Comparison of the K-Means method with and without Principal Component Analysis (PCA) in predicting employee resignation

Iwan Binanto[1*] and *Andrianto* Tumanggor[1]

[1]Informatics Department, Faculty of Science and Technology, Sanata Dharma University, Yogyakarta, Indonesia

Abstract. Employees are individuals who work for a company or organization and receive a salary. Employees are the most important assets that need to be effectively managed by the company in order to maximize their contribution. However, many employees feel dissatisfied with the outcomes of their contributions to the company, as they do not receive the expected rewards. This study utilizes a dataset from Kaggle.com, consisting of a total of 14,999 data rows with 10 attributes. In the first experiment, the dataset was reduced using PCA before applying the K-means clustering method. In the second experiment, the dataset is directly fed into the K-means clustering method without PCA. To evaluate the clusters in the K-means method, this study applies the sum of squared error (SSE) method and the silhouette coefficient method to determine the optimal clusters. The study concludes that there are two dominant factors, last_evaluation and average_monthly_hours, that contribute to employees resigning from a company. The SSE evaluation indicates that both methods have an elbow point at 3 clusters, suggesting that dividing the data into more than 3 clusters does not provide significant additional information. The silhouette coefficient evaluation shows that K-means without PCA obtain the best silhouette coefficient value of 0.5674, while K-means with PCA obtain a silhouette coefficient value of 0.5491. Although K-means with PCA have the advantage of reducing the dimensionality of the dataset, they have a longer execution time compared to K-means without PCA, with an execution time of 181.53 seconds for K-means with PCA and 95.84 seconds for K-means without PCA.

1 Introduction

Employees are assets and the main elements in an organization, playing a crucial role in achieving the organization's goals [1]. The success of a company depends on its employees as human resources [2]. Therefore, a company must be capable of managing its human resources or employees who have a strong commitment to the sustainability and progress of the organization or business. However, managing employees is not easy because employees

* Corresponding author: iwan@usd.ac.id

cannot be equated with factory tools or machines; each of them has different thoughts, feelings, positions, desires, and backgrounds from others [3].

Employees who resign can have serious consequences for the company, such as economic impacts. The economic impact on the company includes the cost of training new employees [4]. Although there are still employees who remain loyal to the company with high levels of satisfaction. Therefore, an analysis of the reasons for employee resignations is needed. The causes here are the factors that influence them.

Avinash Navlani discovered that the factors influencing employees to resign from a company are low job satisfaction, low promotion opportunities, low salary, and working longer hours compared to those who stay in the company [5]. Therefore, it is important for the company to understand what factors influence employees' decisions to resign and seek ways to reduce the resignation rate. These factors were identified by applying the K-means method by grouping the most dominant attributes.

However, in high-dimensional data clustering, conventional K-means algorithms are susceptible to data with large attributes, leading to the "curse of dimensionality" [6]. Reducing dimensions also means reducing data complexity. To address this issue, dimensionality reduction of data is necessary. One of the methods used for dimensionality reduction is the PCA method.

This research will compare the results of the K-means method with Principal Component Analysis (PCA) and compare the results of the K-means method alone. The aim of this comparison between the two methods is to determine the most optimal method for predicting employees who will resign from a company. Both the K-means method and the PCA method are commonly used clustering methods in research [7].

In this research, data sourced from kaggle.com, which was published in 2018 [8], will be used. This study focuses on the use of clustering methods by applying the K-means method with and without PCA to predict resigning employees. The author will test which factors are most dominant in causing employees to resign, as well as the most optimal number of clusters in the dataset, and which is better between K-means with PCA and K-means without PCA in clustering this dataset. Cluster evaluation processes will use the Sum of Squared Error (SSE) and Silhouette Coefficient.

2 Literature reviews

Avinash Navlani predicted resigning employees using the K-means clustering method with a obtained k = 3 and the dataset used was from Kaggle.com. By applying the Gradient Boosting Classifier model, he achieved an accuracy of 97%, a precision of 95%, and a recall of 92% [5].

Ainun Umami conducted a Classification of Factors Influencing Employee Reduction at "XYZ" Company using the Naïve Bayes, SVM, Logistic Regression, MLP, Gradient Boosting, KNN, Random Forest, and Decision Tree methods. The data used was the IBM HR Analytics Employee Attrition & Performance Dataset downloaded from Kaggle. After analysis, the best classification methods were found to be Naive Bayes, SVM, Logistic Regression, and MLP [9].

Susanti and Palupiningdyah conducted research on the direct and indirect effects of job satisfaction, organizational commitment, and turnover intention on employee performance, mediated by turnover intention. This study used a sample of 82 out of 144 employees. Based on the hypothesis testing results, it can be concluded that job satisfaction and organizational commitment have a negative and significant impact on turnover intention. Job satisfaction and organizational commitment have a positive and significant impact on employee performance, and turnover intention has a negative and significant effect. The impact on

employee performance and turnover intention can convey the influence of job satisfaction and organizational commitment on employee performance [2].

The calculation of the Sum of Squared Error (SSE) is performed to determine the optimal number of clusters by analyzing the comparison between the number of clusters and the inertia values generated on a graph, and finding the point on the graph where the decrease in inertia value significantly slows down, forming an "elbow" at that point [10]. The K-means clustering algorithm has limitations in determining the optimal number of clusters out of n trials. However, SSE can help overcome this weakness and improve the quality of the model generated by the K-means algorithm [11]. Therefore, to address the limitations of this coefficient method, Kneelocator is applied. Kneelocator functions to dynamically determine the optimum K value based on the application of SSE [12].

3 Research method

In general, the research method is depicted in Figure 1.

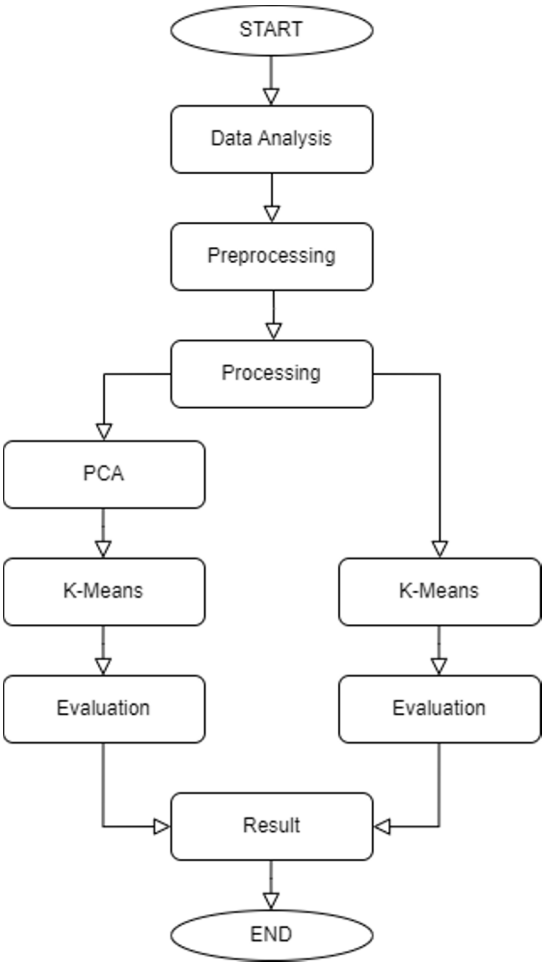

Fig. 1. Diagram of research method.

3.1 Data analysis

Table 1. Attribute and description.

No	Attribute	Description
1	*Satisfaction_level*	This is the point of employee satisfaction, which ranges from 0 to 1.
2	*last_evaluation*	This is the performance evaluated by the employer, which also ranges from 0 to 1.
3	*number_projects*	How many projects are assigned to an employee?
4	*average_monthly_hours*	What is the average number of working hours for an employee in a month?
5	*time_spent_company*	Employee experience. The number of years spent by an employee in the company.
6	*work_accident*	Has an employee ever experienced a work accident or not?
7	*promotion_last_5years*	Has an employee received a promotion in the last 5 years or not?
8	*Departments*	Employee's division.
9	*Salary*	Employee salary levels: low, medium, and high..
10	*Left*	Has the employee left the company or not?

The dataset used in the study consists of 14,999 rows and 10 attributes. The required attributes for this research include: satisfaction_level, last_evaluation, number_project, average_monthly_hours, time_spend_company, work_accident, left, promotion_last_5years, departments, and salary.

The data is stored in comma-separated value (CSV) format. String-type attributes will be converted into numeric values to facilitate the reduction and clustering processes. Data details can be seen in Table 1.

3.2 Preprocessing

This stage is a step for cleaning, standardizing, and normalizing the data so that the data is ready for processing.

This dataset is actually intended for classification, so the target attribute needs to be removed to make this dataset suitable for clustering. Fortunately, there are no missing values in this dataset, so no data needs to be deleted. Non-numeric attributes are converted to enable normalization and standardization.

Additionally, a data correlation process is performed to determine the relationships between variables. The aim is to facilitate data clustering. The correlation method used is the Pearson correlation. The results can be seen in Figure 2.

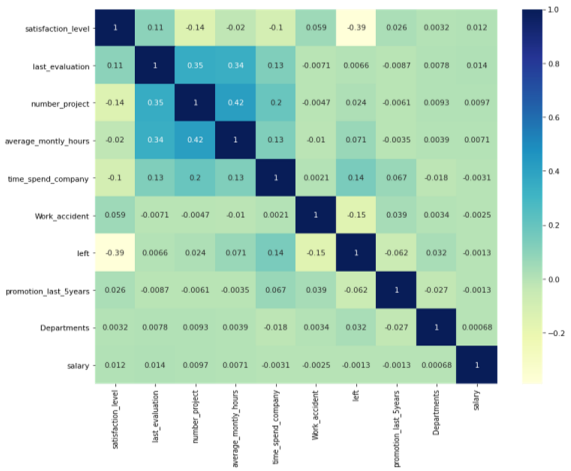

Fig. 2. Visualisation correlation.

Based on the correlation calculation above, there are two variables that have a significant impact based on the correlation scores obtained, namely last_evaluation and average_monthly_hours. Both of these variables will be used in the K-means clustering visualization.

3.3 Processing

The data that has been prepared is then processed using K-Means only, as well as a combination of PCA and K-Means.

3.3.1 K-Means only

In this stage, the K-means method is implemented without dimension reduction. The dataset, which has already been normalized and standardized, is processed using K-Means alone. The results are visualized as shown in Figure 3.

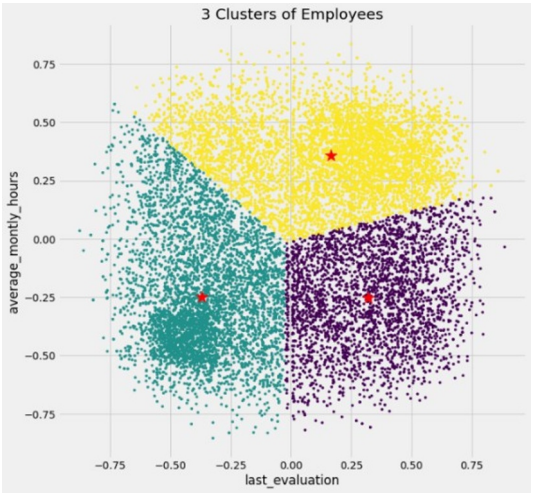

Fig. 3. Cluster visualisation k-means only.

From the visualization, it is evident that the density is quite high, and there is a cluster that is denser than the others, which is the cluster on the left side.

3.3.1 PCA and K-Means

In this stage, the K-means method is implemented with PCA for dimension reduction of the data before clustering. PCA is used to reduce the data dimensions while retaining significant information from the original data. The result of implementing PCA on the dataset is the reduction of dataset attributes to 8 attributes, namely satisfaction_level, last_evaluation, number_project, average_monthly_hours, time_spent_company, work_accident, promotion_last_5years, and departments. These eight attributes will be processed using K-Means. The result can be seen in Figure 4.

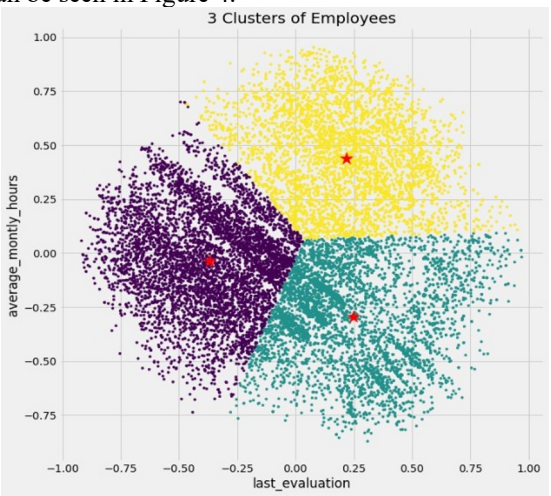

Fig. 4. Cluster visualisation PCA with K-Means.

From the visualization, it can be seen that the density is quite high and more evenly distributed.

3.4 Evaluation

In this stage, clustering evaluation is conducted to determine the quality of the clustering results that have been obtained. Evaluation is carried out to assess how well the clustering method can group data and minimize variance between clusters. Evaluation can also assist in determining the optimal number of clusters and evaluating the effectiveness of the features used in the clustering process.

In this research, two clustering evaluation metrics are used to determine the value of k in the K-means clustering method, namely the Sum of Squared Error (SSE) and the Silhouette Coefficient.

3.4.1 K-Means only

To facilitate the visualization of SSE, the Elbow method is used to find the "elbow" point on the curve, which represents the point of diminishing returns or the optimal number of clusters. The result is visualized by plotting SSE against the number of clusters, with a vertical line indicating the elbow point. The plot helps visualize the "elbow" point and observe the SSE trend as the number of clusters increases. The result can be seen in Figure 5.

Fig. 5. Elbow visualisation K-Means only.

In this visualization, it can be seen that the optimal number of clusters is 3.

```
For n_clusters = 2, the silhouette score is 0.5541
For n_clusters = 3, the silhouette score is 0.5674
For n_clusters = 4, the silhouette score is 0.4362
For n_clusters = 5, the silhouette score is 0.4003
For n_clusters = 6, the silhouette score is 0.3557
For n_clusters = 7, the silhouette score is 0.3388

The best k value is 3 with a silhouette score of 0.5674
```

Fig. 6. *Silhouette Coefficient* Visualisation K-Means only.

Then, the Silhouette Coefficient is calculated to also aid in determining the optimal number of clusters based on the maximum point on the graph because the number of clusters with the highest Silhouette Coefficient value indicates the best number of clusters for separating the data into different groups. The result can be seen in Figure 6.

In above visualization, the highest score is observed at the number of clusters 3.

3.4.2 PCA and K-Means

Just like the previous stage, in this stage, the Elbow method is also applied. The result is as shown in Figure 7.

Fig. 7. Elbow visualisation PCA with K-Means.

In above visualization, the highest score is observed at the number of clusters 3. Then, the Silhouette Coefficient is also calculated. The results can be seen in Figure 8.

```
For n_clusters = 2, the silhouette score is 0.5315
For n_clusters = 3, the silhouette score is 0.5491
For n_clusters = 4, the silhouette score is 0.4248
For n_clusters = 5, the silhouette score is 0.3919
For n_clusters = 6, the silhouette score is 0.3488
For n_clusters = 7, the silhouette score is 0.3334

The best k value is 3 with a silhouette score of 0.5491
```

Fig. 8. *Silhouette Coefficient* Visualisation PCA with K-Means.

In the visualization above, the highest score is observed at the number of clusters 3.

3.5 Results

The experimental results can be summarized as shown in Table 2.

Table 2. Experiment results.

SSE	Cluster	*Silhouette Coefficient*	
		K-means with PCA	**K-means**
Coeffisient point at 3 clusters	2	0.5315	0.5541
	3	0.5491	0.5674
	4	0.4248	0.4362
	5	0.3919	0.4003
	6	0.3488	0.3557
	7	0.3334	0.3388
Execution Time		181.539 seconds	95.843 seconds

4 Discussions

The SSE evaluation indicates that both methods have an elbow point coefficient at 3 clusters, indicating that dividing the data into more than 3 clusters does not provide much additional information.

The Silhouette Coefficient evaluation indicates that both K-means with and without PCA have relatively similar Silhouette Coefficient values for 3 clusters. However, K-means without PCA achieves the best Silhouette Coefficient value, which is 0.5674.

The execution time for K-means with PCA is longer compared to K-means only. The execution time obtained from the K-means with PCA method is 181.53 seconds, while the execution time obtained from the K-means only method is 95.84 seconds.

5 Conclusions

From the obtained data correlation results, it can be concluded that there are two factors that are most dominant in causing employees to resign, namely *last_evaluation* and *average_monthly_hours*.

The difference in execution time between processing with K-Means only and PCA with K-Means is quite significant, so in this case, we recommends using K-Means only for clustering.

References

1. B. Usman, J. Media Wahana Ekon. **17**, 18 (2020)
2. S. Susanti and P. Palupiningdyah, Manag. Anal. J. **5**, (2016)
3. Handoko, *Manajemen Personalia & Sumber Daya Manusia. Manajemen Personalia & Sumber Daya Manusia.*, Edisi 2. (BPFE, Yogyakarta, 2001)
4. R. K. Sari, H. F. Fajar, R. B. Rizqi, and R. A. Putra, ISOQUANT J. Ekon. Manaj. Dan Akunt. **3**, 45 (2019)
5. A. Navlani, (2018)

6. R. Laraswati, M. I. Jambak, and D. Rodiah, Perbandingan Teknik Reduksi Dimensi Antara Algoritma Principal Component Analysis Dengan Fuzzy Association Rule, 2020
7. M. Herlambang, (2019)
8. Kakisama, (2018)
9. A. Umami, Acad. Journal, Surabaya 4 (2018)
10. N. H. Harani, C. Prianto, and F. A. Nugraha, J. Manaj. Inform. **10**, 133 (2020)
11. M. Billah, M. A. Zartesya, and D. S. Prasvita, in *Semin. Nas. Mhs. Ilmu Komput. Dan Apl.* (2021)
12. A. W. A. Ruslam, (2021)