total of "1,610" duplicate instances was discovered, having a proportion of approximately 2.78%. These duplicates were handled by removing it completely from the datasets. The removal process was also applied because the proportion of the duplicate's values were small, therefore removing them would not affect the analysis.

iii. **Outliers**: After the removal of duplicates values, another analysis was made to assert how missing values and outliers were present. Through the examination, no missing values were found, while the depiction in Figure5. shows a substantial proportion of outliers that were identified. Approximately 14.7% of the dataset were observed to exhibit outlier behavior. The use of interquartile Range (IQR) was the method that was used to detect and eventually remove the anomalies from the dataset following an empirical design from the previous dataset (Adult income).

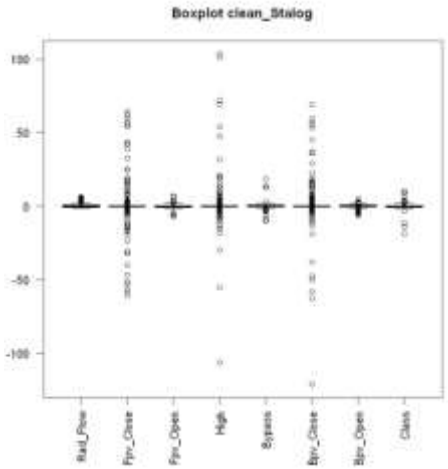

Figure5: Boxplot for Statlog dataset.

iv. **Feature Engineering**: Feature engineering was also conducted on Statlog data. Unlike the web page phishing data that removed some features, the Statlog data created an additional feature. The dataset signified class imbalance. Approximately 80% of the data belong to class 1 which could make mislead the accuracy metric, as the model might not be very good at identifying the less frequency class (class 0). A new feature was then created with a balance split of the data.

v. **Label Encoding and Normalization:** As stated previously the benefits of label encoding and normalization, the target variable of Statlog data was encoded, where the "class" became levels of 0 and 1 formart. Additionally, all variables were then scaled to a normalized range for better analysis processes.

After the preprocessing steps were over, the dataset still consists of 10 attributes but 53,847 instances. It is noteworthy that while retaining attributes such as Fpv_Close, Fpv_Open, High, Bypass, Bpv_Close, Bpv_Open and "class" features like "Class", "Unknown1" and "Unknown2" were not included.

3. Model Selection and Training:

i. **Data Balancing**: In the phase of the analysis, an investigation was carried out to know the number of records in the target variable that would be used for predicting sensor readings. Upon inspection, the dataset was observed to be balanced with 37,694 labeled for class 0 and 16,153 instances labeled for class 1.

ii. **Data Splitting**: Following the previous ratio used earlier, the dataset was split into training and testing sets, with 70% of the data allocated for training and 30% for testing. The training set consisted of 6,975 instances for class 0 and 30,719 instances for class 1

4. Model Evaluation and Comparison

Building upon the model selection and training, this is the last phase for the Statlog shuttle analysis. With the use of the random forest and support vector machine libraries, performance evaluation provided a series of results. Based on their performance metrics for predicting Statlog, all libraries demonstrated a remarkable performance, by achieving high accuracy, recall and F1-Score values. But in comparison to knowing the top performing model for this dataset, Random Forest

technique from randomForest library proved to be more efficient. Having accuracy of 99.27%, recall of 97.62%, and an impressive F1-Score of 98.01%. Table 6 provides a summarized evaluation of their results.

Library	Accuracy %	Recall %	F1- score %
randomForest	99.27	97.62	98.01
Ranger	99.24	97.39	97.93
e1071	98.37	94.45	95.55
Kernlab	98.71	95.82	96.48

Table6: Overview of Model Evaluation in Statlog Shuttle dataset.

VI. CONCLUSION

Based on applying a homogenous empirical design across three datasets with two district machine learning techniques, it is proof that the performance varies depending on datasets characteristics and implementation methodology. Through the exploration of the adult income predictions, web page phishing predictions, and the Statlog shuttle predictions, the random forest model consistently demonstrated strong performance in achieving a high accuracy, recall, and F1 -score. Particularly noteworthy that even though the Support vector model had close range values to the random forest model, the ranger library which consistently outperformed other model most of the time across all three data set highlights its effectiveness in classification tasks. However, notwithstanding, it is important to consider the different datasets characteristics as well as specific problem domains when selecting the most suitable machine learning model. Finally, this study highlights the significance of empirical evaluation and a careful selection of machine learning techniques tailored to the dataset's characteristics using R programming language, as improvement can be made in the future.

VII. REFERENCES

[1] ,. A. N. a. A. K. Jafar Alzubi1, "Machine Learning from Theory to Algorithms: An Overview," Second National Conference on Computational Intelligence , 2018.

[2] P. T. R, "A Comparative Study on Decision Tree and Random Forest Using R Tool," International Journal of Advanced Research in Computer and Communication Engineering, 2015.

[3] D. M. a. K. H. Alexandros Karatzoglou, "Support Vector Machines in R," American Statistical Association, 2006.

[4] J. B. a. D. Lowd, "Machine Unlearning for Random Forests," International Conference on Machine Learning, 2021.

[5] S. A. a. H. O. Atsushi Matsumoto, "Comparison of Random Forest and SVM for Raw Data in Drug Discovery," International Journal of Machine Learning and Computing, 2016.

[6] O. o. N. Statistics, "Phishing attacks – who is most at risk?," Office of National Statistics , September 2022. [Online]. Available: https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/articles/phishingattackswhoismostatrisk/2022-09-26.

[7] H. Wickham, "ggplot2: Elegant Graphics for Data Analysis," Springer International Publishing., New York, 2015.

[8] L. Breiman, "Random forests. Machine Learning," Springer Link, 2001.

[9] C. C. a. V. Vapnik, " Support-Vector Networks," Kluwer Academic Publishers, Boston. Manufactured in The Netherlands., USA, 1995.

[10] R. p. b. A. L. a. M. W. r Fortran original by Leo Breiman and Adele Cutler, "randomForest: Breiman and Cutler's Random Forests for Classification and Regression," CRAN, 2022.

[11] K. B. S. D. N. a. A. M. Muhammad Achirul Nanda, "A Comparison Study of Kernel Functions in the Support Vector Machine and Its Application for Termite Detection," ResearchGate, 2018.

[12] B. S. ¨. a. A. J. Smola, "Learning with Kernels," Massachusetts Institute of Technology Press, Cambridge, Massachusetts, London, England, 2004.

[13] O. S. M. N. a. C. F. N. GONZALO MARISCAL, "A survey of data mining and knowledge discovery process models and methodologies," Cambridge University Press, 2010.