Int. J. Pure Appl. Sci. 9(1);209-222 (2023)

IJPAS
ijpas@munzur.edu.tr
ISSN: 2149-0910

Research article/Araştırma makalesi
DOI:10.29132/ijpas.1278880

Comparison of Performance of Classification Algorithms Using Standard Deviation-based Feature Selection in Cyber Attack Datasets

Ali Şenol
Tarsus University, Faculty of Engineering, Departmant of Computer Engineering, Mersin, Türkiye
alisenol@tarsus.edu.tr

Received date:07.04.2023, Accepted date: 14.06.2023

Abstract

Supervised machine learning techniques are commonly used in many areas like finance, education, healthcare, engineering, etc., because of their ability to learn from past data. However, such techniques can be prolonged if the dataset is high-dimensional, and irrelevant features may reduce classification performance. Therefore, feature selection or feature reduction techniques are commonly used to overcome the mentioned issues. Hence, feature selection approaches are needed to make the algorithms faster without reducing the classification performance. On the other hand, information security for both people and networks is crucial and must be secured without wasting time. In this study, we compare the classification and run-time performances of state-of-the-art classification algorithms using standard deviation-based feature selection in security datasets. For this purpose, we applied standard deviation-based feature selection to KDD Cup 99 (KDD) and Phishing Legitimate datasets for selecting the most relevant features. Then we ran the selected classification algorithms on the datasets to compare the results. According to the obtained results, while the classification performances of all algorithms were satisfying, Decision Tree (DT) was the best among others. On the other hand, while DT, k Nearest Neighbors (kNN), and Naïve Bayes (NB) were sufficiently fast, Support Vector Machine (SVM) and Artificial Neural Networks (ANN) were too slow.

Keywords: Classification, cyber security, feature selection, information security, machine learning.

Siber Saldırı Veri Kümelerinde Standart Sapmaya Dayalı Öznitelik Seçimi Kullanan Sınıflandırma Algoritmalarının Performanslarının Karşılaştırması

Öz

Denetimli makine öğrenimi teknikleri, geçmiş verilerden öğrenme yetenekleri nedeniyle finans, eğitim, sağlık, mühendislik vb. pek çok alanda yaygın olarak kullanılmaktadır. Ancak, veri kümesi çok boyutlu ise bu tür teknikler çok yavaş olabilir ve alakasız özellikler nedeniyle de sınıflandırma başarısı düşebilir. Bu nedenle, bahsedilen sorunların üstesinden gelmek için öznitelik seçme veya nitelik azaltma teknikleri yaygın olarak kullanılmaktadır. Öte yandan, bilgi güvenliği hem insanlar hem de ağlar için çok önemlidir ve zaman kaybetmeksizin güvence altına alınması gerekir. Bu nedenle, sınıflandırma başarısını düşürmeden algoritmaları hızlandırabilen öznitelik seçim yaklaşımlarına ihtiyaç duyulmaktadır. Bu çalışmada, güvenlik veri kümeleri açısından standart sapmaya dayalı öznitelik seçimi kullanan en temel sınıflandırma algoritmalarının hem sınıflandırma başarılarını hem de çalışma zamanı performanslarını karşılaştırdık. Bu amaçla KDD Cup 99 (KDD) ve Phishing Legitimate veri setlerine standart sapma tabanlı öznitelik seçimi uygulayarak en ilgili nitelikleri seçtik ve seçilen sınıflandırma algoritmalarını veri setlerine uygulayarak sonuçları karşılaştırdık. Elde edilen sonuçlara göre, tüm algoritmaların sınıflandırma başarıları tatmin edici iken, Karar Ağacı (DT) diğerleri algoritmalara göre en iyisi olarak dikkat çekmiştir. Bununla birlikte, DT, k En Yakın Komşu (kNN) ve Naïve Bayes (NB) tatmin edici düzeyde hızlıyken, Destek Vektör Makinesi (SVM) ve Yapay Sinir Ağları'nın (ANN) çok yavaş oldukları tespit edilmiştir.

Anahtar Kelimeler: Bilgi güvenliği, makine öğrenmesi, öznitelik seçimi, sınıflandırma, siber güvenlik.

Research article/Araştırma makalesi
DOI:10.29132/ijpas.1278880

INTRODUCTION

Machine learning approaches are commonly used in many areas like bioinformatics, image processing, financial applications, science applications, healthcare systems, information security, etc. (Deiana et al., 2022; Heidari, Jafari Navimipour, Unal, Toumaj and Applications, 2022; Khaire, Dhanalakshmi and Sciences, 2022; Şenol, Canbay and Mahmut; Zhou, Wang and Zhu, 2022). One of these approaches is classification techniques. In classification applications, algorithms learn information from past data and use this information to predict new arrival data (Şenol et al.). NB (Russell, 2010), DT (Fürnkranz, 2017), SVM (Manevitz and Yousef, 2001), kNN (Ali, Neagu and Trundle, 2019), and ANN (Jain, Mao and Mohiuddin, 1996) can be given as primary examples of classification algorithms.

Although classification algorithms are very successful, they may be too slow on high-dimensional datasets (Cheng, Cui, Wang and Zhang, 2023). Various feature selection and feature reduction methods have been proposed to overcome this issue. In the feature reduction approaches, proposed techniques project the dimensions to reduce the dimensionality (Di Mauro, Galatro, Fortino and Liotta, 2021). Linear Discriminant Analysis (LDA) and Principal Component Analysis (PCA) can be basic examples of these methods. On the other hand, in feature selection techniques, a subset of features is selected instead of using all features according to relativity to the actual classes. Correlation-based feature selection (CFS), Mutual Information (MI), Information Gain (IG), Chi-square test (Chi2), and Variance Threshold-based (VT) feature selection methods can be said as examples of them (Şenol, 2022a).

When cyber-attacks against the right of society are considered, the value of any approach that might support people against these kinds of attacks can be understood better. Machine learning-based approaches are one of them. Machine learning-based techniques could successfully detect any attack in real-time (Almaiah, Al-Zahrani, Almomani and Alhwaitat, 2021; Bahaa, Abdelaziz, Sayed, Elfangary and Fahmy, 2021; Uma and Padmavathi, 2013). Therefore, they are commonly used in information security and cyber attack areas (Abdullahi et al., 2022; Ansari, Sharma and Dash, 2022). This study compares state-of-the-art classification algorithms with standard deviation-based feature selection on cyber attack datasets. This study aims to reveal the efficiency of classification performances of state-of-the-art algorithms with a standard deviation-based feature selection approach on cyber attack datasets. For this purpose, we used two cyber attack datasets in the experimental study.

The rest of the paper is organized as follows. The next section provides background information about used techniques and algorithms. The proposed model is described in detail in the third section, while the experimental study is shared in the fourth section. We discuss obtained results in the fifth section, while the study is concluded in the last section.

RELATED WORKS

In parallel with technological developments, the number of cyber-attacks is increasing. So, the need for methods to protect people from such attacks is also increasing. Additionally, since the speed of data growth is enormous, these methods require various feature selection or reduction methods to increase the run-time performance of the models (Lyu, Feng, and Sakurai, 2023). Therefore, there are many studies have been conducted in this area.

One of these studies is proposed by Li et al. (Li, Fang, Chen and Guo, 2006). They used Maximum Entropy Model (ME) with IG and Chi2 feature selection methods to classify the KDD dataset. According to the results, their model's accuracy was 99.82%. Moreover, since the computational complexity of their method was low, it could be used for real-time applications. Similarly, Niguyen et al. (Niguyan, Franke and Petrovic, 2010) used DT and NB classifiers with hybrid versions of CFS to classify the KDD dataset. The DT's accuracy was better than that of NB, which were 99.41% and 98.82%, respectively.

In another study, Eid et al. (Eid, Hassanien, Kim and Banerfee, 2013) used Pearson Correlation with DT to classify the NSL-KDD dataset. According to the experimental results, the accuracy of their model was 99.1%. In addition, their model reduced the number of features from 41 to 17. However, the run-time of their model was high a little bit. Another proposed model to classify KDD-NLS dataset was proposed by Wahba et al. (Wahba, ElSalamouny and ElTaweel, 2015). Their model used CFS and IG as feature selection methods and NB as classifier. According to the results, their classsifier's performance was better than Eid et al's model. The accuracy of their model was 99.3% in

IJPAS
ijpas@munzur.edu.tr
ISSN: 2149-0910

Research article/Araştırma makalesi
DOI:10.29132/ijpas.1278880

accuracy. Similarly, Shahbaz et al. (Shahbaz, Wang, Behnad, Samarabandu, 2016) used DT with CFS and tested it on NSL-KDD. However, the accuracy of their model was too low. Because, the number of selected features in their model was only 4. In addition to the NSL-KDD dataset, Ullah and Mahmoud (Ullah and Mahmoud, 2017) tested their model on the ISCX dataset, which is also an intrusion detection dataset. They integrated IG into J48 classifier. The accuracy of their model on ISCX was 99.70%, while it was 99.90% on NSL-KDD.

Kushwaha et al. (Kushwaha, Buckchash, Raman, 2017) ensembled SVM with IG feature selection method. Then, they tested the model on KDD dataset. According to obtained results, the accuracy of their model was very high, which was 99.91%, although the number of selected features was 5. Another model that used SVM as classifier was proposed by Mohammadi et al. (Mohammadi, Desai and Karimipour, 2018). To improve the accuracy of their model, they used Least Squared SVM (LSSVM). In addition to KDD and NSL-KDD, they tested their model on Kyoto+ 2006 dataset. The accuracies of their model on these datasets were 94.31%, 98.31%, and 99.11%, respectively. Similarly, Wang et al. (Wang, Du and Wang, 2019) applied SVM to KDD and NSL-KDD datasets. However, they used a feature selection method called Efficient CFS which was based on symmetric uncertainty. The accuracy performance of their model was sufficient on both datasets.

On the other hand, Shabudin et al. (Shabudin, Arrifin, Sani and Aliff, 2020) proposed a study combining various feature selection methods with Random Forest (RF), Multi-Layered Perceptrons (MLP), and NB classifiers. Then, they tested their model on the Phishing Website dataset of UCI data repository. According to the experimental study, RF was the most successful classifier. On the other hand, Aljabri and Mirza (Aljabri and Mirza, 2022) proposed Machine Learning and Deep Learning with correlation-based feature selection for phishing attack detection. They used RF, Logistic Regression, Convolutional Neural Network, ANN, and SVM as classifiers while using two phishing datasets from UCI and Kaggle websites. According to the experimental comparisons, the RF was the most successful classifier.

METHODOLOGY

In this section, we provide details about the method and algorithms used in this study.

Classification Algorithms

This study compares some of the most known classification algorithms with standard deviation-based feature selection on two cyber-attack datasets. We used DT, SVM, kNN, NB, and ANN as classification algorithms. Because these algorithms are some of the most renowned classifiers.

DT is a supervised machine-learning algorithm for classification and regression tasks (Lee, Cheang and Moslehpour, 2022; Rivera-Lopez, Canul-Reich, Mezura-Montes, Cruz-Chávez and Computation, 2022). It is a tree-like structure where each internal node represents a feature or attribute, each branch represents a decision rule, and each leaf node represents a class label or a numerical value (Fürnkranz, 2017). It is easy to use, speedy to train and test, and also easy to understand. Furthermore, the resultant tree can be visualized easily.

On the other hand, SVM is a supervised machine-learning algorithm that can be used for classification and regression tasks. In SVM, the algorithm finds a hyperplane that best separates the different classes or predicts the continuous output values based on the input features. The hyperplane is chosen to maximize the margin between the closest points from other classes. SVM can handle linear and non-linearly separable datasets by transforming the input features into a higher-dimensional space where the classes become linearly separable. This is done using a kernel function that calculates the similarity between two instances in the higher-dimensional space. Some popular kernel functions are linear, polynomial, Gaussian (RBF), and sigmoid.

Come to kNN, it is based on the idea that assumes similar things exist nearby, which means that similar things are close to each other (Ali et al., 2019). In addition to being used in many areas, it is commonly used in pattern recognition (Maheswari, Aluvalu and Mudrakola, 2022; Malik, Abu Bakar and Sheikh, 2022; Patil and Patil, 2022). Its most crucial advantage is that it does not require the datasets to be linearly separable.

As for NB, it is a statistical classification algorithm based on the Bayes theorem. The NB algorithm statistically learns from the training dataset. It is a strong classifier in terms of accuracy and speed. NB assumes that the input features are conditionally

IJPAS
ijpas@munzur.edu.tr
ISSN: 2149-0910

Research article/Araştırma makalesi
DOI:10.29132/ijpas.1278880

independent given the class label, which means that the probability of observing a set of features given a class label can be calculated by multiplying the probabilities of each feature given that class label. This assumption makes the algorithm computationally efficient and reduces the need for large training data.

Finally, ANN is a bio-inspired classification algorithm that models the neural nervous system of humans. ANN is a machine learning algorithm based on modeling the structure and function of the human brain. They are composed of interconnected nodes, or neurons, organized into layers. Each neuron takes in one or more input signals, performs a computation, and outputs a signal to the next layer of neurons. It uses a mathematical weighting system to modulate the effect of the associated input signal. ANN can be used for classification and regression tasks and can handle structured and unstructured data such as images and text. They are often used in applications such as computer vision, natural language processing, and speech recognition, where they have achieved state-of-the-art performance. However, they can be computationally expensive to train and require large amounts of data.

Feature Selection Methods

In general, feature selection techniques are approaches that reduce the dataset's number of features. The main aim is to reduce the number of features without reducing the dataset's quality. Furthermore, this process reduces the time complexity and uses less storage while obtain satisfactory results. Mainly, feature selection techniques are divided into three types. These approaches are filters, wrappers, and embedded methods (Çetin and Yıldız, 2022; Khaire et al., 2022).

Filter methods: Unless using all the features, they pick up instinct features through univariate statistics. They are efficient methods in terms of speed. These methods are faster and less computationally when compared with wrapper methods. IG, Chi2 test, Fisher's score, CFS, VT, Mean Absolute Difference (MAD), and Dispersion ratio are some of filter-based feature selection methods (Khaire et al., 2022).

Wrapper methods: In wrapper methods, possible all subsets of features are assessed, and then the subset which produces the best results is assigned as the selected features (Kira and Rendell, 1992; Kohavi and John, 1997). These techniques use classifiers and assessment techniques to find the best subset. In terms of dataset quality, they are better than filtering techniques. But the required time and space complexity of wrapper methods is greater than filter methods. These methods include forward feature selection, backward feature elimination, exhaustive feature selection, and recursive feature selection.

Embedded methods: These methods combine filter and wrapper methods' benefits by including features' interactions while maintaining reasonable computational costs. Embedded methods use an iterative method to determine the best features contributing the most to training for a given iteration. LASSO Regularization (L1) (Tibshirani, 1996) and Random Forest Importance (Breiman, 2001) can be given as examples.

Proposed Method

This study aims to reveal the effect of using a standard deviation-based feature selection method with state-of-the-art classification algorithms on security datasets. First, we used a standard deviation-based feature selection explained in the following subsection to find the best features. Then, we ran the algorithms on selected features. Finally, we compared the performances of the models.

Standard Deviation-Based Feature Selection Method

The standard deviation of any dataset shows how the data is gathered around the center. The larger the standard deviation of a feature, the greater its effect on the result (Şenol, 2022b). Therefore, if a feature has a large standard deviation, it also has more distinguishability on the results, as shown in Fig. 1 (Yousefpour, Ibrahim, Abdull Hamed and Hajmohammadi, 2014). In this example, the *Y* feature has more distinguishability. So, we try to select features with a large standard deviation as much as possible in standard deviation-based feature methods. In this study, we use a user-defined variable, ratio, calculated by dividing the standard deviation value of the selected one by the summation of the standard deviation of all features. Standard deviation is

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})}{n-1}} \qquad (1)$$

where σ is the standard deviation, n is the data size, x_i is each feature's value, and x is the mean of i^{th} feature.

Research article/Araştırma makalesi
DOI:10.29132/ijpas.1278880

After calculating the standard deviation for each feature, all features are sorted in descending order. The feature selection progress is continued until the ratio of the sum of the standard deviations of the selected features to the sum of the standard deviations of all the features is greater than the ratio. When the process is finished, the selected features are the final features that will be processed. Therefore, the algorithm we use for the standard deviation-based feature selection method is given in Algorithm 1.

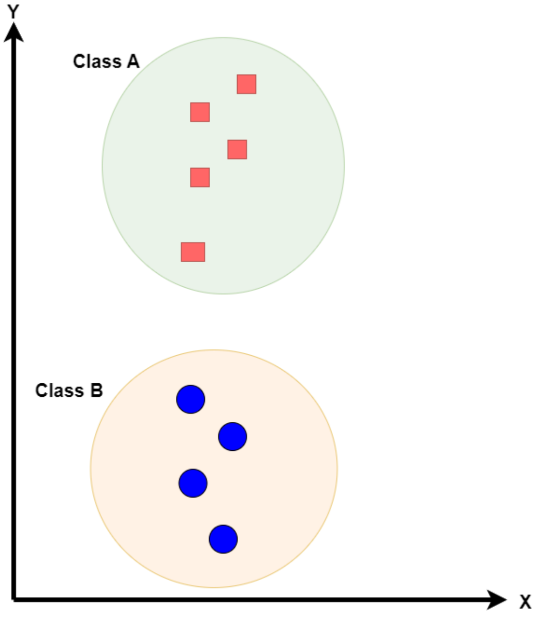


Figure 1. An example of standard deviation with two features and two classes.

Algorithm 1: Standard Deviation based FS

```
Input: Data $X = x_1, x_2, x_3, \ldots, x_n \subseteq \mathbb{R}^d$;
ratio;                              ▷std percentage
Output: SF;   ▷the List of Selected Features
foreach $\sigma_j \in \sigma$ do
    $\sigma_j = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})}{n-1}}$
end
Features ← DesOrder($\sigma$); ▷Descending order
foreach $f \in Features$ do
    if $sum \leq ratio$ then
        sum=sum+$f$;
        SF← $f$;
    end
end
return SF;
```

EXPERIMENTAL STUDY

Experimental Environment

We used Python programming language in the experimental study in Anaconda Spyder environment and required libraries. All the experimental works were performed on a computer that has i7 processor, 16 GB RAM, and on which Microsoft Windows 11 was installed.

Used Datasets

We used two cyber-attack datasets to compare the efficiency of classification algorithms on security datasets. One is the KDD dataset, and the other is the Phishing Legitimate dataset.

KDD Cup 99 dataset

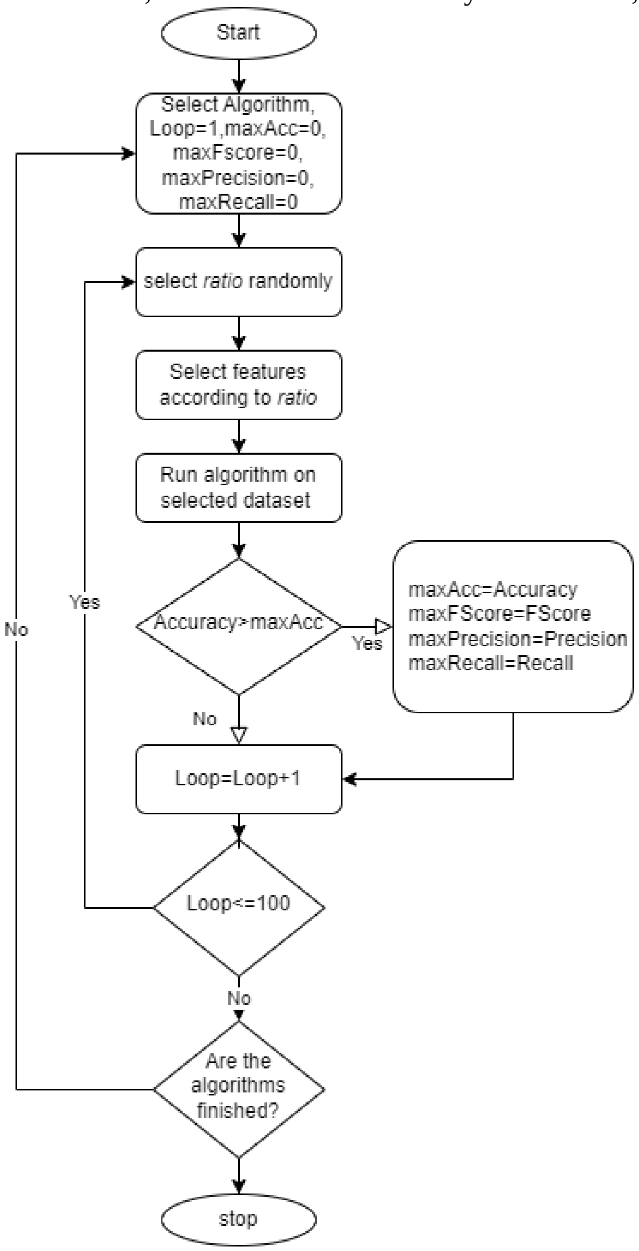
KDD dataset is one of the commonly used datasets in machine learning applications. It is the dataset related to the intrusion detection system. Intrusion detection systems aim to detect any attack carried out via network systems. In its original form, it consists of 5 million records, each consisting of 39 features, and one of these features is the class label that could be one of 24 classes (Dua and Graff, 2023). This study uses a subset of the KDD dataset with 50000 records to speed up the experimental process.

Phishing Legitimate dataset

Phishing is one of the cyber attacks against human information security. These attacks aim to steal people's valuable information like bank account

Research article/Araştırma makalesi
DOI:10.29132/ijpas.1278880

information, account information on any social media, or any other information that could be valuable for the attacker (Ojewumi et al., 2022). Phishing Legitimate dataset contains records related to phishing attacks (Tan, 2018). According to features, it is decided if it was distrustful or malicious. It has 48 features and one for the class label. It has 10 thousand records, 5 thousand are reliable, and 5 thousand are malicious.


Start
Select Algorithm, Loop=1,maxAcc=0, maxFscore=0, maxPrecision=0, maxRecall=0
select *ratio* randomly
Select features according to *ratio*
Run algorithm on selected dataset
Accuracy>maxAcc
Yes
maxAcc=Accuracy
maxFScore=FScore
maxPrecision=Precision
maxRecall=Recall
No
Loop=Loop+1
Loop<=100
Yes
No
Are the algorithms finished?
No
stop

Test Procedure

We performed all selected classification algorithms on both datasets in the experimental study and compared the results. Firstly, we split the datasets as 66% as the train and %34 as the test dataset. Then, to evaluate the results, we run each algorithm on each dataset 100 times with a randomly selected ratio value and parameters. The procedure's objective is to reach the highest value for each algorithm. The highest value of obtained accuracy means the best feature selection ratio is the used one. The procedure that was used to compare algorithms is shared in Fig. 2.

Figure 2. Used procedure to compare the algorithms.

RESULTS AND DISCUSSION

5.1. Comparison of Classifiers in Terms of Classification Performance

All selected algorithms are executed on each dataset according to the procedure in Fig. 2. While obtained, visual results with the DT algorithm are given in Fig. 3, 4, 5, and 6, respectively; all the obtained results are given in Table (1). DT is the most successful one in classification manner on both datasets. According to obtained results, the number of the selected features as the best was 25 for DT in the KDD dataset, while it was 25, 29, 29, and 23 for NB, SVM, kNN, and ANN, respectively. As for the Phishing Legitimate dataset's selected features were 20, 42, 42, 40, and 40 for DT, NB, SVM, kNN, and ANN, respectively. When the results are analyzed, it can be said that the feature selection operation makes DT, NB, and kNN faster. On the other hand, it is not an advantageous way for SVM and ANN.

5.2. Comparison of Classifiers in Terms of Run-time Complexity

As shown in Fig. 7, in terms of execution time, kNN was the fastest one among them. But DT and NB were also sufficiently fast in both datasets. On the other hand, SVM and ANN were very slow when compared to the others. Besides, because of using the feature selection process, the times consumed by SVM and ANN with the feature selection method were even more significant than the times in the methods without the feature selection method in some cases. Therefore, regarding execution time, we can say that DT, NB, and kNN were sufficiently fast, but SVM and ANN were too slow on cyber-attack datasets.

CONCLUSION

Machine learning approaches are commonly used to protect people against cyber-attacks that threaten people's rights. Because machine learning can be used in detecting cyber attacks by training models on large datasets of network traffic and system logs to learn patterns of normal behaviors and abnormal behaviors. This can be done using machine learning techniques such as supervised, unsupervised, and reinforcement learning. Classification algorithms are one of the supervised approaches. On the other hand, feature selection/reduction techniques that are used to select more related features to reduce the execution time of algorithms are also widespread in the machine learning area.

This study compared state-of-the-art classification algorithms with a standard deviation-based feature selection method on cyber-attacks in both run-time complexities and classification performance. According to obtained results, DT is the most successful algorithm in classification performance. Its accuracy on the KDD dataset was 0.9992, and on the Phishing Legitimate dataset was 0.9721, while the accuracy of SVM on the KDD dataset was 0.9982, and the accuracy of ANN on the Phishing Legitimate dataset was 0.9497 that were the second-best results. Besides, the run-time Complexity of DT is also sufficient with kNN and NB algorithms. Therefore, we can say that DT using standard deviation-based feature selection is the most suitable classification algorithm for cyber-attack datasets.

ijpas@munzur.edu.tr
ISSN: 2149-0910

Research article/Araştırma makalesi
DOI:10.29132/ijpas.1278880

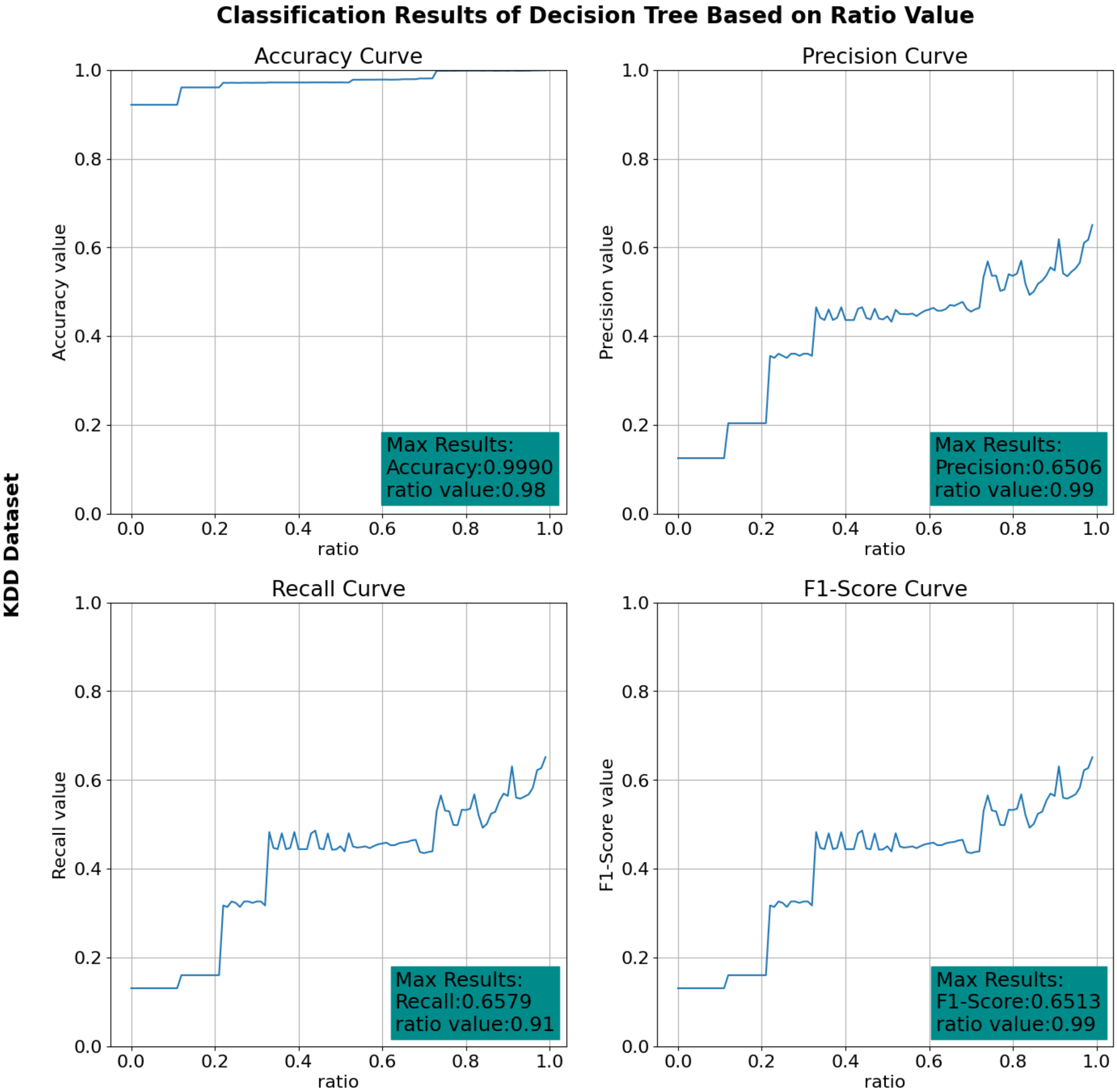
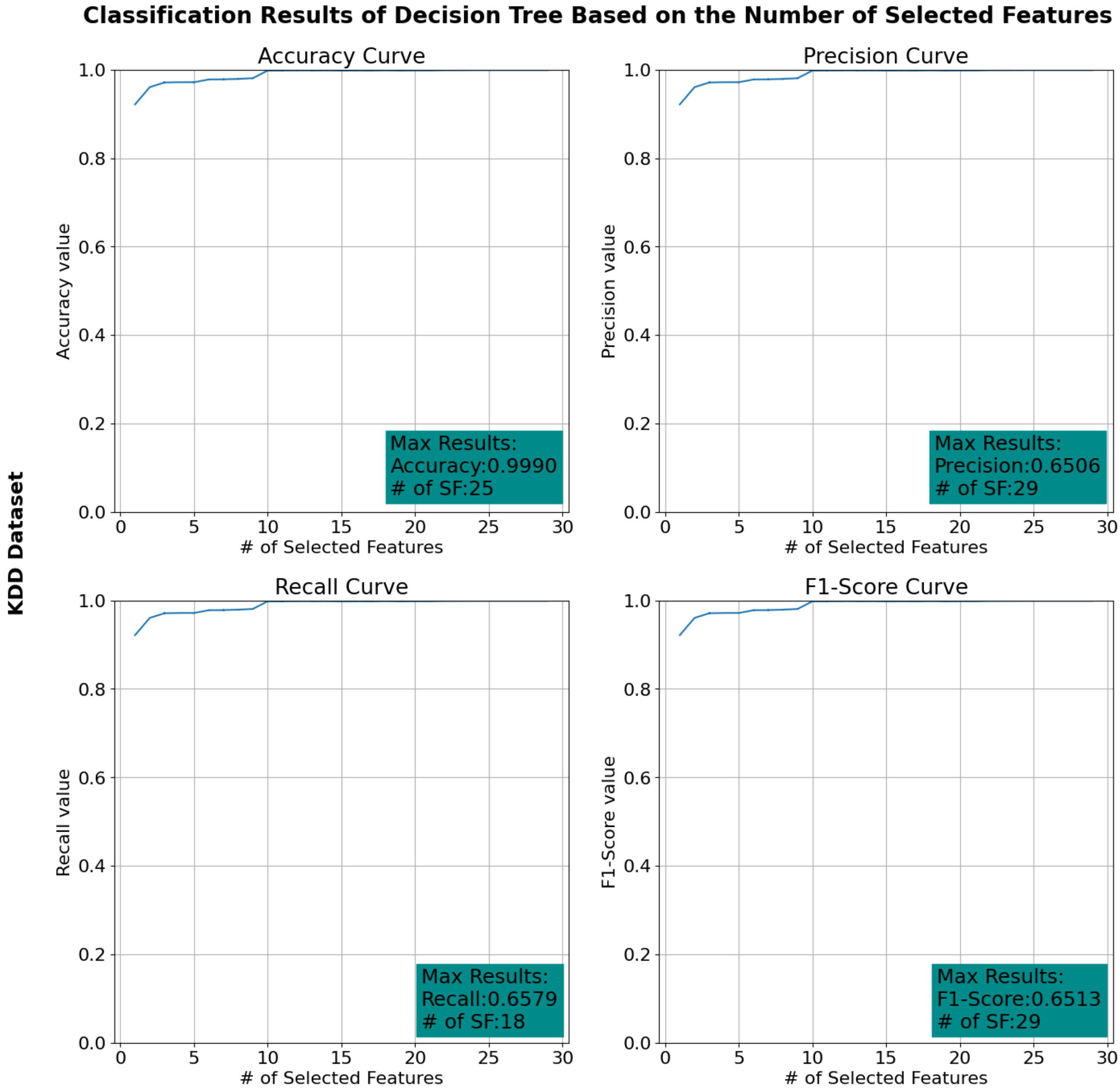


Figure 3. Classification results via ratio value using DT on KDD dataset.

Research article/Araştırma makalesi
DOI:10.29132/ijpas.1278880



# of SF: The number of selected features

Figure 4. Classification results via # of selected features using DT on KDD dataset.

IJPAS
ijpas@munzur.edu.tr
ISSN: 2149-0910

Research article/Araştırma makalesi
DOI:10.29132/ijpas.1278880

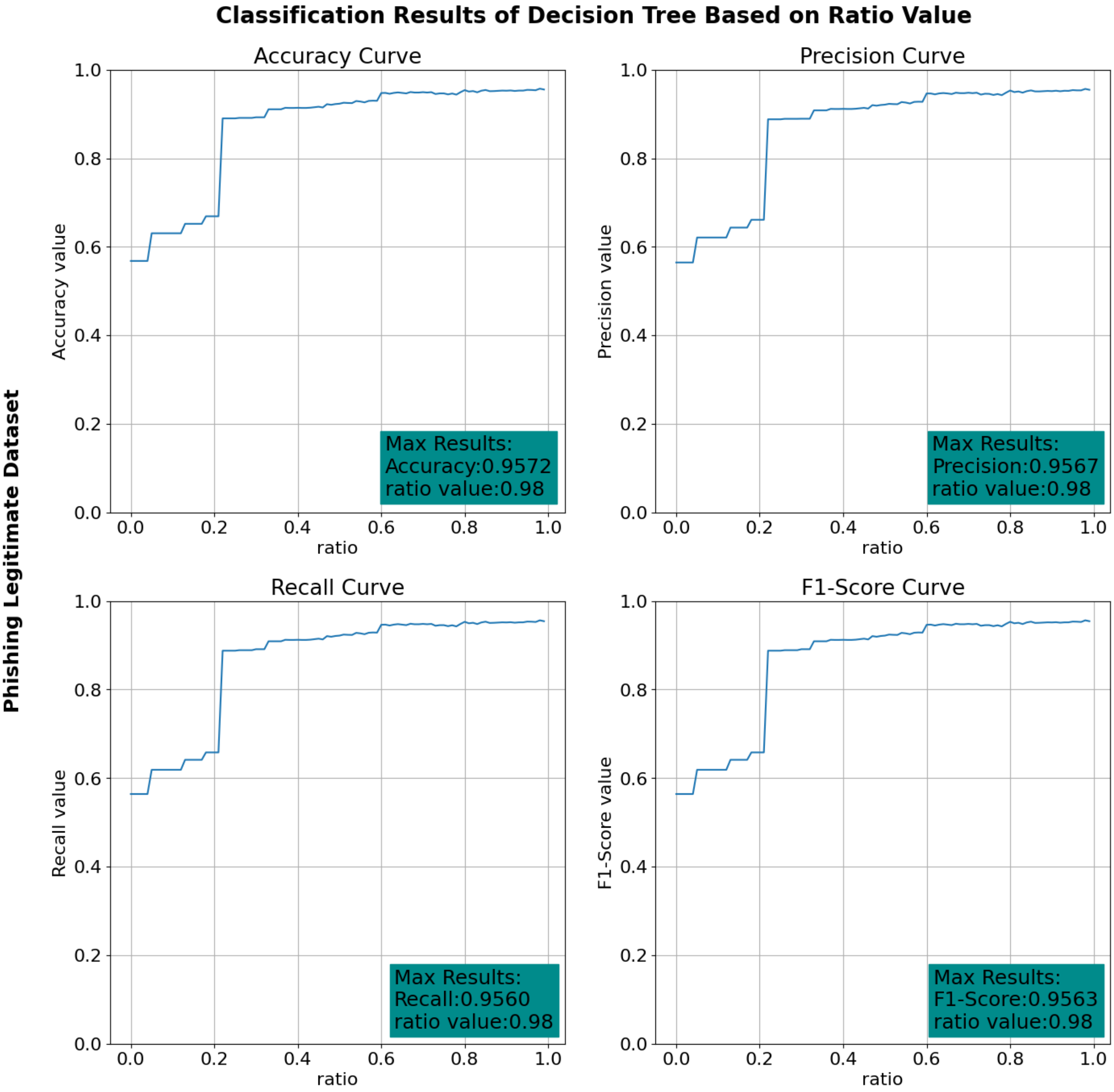
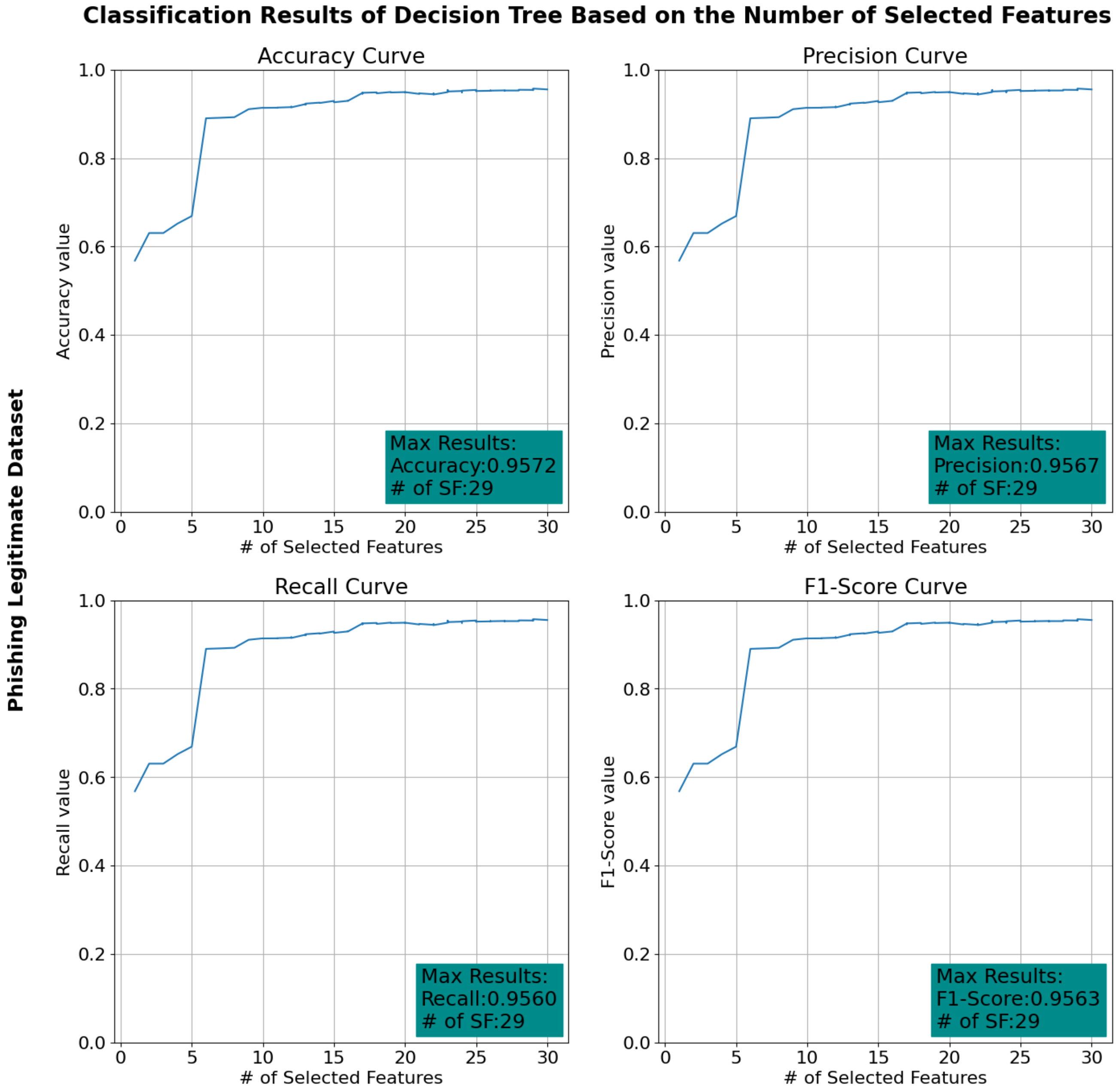


Figure 5. Classification results via ratio value using DT on Phishing Legitimate dataset.

ijpas@munzur.edu.tr
ISSN: 2149-0910



# of SF: The number of selected features

Figure 6. Classification results via # of selected features using DT on Phishing Legitimate.

Table 1. The best results of classifiers according to selected ratio values.

Datasets	Classifiers	ratio	# of Selected Features	Accuracy	F1-Score	Precision	Recall
KDD	DT	0.9800	25	**0.9992**	**0.6614**	**0.7022**	**0.6750**
	NB	0.9800	25	0.9016	0.5194	0.6965	0.5389
	SVM	0.9900	29	0.9982	0.5672	0.5577	0.5622
	kNN	0.9900	29	0.6305	0.3616	0.6369	0.3647
	ANN	0.9700	23	0.9979	0.5218	0.5426	0.5315
Phishing Legitimate	DT	0.7400	20	**0.9721**	**0.9722**	**0.9721**	**0.9721**
	NB	0.9800	42	0.8385	0.8519	0.8401	0.8374
	SVM	0.9800	42	0.9485	0.9484	0.9486	0.9485
	kNN	0.9700	40	0.8136	0.8291	0.8155	0.8120
	ANN	0.9700	40	0.9624	0.9624	0.9625	0.9624

Table 2. The best results of classifiers without feature selection

Datasets	Classifiers	# of Selected Features	Accuracy	F1-Score	Precision	Recall
KDD	DT	38	**0.9992**	**0.7293**	**0.7210**	**0.7250**
	NB	38	0.8737	0.5060	0.6808	0.5199
	SVM	38	0.9982	0.5672	0.5577	0.5622
	kNN	38	0.6407	0.3501	0.6385	0.349
	ANN	38	0.9882	0.2140	0.2382	0.2243
Phishing Legitimate	DT	48	**0.9694**	**0.9694**	**0.9694**	**0.9694**
	NB	48	0.8412	0.8563	0.8430	0.8400
	SVM	48	0.9482	0.9481	0.9482	0.9482
	kNN	48	0.8130	0.8281	0.8148	0.8114
	ANN	48	0.9497	0.9497	0.9497	0.9497

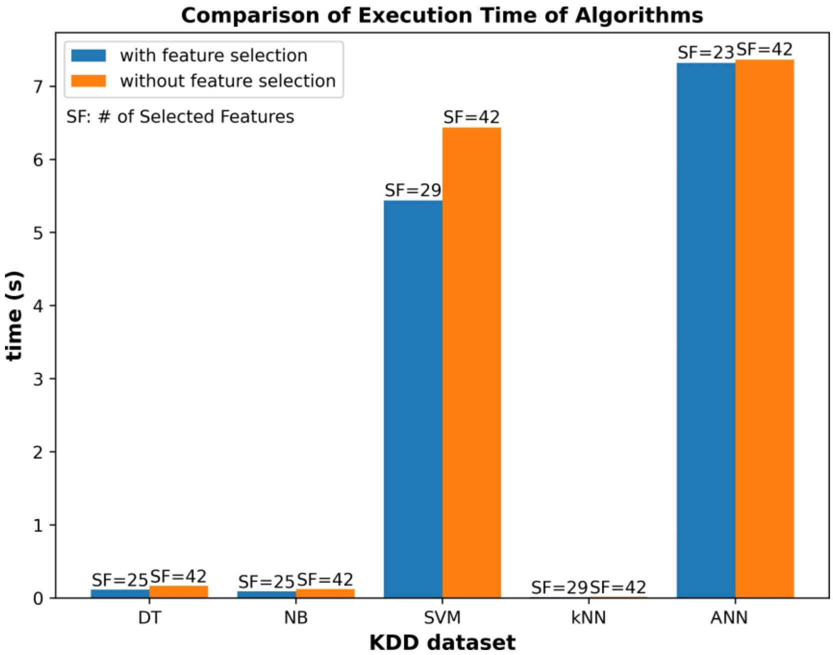


Figure 7. Comparison of algorithms in the aspect of execution time on KDD Dataset. Here, the value of the ratio which produces the highest Accuracy value was selected for each algorithm.

IJPAS
ijpas@munzur.edu.tr
ISSN: 2149-0910

Research article/Araştırma makalesi
DOI:10.29132/ijpas.1278880

CONFLICT OF INTEREST

The Author reports no conflict of interest relevant to this article

RESEARCH AND PUBLICATION ETHICS STATEMENT

The Author declares that this study complies with research and publication ethics.

REFERENCES

Abdullahi, M., Baashar, Y., Alhussian, H., Alwadain, A., Aziz, N., Capretz, L. F. and Abdulkadir, S. J. J. E. (2022). Detecting cybersecurity attacks in internet of things using artificial intelligence methods: A systematic literature review. 11(2), 198.

Ali, N., Neagu, D. and Trundle, P. J. S. A. S. (2019). Evaluation of k-nearest neighbour classifier performance for heterogeneous data sets. 1, 1-15.

Aljabri, M. and Mirza, S. (2022). Phishing Attacks Detection using Machine Learning and Deep Learning Models, 7th International Conference on Data Science and Machine Learning Applications (CDMA), Riyadh, Saudi Arabia, 2022, pp. 175-180, doi: 10.1109/CDMA54072.2022.00034.

Almaiah, M. A., Al-Zahrani, A., Almomani, O. and Alhwaitat, A. K. (2021). Classification of cyber security threats on mobile devices and applications. In Artificial Intelligence and Blockchain for Future Cybersecurity Applications (pp. 107-123): Springer.

Ansari, M. F., Sharma, P. K. and Dash, B. J. P. (2022). Prevention of phishing attacks using AI-based Cybersecurity Awareness Training.

Bahaa, A., Abdelaziz, A., Sayed, A., Elfangary, L. and Fahmy, H. J. I. (2021). Monitoring real time security attacks for IoT systems using DevSecOps: a systematic literature review. 12(4), 154.

Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5-32. doi:10.1023/A:1010933404324

Çetin, V. and Yıldız, O. (2022). A comprehensive review on data preprocessing techniques in data analysis. Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi, 28(2), 299-312.

Cheng, F., Cui, J., Wang Q., and Zhang, L. (2023). A Variable Granularity Search-Based Multiobjective Feature Selection Algorithm for High-Dimensional Data Classification, in IEEE Transactions on Evolutionary Computation, vol. 27, no. 2, pp. 266-280, April 2023, doi: 10.1109/TEVC.2022.3160458.

Deiana, A. M., Tran, N., Agar, J., Blott M.., Di Guglielmo G., Duarte, J. Harris, P., Hauck, S., Liu, M., Neubauer M., S., Ngadiuba J., Ogrenci-Memik, S., Pierini, M., Aarrestad, T., Bähr, S., Becker, J., Berthold A.-S,, Bonventre, R. J., Müller, Bravo, T. E., Diefenthaler M., Dong, Z., Fritzsche, N., Gholami, A., Govorkova, E., Guo, D., Hazelwood, K. J., Herwig, C., Khan, B., Kim, S., Klijnsma, T., Liu, Y., Lo, K. H., Nguyen, T., Pezzullo, G., Rasoulinezhad, S., Rivera, R, A., Scholberg, K., Selig, J., Sen, S., Strukov, D., Tang, W., Thais, S., Unger, K. L., Vilalta, R., von Krosigk, B., Wang, S. and Warburton, T. K. (2022). Applications and Techniques for Fast Machine Learning in Science. Front. Big Data 5:787421. doi: 10.3389/fdata.2022.787421

Di Mauro, M., Galatro, G., Fortino, G. and Liotta, A. (2022). Supervised feature selection techniques in network intrusion detection: A critical review, Engineering Applications of Artificial Intelligence, vol. 101, https://doi.org/10.1016/j.engappai.2021.104216.

Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Eid, H. F., Hassanien, A. E., Kim, T. H., Banerjee, S. (2013). Linear correlation-based feature selection for network intrusion detection model. In Proceedings of the International Conference on Security of Information and Communication Networks 2013, Cairo, Egypt, 3–5 September 2013; Springer: Berlin/Heidelberg, Germany, 2013; pp. 240–248.

Fürnkranz, J. (2017). Decision Tree. In C. Sammut and G. I. Webb (Eds.), Encyclopedia of Machine Learning and Data Mining (pp. 330-335). Boston, MA: Springer US.

Heidari, A., Jafari Navimipour, N., Unal, M., Toumaj, S. J. N. C. and Applications. (2022). Machine learning applications for COVID-19 outbreak management. 34(18), 15313-15348.

Jain, A. K., Mao, J. and Mohiuddin, K. M. (1996). Artificial neural networks: A tutorial. J Computer, 29(3), 31-44. doi:10.1109/2.485891

Khaire, U. M., Dhanalakshmi, R. (2022). Stability of feature selection algorithm: A review, Journal of King Saud University - Computer and Information Sciences, 34(4), https://doi.org/10.1016/j.jksuci.2019.06.012.

Kira, K. and Rendell, L. A. (1992). The feature selection problem: traditional methods and a new algorithm. Paper presented at the Proceedings of the tenth national conference on Artificial intelligence, San Jose, California.

Kohavi, R. and John, G. H. (1997). Wrappers for feature subset selection. Artificial Intelligence, 97(1), 273-324. doi:https://doi.org/10.1016/S0004-3702(97)00043-X

Kushwaha, P., Buckchash, H. and Raman, B. (2017) Anomaly based intrusion detection using filter based feature selection on KDD-CUP 99. In Proceedings of the TENCON 2017—2017 IEEE Region 10 Conference, Penang, Malaysia, 5–8 November 2017; pp. 839–844.

Lee, C. S., Cheang, P. Y. S. and Moslehpour, M. J. A. i. D. S. (2022). Predictive analytics in business analytics: decision tree. Advances in Decision Sciences, 26(1), 1-29.

Li, Y., Fang, B. X., Chen, Y., Guo, L. (2006). A lightweight intrusion detection model based on feature selection and maximum entropy model. In Proceedings of the 2006 International Conference on Communication Technology, Guilin, China, 27–30 November 2006; pp. 1–4.

Lyu Y, Feng Y and Sakurai K. A Survey on Feature Selection Techniques Based on Filtering Methods for Cyber Attack Detection. Information. 2023; 14(3):191. https://doi.org/10.3390/info14030191

Maheswari, V. U., Aluvalu, R. and Mudrakola, S. (2022). An integrated number plate recognition system through images using threshold-based methods and KNN. Paper presented at the 2022 International Conference on Decision Aid Sciences and Applications (DASA).

Malik, N. U. R., Abu Bakar, S. A. R. and Sheikh, U. U. (2022). Multiview human action recognition system based on OpenPose and KNN classifier. Paper presented at the Proceedings of the 11th International Conference on Robotics, Vision, Signal Processing and Power Applications: Enhancing Research and Innovation through the Fourth Industrial Revolution.

Manevitz, L. M., and Malik Y. (2001). One-class svms for document classification. J. Mach. Learn. Res. 2, 139–154.

Mohammadi, S., Desai, V., Karimipour, H. (2018). Multivariate mutual information-based feature selection for cyber intrusion detection. In Proceedings of the 2018 IEEE Electrical Power and Energy Conference (EPEC), Toronto, ON, Canada, 10–11 October 2018; pp. 1–6.

Nguyen, H., Franke K. and Petrovic, S. (2010). Improving Effectiveness of Intrusion Detection by Correlation Feature Selection, 2010 International Conference on Availability, Reliability and Security, Krakow, Poland, 2010, pp. 17-24, doi: 10.1109/ARES.2010.70.

Ojewumi, T. O., Ogunleye, G., Oguntunde, B., Folorunsho, O., Fashoto, S. and Ogbu, N. J. S. A. (2022). Performance evaluation of machine learning tools for detection of phishing attacks on web pages. 16, e01165.

Patil, S. and Patil, Y. (2022). Face Expression Recognition Using SVM and KNN Classifier with HOG Features. In Applied Computational Technologies: Proceedings of ICCET 2022 (pp. 416-424): Springer.

Rivera-Lopez, R., Canul-Reich, J., Mezura-Montes, E., Cruz-Chávez, M. A. J. S. and Computation, E. (2022). Induction of decision trees as classification models through metaheuristics. 69, 101006.

Russell, S. J. (2010). Artificial intelligence a modern approach: Pearson Education, Inc.

Shabudin, S., Samsiah, N., Akram, K. and Aliff, M. (2020). Feature Selection for Phishing Website Classification. International Journal of Advanced Computer Science and Applications, 11.

Shahbaz, M.B., Wang, X., Behnad, A., Samarabandu, J. (2016). On efficiency enhancement of the correlation-based feature selection for intrusion detection systems. In Proceedings of the 2016 IEEE 7th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), Vancouver, BC, Canada, 13–15 October 2016; pp. 1–7

Şenol, A. (2022a). Comparison of Feature Selection Methods in the Aspect of Phishing Attacks. Paper presented at the International Conference on Engineering Technologies, ICENTE'22, Konya.

Şenol, A. (2022b). Standard Deviation-Based Centroid Initialization For K-Means. Paper presented at the 3. International Anatolian Scientific Research Congress, Kayseri.

Şenol, A. , Canbay, Y. and Kaya, M. (2021). Trends in Outbreak Detection in Early Stage by Using Machine Learning Approaches. Bilişim Teknolojileri Dergisi, 14(4), 355-366.

Tan, C. L. (2018). Phishing Dataset for Machine Learning: Feature Evaluation.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society. Series B (Methodological), 58(1), 267-288.

Uma, M. and Padmavathi, G. (2013) A Survey on Various Cyber Attacks and Their Classification. International Journal of Network Security, 15, 390-396..

Wahba, Y., ElSalamouny, E., ElTaweel, G. (2015). Improving the performance of multi-class intrusion detection systems using feature reduction. arXiv:1507.06692

Wang, W., Du, X., Wang, N. (2019). Building a cloud IDS using an efficient feature selection method and SVM. IEEE Access 2018, 7, 1345–1354.

Yousefpour, A., Ibrahim, R., Abdull Hamed, H. N. and Hajmohammadi, M. S. (2014). Feature reduction using standard deviation with different subsets selection in sentiment analysis. Paper presented at the Intelligent Information and Database Systems: 6th Asian Conference, ACIIDS 2014, Bangkok, Thailand, April 7-9, 2014, Proceedings, Part II 6.

Zhou, H., Wang, X. and Zhu, R. J. A. I. (2022). Feature selection based on mutual information with correlation coefficient. 1-18.