Accredited Ranking SINTA 2
Decree of the Director General of Higher Education, Research and Technology, No. 158/E/KPT/2021
Validity period from Volume 5 Number 2 of 2021 to Volume 10 Number 1 of 2026

Published online on: **http://jurnal.iaii.or.id**

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 7 No. 5 (2023) 1064 - 1076 ISSN Media Electronic: 2580-0760

K-Means Algorithm Implementation for Project Health Clustering

Ajeng Arifa Chantika Rindu[1], Ria Astriratma[2], Ati Zaidiah[3]
[1]Informatics, Faculty of Computer Science, Universitas Pembangunan Nasional "Veteran" Jakarta, Jakarta, Indonesia
[2,3]Information System, Faculty of Computer Science, Universitas Pembangunan Nasional "Veteran" Jakarta, Jakarta, Indonesia
[1]ajengacr@upnvj.ac.id, [2]astriratma@upnvj.ac.id, [3]atizaidiah@upnvj.ac.id

Abstract

Indonesia has several companies that are engaged in the telecommunications sector. Various projects run in parallel to support the success of telecommunications companies. A project's potential can boost the company's revenue and productivity. On the other hand, there are some risks that need to be considered for every project when it is about to start. Project data is recorded from start to finish so that the project's progress and improvements can be monitored and analyzed. As the project runs, the project team at one of Indonesia's telecommunication companies, which is responsible for the processes leading to project success, requires a project health category. Therefore, this study is conducted to develop a process for clustering project health, which is included in a type of unsupervised learning that runs on unlabeled data. One of the clustering algorithms is K-Means, which groups data based on similar criteria. Researchers also use dimensionality reduction with the Principal Component Analysis (PCA) method to determine its impact on the clustering process with the K-Means algorithm. From this study, the researcher obtained three clusters or project health categories, consisting of clusters 0, 1, and 2. Evaluation results with the Calinski-Harabasz Index showed that the K-Means model on the dimensionality reduction data with PCA performed better than the standard K-Means model with a Calinski-Harabasz Index value of 55633,12776405707, which is higher than 25914,578262576793.

Keywords: project; project health; clustering; K-Means

1. Introduction

Project management has experienced rapid development in recent years, especially because of digitalization in almost all fields. The number of projects that run in a certain period requires managers to help identify these projects to see their potential and risks. Project health is an important component that shows the status of an overall project running towards project success. This component is important for a company because it can directly affect client satisfaction, productivity, and business success. Project health can be measured using various indicators including the time component, financial success, employee productivity, budgeting funds, and the quality of the project itself [1]. Every project certainly has a baseline or target for project completion in order to make a profit according to the expected time period [2].

Numerous businesses in Indonesia are devoted to the telecommunications industry. One of the main projects of the telecommunications company is in the field of radio access networks (RAN). This project can be in the form of tower construction, network installation, or capacity increases. The number of projects running in parallel, complex data, and a long processing time for manually determining project health make the project team at one of Indonesia's telecommunications companies feel difficult.

Machine learning is a computational paradigm in which the capacity for problem-solving is built from previous examples [3]. The unsupervised learning method is a type of machine learning that handles unlabeled data [4]. In this type of learning, it is assumed that all training examples are unlabeled. Unlabeled examples are learned depending on their similarities [3]. Clustering includes several algorithms that can be used based on the data and problems. In this study, we used the K-Means algorithm to cluster project health, an algorithm that is widely used and has proven results for clustering problems.

Dimensionality reduction is a technique for finding lower-dimensional representations of data while retaining properties that are key to a particular problem. The classic technique for dimensionality reduction is the principal component analysis (PCA).

Accepted: 20-06-2023 | Received in revised: 16-09-2023 | Published: 30-09-2023

Dimensionality reduction is most effective when used on large datasets with a large number of input variables that are also correlated [5]. In this study, because we used a dataset with a sufficiently large dimension, we also tried the use of PCA so that it could be compared with the results of K-Means clustering on data without using PCA.

Clustering is a machine learning technique that is used to see hidden models, relationships, or summarize data. Technically, clustering aims to be a process that relies on the notion of similarity, which is often based on distance measurement [6]. Currently, clustering methods are widely used in studies worldwide. Several researchers have conducted studies on this topic by using various types of data. One of these is research on the quality of power plants. Aksan et al. compared the use of nonhierarchical approaches with the K-Means algorithm and hierarchical approaches using agglomerative clustering. Consequently, the optimal number of clusters is three. Aksan et al. also evaluated the clusters that had been obtained, one of which was the Calinski-Harabasz Index (CH Index). Based on the CH Index value, it is concluded that in general, K-Means performance is better than agglomerative clustering [7]. Not only that, previously, Zubair et al. conducted research related to clustering using the K-Means algorithm combined with the PCA (Principal Component Analysis) technique to determine the centroid. Zubair et al. claimed that their proposed model using K-Means and PCA outperformed the K-Means clustering algorithm in real-world application cases and reduced computing power [8].

Based on the literature reviewed and the problems found, we conducted this study in order to group and find out the project health of the RAN project at one of Indonesia's telecommunications companies. The aim of this research is to help the project team at a telecommunications company monitor the progress of the RAN project and follow up on projects that have poor project health in the hope that the project can later be completed on time. By conducting this study, we hope to facilitate the project team at a telecommunications company in determining project health based on the project baseline in a fairly short time and in a more effective way.

2. Research Methods

To achieve this study objectives, we compiled the research stages presented in the form of a flow chart which can be seen in Figure 1.

2.1 Problem Identification and Formulation

First, we started this study by identifying and formulating problems that exist in one of Indonesia's telecommunications companies, especially in the project team. In accordance with company regulations, hereinafter the name of this company is referred to as PT XYZ. The problem experienced by the project team is the lack of effectiveness in determining project health in a Radio Access Network (RAN) project if it is determined manually by the project team.

Based on the results of interviews with the project team at PT XYZ, there are several stages of the RAN project at PT XYZ, from project planning to project closure. These stages are represented in Figure 2.

Purchase Order (PO): The first stage in the RAN project at PT XYZ is the existence of a purchase order, which is a commercial document issued when the vendor and the company have agreed to work together to build the project. POs are used to initiate purchases and provide a means of ensuring that transactions are covered by the right contract.

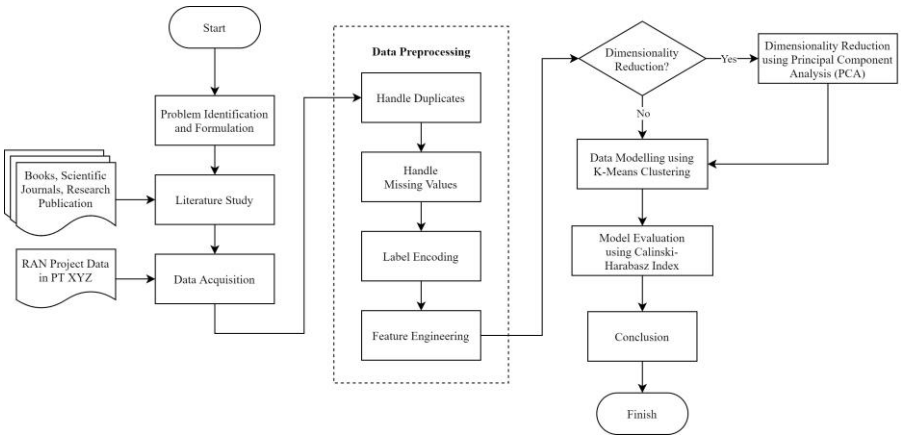

Figure 1. Flow Chart of the Research Stages

Purchases of goods or services are processed through the company's financial system and must be preceded by a purchase order given to the vendor [9].

Project Charter (PC): This step is one that can be done in parallel with the PO and kick-off meeting. A project charter is the process of developing a document that

DOI: https://doi.org/10.29207/resti.v7i5.5181

formally certifies the existence of the project and authorizes the project manager to apply organizational resources to project activities.

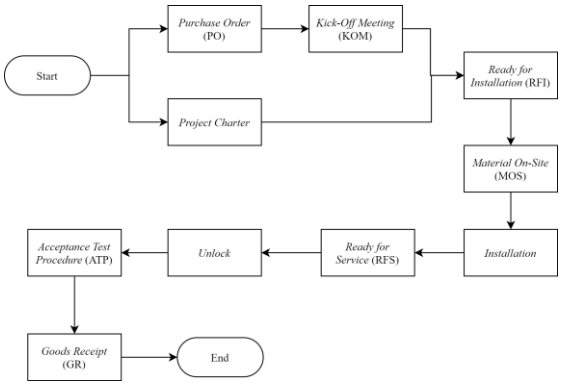

Figure 2. Flow Chart of the Radio Access Network Project Development (PT XYZ's Document)

Kick-Off Meeting (KOM): According to the discussion with the PT XYZ project team, the Kick-Off Meeting (KOM) is a strategy to increase a project's chance of success. KOM also holds the key to ensuring a shared understanding of the project objectives. At this stage, stakeholders begin to determine the target or project baseline based on the contract between the two parties, timeline, organizational structure, business processes, and Responsible, Accountable, Consulted, and Informed (RACI) matrix.

Ready for Installation (RFI): RFI is a project status where the tower is ready for installation of base transceiver station tower equipment.

Material On-Site (MOS): As the name implies, material on site, or MOS, indicates the stage at which construction materials have arrived at the construction site [10].

Installation: Base Transceiver Station (BTS) tower installation stage.

Ready for Service (RFS): Configuration of the installed tower, and if this configuration is successful, it will be connected to the network.

Unlock: When the tower construction is in the unlock stage, the network is ready to use.

Acceptance Test Procedure (ATP): Based on the discussion with project team at PT XYZ, Acceptance Test Procedure (ATP) is a stage that can take the form of checking devices, validating sites with field testing, and verifying device performance to be able to take the project to commercial services. At this stage, there are various conditions and testing criteria that must be met so that the project development results can be accepted and completed. ATP is one of the essential stages because it is the main step in the risk management of a project.

Goods Receipt (GR): Goods Receipt is the stage when the company has received the goods and or services that have been ordered using a purchase order (PO) [11]. GR is not always at the end, but it depends on the vendor contract. There are some vendors who also have GR stages in the middle of the project, for example, a goods receipt when the materials have been received.

Therefore, based on the problems that have been identified and formulated, this research focuses on how researchers can determine the category of project health in RAN projects at PT XYZ based on project baseline using the K-Means clustering algorithm.

2.2 Literature Study

Literature studies were conducted to understand the important terms that were further examined at the research stage. We conducted observations and searches for books, journals, and other research publications related to research topics and problems to obtain information that can support research. With this basic knowledge, it is hoped that we can proceed to the next stage of research and complete it properly.

2.3 Data Acquisition

In this study, the data used and processed in the data modeling were obtained from PT XYZ's website-based Project Management Information System (PMIS), managed by the project team. The data acquired by are part of PT XYZ's 2022 Radio Access Network (RAN) project.

2.4 Data Preprocessing

The input data in the machine learning process has its own standard and structure that depend on the problem and the machine learning work to be performed. Data preprocessing is a pivotal step in both data analytics and machine learning. However, it is crucial to understand that the preprocessing performed for data analytics is significantly different from that of machine learning. [12]. Missing data and noise are some of the issues addressed by data cleaning [13]. Data preprocessing is performed when there are still data that need to be cleaned from the dataset obtained. We also carried out this stage to minimize errors that occurred before entering the data modeling process. This stage consists of handling duplicate data, missing values, label encoding, and feature engineering.

Handling duplicate data: The first stage of data preprocessing involves handling duplicate data and data with the same project identity attributes in more than one data row. The existence of duplicate identities can cause the addition of the same sample weight; therefore, bias can occur, which results in the performance of the machine learning model. The treatment that the researcher did was to delete duplicate data and not impute it because identity is something unique and does not depend on other data or attributes.

DOI: https://doi.org/10.29207/resti.v7i5.5181

Handling missing values: In addition to duplicate data, we checked for missing values. If there are many missing values in the dataset, the missing values must be handled because they can affect the machine learning model that is built. In this study, missing values were removed because the raw data were obtained from PMIS PT XYZ, which could have changed the originality of the data if it was imputed. Similar to duplicate data, if the missing values are not handled properly, the machine learning model can be biased and reduce data representation.

Label encoding: Clustering is a method that can only process numerical data. However, it is normal to find several types of variables or features in data. Not only numerical variables, but categorical variables are also commonly found in data. Categorical variables are variables that can be classified into two main types, namely nominal and ordinal. Nominal variables are variables that have two or more categories that do not consider the order of the categories. On the other hand, ordinal variables are variables with categories that have certain levels. Thus, data preprocessing is required to convert categorical data into numerical. One commonly used method is the label encoding process. The label encoding process encodes all categories into numeric labels so that they can be processed in the clustering stage [14].

In this study, label encoding is required for columns that have text as their data type and that consist of several categorical values. In this stage, we converted categorical columns into numerical data consisting of several values so that it can facilitate the next process, namely, data modeling.

Feature engineering: Features are numerical representations of an aspect of the raw data. Features sit between the data and the model in the machine learning pipeline. The number of features in the data is also essential to machine learning. If it does not have enough informative features, then this can result in the model not being able to perform the final task. If there are many features but they are not relevant, then the model will be more expensive and difficult to train. In this regard, in machine learning, there is a term called feature engineering, which is an activity to extract features from previously acquired raw data and then transform them into a form that is suitable for machine learning models. Therefore, if appropriate to the data, problem, and research objectives, feature engineering can enable machine learning to produce higher quality output [15].

In this stage, we extracted new columns from the existing columns in the raw data, and then transformed the columns into a form that was suitable for machine learning models. After preprocessing the data, we conducted experiments under two conditions. In the first experiment, clustering was conducted using the standard K-Means algorithm. Furthermore, in the second experiment, we reduced the data dimension using the Principal Component Analysis (PCA) method and then continued the clustering process using the K-Means algorithm. These two experiments were conducted to determine the effect of PCA on the clustering process using the K-Means algorithm.

2.5 Dimensionality Reduction using Principal Component Analysis (PCA)

Dimensionality reduction is one of the processes that can reduce the representation of data that were previously high-dimensional into lower dimensions while maintaining the main components of the data [5]. The main goal of PCA is to reduce the dimensionality of a dataset where there are a large number of interrelated variables and also retain as much of the variation in the data as possible. This reduction is achieved by transforming the variable data into a new set of variables called principal components (PCs) that are uncorrelated and ordered so that the first few variables can represent and retain most of the variation in the overall variables [16]. In general, the dimension reduction stage with PCA is described as follows:

Calculate the covariance matrix: The main goal of reducing the dimensions is to obtain the main components without eliminating the data characteristics. In the process of obtaining the principal components, we need a covariance matrix to determine the correlation between the columns.

The simple PCA approach is as follows. Suppose there are data samples$X = [x_1\ x_2 \ldots x_n] \in \mathbb{R}^{d \times n}$, where each sample is in column vector form with the covariance matrix defined as in Formula (1).

$$C = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^T \quad (1)$$

$\bar{x}$ is the mean or average of the sample. After that, a low-dimensional basis that covers most of the data variance can be found by extracting the most significant eigenvectors from the covariance matrix C.

Eigenvalue Decomposition: The covariance matrix obtained from the previous stage was used for the Eigenvalue calculation. Eigenvalue is a constant that indicates the level of representation of a feature or attribute relative to the overall attribute. The Eigenvalue formula is represented in Formula (2).

$$(C - \lambda I)\ v = 0; v\text{^}T\ v = 1 \quad (2)$$

$v \in \mathbb{R}^d$ a is the Eigenvector and λ is the corresponding eigenvalue. The eigenvalues describe the variance retained by eigenvectors.

Sort the Eigen Values and obtaining the Principal Component (PC): Eigenvalues obtained at the previous point are sorted by the largest Eigenvalue. The greater

DOI: https://doi.org/10.29207/resti.v7i5.5181

the Eigenvalue, the more representative the component is of the overall data features.

Transformation of the New Dataset of PCA Results: The principal component (PC) obtained can be used to transform the dataset into a new dataset with smaller dimensions. The PC is the main component resulting from the PCA dimension reduction process, which is representative of other attributes. A new dataset can be obtained by multiplying the Eigenvector with the initial dataset.

We performed dimension reduction on the data using the principal component analysis (PCA) method to compare the performance of the clustering result model using the K-Means algorithm on data through PCA and without PCA (standard K-Means).

2.6 Data Modelling using K-Means Clustering

K-Means is one of the clustering algorithms that groups data based on the average vector or mean of the cluster. Parameter K is the number of clusters that need to be determined before the K-Means clustering process begins. The cluster mean vector is given as a cluster prototype in the algorithm execution. K-Means is a type of unsupervised learning because it belongs to a learning paradigm where data has no label and "learns" based on the similarity of existing data. This learning process involves optimizing the cluster prototype based on the similarity between the prototype and individual items [17].

The K-Means algorithm is an iterative method that consists of partitioning a set of n objects into $k \geq 2$ clusters so that the objects are similar to each other and different from other clusters. In general, the K-Means clustering algorithm consists of four main steps:

Step 1: Determine the desired number of clusters (K) or groups. Then, k points are randomly generated in the field, where k is used as the initial centroid.

Step 2: Calculate the distance from each object to all the centroids. This distance is commonly referred to as the Euclidean distance. The formula for calculating the Euclidean distance between two points is given in Formula (3).

$$d(x, \mu) = \sqrt{(\mu_1 - x_1)^2 + (\mu_2 - x_2)^2} \quad (3)$$

Step 3: Move each object to the cluster with the closest centroid distance.

Step 4: If there is a change, then the process continues to the centroid calculation stage. The new centroid is calculated using the average value of the objects that are members of each cluster. The process will repeat from the second step. Otherwise, if there is no change, the K-Means algorithm will stop running [18].

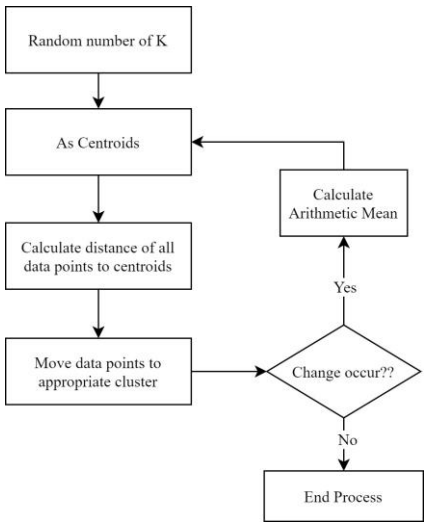

Figure 3. Flow Chart of the K-Means Algorithm [18]

The K-Means algorithm is generally represented in Figure 3. Clean and complete data indicate that the data are ready to be modeled using the K-Means clustering algorithm.

Subsequently, we started the data processing using machine learning. The process carried out in the data modeling stage is clustered with the K-Means algorithm using the Scikit-Learn library and package in the Python programming language.

In this study, the number of clusters is determined using the elbow method. First, we built a machine learning model using clean data (without the PCA process) using K-Means. Then, the study continued by building a model on clean data that went through the PCA process. The results of both the experiments were stored in different data frames to evaluate the performance of each model.

2.7 Model Evaluation using Calinski-Harabasz Index

The Calinski-Harabasz Index is a commonly used measurement index to evaluate the quality of cluster results [19]. Formally, cluster evaluation is defined as quantitatively evaluating clustering results. The motivation for this evaluation process is that almost every clustering algorithm will find clusters in data sets that do not even have clusters naturally. Therefore, a validation measure is needed so that it can be known how well the clustering results have been obtained.

In this study, we built two models, which consisted of the standard K-Means model and the K-Means with PCA dimensionality reduction model. To compare the performance of the two built models, we used the Calinski-Harabasz index based on Formula (4).

$$CH(k) = \frac{SS_B}{SS_W} \times \frac{(N-k)}{(k-1)} \quad (4)$$

DOI: https://doi.org/10.29207/resti.v7i5.5181

CH is the Calinski-Harabasz index, k is the number of clusters, N is the number of data point, SS_B is the sums of squares between-cluster, and SS_W is the sums of squares within-cluster.

SS_B and SS_W are represented in Formula (5) and (6) respectively.

$$SS_B = \sum_{i=1}^{k} N_i * d^2(\mu_i, M) \quad (5)$$

$$SS_W = \sum_{i=1}^{k} \sum_{x=c_i} d^2(\mu_i, x) \quad (6)$$

We evaluated the machine learning models using the Calinski-Harabasz Index, which shows how close the characteristics are between members in the same cluster and how far one cluster is from another.

2.8 Conclusion

In the last stage of the study, we analyzed the characteristics of each cluster and examined the advantages and disadvantages of the machine learning model. From the evaluation and analysis, we can draw conclusions from this study that has been done.

3. Results and Discussions

3.1 Data Acquisition

In this study, the dataset used was obtained from PT XYZ's website-based Project Management Information System (PMIS). The data used were obtained from the RAN 2022 projects. This dataset contains information on the RAN project managed by PT XYZ and related vendors. We checked the number of rows and data in the dataset using the shape function from the Pandas library. After checking, the amount of raw data in this dataset was 65074 rows and 13 columns. Some examples of the data rows are listed in Table 1. The Index column in the table shows the data sequence number or index when the data are first loaded with Python.

Table 1. RAN Project Data
(PT XYZ's Project Management Information System)

Index	PT Index	Region	NE Type	Build Type	PO Date	…
0	22BNHO	EAST	NaN	EXISTING	NaN	…
1	22MS8P	EAST	NaN	EXISTING	NaN	…
2	22UFXL	EAST	NaN	EXISTING	NaN	…
3	22XPQA	EAST	G1800	EXISTING	2022-05-17	…
4	22TMTZ	EAST	G1800	EXISTING	2022-05-17	…
5	226UZ7	WEST	G900	B2S	2022-03-01	…
…	…	…	…	…	…	…
10	22RWIB	EAST	G900	EXISTING	2022-12-20	…
11	22OZWD	EAST	G1800	EXISTING	2022-03-01	…
12	22YKTD	EAST	G900	EXISTING	2022-03-01	…
…	…	…	…	…	…	…
332	22MS3C	CENTRAL	G900	EXISTING	2022-05-17	…
…	…	…	…	…	…	…
335	22JTFJ	CENTRAL	G1800	EXISTING	2022-12-23	…
336	22KLRT	CENTRAL	G900	EXISTING	2022-05-17	…
…	…	…	…	…	…	…
713	22OHQD	EAST	G900	EXISTING	2022-02-10	…
714	22K4WX	EAST	G900	EXISTING	2022-12-23	…
…	…	…	…	…	…	…
3072	22YESL	CENTRAL	G900	EXISTING	NaN	…
3073	22YESL	CENTRAL	G900	EXISTING	NaN	…
…	…	…	…	…	…	…
65067	22WFV8	EAST	L900	EXISTING	2022-12-20	…
…	…	…	…	…	…	…
65073	22RN2B	JABO	G1800	EXISTING	2022-11-24	…

3.2 Data Preprocessing

After acquisition, the dataset from PMIS PT XYZ was found to still have many deficiencies, including duplicate data and missing values. Therefore, we performed a preprocessing stage to obtain clean data and minimize errors before entering the data modeling process. The data preprocessing stage carried out consists of handling duplicate data, missing values, label encoding, and feature engineering.

Handling duplicate data: The obtained RAN project data have identity attributes obtained from four columns, namely PT Index, Region, NE Type, and Build Type, which show the identity of each RAN project running at PT XYZ. Therefore, we need to check whether there is more than one row of data that has the same value in the four attributes in the data, so that the data is clean when processed using machine learning. We checked duplicate data using the duplicated function in Python. The duplicated function returns a Boolean value that is true or false, such that when there is a data row with the same PT Index, Region, NE Type, and Build Type, the output of the function returns a true value. Conversely, if there are no duplicate identity columns, the output of the function is false. To determine the total number of duplicates, we combined the use of the duplicated function with the aggregate sum function, which returns the sum of the true values from the duplicate checking result. After the functions were run, we obtained 4101 rows of duplicate identity attributes. We used the *drop_duplicates* function from the Pandas library with the subset parameters set as the PT Index, Region, NE Type, and Build Type as attributes that determine if duplicates will

DOI: https://doi.org/10.29207/resti.v7i5.5181

be removed, and the *inplace* parameter with a value of True that will change the initial dataframe to reflect the result of removing duplicate values. By default, the *drop_duplicates* function in the Pandas library leaves the first row among other duplicate values. For example, when there are three rows of data with the same PT Index, Region, NE Type, and Build Type, the first row remains in the dataset. Table 2 shows the data without duplicate values for PT Index, Region, NE Type, and Build Type, with a total of 60973 rows.

Table 2. RAN Project Data After Removing Duplicates

Index	PT Index	Region	NE Type	Build Type	PO Date	…
0	22BNHO	EAST	NaN	EXISTING	NaN	…
1	22MS8P	EAST	NaN	EXISTING	NaN	…
2	22UFXL	EAST	NaN	EXISTING	NaN	…
3	22XPQA	EAST	G1800	EXISTING	2022-05-17	…
4	22TMTZ	EAST	G1800	EXISTING	2022-05-17	…
5	226UZ7	WEST	G900	B2S	2022-03-01	…
…	…	…	…	…	…	…
10	22RWIB	EAST	G900	EXISTING	2022-12-20	…
11	22OZWD	EAST	G1800	EXISTING	2022-03-01	…
12	22YKTD	EAST	G900	EXISTING	2022-03-01	…
…	…	…	…	…	…	…
332	22MS3C	CENTRAL	G900	EXISTING	2022-05-17	…
…	…	…	…	…	…	…
335	22JTFJ	CENTRAL	G1800	EXISTING	2022-12-23	…
336	22KLRT	CENTRAL	G900	EXISTING	2022-05-17	…
…	…	…	…	…	…	…
713	22OHQD	EAST	G900	EXISTING	2022-02-10	…
714	22K4WX	EAST	G900	EXISTING	2022-12-23	…
…	…	…	…	…	…	…
3072	22YESL	CENTRAL	G900	EXISTING	NaN	…
…	…	…	…	…	…	…
65067	22WFV8	EAST	L900	EXISTING	2022-12-20	…
…	…	…	…	…	…	…
65073	22RN2B	JABO	G1800	EXISTING	2022-11-24	…

In Table 2, the duplicates were removed. For example, the data with indices 3072 and 3073 initially had the same PT Index, Region, NE Type, and Build Type of 22YESL, CENTRAL, G900, and EXISTING, respectively. Therefore, after the duplicate data removal stage, the data with index 3073 no longer appear, and the data with index 3072 is already unique.

Handling missing values: We removed the missing values from the raw data because the attributes in the data from the PMIS were real data based on the ongoing RAN project. Therefore, these values cannot be replaced. We used the *dropna()* method to remove missing values. We removed missing values gradually, starting with the Region and NE Type columns, which are the identity columns of the project, followed by removing missing values in the date columns of the RAN project stages, which consisted of PO Date, KOM Date, RFI Date, MOS Date, Installation Date, RFS Date, Unlock Date, ATP Date, and GR Date. The total amount of data after cleaning amounted to 36086 rows. Table 3 shows the dataset that was cleaned from missing values.

Table 3. RAN Project Data Without Missing Values

Index	PT Index	Region	NE Type	Build Type	PO Date	…
3	22XPQA	EAST	G1800	EXISTING	2022-05-17	…
4	22TMTZ	EAST	G1800	EXISTING	2022-05-17	…
10	22RWIB	EAST	G900	EXISTING	2022-12-20	…
11	22OZWD	EAST	G1800	EXISTING	2022-03-01	…
12	22YKTD	EAST	G900	EXISTING	2022-03-01	…
332	22MS3C	CENTRAL	G900	EXISTING	2022-05-17	…
335	22JTFJ	CENTRAL	G1800	EXISTING	2022-12-23	…
336	22KLRT	CENTRAL	G900	EXISTING	2022-05-17	…
713	22OHQD	EAST	G900	EXISTING	2022-02-10	…
714	22K4WX	EAST	G900	EXISTING	2022-12-23	…
…	…	…	…	…	…	…
65067	22WFV8	EAST	L900	EXISTING	2022-12-20	…

Label encoding: Label encoding is performed to convert categorical attributes or features into numeric values. These attributes are the Region, NE Type, and Build Type. We used the map method from the Pandas library, in which categorical attributes that were originally of the object data type were mapped to integers. Region, NE Type, and Build Type attributes indicate information on the project development area, development type, and cellular network speed level, respectively. The label encoding results in Table 4 shows the values of the Region, NE Type, and Build Type attributes before and after encoding.

Table 4. Label Encoding Result

Attribute Name	Before Encoding	After Encoding
Region	WEST	0
	JABO	1
	CENTRAL	2
	EAST	3
NE Type	G900	0
	G1800	1
	U900	2
	U2100	3
	L900	4

DOI: https://doi.org/10.29207/resti.v7i5.5181

Attribute Name	Before Encoding	After Encoding
	L1800	5
	L2100	6
Build Type	EXISTING	0
	COLLO	1
	B2S	2

Feature engineering: Feature engineering was started by creating the *get_days_between* function to obtain the number of days between stages. The *get_days_between* function accepts two parameters: the date it enters the first stage and the date it enters the second stage. The function determines the difference in days between the two dates and returns a duration value in days using the *dt.days* function from Pandas Python. This function calculates the difference in days between the two dates, which were previously set as the parameters. After the data were preprocessed, we obtained clean data that could be processed using machine learning. However, the dataset still contains the PT Index identity column and date column at each stage. For the clustering process, we did not use the PT Index and date columns because the clustering process did not require the use of these columns. To observe patterns from the data using machine learning, researchers used the Region column, Build Type, NE Type, and duration column between stages (PO-KOM, KOM-RFI, RFI-MOS, MOS-Install, Install-RFS, RFS-Unlock, Unlock-ATP, and ATP-GR).

Table 5 shows the clean data processed by machine learning.

Table 5. Clean Data After Preprocessing

Index	Region (a)	NE Type (b)	Build Type (c)	PO-KOM (d)	KOM-RFI (e)	…
3	3	1	0	-128	41	…
4	3	1	0	-25	44	…
10	3	0	0	-136	122	…
11	3	1	0	-125	85	…
12	3	0	0	-125	124	…
332	2	0	0	-142	69	…
335	2	1	0	-119	69	…
336	2	0	0	40	22	…
713	3	0	0	29	12	…
714	3	0	0	-120	69	…
…	…	…	…	…	…	…
65067	3	4	0	-108	69	…

3.3 Dimensionality Reduction using Principal Component Analysis (PCA)

After the preprocessing stage, we performed dimensional reduction using the Principal Component Analysis (PCA) method because the dataset used has a large number of attributes, making it difficult to analyze the characteristics of each cluster produced. PCA was used to reduce the dimensionality of this dataset to obtain the principal components of all attributes but still retain as much variation as possible in the data.

Table 6. Covariance Matrix Structure

	a	b	c	d	e	f	g	h	i	j	k
a	var a	cov (a,b)	cov (a,c)	cov (a,d)	cov (a,e)	cov (a,f)	cov (a,g)	cov (a,h)	cov (a,i)	cov (a,j)	cov (a,k)
b	cov (a,b)	var b	cov (b,c)	cov (b,d)	cov (b,e)	cov (b,f)	cov (b,g)	cov (b,h)	cov (b,i)	cov (b,j)	cov (b,k)
c	cov (a,c)	cov (b,c)	var c	cov (c,d)	cov (c,e)	cov (c,f)	cov (c,g)	cov (c,h)	cov (c,i)	cov (c,j)	cov (c,k)
d	cov (a,d)	cov (b,d)	cov (c,d)	var d	cov (d,e)	cov (d,f)	cov (d,g)	cov (d,h)	cov (d,i)	cov (d,j)	cov (d,k)
e	cov (a,e)	cov (b,e)	cov (c,e)	cov (d,e)	var e	cov (e,f)	cov (e,g)	cov (e,h)	cov (e,i)	cov (e,j)	cov (e,k)
f	cov (a,f)	cov (b,f)	cov (c,f)	cov (d,f)	cov (e,f)	var f	cov (f,g)	cov (f,h)	cov (f,i)	cov (f,j)	cov (f,k)
g	cov (a,g)	cov (b,g)	cov (c,g)	cov (d,g)	cov (e,g)	cov (f,g)	var g	cov (g,h)	cov (g,i)	cov (g,j)	cov (g,k)
h	cov (a,h)	cov (b,h)	cov (c,h)	cov (d,h)	cov (e,h)	cov (f,h)	cov (g,h)	var h	cov (h,i)	cov (h,j)	cov (h,k)
i	cov (a,i)	cov (b,i)	cov (c,i)	cov (d,i)	cov (e,i)	cov (f,i)	cov (g,i)	cov (h,i)	var i	cov (i,j)	cov (i,k)
j	cov (a,j)	cov (b,j)	cov (c,j)	cov (d,j)	cov (e,j)	cov (f,j)	cov (g,j)	cov (h,j)	cov (i,j)	var j	cov (j,k)
k	cov (a,k)	cov (b,k)	cov (c,k)	cov (d,k)	cov (e,k)	cov (f,k)	cov (g,k)	cov (h,k)	cov (i,k)	cov (j,k)	var k

Calculate the covariance matrix: Table 6 shows the structure of covariance matrix for this dataset. The column names in Table 6 refer to the column names in Table 5.

For example, we used the Region attribute (a) to calculate the variance and covariance matrix. Table 7 shows an example of data calculation to obtain the covariance matrix based on equation (1).

Table 7. Covariance Matrix Calculation for Region Feature

Component	Calculation
Feature Variance (a)	$C_{(a,a)} = \frac{1}{n-1}\sum_{i=1}^{n}(a_i - \bar{a})^2$
	$C_{(a,a)} = \frac{1}{36086-1} \times 66346{,}316$
	$C_{(a,a)} = \frac{36086}{36085} = 1{,}838612$
Covariance (a,b)	$C_{(a,b)} = \frac{1}{n-1}\sum_{i=1}^{n}(a_i - \bar{a})(b_i - \bar{b})$

DOI: https://doi.org/10.29207/resti.v7i5.5181

Component	Calculation
	$C_{(a,b)} = \frac{1}{36086-1} \times (-1310{,}295)$
	$C_{(a,b)} = \frac{-1310{,}295}{36085} = -0{,}036311$
Covariance (a,c)	$C_{(a,c)} = \frac{1}{n-1}\sum_{i=1}^{n}(a_i - \bar{a})(c_i - \bar{c})$
	$C_{(a,c)} = \frac{1}{36086-1} \times (-996{,}801)$
	$C_{(a,c)} = \frac{-996{,}801}{36085} = -0{,}027624$
Covariance (a,d)	$C_{(a,d)} = \frac{1}{n-1}\sum_{i=1}^{n}(a_i - \bar{a})(d_i - \bar{d})$
	$C_{(a,d)} = \frac{1}{36086-1} \times (-422099{,}505)$
	$C_{(a,d)} = \frac{-422099{,}505}{36085} = -11{,}697367$
Covariance (a,e)	$C_{(a,e)} = \frac{1}{n-1}\sum_{i=1}^{n}(a_i - \bar{a})(e_i - \bar{e})$
	$C_{(a,e)} = \frac{1}{36086-1} \times 83453{,}142$
	$C_{(a,e)} = \frac{83453{,}142}{36085} = 2{,}312682$
Covariance (a,f)	$C_{(a,f)} = \frac{1}{n-1}\sum_{i=1}^{n}(a_i - \bar{a})(f_i - \bar{f})$
	$C_{(a,f)} = \frac{1}{36086-1} \times 349561{,}903$
	$C_{(a,f)} = \frac{349561{,}903}{36085} = 9{,}687180$
Covariance (a,g)	$C_{(a,g)} = \frac{1}{n-1}\sum_{i=1}^{n}(a_i - \bar{a})(g_i - \bar{g})$
	$C_{(a,g)} = \frac{1}{36086-1} \times 5882{,}405$
	$C_{(a,g)} = \frac{5882{,}405}{36085} = 0{,}163015$
Covariance (a,h)	$C_{(a,h)} = \frac{1}{n-1}\sum_{i=1}^{n}(a_i - \bar{a})(h_i - \bar{h})$
	$C_{(a,h)} = \frac{1}{36086-1} \times -19240{,}971$
	$C_{(a,h)} = \frac{-19240{,}971}{36085} = -0{,}533212$
Covariance (a,i)	$C_{(a,i)} = \frac{1}{n-1}\sum_{i=1}^{n}(a_i - \bar{a})(i_i - \bar{i})$
	$C_{(a,i)} = \frac{1}{36086-1} \times 29036{,}631$
	$C_{(a,i)} = \frac{29036{,}631}{36085} = 0{,}804673$
Covariance (a,j)	$C_{(a,j)} = \frac{1}{n-1}\sum_{i=1}^{n}(a_i - \bar{a})(j_i - \bar{j})$
	$C_{(a,j)} = \frac{1}{36086-1} \times 58483{,}516$
	$C_{(a,j)} = \frac{58483{,}516}{36085} = 1{,}620715$
Covariance (a,k)	$C_{(a,k)} = \frac{1}{n-1}\sum_{i=1}^{n}(a_i - \bar{a})(k_i - \bar{k})$
	$C_{(a,k)} = \frac{1}{36086-1} \times 204207{,}169$
	$C_{(a,k)} = \frac{204207{,}169}{36085} = 5{,}659060$

1. Eigenvalue Decomposition

Based on formula (2), we obtained eleven Eigenvalues, as shown in Table 8.

Table 8. Eigenvalue Decomposition Result

Attribute	Eigenvalues
1	$\lambda_1 = 8577{,}688$
2	$\lambda_2 = 6695{,}492$
3	$\lambda_3 = 2523{,}776$
4	$\lambda_4 = 1498{,}168$
5	$\lambda_5 = 171{,}565$
6	$\lambda_6 = 95{,}382$
7	$\lambda_7 = 40{,}100$
8	$\lambda_8 = 25{,}709$
9	$\lambda_9 = 0{,}115$
10	$\lambda_{10} = 1{,}764$
11	$\lambda_{11} = 3{,}28834271$

2. Sort the Eigen Values and obtaining the Principal Component (PC)

Table 9 shows the rank of the Eigenvalues that have been obtained.

Table 9. Eigenvalue Decomposition List

Eigenvalue	Percentage	Rank
$\lambda_1 = 8577{,}688$	$\frac{8577{,}688}{19633{,}04847145} = 43{,}69\%$	1
$\lambda_2 = 6695{,}492$	$\frac{6695{,}492}{19633{,}04847145} = 34{,}103\%$	2
$\lambda_3 = 2523{,}776$	$\frac{2523{,}776}{19633{,}04847145} = 12{,}855\%$	3
$\lambda_4 = 1498{,}168$	$\frac{1498{,}168}{19633{,}04847145} = 7{,}631\%$	4
$\lambda_5 = 171{,}565$	$\frac{171{,}565}{19633{,}04847145} = 0{,}874\%$	5
$\lambda_6 = 95{,}382$	$\frac{95{,}382}{19633{,}04847145} = 0{,}486\%$	6
$\lambda_7 = 40{,}100$	$\frac{40{,}100}{19633{,}04847145} = 0{,}204\%$	7
$\lambda_8 = 25{,}709$	$\frac{25{,}709}{19633{,}04847145} = 0{,}131\%$	8
$\lambda_9 = 0{,}115$	$\frac{0{,}115}{19633{,}04847145} = 0{,}001\%$	11
$\lambda_{10} = 1{,}764$	$\frac{1{,}764}{19633{,}04847145} = 0{,}009\%$	10
$\lambda_{11} = 3{,}28834271$	$\frac{3{,}28834271}{19633{,}04847145} = 0{,}017\%$	9

From Table 9, it can be observed that two principal components can be obtained from the Eigenvalues of λ_1 and λ_2 with percentages of 43.69% and 34.103%, respectively. Eigenvectors can be obtained by multiplying the Eigen values with the covariance matrix.

3. Transformation of the New Dataset of PCA Results

The initial dataset had dimensions of 36086×11, whereas the Eigenvector had dimensions of 11×2. Thus, we obtained a new dataset with dimensions of 36086×2, as shown in Table 10.

Table 10. New Dataset as a Result of Dimensionality Reduction using PCA

Index	Principal Component 1 (PC1)	Principal Component 2 (PC2)
3	133,3249451	-126,6368179
4	30,59927178	30,18983773
10	145,6313574	-66,87745642
11	152,1504787	47,02020671
12	152,5931288	14,46904174
332	173,3323804	54,35417056
335	119,289366	-70,46962289
336	-54,05351183	65,57077309
713	-53,35448999	54,23835902
714	121,5896038	-71,54279649
…	…	…
65067	126,0934894	75,29792295

DOI: https://doi.org/10.29207/resti.v7i5.5181

3.4 Data Modelling using K-Means Clustering

The created model is a machine learning model using the K-Means algorithm. We used the *KMeans* package from the Scikit Learn Python library to utilize functions or methods to process data that have been acquired and cleaned. The K-Means clustering algorithm consists of five steps, as describe:

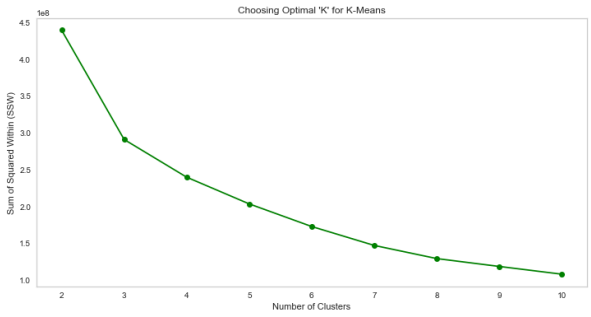

Figure 4. Optimal Number of Cluster Search Results

Determine the number of clusters K and randomly obtain the centroid: The number of clusters is determined using the elbow method. We displayed a line graph visualization in Figure 4. where the coordinate points that appeared showed the sum of squared error values of the data at each number of k.

As shown in Figure 4, an elbow was formed when the number of clusters was three. Therefore, in this study, three clusters k were used.

The K-Means package initializes the centroid randomly in the first iteration. An example of the centroid of each cluster set on the first iteration is presented in Table 11.

Table 11. Centroid on the First Iteration

Cluster	Region	NE Type	Build Type	PO-KOM	…
0	1.74435360	3.0398 2838	0.198818 10	(-) 97.15971 829	…
1	1.64383821	3.4312 6260	0.055758 57	30.79662 300	…
2	1.44790739	2.6857 9061	0.012084 98	48.85752 450	…

Calculate the Euclidean distance from all objects to the centroid of each cluster: After obtaining the Euclidean distance for each centroid, we compared the three Euclidean distances for each row of the data. At this stage, the smallest Euclidean distance is obtained or the one with the closest distance.

Calculate the new centroid: The new centroid was calculated using the average value of the objects that were members of each cluster. This process is repeated in the second step. Otherwise, if there is no change, the K-Means algorithm stops running [18].

Convergence checking: The K-Means algorithm continues to iterate until convergence is achieved. In this case, convergence occurs when there is no movement of the cluster members from one cluster to another.

There are two scenarios for processing data, that consist of making a standard K-Means model and a model that uses the K-Means algorithm and PCA dimensionality reduction.

Standard K-Means Model: Clusters or groups were obtained for each data row. We also displayed the number of iterations, number of members in each cluster, and centroid in each cluster. Figure 5 shows the clustering results using the standard K-Means algorithm with the Scikit Learn library using Python.

```
Cluster:  [0 1 2]
Number of Iterations:  6
Member of Each Cluster:  [12256 13244 10586]
Centroid of Each Cluster:  [[ 1.81575510e+00  3.22457143e+00  2.12408163e-01 -1.05952980e+02
   8.10807347e+01  8.86180408e+01  1.77983673e+00  2.35200000e+00
   6.03151020e+00  4.70980408e+01  1.34149306e+02]
 [ 1.56211321e+00  3.49773585e+00  5.53207547e-02  3.84843019e+01
   1.71849057e-01  2.45178113e+01  1.39177358e+00  2.16694340e+00
   5.91335849e+00  5.37947925e+01  2.15645434e+02]
 [ 1.51898734e+00  2.57689401e+00  9.54090308e-03  4.35165313e+01
   1.19404874e+01  5.54761005e+00  8.81069337e-01  1.13980729e+00
   1.73767240e+00  8.76936520e+01  6.16898734e+01]]
```

Figure 5. Model of Standard K-Means Clustering

K-Means with PCA Dimensionality Reduction Model: Similar to the standard K-Means model, we obtained clusters or groups for each data object. We displayed the cluster name, the number of iterations, members of each cluster, and the centroid of each cluster in Figure 6.

```
Cluster:  [0 1 2]
Number of Iterations:  8
Member of Each Cluster:  [10520 13189 12377]
Centroid of Each Cluster:  [[-55.97213499  92.41755552]
 [-66.90868424 -66.49023183]
 [119.02276894  -7.65166283]]
```

Figure 6. Model of K-Means with PCA Dimensionality Reduction

We also visualized the clustering results using a scatter plot graph with three different colors, as shown in Figure 7. Different colors indicate clusters of data points. Clusters 0, 1, and 2 are represented by green, yellow, and blue, respectively. The symbol 'x' indicates the center point or centroid of each cluster.

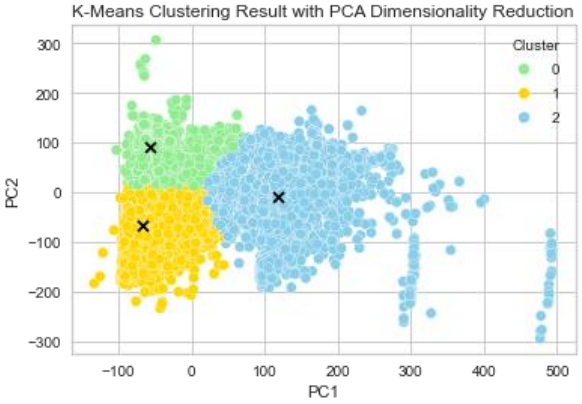

Figure 7. Cluster Visualization of the K-Means with PCA Dimensionality Reduction Model

DOI: https://doi.org/10.29207/resti.v7i5.5181

3.5 Model Evaluation using Calinski-Harabasz Index

First, we calculated the inter- and intra-cluster sums of squares to determine the level of difference in the characteristics of members between clusters and the level of similarity of members within the same cluster. We evaluated both models using the Calinski-Harabasz Index, whether using PCA or not.

Standard K-Means Model:

$$CH(k) = \frac{SS_B}{SS_W} \times \frac{(N-k)}{(k-1)}$$

$$CH(k) = \frac{417709586{,}167725}{290782505{,}515069} \times \frac{36086-3}{3-1}$$

$$CH(k) = 25916{,}646827$$

K-Means with PCA Dimensionality Reduction Model:

$$CH(k) = \frac{SS_B}{SS_W} \times \frac{(N-k)}{(k-1)}$$

$$CH(k) = \frac{416223754{,}996724}{134961949{,}391688} \times \frac{36086-3}{3-1}$$

$$CH(k) = 55640{,}133457$$

3.6 Result Analysis

We analyzed cluster characteristics from categorical attributes such as Region, NE Type, and Build Type to numerical fields that inform durations such as PO-KOM, KOM-RFI, and so on.

Figure 8 shows that in the standard K-Means model, the number of members is evenly distributed with a ratio close to 1:1:1. The members of cluster 0 totaled 12256, cluster 1 totaled 13244, and cluster 2 totaled 10586 rows of data.

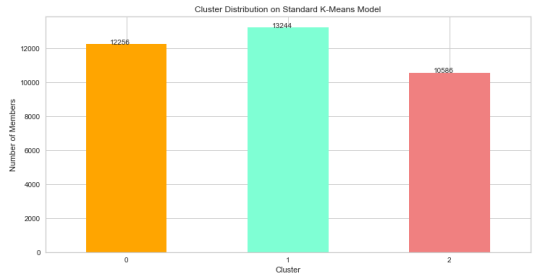

Figure 8. Cluster Members Distribution on the Standard K-Means Model

In addition to the distribution of the number of members in each cluster, we also visualized the number of members in each cluster in the categorical column. As shown in Figure 9, the majoristy of cluster 0 members are in the EAST Region, have NE Type L900, and have Build Type EXISTING. For cluster 1, the majority of its members are in the EAST Region, have NE Type L900, and have Build Type EXISTING. Unlike clusters 0 and 1, cluster 2 members are mostly located in the WEST Region and have NE Type U2100. All three clusters have a majority of Build Type EXISTING, which indicates that the majority of 2022 projects are existing projects and not new sites.

Figure 9. Distribution of Categorical Columns on the Standard K-Means Model

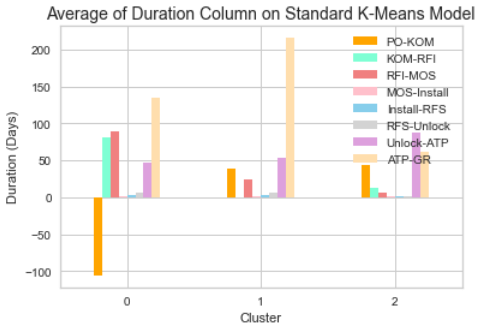

Figure 10. Average of Duration Columns on the Standard K-Means Model

According to Figure 10, from all stages of project development, cluster 2 is a cluster with an average duration that is relatively faster than other clusters. In contrast, cluster 0 had a negative average PO-KOM duration. This shows that projects in cluster 0 carry out the kick-off meeting (KOM) stage before the purchase order (PO).

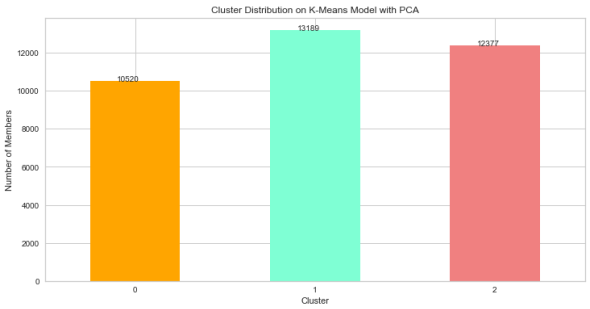

Figure 11. Cluster Members Distribution on the K-Means with PCA Model

Not much different from the standard K-Means model, based on Figure 11, K-Means with PCA Dimensionality Reduction Model also shows a fairly even distribution

DOI: https://doi.org/10.29207/resti.v7i5.5181

of the number of cluster members with a ratio close to 1:1:1. The difference is that in the standard K-Means model, the order of clusters based on the number of members is cluster 1, 0, and 2. Unlike the case with K-Means with the PCA model, which is preceded by clusters 1 and 2, and finally cluster 0. Cluster 0 had 10520 members, cluster 1 had 13189 members, and cluster 2 had 12377 rows of data.

Cluster 0 of the K-Means model on the data resulting from dimension reduction with PCA shows that the majority of its cluster members are in the WEST Region, have NE Type U2100, and have Build Type EXISTING. Furthermore, the categorical attribute distribution of cluster 1 in Figure 12 shows that the majority of its members are in the EAST Region, have NE Type L900, and have Build Type EXISTING. Cluster 2 of the K-Means model on the data resulting from dimension reduction with PCA has the same categorical attribute characteristics as cluster 1, namely that the majority of its members are in the EAST Region, have NE Type L900, and have Build Type EXISTING.

Next, we also analyzed the average duration between stages in days. From Figure 13, it can be said that cluster 0 had a long ATP-GR duration in average. Overall, when compared to other clusters, cluster 1 members have the fastest average duration between stages, relatively, as seen from the duration or aging of each stage to another stage.

Figure 12. Distribution of Categorical Columns on the K-Means with PCA Model

Furthermore, cluster 2 is the only cluster that has a negative average PO-KOM attribute. This indicates that there are many data anomalies in that column because by default, RAN projects are preceded by the PO stage before the KOM.

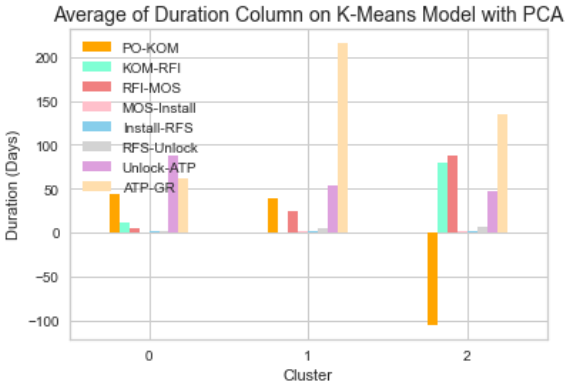

Figure 13. Average of Duration Columns on the K-Means with PCA Model

4. Conclusion

According to the findings of this study, the project health category on the RAN project at PT XYZ was effectively determined using the K-Means clustering algorithm, yielding three clusters, namely clusters 0, 1, and 2. We also discovered that the PCA technique affects the implementation process and performance of K-Means clustering in determining project health in the RAN project at PT XYZ. As a result, data that goes through the PCA dimension reduction process when implemented in K-Means clustering produces a higher Calinski-Harabasz Index (CH Index) value of 55633.12776405707.

Future research should consider several suggestions. First, the research can be developed using the latest RAN project data and a wider scope, for example, all active RAN project data, not limited by the project year. Furthermore, from the aspect of information technology, exploration is needed regarding the parameters used in the K-Means algorithm and the exploration of tools, supporting algorithms, and other approaches to improve the results and performance of the clustering model.

References

[1] M. Hopmere, L. Crawford, and M. S. Harré, "Proactively Monitoring Large Project Portfolios," *Proj. Manag. J.*, vol. 51, pp. 656–669, 2020.

[2] A. A. Wael, A. Elyamany, and A. Elhakeem, "Classification of Evaluation Metrics for Project Baseline Schedules," *Int. J. Eng. Adv. Technol.*, vol. 10, no. 1, pp. 235–239, 2020, doi: 10.35940/ijeat.c5456.1010120.

[3] T. Jo, *Machine Learning Foundations: Supervised, Unsupervised, and Advanced Learning*. 2021.

[4] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning*. The MIT Press, 2018.

[5] M. Garzon, *Dimensionality Reduction in Data Science*. 2022.

[6] O. Nasraoui and C.-E. Ben N'Cir, *Clustering Methods for Big Data Analytics*. 2019.

[7] F. Aksan *et al.*, "Clustering methods for power quality measurements in virtual power plant," *Energies*, vol. 14, no. 18, 2021, doi: 10.3390/en14185902.

[8] M. Zubair, M. Asif Iqbal, A. Shil, E. Haque, M. Moshiul Hoque, and I. H. Sarker, "An Efficient K-Means Clustering Algorithm for Analysing COVID-19," *Adv. Intell.*

DOI: https://doi.org/10.29207/resti.v7i5.5181

Syst. Comput., vol. 1375 AIST, pp. 422–432, 2021, doi: 10.1007/978-3-030-73050-5_43.
[9] Fairleigh Dickinson University, "Purchasing Policies and Procedures," no. October, 2019.
[10] M. Ashika and V. Monisha, "A Material Management in Construction Project Using Inventory Management System," *Int. J. Mod. Trends Sci. Technol.*, vol. 6, pp. 32–40, 2020, doi: 10.46501/IJMTST060506.
[11] One Finance, *How to Receipt Purchase Orders*. The London School of Economics and Political Science, 2019.
[12] R. Jafari, *Hands-On Data Preprocessing in Python*. Packt Publishing, 2022.
[13] E. Antony, N. S. Sreekanth, R. K. Sunil Kumar, and T. Nishanth, "Data preprocessing techniques for handling time series data for environmental science studies," *Int. J. Eng. Trends Technol.*, vol. 69, no. 5, pp. 196–207, 2021, doi: 10.14445/22315381/IJETT-V69I5P227.
[14] Abhishek Thakur, *Approaching (Almost) Any Machine Learning Problem*. Abhishek Thakur, 2020.
[15] A. Zheng and A. Casari, *Feature Engineering for Machine Learning: Principles and Techniques for Data Scientists*. O'Reilly Media, 2018.
[16] G. R. Naik, *Advances in principal component analysis: Research and development*. 2018.
[17] K. Sud, P. Erdogmus, and S. Kadry, *Introduction to Data Science and Machine Learning*. Rijeka: IntechOpen, 2020.
[18] I. Zada *et al.*, "Performance Evaluation of Simple K-Mean and Parallel K-Mean Clustering Algorithms: Big Data Business Process Management Concept," *Mob. Inf. Syst.*, vol. 2022, 2022, doi: 10.1155/2022/1277765.
[19] N. Zumel, J. Mount, J. Howard, and R. Thomas, *Practical Data Science With R*. Manning, 2020.

DOI: https://doi.org/10.29207/resti.v7i5.5181