IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.134112

Enhancing Intrusion Detection System with Machine Learning Algorithms

Janardhan K[1], Udaykiran S[2], Harish K[3], Pujiita T[4], Rupesh B[5]

Assistant Professor, Department of Computer Science and Engineering, Rajeev Gandhi Memorial College of Engineering and Technology, Nandyal, Andhra Pradesh.[1]

Department of Computer Science and Engineering, Rajeev Gandhi Memorial College of Engineering and Technology, Nandyal, Andhra Pradesh.[2-5]

Abstract: The rapid growth of online data transmission has increased the demand for stronger data security. Intrusion Detection Systems (IDS) are essential for identifying virtual security threats by using advanced technologies, especially Machine Learning Algorithms, to swiftly detect and categorize attacks in real-time and determine the most accurate algorithm for attack classification.

The current setup uses various intrusion detection algorithms, with a focus on improving performance through advanced algorithms like the Ensemble Learning and Discriminate Analysis. Unlike the existing approach that primarily relies on accuracy, we have used performance parameters such as Accuracy, Precision, Recall, and F1-Measure for evaluating the performance of the models. This comprehensive analysis aims to improve intrusion detection, offering a deeper understanding of algorithm effectiveness, and increasing confidence in the system's intrusion detection capabilities.

Keywords: Machine Learning, Datasets, Feature Selection, Machine Learning algorithms, Intrusion Detection System

I. INTRODUCTION

Now-a-days internet has been playing a vital role in different domains that we are using everyday to day life many of us familiar with wide range of topics such as business, education, and entertainment which insists lot of Technology Stack. Along with this there is rigorous increment in Network Attacks and threats.

Intrusion Detection System (IDS) mainly upholds in Filtering all incoming and outgoing traffic (while transferring packets from source to destination) based on predefined inbound and outbound rules for access control list (ACL's), in which IDS just observes the network flow and intimates an alert message to the administrator of the network. If any unusual activity or suspicious activities detected, but when compared to Firewalls, Intrusion Detection System is more secure and performs better.

The anomaly-based intrusion detection can be also known as behaviour-based detection because mainly it focuses on behaviour of the network, Host systems which results in generating alarm or alert. Hybrid based detection system is combination of signature-based Intrusion detection and anomaly-based intrusion detection system. For taking consideration with the parameters that result higher performance and more optimal we use Hybrid IDS.

A NIDS is a network intrusion detection system that must be installed on a piece of hardware in order to be used. When set up, a NIDS will collect data from each packet (group of data) that goes through it. The conventional NIDS can inspect all of the data passing through it. But you should be wary of analysing every single alert from your NIDS for better protection, since doing so might cause you to miss an actual intrusion attempt.

Most NIDSs provide a rule making feature that lets you specify the data packets you want to keep in order to counteract this problem. In order to target certain sorts of traffic, rules are useful, but they also need familiarity with the NIDS' syntax.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.134112

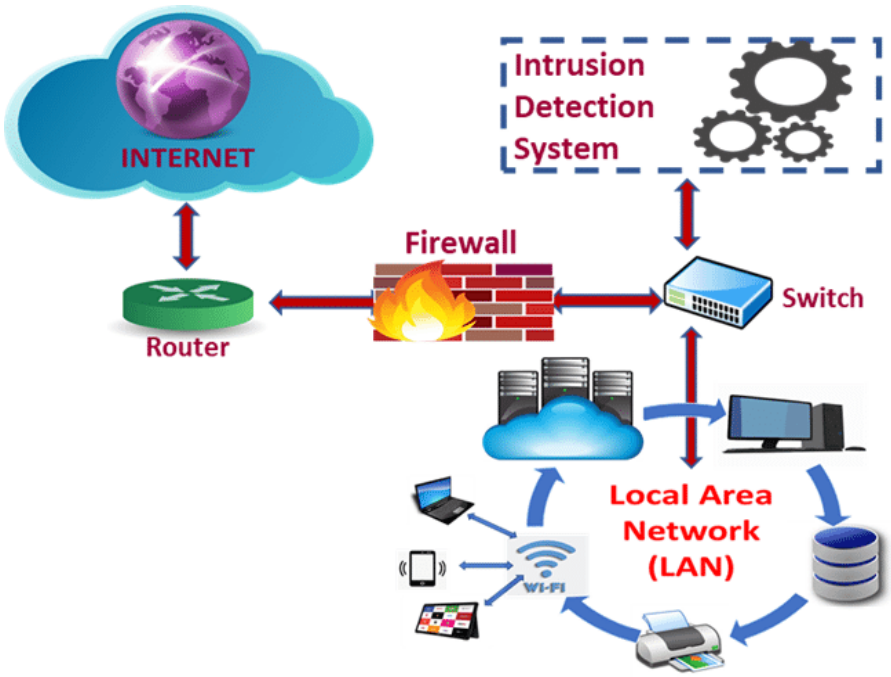


Fig.1 Intrusion Detection System

This research work focuses on KDD99 Dataset by using intelligent Machine Learning algorithms. One of the major advantages of machine learning in IDS is its ability to detect both known and unknown threats. By training on large datasets that encompass diverse attack scenarios, machine learning models can recognize subtle anomalies and identify potential threats that may not have been previously encountered. This enables IDS to stay ahead of attackers who continuously evolve their tactics.

ML demonstrates the improving the learning process of computers based on their experiences without being actually programmed. The work can compare with the performance of various ML algorithms such as K-means clustering, Support Vector Machine (SVM), Logistic Regression, Decision Tree and Artificial Neural Networks for IDS. In addition to the above we are also implemented algorithms like Linear Discriminant Analysis (LDA, Classification and Regression Trees and Random Forest algorithms for Classifying the Enhanced Intrusion Detection. The Performance of the algorithms was compared using metrics like accuracy, precision, recall and F1 Score.It can be used to make predictions on new data points by evaluating which side of the hyperplane they fall on. Data points on one side of the hyperplane are classified as belonging to one class, while data points on the other side of the hyperplane are classified as belonging to another class.

1.LogisticRegression:

Logistic regression is a statistical algorithm used for binary classification, which means it predicts outcomes that belong to one of two possible classes. These classes can be as diverse as “yes/no,” “true/false,” “spam/not spam,” or any other dichotomous categorization. Despite its name, logistic regression is not used for regression (predicting continuous values) but rather for classification problems.

More specifically, it is used for predicting the categorical dependent variable using a given set of independent variables. Logistic regression predicts the output of a categorical dependent variable. Therefore, the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.

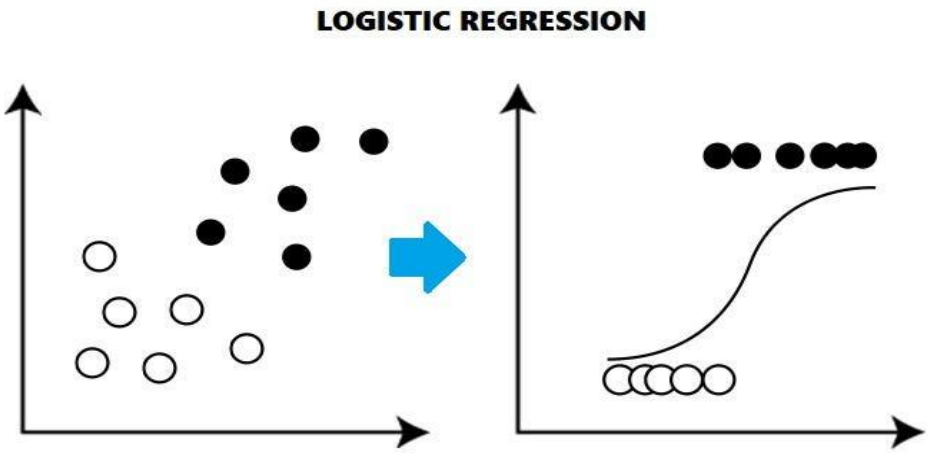
It is similar to the Linear Regression except in how they are used. Linear Regression is used for solving Regression problems, whereas Logistic regression is used for solving classification problems. In this, instead of fitting a regression line, we fit an “S” shaped logistic function, which predicts two maximum values (0 or 1). The curve from the logistic function indicates the likelihood of something such as whether the cells are cancerous or not, a mouse is obese or not based on its weight, etc.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.134112



Logistic Regression is a significant machine learning algorithm because it has the ability to provide probabilities and classify new data using continuous and discrete datasets. It can be used to classify the observations using different types of data and can easily determine the most effective variables used for the classification.

2.Support Vector Machine

Support Vector Machines (SVM) output an optimal line of separation between the classes, based on the training data which was entered as input. We have different groups of points scattered in space, and you want to draw a line to separate them as best as possible. This line is like a boundary between the groups, and it's called a hyperplane. When we use Support Vector Machines (SVM), we're careful about points that are really close to the boundary between groups, even if they're not exactly in one group or the other. These points are called outliers. Once we've figured out the best line (hyperplane) to separate the groups, we can use it to predict where new points belong. We just check which side of the line the new point falls on, and that tells us which group it is belongs to. While analysing the predicted output list, we see that the accuracy of the model is at 95%. A comparative between the actual and predicted values is also shown.

3. Decision Tree

Decision tree algorithm falls under the category of supervised learning. They can be used to solve both regression and classification problems. It uses the tree representation to solve the problem in which each leaf node corresponds to a class label and attributes are represented on the internal node of the tree. We can represent any Boolean function on discrete attributes using the decision tree.

some assumptions that we made while using decision tree:

At the beginning, we consider the whole training set as the root.

Feature values are preferred to be categorical. If the values are continuous then they are discretized prior to building the model. On the basis of attribute values records are distributed recursively. We use statistical methods for ordering attributes as root or the internal node.

The main goal of decision trees is to make the best splits between nodes which will optimally divide the data into the correct categories. To do this, we need to use he right decision rules. The rules are what directly affect the performance of the algorithm. Decision trees take very little time in processing the data when compared to other algorithms. Few preprocessing steps like normalization, transformation, and scaling the data can be skipped. Although there are missing values in the dataset, the performance of the model won't be affected. A Decision Tree model is intuitive and easy to explain to the technical teams and stakeholders, and can be implemented across several organizations.

4.KNN Classifier:

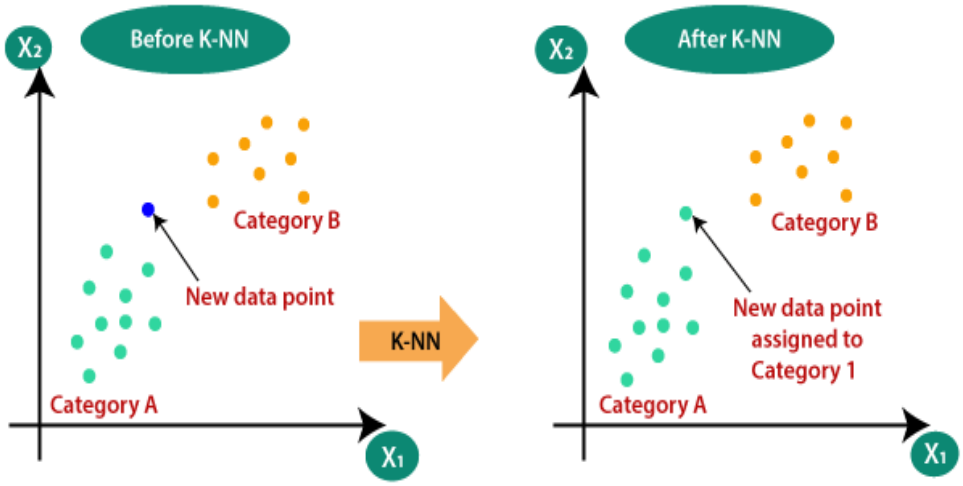
The KNN algorithm is a supervised machine learning model. That means it predicts a target variable using one or multiple independent variables. K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a good suite category by using K- NN algorithm.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.134112



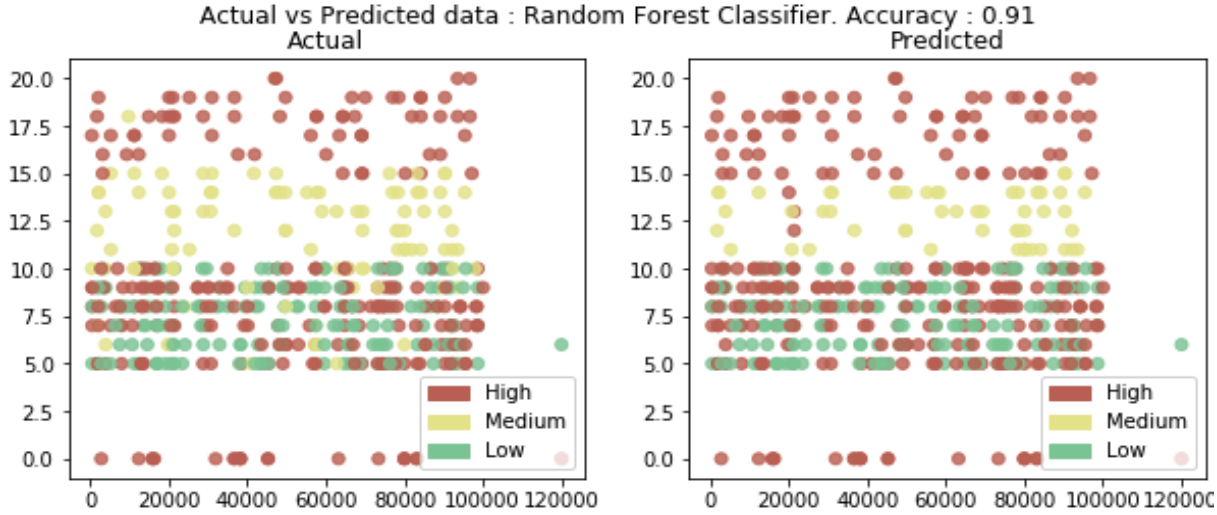
This algorithm works by classifying the data points based on how the neighbors are classified. Any new case is classified based on a similarity measure of all the available cases. Technically, the algorithm classifies an unknown item by looking at k of its already -classified, nearest neighbor items by finding out majority votes

from nearest neighbours that have similar attributes as those used to map the items. Selecting the value of K depends on individual cases and sometimes the best method of choosing K is to run through different values of K and verify the outcomes. Using cross-validation, the KNN algorithm can be tested for different values of K and the value of K that results in good accuracy can be considered as an optimal value for K.

5. Random Forest

Decision tree algorithms are efficient in eliminating columns that don't add value in predicting the output. In some cases, we are even able to see how a prediction was derived by backtracking the tree.

When that happens, the predictions might not be very reliable. We call these messy trees "weak models" because they don't do a great job on their own. That's where random forests come in. Instead of relying on one big, messy tree, we make lots of smaller, simpler trees. Each tree looks at a random subset of the data and makes its own predictions. Then, we combine all these predictions together to make a more accurate guess. It's like asking a bunch of different people for their opinions and then averaging them out to get a better answer. The model performance is improvised by taking an average of several such decision trees derived from the subsets of the training data. This approach is called the *random forest* classification.



IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.134112

II. PROPOSED METHODOLOGY

Ensemble learning

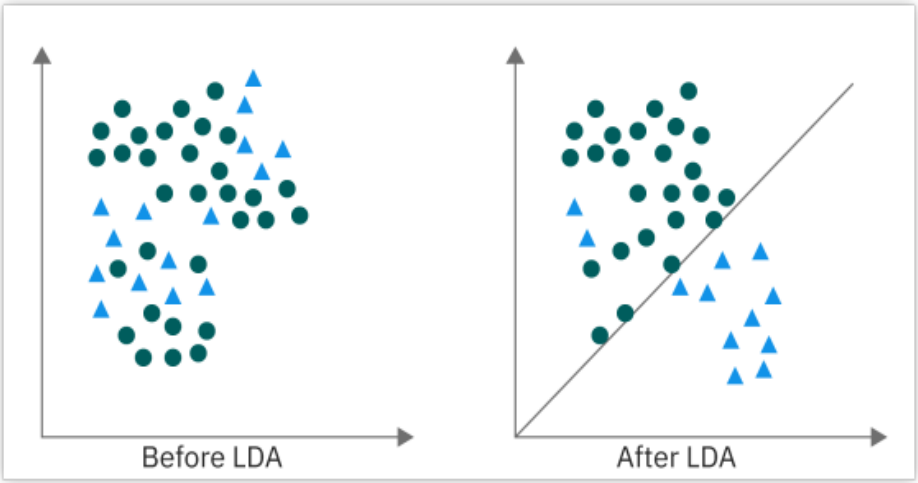
Ensemble learning refers to the type of machine learning algorithms where more than one algorithm is combined to produce a better model. When two or more same algorithms are repeated to achieve this, it is called a *homogenous ensemble* algorithm. we'll look at how we can combine a decision tree-based model into a random forest and gradient boosted tree to get a higher accuracy level

1.Gradient boosted trees

Gradient boosted trees are also a type of ensemble learning. They are based on the method called boosting, which involves training a model one after another based up on the outputs from the previous models. In gradient boosted trees, we calculate the error from the previous model, also known as *residuals.* Now we define another model that is trained on this residual. The resulting model is the sum of previous model and the model trained on residuals. This process is repeated until our condition was able to meet. Even though gradient boosted trees outperform random forest models, they are computationally expensive because they are built sequentially. A specific implementation called XGBoost is used to overcome this issue. XGBoost is a machine learning algorithm that belongs to the ensemble learning category, specifically the gradient boosting framework. It utilizes decision trees as base learners and follows repeating techniques to enhance model generalization. Known for its computational efficiency, feature importance analysis, and handling of missing values, XGBoost is widely used for tasks such as regression, classification, and ranking.

2.Linear Discriminant Analysis (LDA):

It is a supervised technique and tries to predict the class of Dependent Variable using the linear combination of Independent Variables. It assumes that the independent variables are normally distributed (continuous and numerical) and equal variance/ covariance for the classes. This technique can be used both for classification and dimensionality reduction. When these assumptions are satisfied, LDA creates a Linear Decision Boundary.



Linear Discriminant Analysis

This flexibility ensures that LDA can be used for multi-class data classification problems, unlike logistic regression, which is limited to binary classification. LDA is so often applied to enhance the operation of other learning classification algorithms such as decision tree, random forest, or support vector machines (SVM).

Algorithms:

Algorithm (XGBoost)

Input:

Training dataset with features and labels

Output:

Trained model for predictions

Data Preparation:

- Collect dataset with features and labels.
- Preprocess data: handle missing values, encode categorical variables, and scale numerical features.

International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024
DOI: 10.17148/IJARCCE.2024.134112

Initialization:
Initialize base model (e.g., decision tree) for the first prediction.
Set dataset weights uniformly.

Iterative Training:
- **For each boosting round:**
- Fit base learner to training data, emphasizing misclassified instances or those with higher residuals.
- Compute residuals for each instance.
- Combine base learner's predictions with previous iterations, minimizing overall loss.
- Regularize base learner to prevent overfitting.
- Update dataset weights, giving more importance to instances with higher residuals.

Stopping Criteria:
Stop when predefined iteration limit is reached or validation performance stalls.

Output Formation:
Aggregate predictions of all base learners for final output.

Utilization:
Utilize trained XGBoost model for tasks like classification, regression, and ranking.
Applications include predicting customer churn, fraud detection, and search engine optimization

Algorithm (Discriminant Analysis)

Input:
Features and corresponding class labels

Output:
Trained model for classifying new instances

Data Preparation:
- Gather features and class labels.
- Ensure balanced dataset if possible.
- Handle missing values and scale features if needed.

Model Training:
- Estimate class-conditional densities.
- Calculate class prior probabilities.
- Optionally, perform dimensionality reduction (e.g., PCA).

Discriminant Function:
- Calculate discriminant function for each class.

Classification:
- Compute discriminant score for each class.
- Assign instance to class with highest score.

Model Evaluation:
- Assess model performance using metrics like accuracy, precision, recall.
- Optionally, validate with cross-validation.

Utilization:
- Use model to classify new instances.
- Applications include pattern recognition, medical diagnosis, fraud detection.

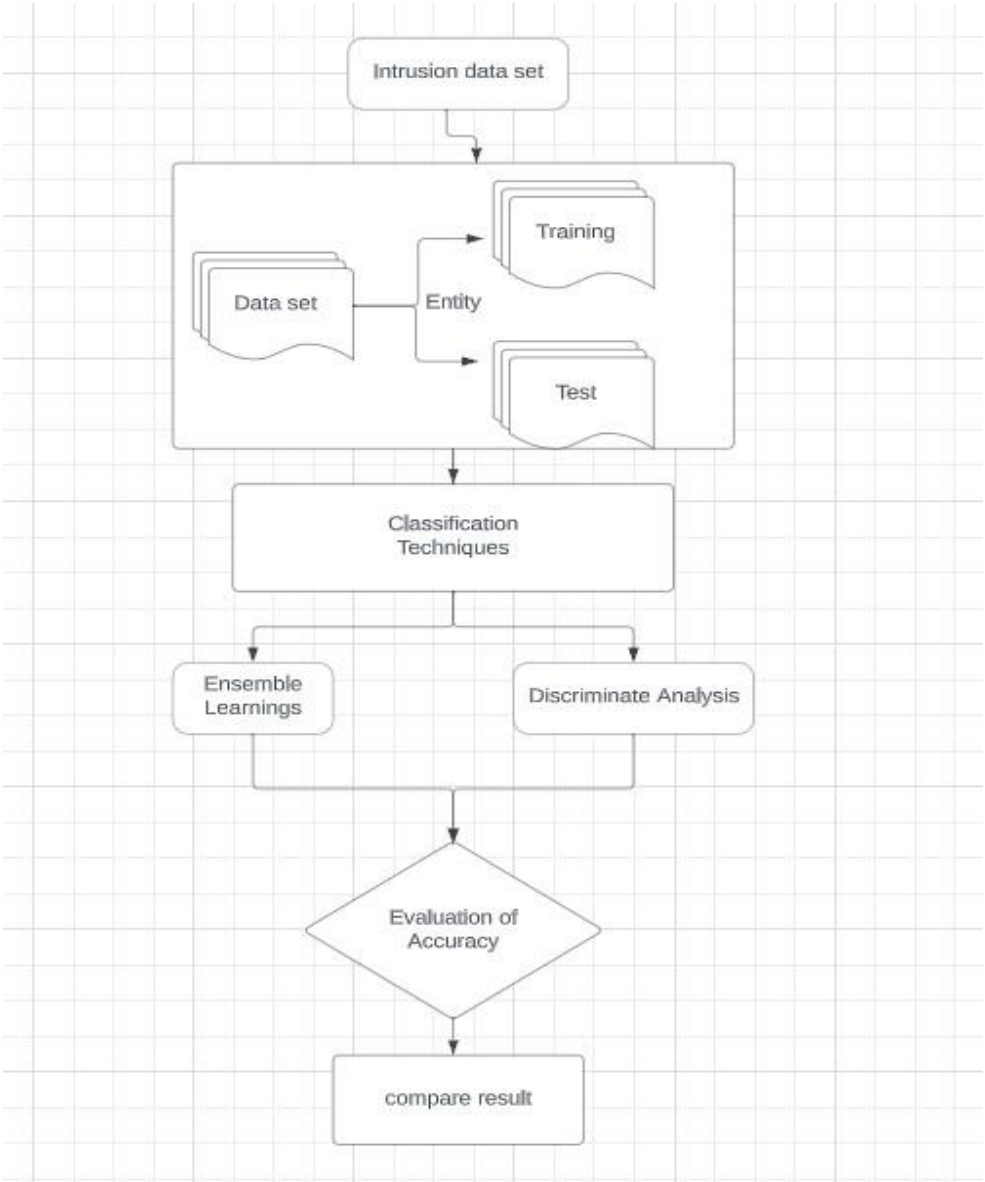
IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.134112

Architecture:



The diagram in the image depicts a text classification architecture. Text classification is a type of machine learning task that involves automatically assigning text data to a specific category or set of categories. Here's a breakdown of the architecture:

- **Intrusion Data Set:** This is the data that the model will be trained on. It likely consists of text samples that have already been categorized as intrusions or non-intrusions.
- **Training**: This is the phase where the model learns from the intrusion data set. The model is able to identify patterns that differentiate intrusions from non-intrusion data.
- **Data Set**: This refers to the data that will be used to test the model's performance.
- **Entity**: This stage possibly refers to the identification of individual data points within the data set.
- **Test:** This is the phase where the model's performance is evaluated using the test data set.
- **Classification**: This is the core function of the model. Here, the model assigns a category label (intrusion or non-intrusion) to a new piece of text data.
- **Ensemble Learnings:** This likely refers to a machine learning technique that involves training multiple models and then combining their predictions. This can improve the overall accuracy of the model.

IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

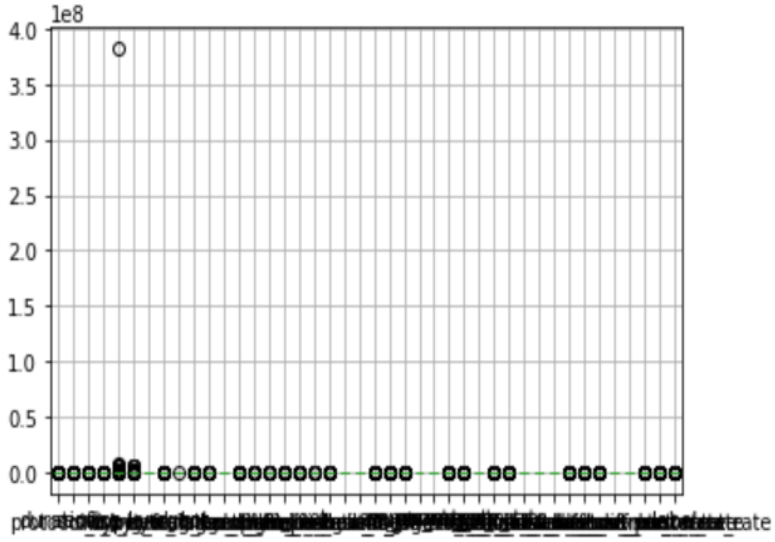
DOI: 10.17148/IJARCCE.2024.134112

- **Discriminate Analysis:** This is a statistical technique that is commonly used for classification tasks. It works by identifying characteristics that differentiate between different categories.
- **Evaluation of Accuracy:** This is the process of measuring how well the model performs on the test data set. This is typically done by calculating metrics like precision, recall, and F1 score. Overall, the architecture in the image seems to depict a system for classifying text data as intrusions or non-intrusions. The system trains on a dataset of labeled intrusion data, and then uses this knowledge to classify new text data

III. RESULT & ANALYSIS

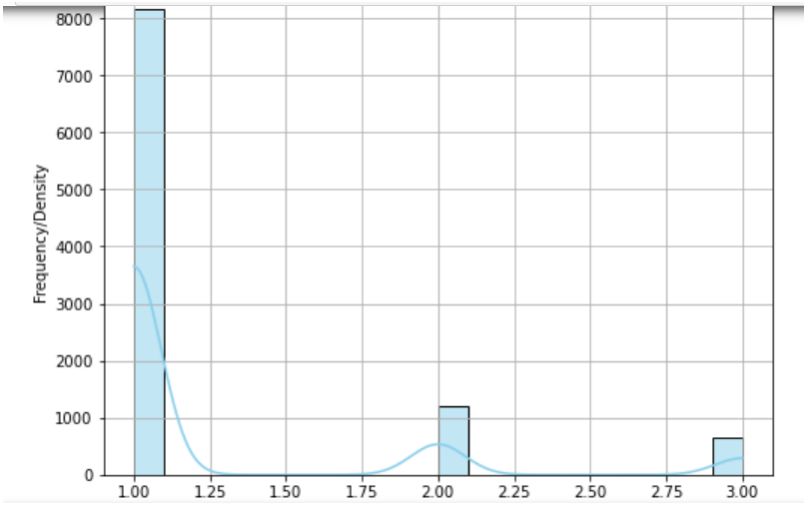
Outlier Detection:

Outlier detection, also known as anomaly detection, is a process used in data mining and statistical analysis to identify unusual patterns or observations within a dataset. These outliers are data points that deviate significantly from the rest of the data, potentially indicating errors in data collection, measurement, or interesting phenomena worthy of further investigation.



After conducting outlier detection analysis on our dataset, it has been observed that the majority of data points conform to expected patterns. However, a single outlier was identified, which exhibited significant deviation from the rest of the dataset.

Understanding Distribution of Each Attributes:



IJARCCE ISSN (O) 2278-1021, ISSN (P) 2319-5940

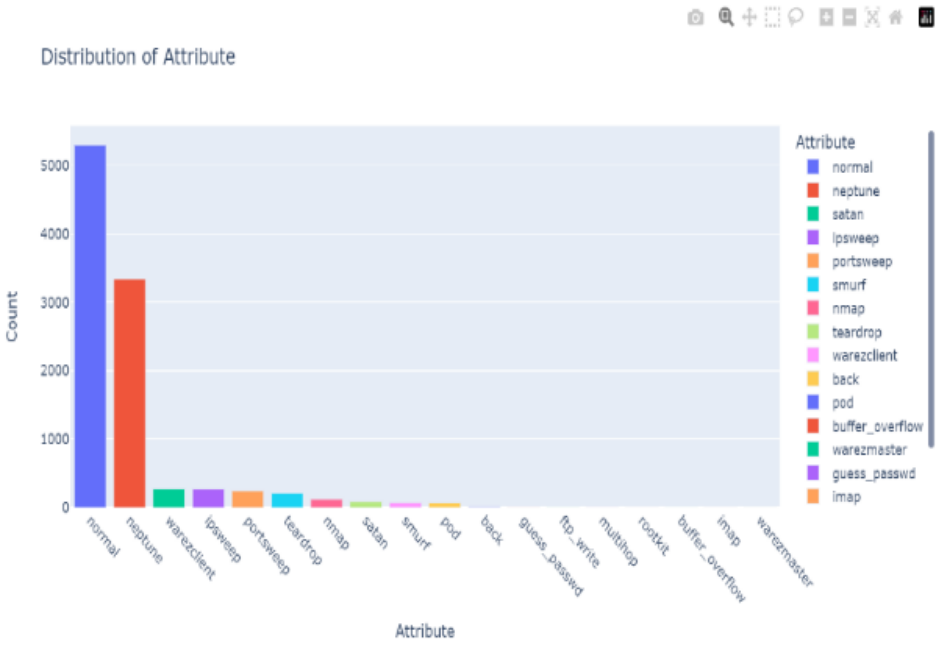
International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.134112

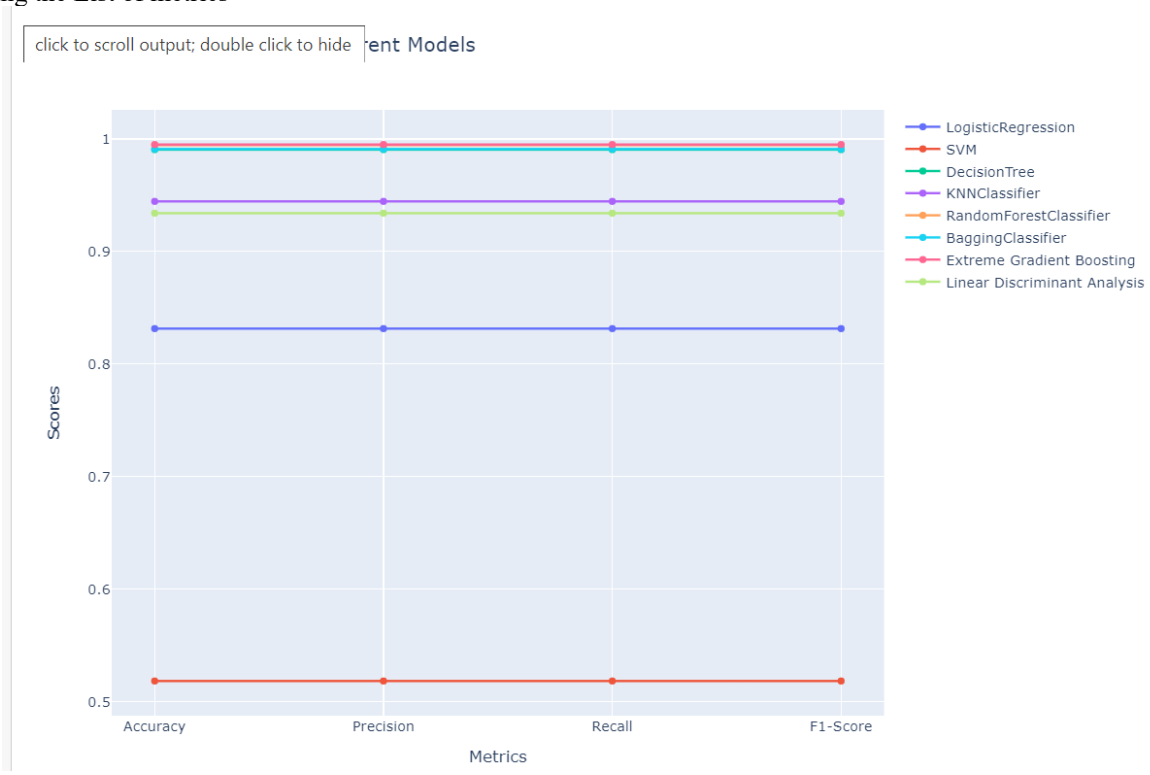
understanding the distribution of each attribute in a dataset is essential for ensuring data quality, selecting appropriate features for analysis or modeling, meeting statistical assumptions, exploring data patterns, selecting suitable modeling techniques, and effectively communicating insights derived from the data. It forms the foundation for rigorous data analysis and decision-making in various domains.

Distribution Of Attributes:



The image is a bar graph showing the distribution of attributes across different types of content. The x-axis of the graph shows the attribute, and the y-axis shows the count. . In cyber security, attributes are used to describe the characteristics of a network event or intrusion attempt

Making the List of metrics



International Journal of Advanced Research in Computer and Communication Engineering
Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024
DOI: 10.17148/IJARCCE.2024.134112

In proposed methodology we have taken two alogrithms one is XGBoost and LDA. While XGBoost achieved a higher accuracy score of 0.995 compared to Linear Discriminant Analysis (LDA) with an accuracy of 0.934, here's a detailed description of both models and potential considerations:

XGBoost is a popular and powerful machine learning algorithm known for its efficiency and effectiveness in various types of data.

It is based on gradient boosting framework and uses decision trees as base learners.Discriminant Analysis is not only used improving accuracy.It can be used for overcoming overfitting in data set.LDA is a classification technique used for dimensionality reduction and classification tasks.

It finds linear combinations of features that best separate different classes.

LDA includes regularization techniques to prevent overfitting, which can be particularly useful when dealing with high-dimensional data or datasets with a small number of samples.

IV. CONCLUSION

The existing system prioritizes accuracy and employs SVM, KNN, Random Forest, and K-means clustering. The proposed system not only emphasizes accuracy but also incorporates performance metrics such as decision rate, false accuracy rate, F1-score, recall, and precision rate.

To enhance accuracy and mitigate overfitting, the proposed system will leverage Discriminant Analysis and Ensemble Learning Techniques. In future you can use latest alogrithms and have chances of increasing accuracy.

REFERENCES

[1]. Tsehay Admassu Assegie, An optimized KNN model for signature-based malware detection, Int. J. Comput. Eng. Res. Trends 8 (2021) 46–49.
[2]. S.G. Kene, D.P. Theng, A review on intrusion detection techniques for cloud computing and security challenges, in: 2nd International Conference on Electronics and Communication Systems (ICECS), 2015, pp. 227–232, https://doi.org/ 10.1109/ECS.2015.7124898.
[3]. S.A. M, P. G, A survey on various intrusion detection system tools and methods in cloud computing, in: 6th International Conference on Computing for Sustainable Global Development (INDIACom), 2019, pp. 439–445.
[4]. K. Kulkarni, G. Ahn, H. Hu, Detecting and resolving firewall policy anomalies, IEEE Trans. Dependable Secure Comput. 9 (2012) 318–331.
[5]. M. Bhavsingh, M.S. Lakshmi, S.P. Kumar, N. Parashuram, "Improved trial division algorithm by Lagrange‟ s, Interpol. Funct. 5 (2017) 1227–1231.
[6]. Tsehay Admassu Assegie, An optimized KNN model for signature-based malware detection, Int. J. Comput. Eng. Res. Trends 8 (2021) 46–49.
[7]. S.G. Kene, D.P. Theng, A review on intrusion detection techniques for cloud computing and security challenges, in: 2nd International Conference on Electronics and Communication Systems (ICECS), 2015, pp. 227–232, https://doi.org/ 10.1109/ECS.2015.7124898.
[8]. K. Kulkarni, G. Ahn, H. Hu, Detecting and resolving firewall policy anomalies, IEEE Trans. Dependable Secure Comput. 9 (2012) 318–331.
[9]. M. Bhavsingh, M.S. Lakshmi, S.P. Kumar, N. Parashuram, "Improved trial division algorithm by Lagrange‟ s, Interpol. Funct. 5 (2017) 1227–1231.
[10]. J. Yick, B.Mukherjee, and D. Ghosal, “Wireless sensor networsurvey,” ComputerNetworks, vol. 52, no. 12, pp. 2292–2330, 2008.
[11]. W. Sun, Z. Cai, Y. Li, F. Liu, S. Fang, and G. Wang, “Security and privacy in themedical internet of things: a review,” Security and Communication Networks, vol. 2018, Article ID 5978636, 9 pages, 2018.
[12]. Z. Pan, J. Lei, Y. Zhang, and F. L. Wang, “Adaptive fractional-Pixel motion estimation skipped algorithm for efficient HEVC motion estimation,”
[13]. C. Karlof and D. Wagner, “Secure routing in wireless sensor networks: Attacks and countermeasures,” AdHoc Networks, vol.1, no. 2-3, pp. 293– 315, 2003.
[14]. P. Li, X. Yu, H. Xu, J. Qian, L. Dong, and H. Nie, “Research on secure localization model based on trust valuation in wireless sensor networks,” Security and Communication Networks, vol.2017, Article ID 6102780, 12 pages, 2017.

International Journal of Advanced Research in Computer and Communication Engineering

Impact Factor 8.102 ⁜ Peer-reviewed & Refereed journal ⁜ Vol. 13, Issue 4, April 2024

DOI: 10.17148/IJARCCE.2024.134112

[15]. Aburomman, A. A., &Reaz, M. B. I. (2016) "Ensemble of binary SVM classifiers based on PCA and LDA feature extraction for intrusion detection."Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC): 636-640.

[16]. Al-Jarrah, O. Y., Al-Hammdi, Y., Yoo, P. D., Muhaidat, S., & Al-Qutayri, M. (2018) "Semi-supervised multi-layered clustering model for intrusion detection." Digital Communications and Networks 4(4): 277-286.

[17]. Al-Yaseen, W. L., Othman, Z. A., &Nazri, M. Z. A. (2017) "multi-level hybrid support vector machine and extreme learning machine based on modified K-means for intrusion detection system." Expert Systems with Applications 67(1): 296-303.

[18]. An, X., Su, J., Lü, X., & Lin, F. (2018) "Hypergraph clustering model-based association analysis of DDOS attacks in fog computing. intrusion detection system." EURASIP Journal on Wireless Communications and Networking 249 (1): 1-9.

[19]. Belavagi, M. C., &Muniyal, B. (2016) "Performance evaluation of supervised machine learning algorithms for intrusion detection." Procedia Computer Science 89(1): 117-123

[20]. F. Bonomi, R. Milito, J. Zhu, and S. Addepalli, "Fog computing and its role in the internet of things," in Proceedings of the 1st ACM Mobile Cloud Computing Workshop, MCC '12, pp. 13–15, ACM, Helsinki, Finland, August 2012.

[21]. L. M. Vaquero and L. Rodero-Merino, "Finding your way in the fog: towards a comprehensive definition of fog computing," ACM SIGCOMM Computer Communication Review Archive, vol. 44, no. 5, pp. 27–32, 2014.

[22]. X. Xu, X. Zhang, M. Khan, W. Dou, S. Xue, and S. Yu, "A balanced virtual machine scheduling method for energy-performance trade-offs in cyber-physical cloud systems," Future Generation Computer Systems, 2017.

[23]. T. H. Luan, L. Gao, Z. Li, Y. Xiang, G. Wei, and L. Sun, "Fog computing: focusing on mobile users at the edge," 2015, https:// arxiv.org/abs/1502.01815.

[24]. G.I. Klas, Fog computing and mobile edge cloud gain momentum open fog consortium, etsimec and cloudlets, 2015.