Intrusion Detection System using Machine Learning with Stacking

BACHELOR OF TECHNOLOGY

In

INFORMATION TECHNOLOGY

By

Rohit Kumar (20118078)
Saniya Rahmani (20118086)
Kumarapu Hemram (20118050)

Under the guidance of

Dr.Tirath Prasad Sahu

Assistant Professor

Department of Information Technology
NIT Raipur, Chhattisgarh

NATIONAL INSTITUTE OF TECHNOLOGY, RAIPUR (Chhattisgarh)

ACKNOWLEDGEMENT

We would like to thank our project guide **Dr.Tirath Prasad Sahu, Assistant Professor,** Department of Information Technology,NIT Raipur for his proficient supervision on the project “**Intrusion Detection System using Machine Learning with Stacking**”. We would also like to thank him for his constant support,help, motivation and guidance. His constant encouragement,constructive criticism and valuable guidance has been the cause for achieving the results. We gratefully acknowledge **Dr. Sanjay Kumar, Associate Professor & Head of Department**, Information Technology , NIT Raipur for allowing us to work in this field and project in particular. We would like to express our gratitude towards UCI for providing us with the KDD Cup 99 dataset. Finally we wish to convey our warmest and deepest gratitude to our family members to whom we all owe our achievements and to our friends for their endless support and encouragement.

Rohit Kumar (20118078)

Saniya Rahmani (20118086)

Kumarapu Hemram (20118050)

Intrusion Detection System using Machine Learning with Stacking

1.Abstract

In the world of computers and interconnected systems, there's a constant worry about unwanted guests – we call them intrusions.These are instances where someone or something tries to get into computer networks without permission, causing problems like stealing information or disrupting normal operations.Sneaky individuals may find ways to get into computer systems without permission, potentially causing harm or stealing important information. Harmful software, like viruses or ransomware, can slip into systems, causing damage, stealing data, or demanding money to fix the mess. These attacks try to overwhelm a network with too much traffic, making it hard for regular users to access services. Intrusion Detection Systems (IDS) are critical for safeguarding computer networks against unauthorized access and malicious activities. An IDS constantly watches the flow of data within a network. It pays attention to who's trying to access what, what data is being sent or received, and if there are any irregular patterns. IDS is like a vigilant detective. It learns what normal, everyday network activity looks like. When something deviates from the usual – an anomaly – it raises a red flag. This project presents an Intrusion Detection System based on Machine Learning with a stacking ensemble approach. The ensemble consists of base learners including Random Forest, Decision Tree, Logistic Regression, and Support Vector Machine (SVM), with a Random Forest meta-learner.
This proposed IDS is subjected to a rigorous evaluation, assessing its performance across key metrics such as F1 score, precision, recall, and accuracy. The proposed system demonstrates exceptional accuracy with an achieved rate of 99.66%. The report elaborates on the feature selection methodology, classification models, experiments, related works, and detailed result analysis.

2.Keywords

Intrusion Detection, Machine Learning, Stacking Ensemble, Random Forest, Decision Tree, Logistic Regression, SVM, Feature Selection, Accuracy.

3.Introduction

In a world filled with computers and networks, there are hidden dangers online. Hackers and malicious software constantly attempt to break into computer systems and steal sensitive information. To protect against these threats, we use Intrusion Detection Systems (IDS).

Think of an IDS as a digital security guard for your computer network. It keeps a watchful eye on all the data flowing through the network and raises an alarm if it spots anything suspicious. Traditional IDS, however, can sometimes miss new and clever attacks because they rely on fixed rules.
As communication technology advances, an increasing array of devices are joining the interconnected realm of the Internet. This surge in connectivity, however, brings with it a concerning trend – a rise in the sophistication of network attackers. These adversaries employ diverse tactics to take command of numerous network devices, orchestrating large-scale distributed denial of service (DDoS) attacks that lead to pervasive network congestion. Consequently, the implementation of a Network Intrusion Detection System (NIDS) emerges as a crucial measure to identify and thwart these malicious activities, thereby safeguarding the integrity and efficiency of the network.A fast and accurate Intrusion detection system is needed because the bad guys on the internet are always coming up with new tricks to break into our digital stuff. That's the reason intrusion detection has become an important topic among researchers. Detection of Intrusion is basically a classification problem and machine learning can be used for such classification problems as machine learning is used to extract information from data.

In our project, We've used the power of Machine Learning, a type of computer intelligence, to create a smarter IDS. This system can learn and adapt to new threats, making it much more effective at keeping your network safe from attackers.
The machine learning model designed in this project includes different machine learning algorithms which are used for classification such as Random Forest, Decision Tree, Logistic Regression, and Support Vector Machine (SVM), each with unique skills and a stacked model with the above mentioned base learners and random forest as meta learner is developed which has an excellent accuracy of 99.66% .This proposed model can work efficiently to catch any unusual activities on the network.

In this report, we'll dive deep into how our Intrusion Detection System works, explaining the smart techniques we used and how well it performs. As online threats keep evolving, our system is here to provide better protection, ensuring that your computer network stays safe and secure.

4.Related Work

Creating a quick and effective system to catch and prevent cyber threats is a big focus in the world of network security. The research and development of a high-performance and efficient cybersecurity intrusion detection system have been a prominent area of interest for the researchers. In reference [2], Lee and his colleagues proposed a intrusion detection system based on signature. The dataset used by them had 41 features and they formulated rules derived from them. In [1] Siamak Layeghy, Marius Portmann conducted research on various Machine Learning based Network Intrusion Detection Systems and then they applied their (NIDSs) work to various domains. Their research involved eight learning models and they analyzed the performance of those models, which include both supervised and unsupervised approaches. Li Yang, Abdallah Shami in [4] observed the public network traffic datasets and used advanced and traditional Machine Learning algorithms.In intrusion detection, feature selection is crucial. Sung and Mukkamala [6] experimented by removing one feature at a time and used machine learning methods to rank the importance of these features in detecting cyber threats from DARPA data. This helped identify the most significant data points for intrusion detection. [13], Patrick Vanin introduces the concept of Intrusion Detection Systems (IDS) and categorizes various machine learning approaches. The paper also discusses the key performance metrics for evaluating IDS and offers a comprehensive review of recent IDS implementations employing machine learning techniques. Within this review, the paper highlights the pros and cons of each solution. [10], Jafferson and colleagues utilize several datasets for intrusion detection, including KDD Cup 99, NSL-KDD, UNSW-NB15, Kyoto, and CSCIDS 2017. They conduct a thorough analysis of these datasets. Additionally, the paper covers the performance metrics employed to assess the effectiveness of machine learning algorithms in their intrusion detection experiments. [21] Lianming Zhang, Kui Liu, Xiaowei Xie, Wenji Bai, Baolin Wu, Pingping Dong proposed a data-driven method called FS-DL, which combines multiple feature selection strategies to remove a large number of redundant features from the dataset, greatly reducing the amount of data used, and ensuring the quality of the input data to improve the detection accuracy and efficiency. Sara Mohammadi, Hamid Mirvaziri, Mostafa Ghazizadeh-Ahsaee, Hadis Karimipour [22] used Cuttlefish Algorithm (CFA): Searches for the best feature subset to enhance classifier accuracy. The Proposed method PLSSVM (Least Squares Support Vector Machine) for modeling the IDS is a novel approach, and the addition of a preprocessing phase for feature selection (LCFS and MMIFS) is a valuable contribution[23] by Fatemeh Amiri, MohammadMahdi Rezaei Yousefi, Caro Lucas, Azadeh Shakery, Nasser Yazdani

Until now, the previous research we've reviewed has prioritized accuracy when it comes to detecting cyberattacks but hasn't emphasized speed and efficiency. In contrast, our paper introduces a framework designed specifically for swift and efficient cybersecurity intrusion detection. In our proposed approach, we deal with a large volume of data.The focus of our research extends beyond just choosing a fast and efficient method; we also evaluate its accuracy and scalability as important factors.

5.Dataset Description

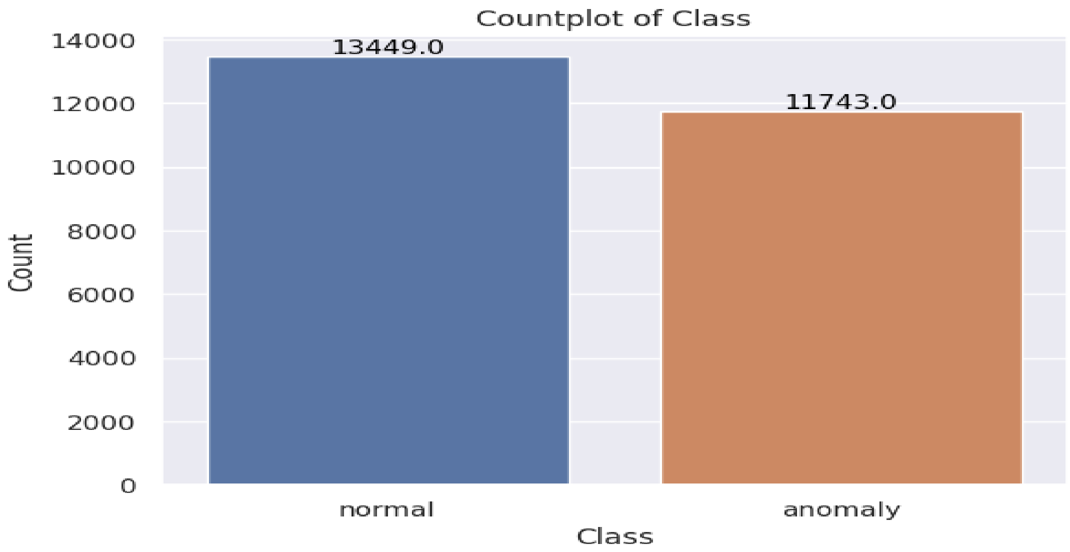
The data set which we have used to develop this Intrusion detection system is KDD Cup 1999 Data set. This dataset was part of a competition associated with the KDD-99, a conference on Knowledge Discovery and Data Mining. The challenge was to create a system that could detect network intrusions. In simpler terms, the goal was to build a smart model that could tell the difference between harmful connections (intrusions or attacks) and regular, safe connections. The dataset comprises a set of information mimicking real-world scenarios in a military network, providing a standardized basis for testing and evaluating intrusion detection systems.The dataset contains 42 features and 25192 instances. For each data transfer, we collect 41 pieces of information, some are numbers that tell us about the data, and some are descriptions of what's happening. The main thing we want to know is whether it's normal network activity or something unusual (an attack). So, we label each data transfer as either "Normal" or "Anomalous" to understand what's happening in this simulated network.

5.1 Data Cleaning

The KDD CUP-99 dataset is first checked if it is having some null values and it was found out that it was not having any null value.Then it was checked if some of the rows might be repeating and it was not so. The column num_outbound_cmds was found to be having all the values in it as 0. So it was dropped from the dataset.

5.2 Data Distribution

Data distribution in binary classification is primarily about the balance between the classes rather than the shape of the overall dataset's distribution. If the balance is adequate, one can proceed with building and evaluating your classification model.In KDD CUP-99 dataset, there are two classes which contain almost equal instances.The ratio of the anomaly class to the normal class is about 0.87 which is above 0.8. So, it can be considered that the data is almost balanced.



5.3 Correlation Coefficient based Feature Selection

Feature selection is a crucial step in building an effective IDS. In this project, the correlation coefficient-based Feature Selection method is being employed to identify the most relevant features for intrusion detection. This Feature Selection method measures the relation between features and the target variable which helps in selecting the most informative features.In the proposed model , there are 9 features which have the highest correlation above 0.8 (threshold value).

The correlation coefficient measures the degree of linear association between two variables. In the context of feature selection:

1) **Positive Correlation:** A positive correlation coefficient indicates that as one feature increases, the other feature also tends to increase. In the context of feature selection, if two features are highly positively correlated, including both in a model might not provide additional information. Thus, one of them could be omitted to reduce redundancy.

2) **Negative Correlation:** A negative correlation coefficient signifies that as one feature increases, the other tends to decrease. In feature selection, if two features are highly negatively correlated, including both might not add much value, and excluding one could simplify the model.

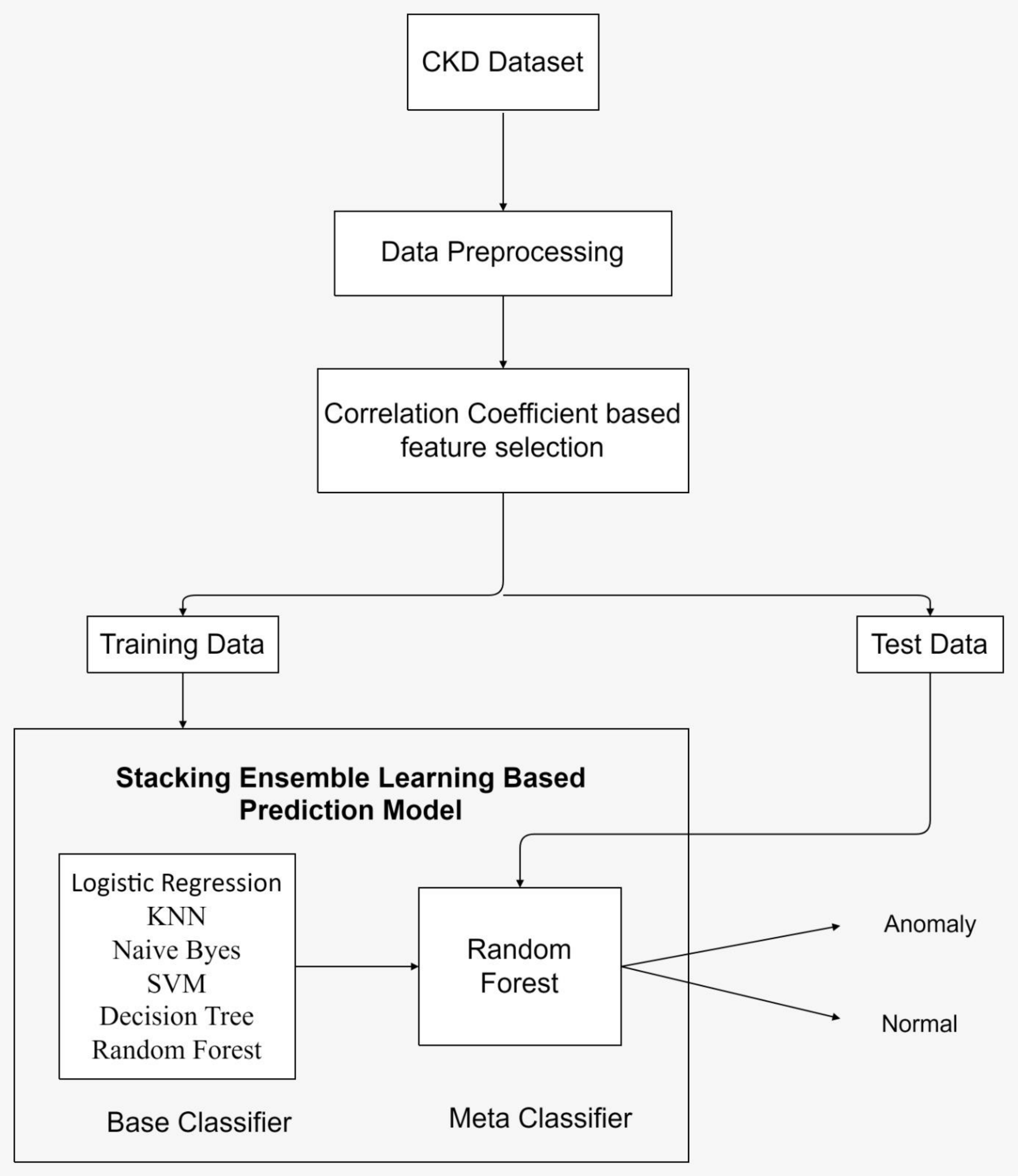
3) **Weak or No Correlation:** If the correlation coefficient is close to zero, there is a weak or no linear relationship between the features. In such cases, both features might offer unique information and could be valuable for the model.

Splitting Dataset

Data splitting is a fundamental practice in machine learning model development, involving the partitioning of a dataset into training and testing sets. The training set serves as the classroom where the model learns patterns, while the testing set acts as an examination room, assessing the model's ability to generalize its learnings to new, unseen data. Data science library scikit-learn is used to split the data into test data and train data. 80% of the data is splitted into training data and 20 percent of the data is splitted into testing data and the random state is set to 42. The random_state is like a magic number in machine learning that lets you recreate a specific train-test split every time you run your code.

Proposed Framework

The proposed framework for an Intrusion Detection System (IDS) using Machine Learning (ML) algorithms aims to enhance network security through intelligent threat detection. This framework is designed to be effective in identifying cybersecurity threats. The proposed framework of machine learning algorithms for intrusion detection begins with data visualization,data cleaning, feature selection(correlation coefficient based analysis). Multiple ML models, including RandomnForest, SVM, Logistic Regression, Naive Bayes, Decision Tree and KNN are trained on the data and the model is assessed using key metrics such as F1 score, recall, precision, and accuracy. In simpler terms, these measurements help us understand how well the model is at correctly identifying threats (precision), capturing all actual threats (recall), achieving a balance between precision and recall (F1 score), and overall accuracy in distinguishing between harmful and normal network connections.


CKD Dataset
Data Preprocessing
Correlation Coefficient based feature selection
Training Data
Test Data
Stacking Ensemble Learning Based Prediction Model
Logistic Regression
KNN
Naive Byes
SVM
Decision Tree
Random Forest
Random Forest
Anomaly
Normal
Base Classifier
Meta Classifier

Classification Machine Learning Models

Classification models which are used as base classifiers for the stacking ensemble learning based intrusion detection model are as follows:

Logistic Regression

Logistic regression is a machine learning algorithm used for classifying data into one of two categories, often labeled as 0 and 1. It does so by estimating the probability that a given input belongs to the positive class (1). Instead of directly giving binary outputs of 0 or 1, logistic regression provides probabilistic values between 0 and 1.

K Nearest Neighbour

K-Nearest Neighbors (KNN) is a machine learning algorithm used for both classification and regression tasks. It is a non-parametric and instance-based learning algorithm, meaning it makes predictions based on the similarity between data points in the feature space.A labeled dataset, where each data point has features (attributes) and a corresponding target label (in classification) or a target value (in regression) is chosen. KNN operates in a multi-dimensional feature space where each dimension corresponds to a feature.The "K" in KNN represents the number of nearest neighbors to consider when making predictions. This is a user-defined parameter, and selecting an appropriate value for K is important. A smaller K can make the algorithm more sensitive to noise, while a larger K can smooth out the decision boundaries.KNN uses a distance metric (usually Euclidean distance, but other metrics like Manhattan or cosine similarity can be used) to measure the similarity between data points. The distance between two data points is calculated in the feature space.
For a classification task, when you want to predict the class label of a new data point, KNN identifies the K-nearest data points in the training set based on the chosen distance metric. It then counts the occurrences of each class among these neighbors and assigns the class that appears most frequently as the predicted class for the new data point.

Naive Bayes

Naive Bayes is a classification algorithm that calculates the probability of an object belonging to a particular category based on its characteristics. It assumes that these characteristics (or features) are independent of each other, simplifying calculations. During training, it learns the probabilities associated with each feature and class, and then, for a new data point, computes the probability of it belonging to each class. The class with the highest probability is predicted as the outcome. Naive Bayes is commonly used in text classification and spam detection due to its simplicity and efficiency, though its assumption of feature independence may not always hold true in real-world data.

Decision Tree

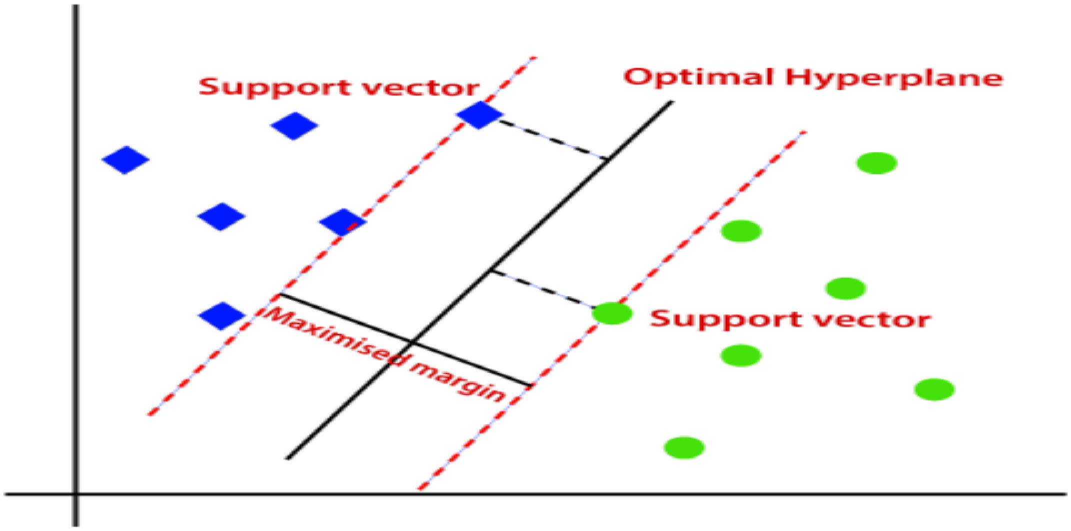
The Decision Tree algorithm is used for both classification and regression tasks.It is a tree-like structure which has branches, internal nodes,root nodes and leaf nodes.Internal nodes are also known as decision nodes. All the possible outcomes within the dataset are represented by the leaf node. Depending on the decision, we will move to the right or left branch when we are on an internal node based on a decision rule.

Random Forest

Random Forest is a powerful machine learning algorithm used for both classification and regression tasks. It's like a group of decision trees working together to make more accurate predictions. Each tree in the "forest" is trained on a different subset of the data, and they vote or average their predictions to reach a final result. This helps reduce overfitting and increases the model's robustness.

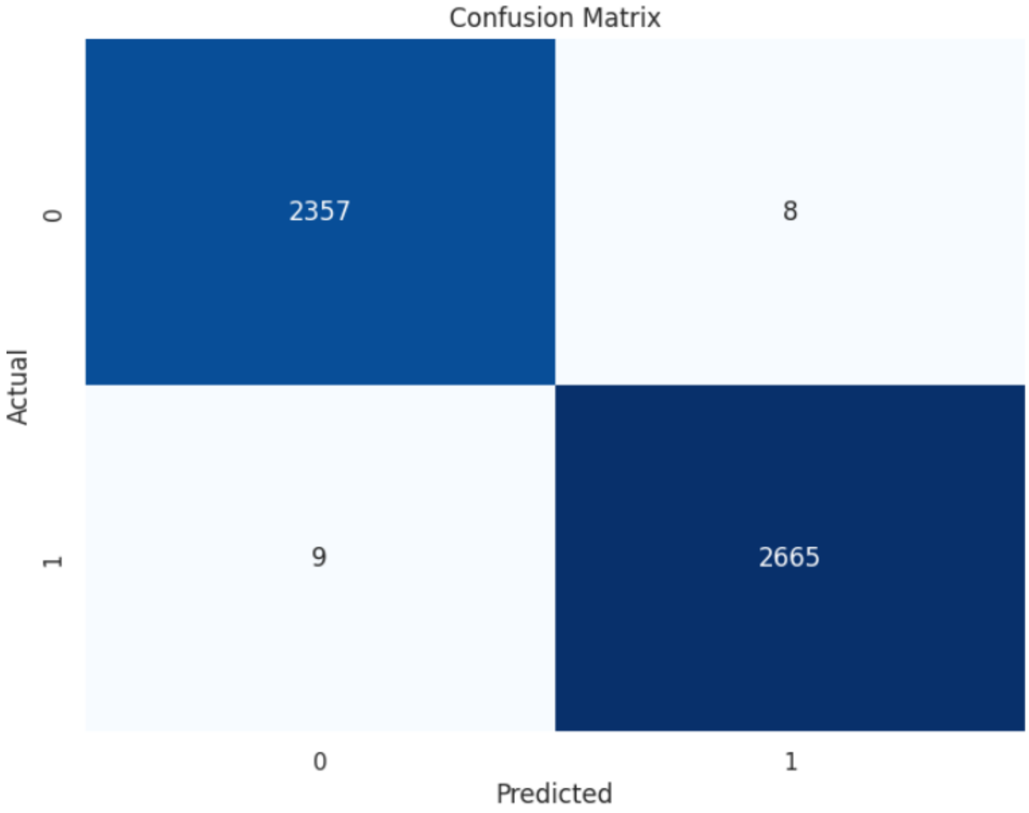
Support Vector Machine

A Linear Support Vector Machine (Linear SVM) is a machine learning algorithm used for classification tasks. It's like a smart line-drawing tool that separates data points into different categories. The goal is to find the best straight line (or hyperplane) that maximizes the gap between different classes of data. This "gap" is called the margin, and Linear SVM strives to create the widest possible margin while correctly classifying data points.
Random forest classifier is used as the meta classifier in stacking of the above mentioned algorithms.



Building Confusion Matrix

Confusion matrix is basically a performance indicator for classification problems. As we are having a binary classification problem,So the matrix formed is of order 2x2.



Experiments and Result Analysis

Our experiments were conducted on a real-world dataset containing both normal and malicious network traffic.The stacking ensemble was trained using a portion of the dataset, and the remaining data were used for testing.

The results demonstrate a remarkable accuracy rate of 99.66% for our Intrusion Detection System. This high accuracy indicates the system's effectiveness in identifying both known and previously unseen attacks. The performance was evaluated using metrics such as precision, recall, and F1-score.

Classifier	Precision	Recall	F1 Score	Accuracy
Random Forest	0.995888	0.996260	0.996074	0.995833
KNN	0.991766	0.991025	0.990871	0.990871
Decision Tree	0.995126	0.992521	0.993821	0.993451
SVM	0.976952	0.982797	0.978563	0.978567
Naive Bayes	0.902629	0.911743	0.907163	0.900972
Logistic Regression	0.951381	0.965969	0.958619	0.955745
Stacked Model	**0.997007**	**0.996634**	**0.996821**	**0.996626**

Conclusion

This project presents a robust Intrusion Detection System based on Machine Learning with a stacking ensemble approach. The utilization of multiple base learners, including Random Forest, Decision Tree, Logistic Regression, and SVM, in combination with feature selection, leads to exceptional accuracy in identifying network intrusions. This system provides an essential layer of security for safeguarding critical computer networks.

References

[1] Siamak Layeghy, Marius Portmann,2023, 'Explainable Cross-domain Evaluation of ML-based Network Intrusion Detection Systems' .
[2] Lee, W., Stolfo, W. 'A framework for constructing features and models for intrusion detection systems', ACM Trans. Inf. Syst. Sec., 2000, 3, (4), pp. 227–261.
[3] A. Lazarevic, L. Ertoz, A. Ozgur, J. Srivastava& V. Kumar, "A Comparative Study of Anomaly Detection Schemes in Network Intrusion Detection", Proceedings of Third SIAM Conference on Data Mining, San Francisco, May 2003.
[4] Li Yang, Abdallah Shami,2022,'IDS-ML: An open source code for Intrusion Detection System development using Machine Learning'.
[5]Sowmya T,Mary Anita E.A.,'A comprehensive review of AI based intrusion detection system',2023.
[6]Andrew H. Sung, Srinivas Mukkamala , 'Identifying important features for intrusion detection using support vector machines and neural networks',Proceedings of the 2003 Symposium on Applications and the Internet.

[7]Mr Mohit Tiwari,Raj Kumar, Akash Bharti, Jai Kishan,'INTRUSION DETECTION SYSTEM',International Journal of Technical Research and Applications,2017.
[8]Ibraheem, Ibraheem. (2022). Enhancing Intrusion Detection Systems using Ensemble Machine Learning Techniques. Data and Metadata. 1. 33. 10.56294/dm202271.
[9]Kalinin, Maxim & Krundyshev, Vasiliy. (2022). Security intrusion detection using quantum machine learning techniques. Journal of Computer Virology and Hacking Techniques. 19. 1-12. 10.1007/s11416-022-00435-0.
[10]Jafferson, Prethija & Subbulakshmi, V & Devi, K. (2023). Datasets used for Intrusion Detection using Machine Learning: A Survey. International Journal of Research in Engineering and Technology.
[11]Wu, Weifei & Zhang, Yanhui. (2023). An efficient intrusion detection method using federated transfer learning and support vector machine with privacy-preserving. Intelligent Data Analysis. 27. 1-21. 10.3233/IDA-226617.
[12]Prajapati, Pratik & Singh, Ishanika & Subhashini, N.. (2023). Network Intrusion Detection Using Machine Learning. 10.1007/978-981-19-8338-2_4.
[13]Vanin, Patrick & Newe, Thomas & Dhirani, Lubna & O'Connell, Eoin & O'Shea, Donna & Lee, Brian & Rao, Muzaffar. (2022). A Study of Network Intrusion Detection Systems Using Artificial Intelligence/Machine Learning. Applied Sciences. 12. 11752. 10.3390/app122211752.
[14]Abdallah, Emad & Eleisah, Wafa' & Otoom, Ahmed. (2022). Intrusion Detection Systems using Supervised Machine Learning Techniques: A survey. Procedia Computer Science. 201. 205-212. 10.1016/j.procs.2022.03.029.
[15]Xu, Hao & Sun, Zihan & Cao, • & Bilal, Hazrat. (2023). A data-driven approach for intrusion and anomaly detection using automated machine learning for the Internet of Things. Soft Computing. 10.1007/s00500-023-09037-4(.
[16]Somashekar, Harshitha & Boraiah, Ramesh. (2023). Network intrusion detection and classification using machine learning predictions fusion. Indonesian Journal of Electrical Engineering and Computer Science. 31. 1147. 10.11591/ijeecs.v31.i2.pp1147-1153.
[17]Čiurlienė, Karina & Stankevičius, Denisas. (2023). Network intrusion detection using hybrid machine learning methods. Mokslas - Lietuvos ateitis. 15. 1-9. 10.3846/mla.2023.19385.
[18]Malhotra, Heena & Sharma, Prabha. (2019). Intrusion Detection using Machine Learning and Feature Selection. International Journal of Computer Network and Information Security. 11. 43-52. 10.5815/ijcnis.2019.04.06.
[19]Surbhi Solanki,Chetan Gupta,Kalpana Rai(2020). "A Survey on Machine Learning based Intrusion Detection System on NSL-KDD Dataset."International Journal of Computer Applications.
[20]James Cannady, Jay Harrell," A Comparative Analysis of Current Intrusion Detection Technologies".
[21]Lianming Zhang, Kui Liu, Xiaowei Xie, Wenji Bai, Baolin Wu, Pingping Dong,
A data-driven network intrusion detection system using feature selection and deep learning, Journal of Information Security and Applications,Volume 78,2023,103606,ISSN 2214-2126, https://doi.org/10.1016/j.jisa.2023.103606.
[22]Sara Mohammadi, Hamid Mirvaziri, Mostafa Ghazizadeh-Ahsaee, Hadis Karimipour,
Cyber intrusion detection by combined feature selection algorithm,Journal of Information Security and Applications,Volume 44,2019,Pages 80-88,ISSN 2214-2126,

https://doi.org/10.1016/j.jisa.2018.11.007.
[23]Fatemeh Amiri, MohammadMahdi Rezaei Yousefi, Caro Lucas, Azadeh Shakery, Nasser Yazdani,Mutual information-based feature selection for intrusion detection systems,Journal of Network and Computer Applications,Volume 34, Issue 4,2011,Pages 1184-1199,ISSN 1084-8045,https://doi.org/10.1016/j.jnca.2011.01.002.
[24]Chaouki Khammassi, Saoussen Krichen,A GA-LR wrapper approach for feature selection in network intrusion detection,Computers & Security,Volume 70,2017,Pages 255-277,ISSN 0167-4048,https://doi.org/10.1016/j.cose.2017.06.005.