International Journal of Research in Engineering, Science and Management
Volume 7, Issue 6, June 2024
https://www.ijresm.com | ISSN (Online): 2581-5792

Comparative Analysis of Pearson and Euclidean Methods for Candidate Recruitment Optimization

Daniel Pratama[1*], Cindy Himawan[2], Ivan Michael Siregar[3]
[1]Student, Department of Industrial Engineering, Harapan Bangsa Institute of Technology, Bandung, Indonesia
[2]Lecturer, Department of Supply Chain Management, Harapan Bangsa Institute of Technology, Bandung, Indonesia
[3]Department of Information System, Harapan Bangsa Institute of Technology, Bandung, Indonesia

***Abstract*: In the era of globalization and intense business competition, companies desperately need human resources (HR) that fit their culture and goals. High-quality human resources increase productivity and contribute to the innovation and development of the company. Therefore, effective and efficient recruitment is essential to get the best candidates. This research aims to determine the optimal method between Pearson correlation and Euclidean distance in candidate selection. The stages include several systematic steps. First, a simulated dataset with 100,000 candidates is created, each having ten attributes such as skills, experience, education, performance score, age, and certifications. Second, algorithms were applied using Google Colaboratory, with Pearson correlation to measure linear relationships between attributes and Euclidean distance to measure absolute distances between candidates. Third, the similarity results from both techniques were compared using a heatmap table. Analysis showed that Pearson correlation was more consistent and relevant in measuring similarity between candidates. In the visualization of the first 20 datasets, Pearson correlation performed better with a high and stable distribution of correlation values, the lowest value being 0.86. In contrast, Euclidean distance performed less optimally with lower and unstable values, the lowest being 0.26. Considering the stability and consistency of the results, Pearson correlation proved to be more effective and reliable for candidate selection. It provides a clearer and more consistent representation of the similarity between candidates, while Euclidean distance is more suitable as a complementary method or in situations that focus on absolute differences between attributes. Overall, Pearson correlation is recommended as the primary method of candidate selection.**

***Keywords*: candidate recruitment, Euclidean distance, human resources, Pearson correlation.**

1. Introduction

In the era of globalization and intense business competition, the company's need for the right human resources that are suitable for the company's culture and goals has become more crucial. High-quality human resources (HR) are not only able to increase productivity but also contribute to innovation and overall company development. Therefore, an effective and an efficient recruitment process is a must to ensure companies get the best candidates available in the labor market.

Traditionally, the recruitment process is performed manually, which is not only time-consuming but also entails significant costs. The process involves various stages, from CV collection, initial screening, interviews, to final selection. However, with the development of technology, especially in the field of Artificial Intelligence (AI) and Machine Learning (ML), the recruitment process can now be done more effectively and efficiently. These technologies allow companies to process and analyze candidate data automatically, reducing human involvement in the early stages of selection.

The application of technology-based recruitment is not only able to reduce the time and costs incurred by companies, but also increase the accuracy in selecting suitable candidates. By using Machine Learning (ML) algorithms, companies can screen candidates based on criteria that are more specific and relevant to their needs. This of course has a positive impact on the quality of recruitment and ultimately, company performance.

In the context of candidate selection, various similarity-based techniques have been used to help identify the most compatible candidates. These techniques involve measuring the similarity between an existing candidate and the ideal profile that the company desires. Two popular and effective methods include Pearson correlation and Euclidean distance. Each of these methods has its own approach and advantages in measuring data similarity.

The linear link between two variables is measured using Pearson correlation, which is highly helpful in determining the strength of the association between the candidate's traits and the necessary criteria. Using this method, a correlation coefficient is calculated, with values ranging from -1 to 1, to show how closely two variables move together. A positive correlation is indicated by positive values, a negative correlation by negative values, and no meaningful linear association is indicated by values near zero.

Euclidean distance, on the other hand, measures the 'straight' distance between two points in a multidimensional space, which is particularly useful in assessing how close a candidate's profile is to the ideal profile a company wants. This method calculates the distance between two vectors, representing the candidate and the ideal profile, in the same feature space. Smaller the Euclidean distance, more similar the two profiles are.

This study aims to analyze these two similarity-based

*Corresponding author: daniel.1224pratama@gmail.com

techniques in the context of candidate selection for corporate recruitment. By comparing the performance of Pearson correlation and Euclidean distance, it is expected to provide clearer guidance for companies in choosing the technique that best suits their needs. This research will also present case studies and data analysis to measure the effectiveness and efficiency of each method.

2. Methodology

This research aims to determine the most optimal method between Pearson correlation and Euclidean distance in the candidate recruitment process. The stages of this research consist of several systematic steps.

First, the determination of the dataset and relevant attributes is done by creating a simulated dataset consisting of 100,000 candidates. Each candidate will have ten attributes covering various skills, experience, education level, performance score, age, and certifications. These attributes are chosen to reflect relevant characteristics in the candidate selection process.

Secondly, the algorithm implementation is done using Google Collaboratory. The two similarity techniques to be applied are Pearson correlation and Euclidean distance. Pearson correlation is used to measure the linear relationship between candidate attributes, while Euclidean distance is used to measure the absolute distance between candidates based on their attributes. The Pearson correlation calculation will be done with functions from the scipy library, while the Euclidean distance will be calculated using functions from the scipy spatial distance library.

Third, the evaluation is done by comparing the similarity results of the two techniques. Data visualization using a heatmap table will be done to display the comparison between Pearson correlation and Euclidean distance. This heatmap table will help in identifying which method is more consistent and relevant in measuring the similarity between candidates.

Finally, through visual and statistical analysis, this research will determine the most optimal method to use in candidate selection. The results of this research are expected to provide clear recommendations regarding more effective and efficient similarity techniques in the candidate selection process, taking into account the type of data and the specific needs of the company.

3. Literature Study

Human resources (HR) is an important part of the company, as it consists of all the individuals who work to achieve the company's goals. HR is useful for ensuring the company has the right employees with the skills and competencies required for optimal performance [1]. Recruitment is the process of finding candidates based on the criteria of the company. Appropriate recruitment can have an effective impact on company performance. Research [1], provides knowledge related to the importance of effective recruitment and selection procedures and their impact on company performance.

Research [2], states that recruitment is one of the important things for companies to ensure that companies get qualified employees, have the skills and competencies that are needed. Effective recruitment can reduce employee turnover rates, save training and orientation costs for new employees, and create a more productive and harmonious work environment [2].

Research [3], developed a more efficient and interpretative model in matching candidates with the right job position in the recruitment process. Improving quality in the recruitment process as well as ensuring that the most appropriate candidates are positioned in the right position and reliable for recruiters to make better decisions according to the needs for the company.

Research [4], digitization in human resource management (HRM) has resulted in the increased use of Artificial Intelligence (AI) in human resource management systems (HRMS). Research [4], explores the sources of publications and literature that feature AI applications in HRM. Research [5], shows that smart technology offers a new approach to managing employees and improving company performance, with smart technology expected to encourage progress in the field of human resources, especially candidate recruitment. Research [6], developed a comprehensive analytical framework that combines Machine Learning and mathematical modeling to support employee recruitment and placement decisions, with the presence of Machine Learning (ML) and assisted by mathematical modeling is expected to minimize the time and costs incurred against inaccuracies in recruitment for companies.

Research [7], shows that Artificial Intelligence (AI) can increase efficiency in performing routine processes through automation, and allow recruiters to focus more on strategy. Research [8], explains that recruitment innovation affects workforce diversity and provides opportunities for companies to apply technology to the recruitment field to recruit qualified candidates.

The direction and strength of a linear relationship between two variables are evaluated statistically using a metric known as the Pearson correlation coefficient (PCC). The Pearson correlation coefficient (PCC) [9] indicates that there is no linear relationship between the two variables at a value of 0 and a value of -1 indicating a highly negative linear association and a value of 1 indicating a significantly positive linear link. The effectiveness of an embeddedness characterisation approach was evaluated in a range of circumstances using research [9] and this methodology was extended to higher dimensions using the Pearson correlation coefficient (PCC).

Research [10], developed a daily activity feature selection strategy in smart homes using Pearson correlation coefficient. The approach using Pearson correlation coefficient (PCC) not only improves activity recognition performance, but also simplifies the use of sensors in smart homes, thereby supporting energy management and detecting health problems in residents. Through the use of significant feature selection and feature algorithms based on Pearson correlation coefficient, research [11] increases the effectiveness of intrusion detection by minimizing the volume of data that needs to be analyzed. According to research [11], using this technique efficiently lowers the complexity of the data and boosts intrusion detection effectiveness, allowing for quicker and more effective data

processing. Research [12] finds that convolutional neural networks (CNNs) perform better when the appropriate image coding technique is chosen, which addresses the problem of detecting abnormalities in time series data. Research [12] investigates the association between statistical aspects in encoded data and CNN accuracy utilizing descriptive statistical analysis techniques and the Pearson correlation coefficient.

A metric called Euclidean distance is used to calculate the separation between two points in Euclidean space. The square root of the sum of the squares of the variations in the point coordinates is used to compute the distance [13]. Because the Euclidean distance is scale-invariant, research [13] outlines methods for creating and assessing Euclidean Distance-Optimized (EDO) data changes intended to increase the precision and efficacy of biological data cluster analysis. Studies [14], [15] have demonstrated how crucial it is to choose the K-Means algorithm's number of centroids and distance measurement technique in order to increase clustering accuracy. One way to assess the impact of differences in the number of centroids in clustering findings is to compare them using the Euclidean distance.

Research [16] created a location-based service application that uses the Euclidean method and the Haversine formula, two distance measurement techniques, to determine the closest public facility. Research [16] was utilized to assess and contrast the two approaches' accuracy when calculating distance in the actual world. Through the use of a resampling technique that takes into account similarity with historical data based on Euclidean distance and guarantees physical consistency throughout the disaggregated climate elements, research [17] developed and tested a Euclidean Distance Model (EDM) that can break down daily climate data into hourly data.

4. Implementation

The implementation uses Google Collaboratory using the Python programming language, and in the end, tables and graphs will be displayed to evaluate which method is optimal.

A. Dataset

The dataset uses 100,000 candidate data with ten attributes used. The ten attributes consist of Skill1, Skill2, Skill3, Skill4, Skill5, Experience, Education, Performance, Age, Certifications. In Figure 1, displays the dataset that will be used in the implementation which displays the top 5 datasets and the last 5 datasets.

	Skill1	Skill2	Skill3	Skill4	Skill5	Experience	Education	Performance	Age	Certifications
0	0	0	1	0	1	5	5	63	53	3
1	1	1	1	1	0	11	2	60	59	3
2	0	1	1	1	1	12	2	84	36	3
3	0	1	0	1	1	14	2	60	54	5
4	0	1	0	1	0	4	1	94	57	0
...	...	...	...	...	...	...	...	...	...	...
99995	0	1	1	1	0	13	3	53	41	5
99996	1	0	0	0	1	15	5	80	48	0
99997	0	1	1	1	0	1	2	69	33	1
99998	1	0	1	0	1	19	4	71	50	0
99999	1	1	1	1	1	4	5	63	60	4

100000 rows × 10 columns

Fig. 1. Dataset

B. Pearson Correlation Implementation

The stage of Pearson correlation implementation will come once the dataset has been designed. The technique of applying Pearson correlation is shown in Figure 2. First, import the "pearsonr" function from the "scipy" library. This function is used to calculate the Pearson correlation coefficient between two data vectors. Next, take two rows of data and calculate the Pearson correlation coefficient between them, returning the correlation value. Finally, calculate the Pearson value for each row in the dataset. This is how Pearson correlation is implemented, as shown in Figure 2.

```
from scipy.stats import pearsonr

def pearson_correlation(row1, row2):
    corr, _ = pearsonr(row1, row2)
    return corr

pearson_scores = dataset.apply(lambda x: pearson_correlation(dataset.iloc[0], x), axis=1)
```

Fig. 2. Pearson Correlation implementation

C. Euclidean Distance Implementation

The application of Euclidean distance will come next, following the introduction of Pearson correlation. The steps involved in creating Euclidean distance are shown in Figure 3. In Figure 3, the steps involved in implementing Euclidean distance are as follows: first, import the "Euclidean" function from the "scipy" library to compute the Euclidean distance between two data vectors; second, take two rows of data, compute the Euclidean distance between them, and return the distance value; third, compute the Euclidean value for each row in the dataset; and fourth, normalize the Euclidean results useful for a direct comparison with the Pearson value.

```
from scipy.spatial.distance import euclidean

def euclidean_distance(row1, row2):
    return euclidean(row1, row2)

euclidean_scores = dataset.apply(lambda x: euclidean_distance(dataset.iloc[0], x), axis=1)

# Normalize the Euclidean result to make it easier to compare
euclidean_scores_normalized = 1 - (euclidean_scores - euclidean_scores.min()) / (euclidean_scores.max() - euclidean_scores.min())
```

Fig. 3. Euclidean Distance implementation

D. Visualization of Results

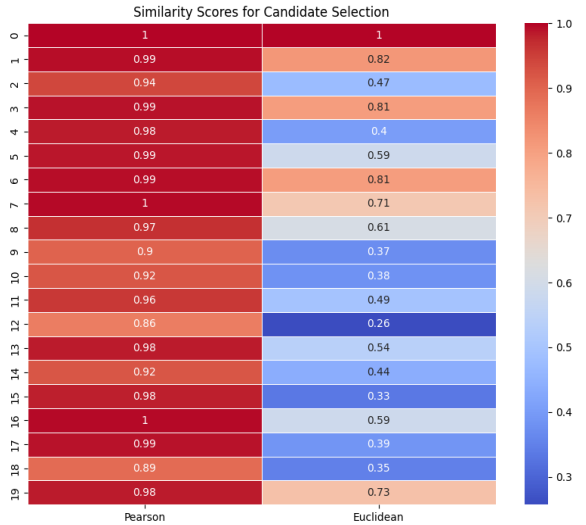

Fig. 4. Comparison table between Pearson correlation and Euclidean distance

After the implementation of Pearson correlation and

Euclidean distance, the next steps will be to visualize the results between Pearson correlation and Euclidean distance by displaying a heatmap table taken from only the first 20 candidate datasets. In Figure 4, displays the results in the form of a heatmap table between Pearson correlation and Euclidean distance.

5. Conclusion

The visualization of the first 20 datasets shows that Pearson correlation and Euclidean distance give a different representation of the similarity between candidates. Pearson correlation performs better in measuring candidate similarity. This can be seen from the high and stable distribution of Pearson correlation values, with the lowest value being 0.86. This value indicates that all candidates in the first 20 datasets have a relatively high degree of similarity with the reference candidate.

In contrast, the Euclidean distance shows less than optimal performance. The values generated by Euclidean distance are lower and unstable, with the lowest value of 0.26. This indicates that Euclidean distance tends to produce greater variation in measuring similarity between candidates, which may be due to the sensitivity of this method to absolute differences in the attributes used.

Considering the stability and consistency in the results shown, Pearson correlation proved to be more effective and reliable for use in the candidate selection process. It provides a clearer and more consistent picture of the similarity between candidates, which is very important in the context of selection optimization. Euclidean distance, while it can provide some additional information, may be better suited as a complementary method or in situations where absolute differences between attributes are the main focus.

Overall, this analysis supports the selection of Pearson correlation as the primary method for measuring similarity in the candidate selection process, due to its ability to provide higher and more stable results compared to Euclidean distance.

References

[1] P. Abdalla Hamza *et al.*, "Recruitment and Selection: The Relationship between Recruitment and Selection with Organizational Performance," *International journal of Engineering, Business and Management*, vol. 5, no. 3, pp. 2456–8678, 2021.

[2] S. G. Abbasi, M. S. Tahir, M. Abbas, and M. S. Shabbir, "Examining the Relationship between Recruitment & Selection Practices and Business Growth: An Exploratory Study," *J Public Aff*, vol. 22, no. 2, May 2022.

[3] C. Qin *et al.*, "An Enhanced Neural Network Approach to Person-Job Fit in Talent Recruitment," *ACM Trans Inf Syst*, vol. 38, no. 2, Feb. 2020.

[4] A. M. Votto, R. Valecha, P. Najafirad, and H. R. Rao, "Artificial Intelligence in Tactical Human Resource Management: A Systematic Literature Review," *International Journal of Information Management Data Insights*, vol. 1, no. 2, Nov. 2021.

[5] D. Vrontis, M. Christofi, V. Pereira, S. Tarba, A. Makrides, and E. Trichina, "Artificial Intelligence, Robotics, Advanced Technologies and Human Resource Management: a Systematic Review," *International Journal of Human Resource Management*, vol. 33, no. 6, pp. 1237–1266, 2022.

[6] D. Pessach, G. Singer, D. Avrahami, H. C. Ben-Gal, E. Shmueli, and I. Ben-Gal, "Employees Recruitment: A Prescriptive Analytics Approach via Machine Learning and Mathematical Programming," *Decis Support Syst*, vol. 134, Jul. 2020.

[7] O. Ore and M. Sposato, "Opportunities and Risks of Artificial Intelligence in Recruitment and Selection," *International Journal of Organizational Analysis*, vol. 30, no. 6, pp. 1771–1782, Dec. 2022.

[8] F. A. Ajayi and C. A. Udeh, "Innovative Recruitment Strategies in the IT Sector: A Review of Successes and Failures," *Magna Scientia Advanced Research and Reviews*, vol. 10, no. 2, pp. 150–165, Apr. 2024.

[9] C. Jebarathinam, D. Home, and U. Sinha, "Pearson Correlation Coefficient as a Measure for Certifying and Quantifying High-Dimensional Entanglement," *Phys Rev A (Coll Park)*, vol. 101, no. 2, Feb. 2020.

[10] Y. Liu, Y. Mu, K. Chen, Y. Li, and J. Guo, "Daily Activity Feature Selection in Smart Homes Based on Pearson Correlation Coefficient," *Neural Process Lett*, vol. 51, no. 2, pp. 1771–1787, Apr. 2020.

[11] P. Chen, F. Li, and C. Wu, "Research on Intrusion Detection Method Based on Pearson Correlation Coefficient Feature Selection Algorithm," *J Phys Conf Ser*, vol. 1757, no. 1, Feb. 2021.

[12] H. Rahadian, S. Bandong, A. Widyotriatmo, and E. Joelianto, "Image Encoding Selection Based on Pearson Correlation Coefficient for Time Series Anomaly Detection," *Alexandria Engineering Journal*, vol. 82, pp. 304–322, Nov. 2023.

[13] A. Ultsch and J. Lötsch, "Euclidean Distance-Optimized Data Transformation for Cluster Analysis in Biomedical Data (EDOtrans)," *BMC Bioinformatics*, vol. 23, no. 1, Dec. 2022.

[14] R. Suwanda, Z. Syahputra, and E. M. Zamzami, "Analysis of Euclidean Distance and Manhattan Distance in the K-Means Algorithm for Variations Number of Centroid K," *J Phys Conf Ser*, vol. 1566, no. 1, Jul. 2020.

[15] M. Faisal, E. M. Zamzami, and Sutarman, "Comparative Analysis of Inter-Centroid K-Means Performance using Euclidean Distance, Canberra Distance and Manhattan Distance," *J Phys Conf Ser*, vol. 1566, no. 1, Jul. 2020.

[16] E. Maria, E. Budiman, Haviluddin, and M. Taruk, "Measure Distance Locating Nearest Public Facilities using Haversine and Euclidean Methods," *J Phys Conf Ser*, vol. 1450, no. 1, Mar. 2020.

[17] C. Görner, J. Franke, R. Kronenberg, O. Hellmuth, and C. Bernhofer, "Multivariate non-Parametric Euclidean Distance Model for Hourly Disaggregation of Daily Climate Data," *Theor Appl Climatol*, vol. 143, pp. 241–265, 2020.