ISSN: 2723-9535

Available online at www.HighTechJournal.org

HighTech and Innovation
Journal

Vol. 5, No. 4, December, 2024

Outlier Detection in VPN Authentication Logs for Corporate Computer Networks Access Using CRISP-DM

Nilo Legowo [1]*, Wilyu Mahendra Bad [1]

[1] Information Systems Management Department, Binus Graduate Program – Master of Information Systems Management, Bina Nusantara University, Jakarta 11480, Indonesia.

Received 14 June 2024; Revised 27 October 2024; Accepted 09 November 2024; Published 01 December 2024

Abstract

A Virtual Private Network (VPN) serves as a critical network access solution widely employed by corporations, enabling users to connect to company computer networks via a global infrastructure. Amid the ongoing Covid-19 pandemic, heightened reliance on computer network access has increased the vulnerability to data breaches by unauthorized parties. This necessitates a proactive approach from companies to safeguard data integrity, particularly by identifying abnormal access patterns and timestamps. This study aims to develop a model for detecting anomalous activities within authentication log data obtained from VPN usage. The dataset comprises log entries from September to November 2022, totaling 36,807 records, selected via a systematic sampling approach. Two key attributes, namely user ID and access time, are analyzed to trace access patterns. Employing the CRISP-DM method ensures a structured and efficient research process. The selection of the k value in the K-Nearest Neighbors (K-NN) method significantly impacts outlier detection and can be tailored to suit organizational requirements. By utilizing the K-Means algorithm for data clustering and K-NN for measuring inter-point distances, the study identifies outliers that warrant further investigation by the company. Integration of the proposed model into the company's big data platform facilitates real-time monitoring, enabling the security team to preemptively address potential threats and mitigate network access misuse. By enhancing awareness and responsiveness to information security risks, the model contributes to fortifying the company's cyber security posture amidst evolving digital landscapes.

Keywords: Outlier Detection; Log VPN; K-Nearest Neighbors, K-Means; Data Mining; CRISP-DM.

1. Introduction

All companies engaged in the field of service provisions, including telecommunications, require information technology support to deliver quality services to stakeholders. Maintaining the quality of service to stakeholders in the use of data, information in application systems, and computer networks that can be accessed by users from various places without being limited by distance and time is the company's main priority for sustaining its business presence. The involvement of many parties is unavoidable to achieve this goal, so a secure solution is needed for accessing the company's computer network. One solution that can be used is to implement virtual private networks (VPN) technology. VPNs are reliable and cost-effective because the communication media is built using the internet for accessing the company's network [1].

* Corresponding author: nlegowo@binus.edu

http://dx.doi.org/10.28991/HIJ-2024-05-04-016

Large-scale companies frequently use computer networks to access application systems on servers. The use of information data facilities in applications accessible via the internet network often encounters various vulnerabilities, including unauthorized user access that can lead to data theft. These security risks pose significant threats to information security and can impact business continuity. Access via VPN computer networks requires vigilance against unauthorized access as a precaution against cybercrime. Employees or third parties outside the office who need access to network devices or servers from applications can use a VPN [2]. VPN is a technology that allows you to connect to a local network using a public network, providing the same rights and settings as an office network or local area network (LAN) [1].

The use of VPN to access the application system requires supervision of users or employees who access the company's computer network from both inside and outside the office, without any time and distance limitation [1]. The company appoints a network administrator to implement general preventive measures against possible attacks and fraud from users or outside parties. However, there may still be vulnerabilities or deliberate actions by those with access rights.

Based on the previously described information, an initial step to minimize losses from cybercrime is to implement early detection of anomalous activities. This process involves several stages that must be completed. The first step is to collect data from activities, which will then be processed to detect anomalous activities. This requires a process to obtain readable information for decision-making. However, implementing a VPN doesn’t eliminate the potential for access abuse activities. Such abuse can originate from within the company itself; this is illustrated in 5 basic security threats shown in Figure 1 [3].

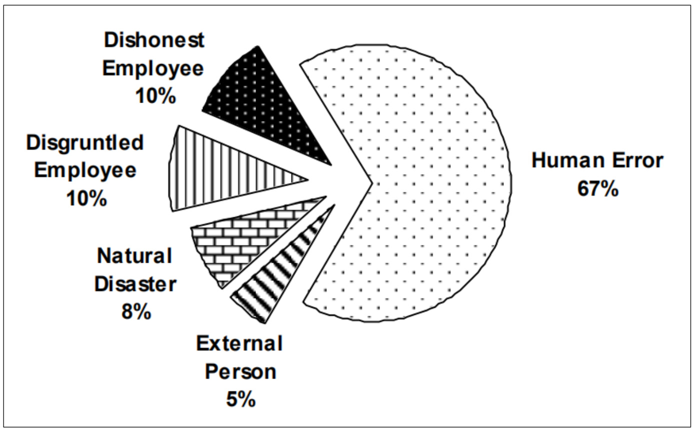

Figure 1. Basic Threats to Security [3]

Based on Figure 1, approximately 10% of security comes from dishonest employees, while only 5% comes from outside parties or third parties. The largest percentage involves unintentional human errors, although this data can vary depending on the level of personal or business activities conducted online [3]. A recent case of access abuse involved public figure Denny Siregar, where personal data that should have been protected for privacy was actually disseminated by irresponsible individuals, who were third parties associated with the company [4]. Access rights granted to employees are often misused due to negligence, including the provision of freedom of access for employees and third parties [1]. This creates a significant potential for misuse of these access rights when using the company's internal network via VPN [2].

Therefore, based on this case, a method is needed to prevent or minimize the potential for misuse of access by individuals within the company. This aims to design a model for detecting anomalous activity using VPN authentication logs by applying the K-Nearest Neighbors method and the CRISP-DM methodology.

This study attempts to utilize VPN authentication log data to propose a method for detecting activities that deviate from the typical behavior of registered parties or those with VPN access. This method aims to identify and prevent potential access misuse activities, whether by third parties or internal company personnel.

The theoretical concept of Knowledge Discovery in Databases (KDD) involves methods for transforming raw data into usable information. This paper employs several stages of KDD, including variable selection, data cleaning, and data transformation. Once the data is normalized, it is used as a dataset for training purposes.

Detecting activities outside the established pattern requires many steps and the use of various methods. One effective method is the K-Nearest Neighbor algorithm, which is a popular method and widely used by researchers to detect global outliers [5]. Before detecting outliers, it is important to group the data, and K-Nearest Neighbor is known for its simplicity and efficiency [6]. The author employs the Cross-Industry Standard Process for Data Mining (CRISP-DM) to ensure a more targeted approach in building an outlier detection model. CRISP-DM, a widely used methodology in data mining, involves identifying, validating, and analyzing various data sources to obtain information [7].

By combining the K-Nearest Neighbor method, K-Means clustering, and outlier detection techniques with the CRISP-DM framework, researchers can design a model to detect anomalous activity using VPN authentication logs. This approach enables companies to identify unusual activities and mitigate potential access abuse. The reason for using the KNN and K-Mean methods by applying the K-Means method for the data grouping process [8] and K-NN for scoring the distance between data points [9] is that the author successfully identified outliers that warranted further investigation by the company. This presents a challenge for an organization on how to utilize log data to be processed and useful for detecting anomalies through business intelligence methods. The processed results can assist management with decision-making and help determine whether activities lead to potential fraud.

2. Literature Review

2.1. VPN (Virtual Private Network)

A VPN is a solution that provides a secure network architecture over a public network, reducing huge infrastructure costs [1]. It can be described as an authenticated and encrypted tunnel that functions as a virtual leased line over a public infrastructure (see Figure 2) [2].

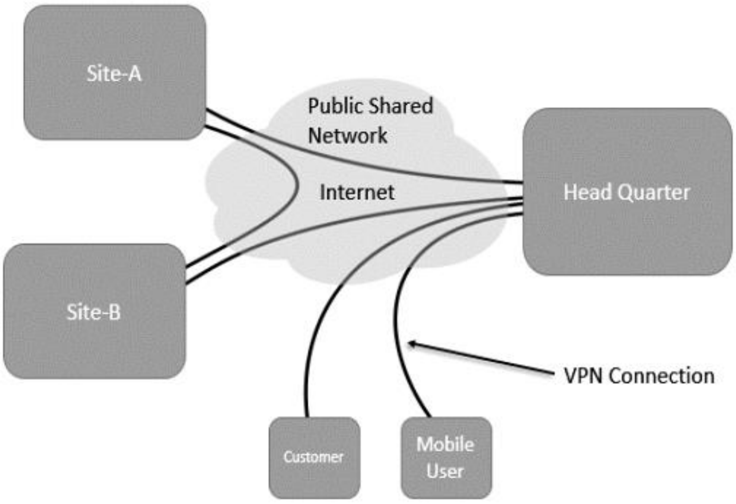

Figure 2. Typical VPN Scenario [1]

2.2. Outlier Detection

Outliers are types of data that have different, inconsistent, and irrelevant characteristics in the data set. The process of identifying the given data is known as outlier detection [7]. Outlier or anomaly detection involves identifying patterns that deviate from regular, defined behavior [10]. Several approaches can be used to detect outlier data, including statistical-based, depth-based, deviation-based, distance-based, density-based, and high-dimensional approaches [7]. The comparison between these techniques can be seen in Table 1:

Table 1. Comparison Table

Parameters	Techniques algorithms		
	Cluster based	Distance based	Density based
Computational Cost	Low	Low	High
Efficiency	Very efficient	Efficient	Efficient
High-dimensional data	Applicable	Applicable	Applicable
Complexity	Less complex	Moderately Complex	Highly complex

2.3. K-Nearest Neighbors Method

This distance-based method performs anomaly detection by calculating the distance between data points. A data point is considered an outlier if it is far from its nearest neighbor. K-Nearest Neighbor is a distance-based method and is widely used in research. The initial stage of this method involves searching for the nearest neighbors for each data point. The distances to these nearest neighbors are then used to calculate the outlier value. In essence, this method examines the surrounding data to determine the distance and density, which serve as a reference for detecting data outliers [5]. The K-NN method can be defined as follows [9]:

1. The first definition (k - distance). For point p (x, y), distance between p and k-the nearest neighbor P_k (X_k, Y_k) is k-distance from p and denotes D^k (p):

$$D^k(p) = \sqrt{(y - y_{k)})^2 + (x - x_k)^2} \tag{1}$$

2. The second definition (maximum value of D^k). Given k and m values, p is considered an outlier if it gives a value less than m-1 from a data point that has a greater magnitude value D^k than the p value.

3. The third definition (outlier). A given k value, for any data point p, if Dk (p)> t, then p will be considered as an outlier with t as the threshold.

4. The fourth definition (k-dist histogram). There is a clear k-distance value for each data point in unsupervised approach procedures. Distribution of all distances described by the k-dist histogram in different intervals. Each pillar presents the k-dist quantity at k in a certain range in the histogram.

The K-Nearest Neighbor (KNN) algorithm categorizes objects or data based on the training data points nearest to them [11]. The algorithm identifies the closest k training samples and predicts the class of a given test sample based on the majority class among these nearest neighbors [12]. The selection of the k value, which represents the number of nearest neighbors to consider, can be optimized through parameter tuning techniques such as cross-validation. Increasing the k value can help mitigate the impact of noise on the classification process.

2.4. K-Means

K-Means is a traditional clustering technique that uses the K parameter to determine the number of clusters to form. Initially, k point is randomly selected as the cluster center [8]. K-Means is a non-hierarchical data grouping method where data with similar characteristics are grouped in one cluster, while data with differing characteristics are assigned into different clusters [6]. The steps in K-Means can be described as follows [6]:

1. Determine k (number of clusters to be formed), using the elbow criterion method with the following equation:

$$SSE = \sum_{K=1}^{K} \sum_{Xi=Sk} \|Ni - Ck\| \tag{2}$$

2. Determine the initial k point of the cluster center (centroid) which is done randomly. Determination of the initial centroid is conducted randomly from the available objects as many as k clusters to calculate the next centroid of the with cluster, the following is the used formula:

$$V = \frac{\sum_{i=1}^{n} X\,1}{n} \quad , \qquad i = 1, 2, 3, \dots, n \tag{3}$$

3. Calculate the distance from each object to each centroid of each cluster using Euclidean Distance, with the following equation:

$$d(x, y) = \|x - y\| \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2 : i = 1,2,3, \dots\dots n)} \tag{4}$$

4. Allocate each object into the nearest centroid. The allocation of objects into each cluster during iteration is generally conducted by means of hard k-means in which each object is clearly stated as a member of the cluster by measuring its proximity to the cluster’s center point.

5. Iterate and then determine the position of the new centroid using the equation.

6. Repeat step three if the new centroid positions are not the same.

2.5. Knowledge Discovery in Databases

Knowledge Discovery in Databases (KDD) is a concept that describes the non-trivial process of identifying novel, valid, potentially useful patterns in data and explaining their significance [13]. The term "pattern" refers to a subset of data expressed in some language or model, used to convey meaningful information about the data.

The purpose of KDD according to Gullo [14] is to find patterns that:

a. Does not result in a direct (i.e., non-trivial) count of a predetermined quantity.

b. Can be applied to new data with some degree of certainty (i.e., valid),

c. Unknown so far (i.e., novel),

d. Provide several benefits for users or for further (i.e., potentially useful) assignments, and

e. Lead to useful insights, immediately or after some post-processing (i.e., understandable) concept.

2.6. CRISP-DM

CRISP-DM is an extension of the original Knowledge Discovery Database process, consisting of the stages of business understanding, data understanding, data preparation, modeling, evaluation, and application. Figure 3 shows the 6 stages in this method [15]:

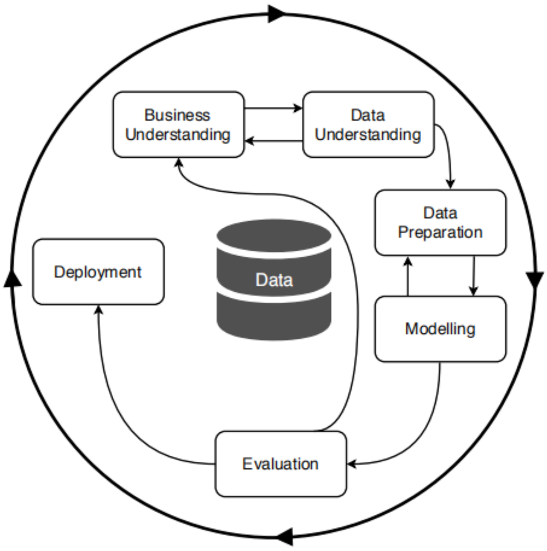

Figure 3. CRISP-DM Process [15]

The stages in CRISP-DM do not necessarily follow a strict sequence as depicted in Figure 3; the arrows indicate the dependencies that often occur between phases [16]. Generally, the tasks in each stage build upon those of the previous phases, and iterative cycles are common, meaning the output of one phase can influence the subsequent stages [17].

Data mining involves extracting patterns and trends from vast datasets to generate new knowledge that aids researchers and informs decision-making processes [18]. This methodological process relies on sequential steps to achieve optimal outcomes, as outlined in the Cross-Industry Standard Process for Data Mining (CRISP-DM) [17]. CRISP-DM serves as an open standard process model, guiding data mining experts through six distinct stages:

1. ***Business Understanding***

   Researchers utilize data mining to derive insights that address business challenges and drive organizational growth. Therefore, establishing clear business objectives is pivotal, as they influence subsequent processes and enhance the effectiveness of outcomes. Defined objectives also inform algorithm selection and strategy development for the following stages.

2. ***Data Understanding***

   Data serves as the cornerstone of data mining endeavors. Once business objectives are defined, researchers must thoroughly examine the dataset to guide subsequent data preparation steps. This stage involves identifying missing values, outliers, and data inconsistencies and ensuring data uniformity. Visualization techniques are often employed to gain insights for subsequent modeling and analysis phases.

3. ***Data Preparation***

   Following data understanding, researchers refine the dataset to make it suitable for modeling. This entails addressing missing values, outliers, inconsistencies, and performing feature engineering and selection. Categorical data is often converted into numerical formats using techniques like one-hot encoding or label encoding. Additionally, the dataset is split into training and testing sets, ensuring balance for classification methods through oversampling or under sampling.

4. ***Modeling***

   Prepared training data is utilized to train models that align with predefined business objectives. Machine learning algorithms are applied to the dataset, typically including classification, clustering, or regression techniques.

5. ***Evaluation***

   The model's performance is assessed using testing data, with metrics such as accuracy, F1-score, recall, and precision providing evaluation criteria. This stage informs algorithm selection and feature performance analysis. If the model falls short of expectations, researchers iterate the process until desired outcomes are achieved.

6. *Deployment*

> The dynamic model obtained requires ongoing monitoring to ensure continued optimal performance. The deployment phase may also encompass final reports, software components, implementation planning, and maintenance efforts.

In their 2015 research on network communication, Kohout & Pevný explored the potential of detecting persistent malware by analyzing its low variability. They developed a novel technique for identifying statistical patterns in network connections and employing outlier detection to recognize malicious activities. By leveraging minimal data, their method is efficient and easy to use. Anomaly detection plays a crucial role in network security. The researchers showed that malicious persistent connections are more consistent compared to those of legitimate users, making them easily detectable as outliers [19].

The research involved observing specific features of Intrusion Detection Systems (ICS) network communication, such as packet arrival time, to create statistical profiles based on patterns in normal traffic. Cyber-attacks on smart grid communications can have serious consequences for energy production and distribution. Since attacks can originate both inside and outside the network, traditional security tools like firewalls and Intrusion Detection Systems (IDS), typically positioned at the network's edge, are inadequate for detecting internal threats. Therefore, it is necessary to analyze the behavior of internal ICS communications as well. This approach is effective, quick, and easy to implement. Our experiments demonstrate that statistical-based anomaly detection can successfully identify common security incidents in ICS communications [20].

This paper presents a novel algorithm called FilterK, developed to improve the purity of k-means clusters derived from physical activity data by reducing the influence of outliers. The data, gathered via body-worn accelerometers, undergoes k-means clustering. The algorithm's effectiveness is evaluated against three existing outlier detection techniques: Local Outlier Factor, Isolation Forest, and KNN, using ground truth (class labels), average clustering, and event purity (ACEP). While the main objective of this new method is to enhance cluster purity for accelerometer data from physical activities, it also shows potential for application to other datasets that utilize k-means clustering [21].

Outlier detection has gained significant attention in various disciplines, especially those related to machine learning and artificial intelligence. Anomalies, regarded as distinct outliers, are classified into point, contextual, and collective outliers. Significant challenges in this area include the indistinct boundary between distant points and natural clusters, the propensity for new data and noise to mimic authentic data, the lack of labeled datasets, and the diverse definitions of outliers across various domains. A universal, domain-agnostic approach is proposed to identify these anomalies in both unsupervised and supervised datasets. To tackle these challenges, we introduce new types of anomalies, named Collective Normal Anomaly and Collective Point Anomaly, to better delineate the subtle boundaries between different anomaly types [22].

In wireless sensor networks, measurements that significantly deviate from the usual pattern of sensed data are identified as outliers. These outliers can be caused by disturbances, errors, specific events, or malicious attacks on the network. Traditional outlier detection methods are often unsuitable for wireless sensor networks due to the unique characteristics of sensor data and the particular requirements and constraints of these networks. This survey provides a comprehensive overview of current outlier detection techniques specifically tailored for wireless sensor networks. The paper addresses the challenge of outlier detection in WSNs and presents a taxonomy framework to categorize the current outlier detection methods designed for these networks [23].

The statistical community has extensively studied outlier detection in time series data, with several surveys underscoring the research conducted in this area. Likewise, the computer science community has made substantial contributions to temporal outlier detection from a computational perspective. Advances in hardware technology have enabled various mechanisms for collecting temporal data, while software innovations have led to diverse data management techniques. Consequently, numerous types of datasets—such as data streams, spatiotemporal data, distributed flows, temporal networks, and time series data—are now generated by many applications. There is an increasing need for a thorough and organized exploration of outlier detection methods as they apply to these temporal datasets.

In computer networks, techniques for detecting outliers in temporal data are extensively used for intrusion detection. These methods leverage multivariate time series data that monitor metrics such as the number of bytes, packets, IP-level flows, protocol types, and the amount of data transferred in TCP connections. These techniques are employed to identify anomalies like high-speed point-to-point byte transfers, denial of service (DOS) attacks, distributed denial of service (DDOS) attacks, and network scans targeting specific ports. Moreover, outliers are detected using subspace methods derived from the multivariate statistical process control literature [24].

This research aims to develop a personalized classifier to detect four categories of human activities: light-intensity activities, moderate-intensity activities, vigorous-intensity activities, and falls. To address the challenges posed by the varying inertial sensor signal distributions among users, a user-adaptive algorithm is proposed that combines K-Means clustering, the local outlier factor (LOF), and the multivariate Gaussian distribution (MGD). An improved K-Means algorithm with an innovative initialization method has been developed to automatically cluster and label activity data for individual users. Since the inertial sensor data distribution varies considerably between users, an activity recognition classifier trained on one user's data may not perform well when applied to other users. Experimental results show that the proposed model can effectively adapt to new users while maintaining high recognition accuracy [25].

3. Research Methodology

In conducting this research, the author followed a scientific research methodology that integrates practical field stages with theoretical frameworks. The research methodology is outlined through the flow of the framework used as a reference. The stages of the research, based on these frameworks, are illustrated in Figure 4.

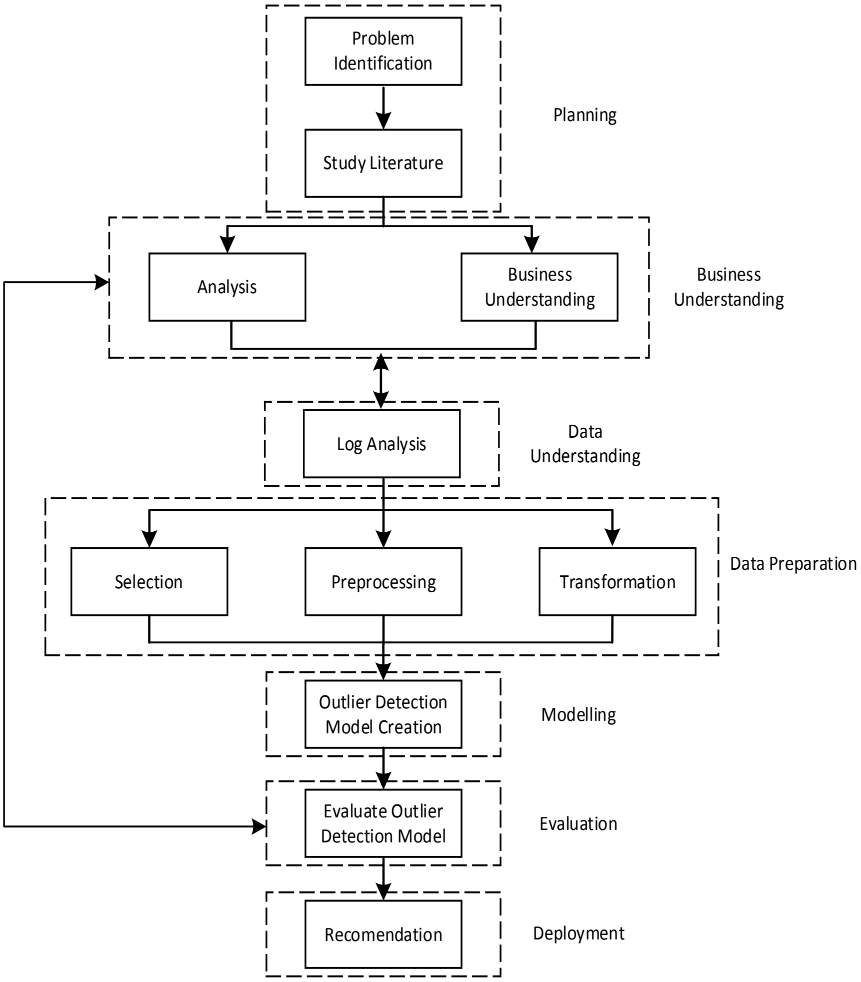

Figure 4. Research Methodology

In the initial stage of the research, the author planned the study by investigating access issues encountered by company employees with the VPN network. Additionally, a literature review is conducted on previous research related to outlier detection and the use of the CRIPS-DM method.

The second stage involves understanding the business through the Business Understanding phase by analyzing the company's business process. The third stage, the Data Understanding, includes analyzing the computer network system by examining Log data from all transaction access through the company's internal VPN network.

The fourth stage is Data Preparation, which includes data selection, Preprocessing, and Transformation. The fifth stage is Modeling, where outlier detection and model creation are performed. The sixth stage is Evaluation, involving the assessment of Evaluate Outlier and Detection Model. The final stage is Deployment, which includes Recommendations.

In this paper, the author employs the CRISP-DM method as a framework stage to address the problem. The K-Means method is used for data clustering, and KNN is utilized to calculate data points. This approach aligns with the case described in the previous chapter. The literature review revealed that many studies focus on similar topics, such as outlier detection or anomalous activity. Based on this review, the author selected the CRISP-DM method as a reference for solving the problem.

Initially, the company faced significant issues due to the broad access rights granted to employees and third parties, which increased the potential for misuse. This situation aligns with the case described in the previous chapter. Based on this, the authors found that many studies took the same topic, namely outlier detection or anomalous activity, with various methods, including using KNN and K-Means; then the author also used the same method, but in the discussion the author used the CRISP-DM method as a reference stage to solve this problem.

Under current conditions, employees or third parties must meet the company's requirements to gain access to the company's internal network. This includes an agreement that can be enforced legally. These requirements help protect the company from both potential fraud by employees and third parties.

After analyzing the current conditions (business understanding), employees and third parties must log into the VPN to access the corporate network. The VPN logs provide data that can be used as references for detecting outliers or anomalies. Outlier detection focuses on two main attributes: user ID and activity timestamp.

To utilize the data from the company's Big Data platform, the following steps were undertaken: masking the original data to maintain the confidentiality and collecting sample data using simple random sampling techniques. Log data from September to November totaling 122,691 records were analyzed, with 36,807 records selected in this study. The author will use the K-Means method to group the data.

The author employs a distance-based K-Nearest Neighbors method to detect outliers. However, prior to modeling, the data will be further refined to ensure it is suitable for the chosen method. This process will be conducted using the RapidMiner tool, which serves as the primary tool for this study.

The designed model shows potential for application within the company; however, the results of this outlier detection model certainly still require evaluation. To effectively implement this model, it is necessary to correlate the findings with the data related to the company. This research aims to provide recommendations that will help companies mitigate the risk of misuse of access to corporate networks.

At this modeling stage, the authors refer to the sources of previous research that utilized similar modeling techniques to predict anomalies in VPN authentication logs at a university [26]. The following are the stages that will be undertaken in the modeling process:

1. *Cluster Development*

In this research, the authors use 2 groups of data, including:

a. Working Day Group (Weekday): This group consists of data groups where employees or third parties successfully authenticate to the VPN to access company devices. It includes VPN authentication activities performed on weekdays (Monday, Tuesday, Wednesday, Thursday, and Friday). National holidays that fall on weekdays are not considered in this thesis.

b. Group Holidays (Weekend): This group consists of successful VPN authentication on Saturdays and Sundays.

c. Working Hours Group (Office Hour): This group consists of VPN authentication data conducted on working days and hours. In this study, the working hours used as a reference are from 08:00 am to 17:00 pm.

d. Group Outside of Working Hours (Non-Office Hour): This group consists of authentication data outside of point

2. *Outlier Detection*

At this stage, the author will detect outliers in each group. The method used is the distance-based method. The expected output in this stage, the author will obtain unusual or anomaly data.

4. Results and Discussions

The author employs the CRISP-DM method, which consists of 6 main stages to develop recommendations for the company. The stages in the research are described as follows:

4.1. Business Understanding

XYZ Company, a service provider company for its customers, involves many parties to maintain its services. Consequently, many parties access the company's local network daily, making the role of a VPN crucial for restricting third-party access.

The solution applied still carries the potential for misuse of access. For example, a single user account can be shared among multiple individuals, particularly since not all users are in the company for an extended period.

Because of this, a method or model for outlier detection is needed as a precaution against misuse of this access, in addition to the standard procedures applied. One of these SOPs is that users must sign an NDA (Non-Disclosure Agreement), which is protected by a legal.

4.2. Data Understanding

This study uses data derived from VPN activity logs generated by parties who have or gain access to the VPN. Based on the information that can be found in the log data, the authors use 2 attributes to support this research: the user ID and the timestamp of the activity.

With these 2 attributes, the authors aim to detect anomalous activities early using the proposed method. This log data is generated every time a user logs into the VPN to access the corporate network.

4.3. Data Preparation

Before making a model, data preparation is essential. This study uses 2 main attributes taken from the VPN log, namely user ID and access time. The initial stage involved extracting data from the available logs. In this study the authors used VPN authentication log data from September to November, which has a total of 122,691. However, only 30% of this data, amounting to 36,807 entries, was utilized in the study.

Data sampling was performed using random sampling techniques with the split data operator in RapidMiner. With this operator, the distribution between data periods will be balanced. The distribution of the data from September to November was as follows: before sampling, the percentages were 40%, 24%, and 36%. The data distribution can be seen in the Figure 5.

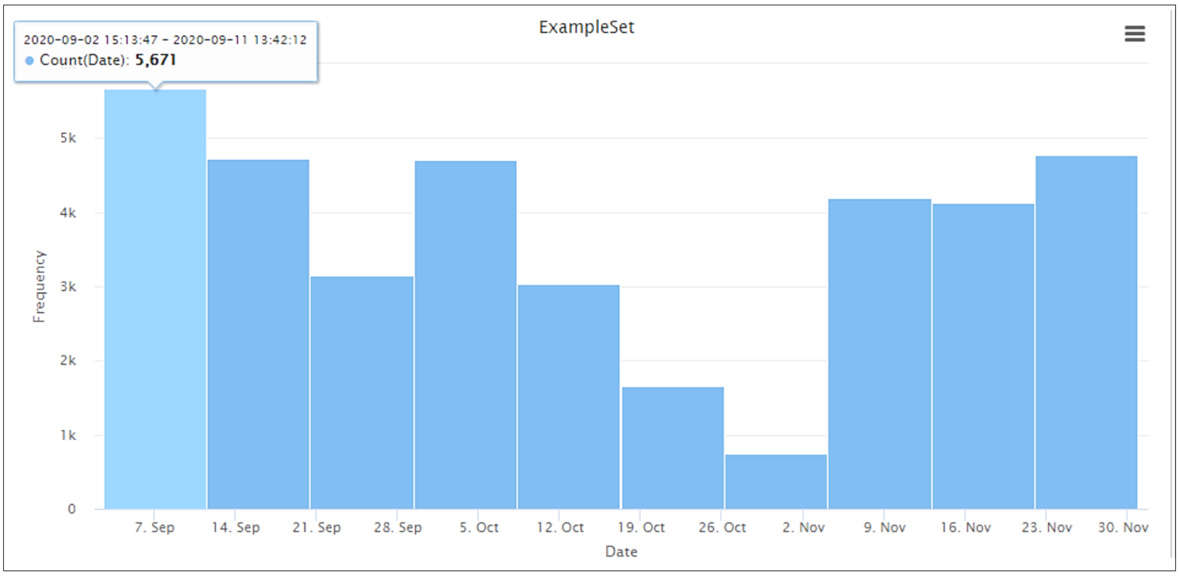

Figure 5. Data Sampling Process

After obtaining the dataset for this study, the authors masked the data to protect company privacy. The author mapped the original attributes and the replacement data on the user attributes to ensure that actual user information could not be retrieved from the dataset.

4.4. Modelling

In this modeling process, there are 4 main stages in the model that the author proposes: preprocessing, clustering, outlier detection and post-processing. In general, it can be described in Figure 6.

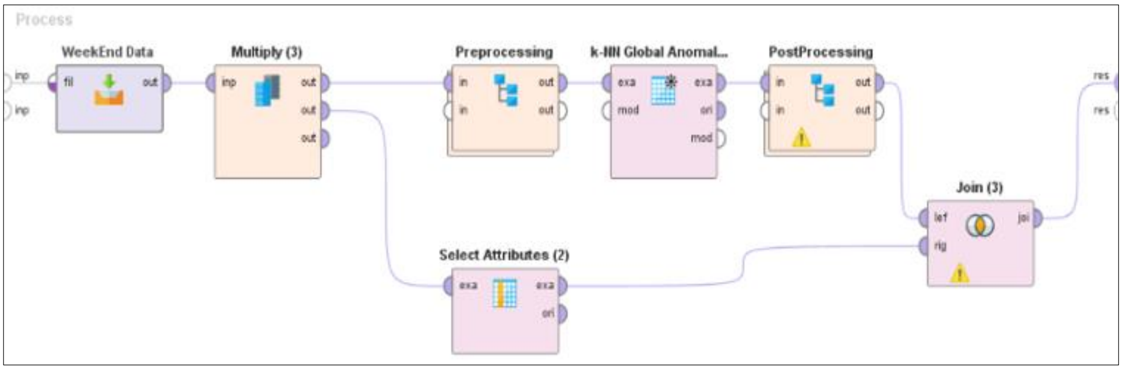

Figure 6. Outlier Detection Model

At the preprocessing stage, the authors extracted attributes related to the timing of activities. This was done using the RapidMiner function, resulting in 4 new attributes: day, time: hour, time: minute, and time: second. Each attribute represents the day, hour, minute, and second the activity occurred. With this data extraction process, the authors remove the date attribute, where the attribute has been extracted to become the attribute displayed in the Table 2.

Table 2. Example Extracted Date Data

Id	Week: Day	Time: Hour	Time: Minute	Time: Second
1	6	0	9	54
2	6	0	9	2
3	6	0	9	28
4	6	0	9	6
5	6	0	9	40

The complete process in the preprocessing stage can be seen in the RapidMiner process below (see Figure 7).

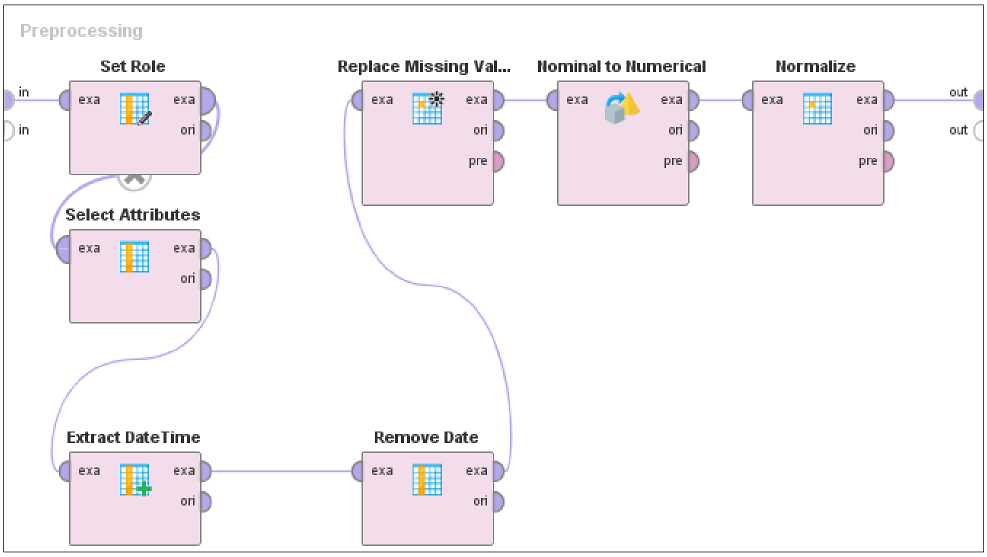

Figure 7. Preprocessing Process

After completing attribute extraction, the next step is to address any missing data points. Missing points are filled with the average of the existing values. Following this, nominal data is converted to numerical format and then normalized to ensure a balanced threshold.

The author uses the K-Means method for grouping. To determine the number of clusters, the centroid values of each attribute are analyzed, considering their proximity to other attributes.

In this study, the authors chose to divide the number of groups into 4 groups. This is based on the centroid value, which is shown in Table 3:

Table 3. Comparison Centroid Score

Attribute	Cluster_0	Cluster_1	Cluster_2	Cluster_3
Week: Day	-0.987	0.758	-0.977	0.677
Time: Hour	-0.038	-0.801	-0.052	1.146
Time: Minute	0.934	0.006	-0.954	0.002
Time: Second	0.042	0.000	-0.052	0.008

The author divides the groups into 4 because the centroid distance between these groups is far, and there is only one attribute that has a short distance between groups. This occurs in the week: the day attribute for groups cluster_0 and cluster_2. Dividing the data into 5 or 6 groups results in centroid values that are closer to more than one attribute, which is less optimal.

There is also the distribution of data from group 4 shown in the Figure 8.

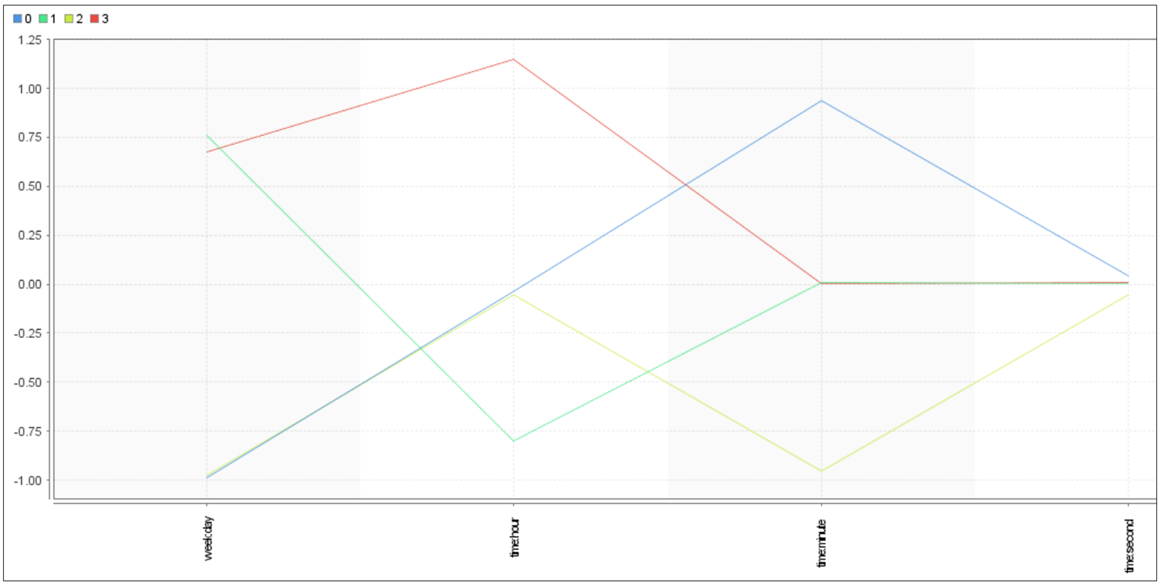

Figure 8. Centroid Per Group

Based on the graph 8 above, the authors can see that group 3 is the group that has the highest centroid value in the time-hour attribute, while the smallest is in group 2. For the week: day attribute with the highest value belongs to group 1, while the lowest value is in group 2. Small, include groups 0 and 2. As for the attribute time: minute, the group with the highest value was group 0, while the lowest was group 2. For time: second the highest was group 0, and the lowest was group 2.

With the division into 4 groups, the distribution for the data is as follows: group 1 has the greatest number of data, namely 12,120 data, followed by group 3 with 9,077 data. While group 0 and group 2 have the total data of 7,840 and 7,770 of the total data of 36,807 data. The amount of data from each group is illustrated in the Figure 9.

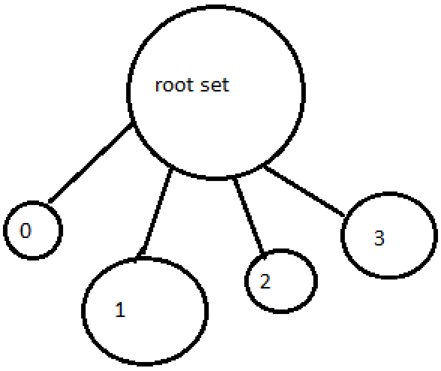

Figure 9. Tree Graph Centroid

The K-NN method was applied to evaluate each data, yielding values ranging from 0.1 to 0.5. Based on these results, this study classified the data into 5 groups according to the obtained distance. This grouping is shown in Table 4:

Table 4. Grouping Based on Anomaly Score

Group	Range
1	0.1 ≤ Score < 0.2
2	0.2 ≤ Score < 0.3
3	0.3 ≤ Score < 0.4
4	0.4 ≤ Score < 0.5
5	Score ≥ 0.5

Grouping based on the anomaly value is utilized for the next process at the post-processing stage. One of these steps is determining the threshold value of a group that is considered an outlier. In this study, the outlier data is derived from the group with the fewest members. The process for determining this threshold is illustrated in Figure 10.

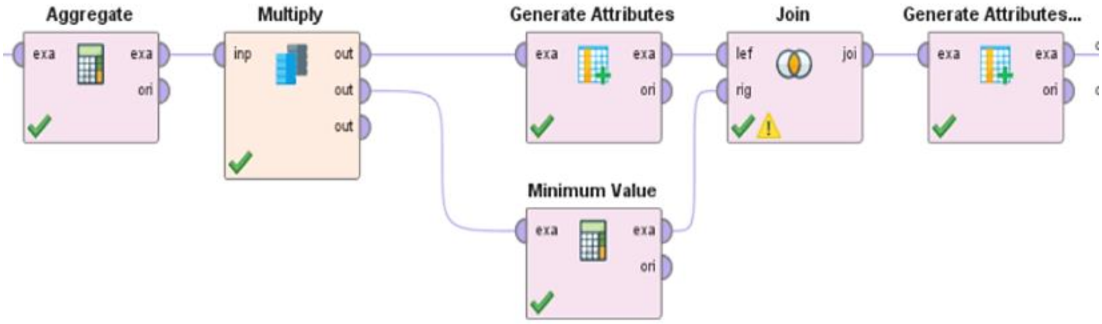

Figure 10. Postprocessing Process (Threshold)

The author obtains outlier data with a value of k = 5 as follows:

a. Cluster 0, with a total data of 7,840, the authors get 1 data detected as an outlier as follows (Table 5):

Table 5. Outlier Data Cluster 0

User	Date Time	outlier_flag	Outlier
U0195	Wed,16 Sep 2020 01:14:59	true	0.405

This anomalous activity conducted by U0195 in the early hours of the morning outside working hours resulted in an outlier value of 0.405. From this, it can be concluded that the data can be grouped into 4 categories based on the outlier values obtained for each row.

b. Cluster 1. Aligning with the data obtained in Cluster 0, in Cluster 1, there were 2 users who were detected doing anomalous activity. This activity is performed on Sundays in the sense that it is performed outside working hours. This can be seen in the Table 6:

Table 6. Outlier Data Cluster 1

User	Date Time	outlier_flag	Outlier
U3783	Sun,20 Sep 2020 08:39:47	true	0.423
U3849	Sun,6 Sep 2020 15:59:02	true	0.454

c. Cluster 2. Based on the outlier detection model using the K-Means and K-NN clustering methods, for cluster 2 the following data obtained (Table 7):

Table 7. Outlier Data Cluster 2

User	Date Time	outlier_flag	Outlier
U1224	Fri,2 Oct 2020 03:13:25	true	0.421
U5203	Sun, 8 Nov 2020 10:13:31	true	0.420

This data shows the activities carried out by U1224 and U5230 users, namely outside working hours. User U1224 conducted these activities on weekdays, while U5203 conducted these activities outside of working days.

d. Cluster 3. The last group with the number of data 9,077, the author detected an outlier data. Similar to the previous results, outlier data is an activity conducted by U5369, which is carried out on Saturdays around 3 in the morning, which is considered out of working hours. This is illustrated in the Table 8:

Table 8. Outlier Data Cluster 3

User	Date Time	outlier_flag
U5367	Sat, 19 Sep 2020 03:24:01	true

Evaluation

The grouping was conducted using the K-Means method. The authors identified that the anomalies detected were associated with VPN user activity outside of working hours, whether on a regular working day or an open working day.

From the results of the outlier values based on the K-NN method, the authors identified outlier' data in each group. To determine the threshold for identifying outliers, the authors grouped the range of values from these outliers and selected the group with the fewest members as outliers. Additionally, the authors can observe the outliers based on the Figures 11 to 14.

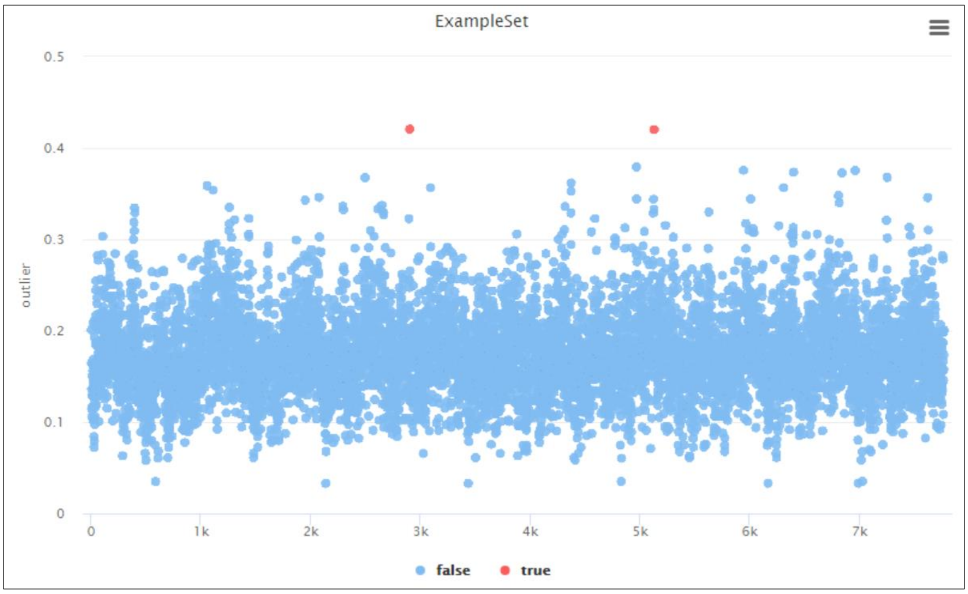

Figure 11. Graph Outlier Data Cluster 0

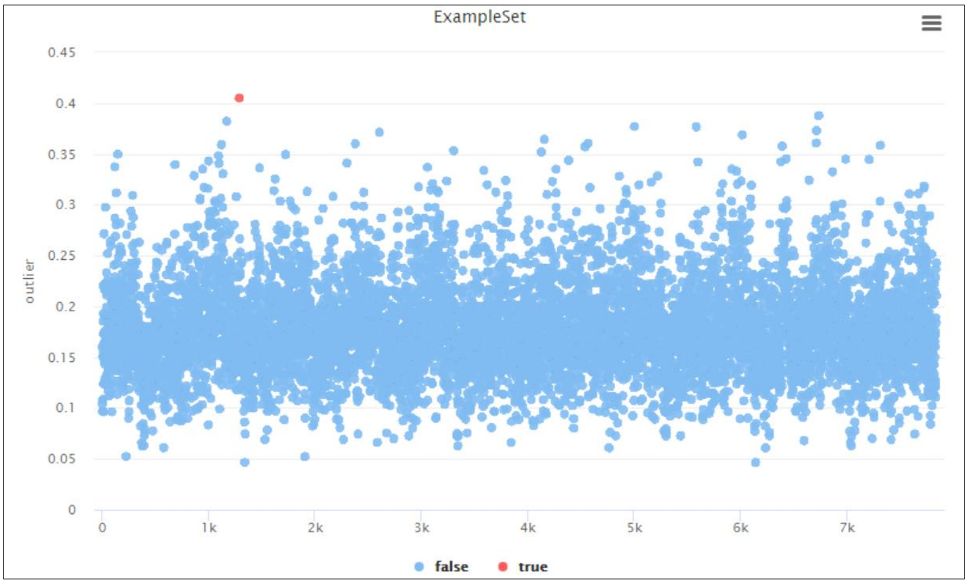

Figure 12. Graph Outlier Data Cluster 1

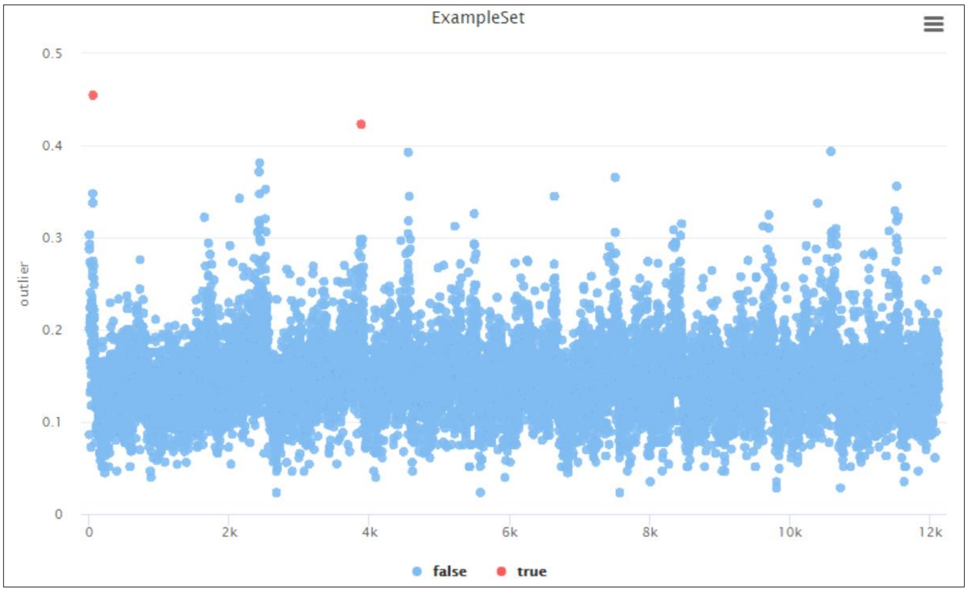

Figure 13. Graph Outlier Data Cluster 2

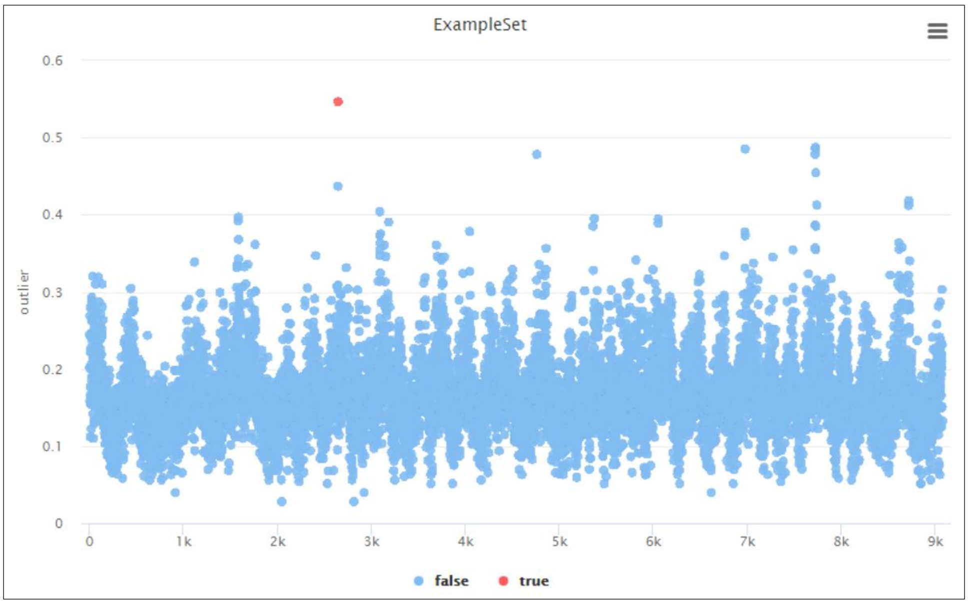

Figure 14. Graph Outlier Data Cluster 3

Figures 11 to 14 depict the position of the outlier data in relation to other data, arranged sequentially from groups 0 to 3. Outlier data is marked with a red circle, indicating that the group has fewer data. In contrast, the blue circle represents data points from other groups that are considered outliers.

Based on the graph above, outlier data identified by the model in this study is positioned significantly distant from other groups. This positioning suggests that the outlier data warrants further investigation by the company.

4.5. Deployment

The model proposed in this study, based on the evaluation results, is effective in detecting data points considered as outliers. These data points represent unusual activities that are appropriately categorized for further investigation by the company. The workflow can be seen in Figure 15.

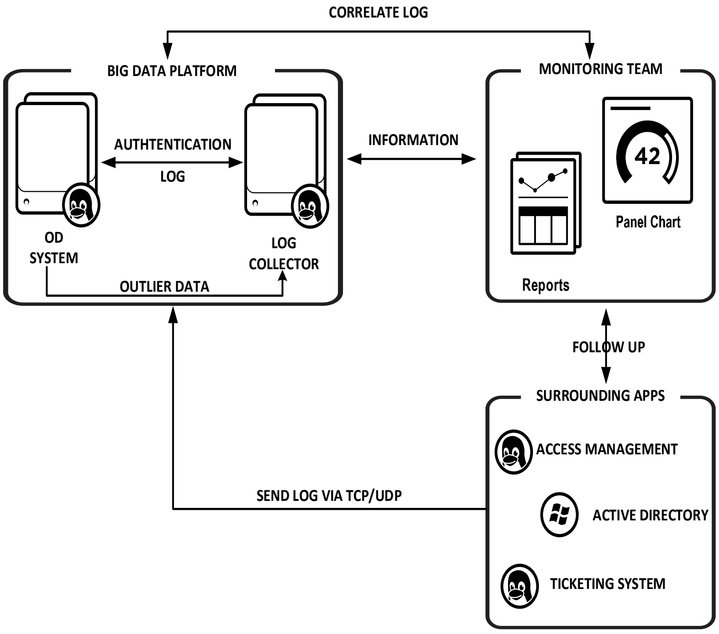

Figure 15. Outlier Detection Flow

Figure 15 illustrates the deployment of the proposed model as a service or additional application within the company’s big data platform. This certainly has an effort to convert the model in the form of RapidMiner into a programming language supported by the platform, specifically Python. By doing integration, companies can easily make adjustments according to the conditions of the company. By integrating the model in this way, companies can easily make adjustments according to their specific needs.

Outlier Detection, which is deployed into the big data platform, can retrieve VPN log data directly periodically according to the needs of the company and then process it by the system. Outlier data that is found will be informed to the monitoring team, either in the form of a dashboard or alert. Thus, the monitoring team can carry out further investigations related to this finding. The monitoring team may conduct investigations based on logs from the company's existing applications related to access management, which in this case are the active directory log and ticket application. With this investigation, the monitoring team can ascertain whether the activity is permitted under certain conditions, such as a change request for certain applications that requires access to the local network during predetermined hours.

For further integration, it would be beneficial if the results of the outlier data could be directly correlated with logs from the relevant applications, provided the company has access to such logs to aid in the investigation process. Therefore, the Monitoring Team can immediately carry out an investigation directly from the dashboard or alerts provided by the big data platform, without having to follow up on each pic of the related application.

The results of the analysis and discussion that have been conducted can be comprehensively elaborated. Below is an explanation of the differences in the results of the research conducted with previous research: Consideration of the use of the method that can be used is K-Nearest Neighbor, which is the most popular method for detecting outliers in general, which functions to provide scoring on the distance between data points that are widely adopted by researchers. The use of the K-Means method for the process of grouping data that will be processed from the existing data set.

This research was conducted on the research object to design a model for detecting anomalous activity using VPN authentication logs and computer networks accessed via an internet connection. The results obtained from the modeling process involve 4 steps: identifying access activities based on the outlier detection model, which leads to 4 new attributes: week:day, time:hour, time:minute, and time:second. The results of the modeling process include preprocessing, clustering, outlier detection, and post-processing. In general, it can be described as follows: Outlier detection model, Preprocessing Process with Centroid score comparison data table, Centroid graph per group, and Tree Graph Centroid image. Postprocessing Process (Threshold) with Outlier Data table Data Cluster 0 (zero), Cluster 1 (one), Cluster 2 (two), and 3 (three).

The next step involves evaluating with the K-Means method for the process of grouping data to be processed from the computer network VPN log data set and the output results using the K-Means method for grouping data from the computer network VPN log dataset. The test data results can be visualized through graphic images: Graph Outlier Data Cluster 0 (zero), outlier data Cluster 1 (one), outlier data Cluster 2 (two), and outlier data Cluster 3 (three).

The findings of this study are consistent with earlier research by Zhang et al. [23], who investigated outlier detection in wireless sensor networks. Their survey defined outliers as measurements that significantly deviate from the typical pattern of sensed data. Potential sources of these outliers include interference, errors, specific events, and malicious attacks on the network. While their research shares similar objectives, it employs different methodologies. Additionally, research conducted by Gupta et al. [24] focused on outlier detection in time series data, surveying temporal outlier detection from a computational perspective within the computer science community. Advances in hardware have facilitated the development of various temporal data collection methods, while software advancements have led to diverse data management techniques. The findings of this research show significant differences from previous studies, underscoring its unique contributions and novelty in the field.

5. Conclusion

Based on the results of research conducted in developing a model to detect anomalous activity in accessing authentication log data in a network via VPN, after testing the model using existing data and using the methods described in the previous chapter, the following conclusions can be drawn:

According to the proposed model design in this study, the author found that by using the K-Mean method, it can group access to company computer network authentication log data, which can help classify data that has the same characteristics. With this grouping, anomalous activity access can be detected from each characteristic based on sample data of two attributes: user ID and timestamp activity. Therefore, it increases awareness of activities performed by users outside the norm.

By applying the K-NN method, the author can classify the data based on the score of each data row to measure the distance between points. The group that has the least total data is defined as outlier data. This can be seen from the evaluation results based on the graph, where the data points that are considered, outliers are located far from other data points. The K-NN method is influenced by the given k value. In this case the determination of the k value can be adjusted to the results of the company's investigation.

The CRISP-DM method used in this study can make this study more systematic according to the stages described in the methodology, and further discussion focuses on what will be produced according to business needs based on data owned by the company. With this method, the research process becomes more effective in detecting users who access the company's internal network log data, so that the company can detect outlier users earlier and reduce the potential risk of losing company data.

6. Declarations

6.1. Author Contributions

Conceptualization, N.L. and W.B.; methodology, N.L.; software, W.B.; validation, N.L. and W.B.; formal analysis, N.L.; investigation, N.L. and W.B.; resources, N.L. and W.B.; data curation, N.L. and W.B.; writing—original draft preparation, N.L. and W.B.; writing—review and editing, N.L.; visualization, W.B.; supervision, N.L.; project administration, W.B.; funding acquisition, N.L. All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding and Acknowledgements

This research can be completed with the support of all parties who have assisted in data collection about user access data through computer networks at the Telecommunication company. As well, we would like to express my sincere gratitude to the Research Interest Group on Quantitative & Data Sciences (RIG Q&DS) of Bina Nusantara University for their invaluable support and facilitation in fostering collaboration and stimulating insightful discussions throughout the course of this research work.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

7. References

[1] Singh, K. K. V. V., & Gupta, H. (2016). A new approach for the security of VPN. ACM International Conference Proceeding Series, 04-05-March-2016(2016). doi:10.1145/2905055.2905219.

[2] Alshalan, A., Pisharody, S., & Huang, D. (2016). A Survey of Mobile VPN Technologies. IEEE Communications Surveys and Tutorials, 18(2), 1177–1196. doi:10.1109/COMST.2015.2496624.

[3] Smith, K. T., Martin, H. M., & Smith, L. M. (2014). Human trafficking: A global multi-billion dollar criminal industry. International Journal of Public Law and Policy, 4(3), 293-308. doi:10.1504/IJPLAP.2014.063006.

[4] CNN Indonesia. (2020). The Difference between the Cases of Denny Siregar - Telkomsel vs. Tokopedia - Bukalapak. CNN Indonesia, Jakarta, Indonesia. Available online: https://www.cnnindonesia.com/teknologi/20200720073117-185-526519/beda-kasus-denny-siregar-telkomsel-vs-tokopedia-bukalapak (accessed on November 2024).

[5] Wang, H., Bah, M. J., & Hammad, M. (2019). Progress in Outlier Detection Techniques: A Survey. IEEE Access, 7, 107964–108000. doi:10.1109/ACCESS.2019.2932769.

[6] Gustientiedina, G., Adiya, M. H., & Desnelita, Y. (2019). Application of K-Means Algorithm for Drug Data Clustering. National Journal of Technology and Information Systems, 5(1), 17–24. doi:10.25077/teknosi.v5i1.2019.17-24.

[7] Mandhare, H. C., & Idate, S. R. (2017). A comparative study of cluster based outlier detection, distance based outlier detection and density based outlier detection techniques. Proceedings of the 2017 International Conference on Intelligent Computing and Control Systems, ICICCS 2017, 931–935. doi:10.1109/ICCONS.2017.8250601.

[8] Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, Massachusetts, United States.

[9] Dang, T. T., Ngan, H. Y. T., & Liu, W. (2015). Distance-based k-nearest neighbors outlier detection method in large-scale traffic data. International Conference on Digital Signal Processing, DSP, 2015-September, 507–510. doi:10.1109/ICDSP.2015.7251924.

[10] Radovanović, M., Nanopoulos, A., & Ivanović, M. (2015). Reverse nearest neighbors in unsupervised distance-based outlier detection. IEEE Transactions on Knowledge and Data Engineering, 27(5), 1369–1382. doi:10.1109/TKDE.2014.2365790.

[11] Andrian, B., Simanungkalit, T., Budi, I., & Wicaksono, A. F. (2022). Sentiment Analysis on Customer Satisfaction of Digital Banking in Indonesia. International Journal of Advanced Computer Science and Applications, 13(3), 466–473. doi:10.14569/IJACSA.2022.0130356.

[12] Ranjan, S., & Mishra, S. (2020). Comparative Sentiment Analysis of App Reviews. 2020 11th International Conference on Computing, Communication and Networking Technologies, ICCCNT 2020. doi:10.1109/ICCCNT49239.2020.9225348.

[13] Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., & Uthurusamy, R. (Eds.). (1996). Advances in knowledge discovery and data mining. American Association for Artificial Intelligence, United States. doi:10.2307/1271414.

[14] Gullo, F. (2015). From patterns in data to knowledge discovery: What data mining can do. Physics Procedia, 62, 18–22. doi:10.1016/j.phpro.2015.02.005.

[15] Martinez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernandez-Orallo, J., Kull, M., Lachiche, N., Ramirez-Quintana, M. J., & Flach, P. (2021). CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. IEEE Transactions on Knowledge and Data Engineering, 33(8), 3048–3061. doi:10.1109/TKDE.2019.2962680.

[16] Wirth, R. (2000). CRISP-DM: Towards a Standard Process Model for Data Mining. Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining, 24959, 29–39.

[17] Bošnjak, Z., Grljević, O., & Bošnjak, S. (2009). CRISP-DM as a framework for discovering knowledge in small and medium sized enterprises' data. Proceedings - 2009 5th International Symposium on Applied Computational Intelligence and Informatics, SACI 2009, 509–514. doi:10.1109/SACI.2009.5136302.

[18] Osman, A. S. Data mining techniques: Review. International Journal of Data Science Research, 2(1), 1–4.

[19] Kohout, J., & Pevny, T. (2015). Unsupervised detection of malware in persistent web traffic. ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings, 1757–1761. doi:10.1109/ICASSP.2015.7178272.

[20] Mutua, N. M., & Matoušek, P. (2021). Outlier Detection in Smart Grid Communication. arXiv, preprint arXiv:2108.12781. doi:10.48550/arXiv.2108.12781.

[21] Jones, P. J., James, M. K., Davies, M. J., Khunti, K., Catt, M., Yates, T., Rowlands, A. V., & Mirkes, E. M. (2020). FilterK: A new outlier detection method for k-means clustering of physical activity. Journal of Biomedical Informatics, 104, 103397. doi:10.1016/j.jbi.2020.103397.

[22] Kiani, R., Keshavarzi, A., & Bohlouli, M. (2020). Detection of Thin Boundaries between Different Types of Anomalies in Outlier Detection Using Enhanced Neural Networks. Applied Artificial Intelligence, 34(5), 345–377. doi:10.1080/08839514.2020.1722933.

[23] Zhang, Y., Meratnia, N., & Havinga, P. (2010). Outlier detection techniques for wireless sensor networks: A survey. IEEE Communications Surveys and Tutorials, 12(2), 159–170. doi:10.1109/SURV.2010.021510.00088.

[24] Gupta, M., Gao, J., Aggarwal, C. C., & Han, J. (2014). Outlier Detection for Temporal Data: A Survey. IEEE Transactions on Knowledge and Data Engineering, 26(9), 2250–2267. doi:10.1109/TKDE.2013.184.

[25] Zhao, S., Li, W., & Cao, J. (2018). A user-adaptive algorithm for activity recognition based on K-means clustering, local outlier factor, and multivariate gaussian distribution. Sensors (Switzerland), 18(6), 1850. doi:10.3390/s18061850.

[26] Chapple, M. J., Chawla, N., & Striegel, A. (2007). Authentication anomaly detection: A case study on a virtual private network. MineNet'07: Proceedings of the Third Annual ACM Workshop on Mining Network Data, 17–22. doi:10.1145/1269880.1269886.