Penerbit Akademia Baru

Journal of Advanced Research in Computing and Applications

Journal homepage:
https://akademiabaru.com/submit/index.php/arca/index
ISSN: 2462-1927

Leveraging Correlation and Clustering: An Exploration of Data Scientist Salaries

Chandra Agoeng[1], Nurul Dini Faqriah Miza Azmi[2], Hakimah Mat Harun[2], Nurzulaikha Abdullah[2], Wan Azani Mustafa[3], Fakhitah Ridzuan[2,*]

[1] Faculty of Computer Science, Universitas Mercu Buana, 11650 Jakarta, Indonesia
[2] Faculty of Data Science and Computing, Universiti Malaysia Kelantan, 16100 Kota Bharu, Kelantan, Malaysia
[3] Faculty of Electrical Engineering & Technology, Universiti Malaysia Perlis, Pauh Putra Campus, 02600 Arau, Perlis, Malaysia

ARTICLE INFO

Article history:
Received 12 February 2024
Received in revised form 1 March 2024
Accepted 2 April 2024
Available online 17 May 2024

Keywords:
Salary; Exploratory data analysis; Data scientist; Clustering; Correlation

ABSTRACT

Data science is a dynamic field with ever-evolving job descriptions and salary structures. While data science offers high earning potential, the factors influencing data scientist salaries remain unclear. This lack of clarity makes it challenging for both employers to determine competitive compensation packages and for employees to understand how career choices like experience level and job title can impact their earning potential. Thus, this study aims to explore the interrelationship between related variables with salary. To achieve the objectives of this research, correlation analysis was employed to identify the strength and direction of linear relationships between these attributes in the dataset. Additionally, k-means clustering was utilized to group data scientists with similar characteristics, allowing for the exploration of potential salary segments within the data science field. It was found that there was a very strong correlation between employee residence and company location (r=0.90). There was a significant moderate positive correlation between salary with company location (r=0.46), residence (r=0.48) and experience level (r=0.41) respectively. Based on the clustering analysis, the group was divided into four different popular roles in data science salary group. Therefore, employers can leverage this knowledge to design the salary packages considering location and experience.

1. Introduction

Understanding the connections between different factors and salaries is essential for employers and employees in the field of data science. An understanding of the relationships between experience levels, job titles, and salaries is crucial for making correct choices in the rapidly evolving field of data science. Companies and professionals are looking for thorough insights to effectively navigate the complexities of compensation structures as the demand for data science expertise increases.

[*] *Corresponding author*
Email address: fakhitah.r@umk.edu.my (Fakhitah Ridzuan)

https://doi.org/10.37934/arca.35.1.1020

Determining the connections within the data can be very beneficial to both employers and employee. Employees gain valuable salary insights, while employers can leverage this analysis to understand current recruitment trends, aiding in improved budget planning for human resource compensation and benefits [1]. This empowers businesses to make informed decisions regarding employee compensation and retention strategies.

In Industry 4.0, driven by IoT and big data, there's a rising demand for data scientists to collect, process, and effectively report on vast amounts of data for organizational management [1]. These professionals are essential for gathering extensive data sets, employing advanced techniques such as machine learning or statistical analysis for data processing, and ultimately delivering comprehensive reports to inform organizational management decisions. Thus, analyzing salaries becomes imperative to assist employers in understanding market trends, ensuring competitive compensation, and strategically managing human resources to attract and retain top talent.

In terms of salary prediction, numerous research has been conducted. For example, Zhang and Cheng [2] used k-nearest neighbour classifier to predict salary for Java back-end engineers, using Java specialized skills as input variables. Similarly, Chen *et al.,* [3] utilized random forest for salary prediction that enables the estimation of income ranges over a specific period. Additionally*, Saeed et al.,* [4] compare three classification approach which are support vector machine, random forest and naïve bayes. However, these studies primarily utilize supervised learning methods.

Aside from supervised learning, unsupervised approaches such as clustering also offer valuable insights for analysis. analytical tools for analysing high-dimensional data by uncovering latent patterns and hidden structures, simplifying complex datasets in the process [5]. Clustering is one of unsupervised machine learning technique used to group similar data points together based on their inherent characteristics or features.

The primary objective of unsupervised algorithms is to minimize within-cluster variation, which measures the extent to which observations within a cluster differ from each other [5]. Clustering analysis can be categorized into three main approaches: hierarchical clustering, centroid-based clustering, and density-based clustering.

Hierarchical clustering is a technique for identifying cluster structures in a dataset, where similarity within the same cluster is maximized, and dissimilarity between different clusters is maximized [5]. Hierarchical clustering does not require pre-specification of the cluster count. Instead, it necessitates the specification of the dissimilarity measure to drive cluster formation [5].

Centroid-based clustering is a type of clustering algorithm where clusters are formed around central points called centroids. These centroids represent the mean or average position of the data points within each cluster. On the other hand, density-based clustering aims to detect dense regions of arbitrary shape, typically identified by the density of points within them [6]. A cluster is defined as a region with high density, while outliers are characterized by low densities.

Montano and Sobrejuanite [7] utilized K-means clustering algorithm to categorize salaries according to their similarities, with the number of clusters was determined through salary profiling and exploratory data analysis [7]. On the other hand, Harahap *et al.,* [8] used k-means clustering to address potential challenges that may arise in the process of determining employee salaries, which could ultimately impact employee performance at work.

Clustering analysis is an important component of this study where Data Science Salary Dataset was used to identify various factors that impact the salary of data scientist while also incorporating clustering analysis. This study is essential to unveil complex patterns and relationships within the dataset, shedding light on the essential dynamics shaping data scientist salaries.

2. Methodology

General methodology of data science project consists of three main phases which are Data Preprocessing, Data Modelling and Data Evaluation (as shown in Figure 1). Upon acquiring a dataset, it undergoes data preprocessing, the initial phase that prepares the raw data for subsequent analysis. Raw data often contains errors, inconsistencies, and missing values. Data preprocessing ensures the validity and reliability of data analysis results by removing outliers and filling missing values [9]. Data modelling involves building the model from the data and make predictions. Subsequently, the results obtained from the model are evaluated, interpreted, and conclusions are drawn from the data.

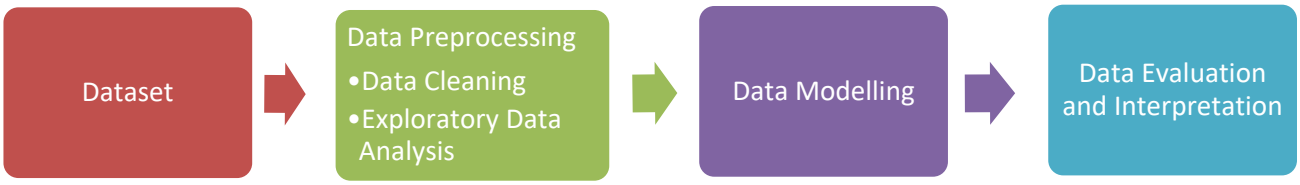

Fig. 1. Data Science Methodology

2.1 Dataset

The Data Science Job Salaries data was obtained from Kaggle [10]. The dataset contains information about salaries and related factors for various data science and related roles. There are 11 attributes in the dataset which are work_year, experience_level, employment_type, job_title, salary, salary_currency, salary_in_usd, employee_residence, remote_ratio, company_location and company_size. A description of the data can be found on Table 1.

Table 1
Data Description

No	Column	Data Type	Description
1	work_year	Integer	The calendar year in which the salary was paid
2	experience_level	String	Experience level for the specific job: EN: Entry-level / Junior MI: Mid-level / Intermediate SE: Senior-level / Expert EX: Executive-level / Director
3	employment_type	String	Employment arrangement: PT: Part-time FT: Full-time CT: Contract FL: Freelance
4	job_title	String	The specific role held by the employee during the pay year.
5	salary	Integer	The total gross salary amount paid.
6	salary_currency	String	Total monetary compensation paid before taxes and deductions, denominated in the original currency of the employer.
7	salary_in_usd	Integer	Gross salary converted into US Dollars (USD)
8	employee_residence	String	The primary country of residence for the employee
9	remote_ratio	Integer	Percentage of work performed remotely: 0%: No Remote Work 50%: Partially Remote 100%: Fully Remote
10	company_location	String	The country where the main office or contracting branch of the employer is located
11	company_size	String	Estimated average number of employees at the company during pay year: Small (S): Companies with less than 50 employees. Medium (M): Companies with 50 to 250 employees. Large (L): Companies with more than 250 employees.

2.2 Data Preprocessing

In the dataset, columns such as 'salary' and 'salary_currency' are unnecessary since the analysis was primarily focus on 'salary_usd'. Therefore, these columns can be removed to streamline the dataset and avoid redundancy, allowing for a more concise and efficient analysis.

To detect outliers in the 'salary_in_usd' column, the interquartile range (IQR) method was employed. In cases where the data distribution deviates significantly from a Gaussian (normal) distribution, the IQR method is a suitable statistic for identifying outliers [11]. The IQR is calculated as the difference between the upper quartile and the lower quartile of the dataset. Figure 2 shows the outlier exist in the dataset.

```
Q1: 25 percentile of the salary_in_usd values is, 62726.0
Q2: 50 percentile of the salary_in_usd values is, 101570.0
Q3: 75 percentile of the salary_in_usd values is, 150000.0
Interquartile range is 87274.0
low_limit is -68185.0
up_limit is 280911.0

Outliers in the dataset is [324000, 325000, 380000, 405000,
412000, 416000, 423000, 450000, 450000, 600000]
```

Fig. 2. Outliers existed in the dataset

A common yet straightforward approach employed in practice to remove outlier is the "detect-and-forget" strategy which involves identifying outliers within the dataset and subsequently removing them [12]. Due to the limited number of data points identified as outliers, thus the affected rows were removed from the dataset. Further verification was conducted to ensure the absence of null values and duplicate entries in the dataset. As no issues were detected during this examination, the dataset is deemed satisfactory for further analysis.

Next, data transformation on the categorical variables within the dataset were performed. Data transformation is essential for analysis as this stage will modifies data structure and format to enhance quality and suitability of the data [13]. This process involves converting these non-numerical values, such as job titles or geographic locations, into a format suitable for statistical analysis. Label encoding is a data transformation technique commonly used for categorical variables. It involves converting each unique category within the variable into a corresponding integer value [13]. All categorical data within the dataset underwent label encoding transformation to facilitate further analysis.

2.3 Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a process to explore the data which can be investigate using graphical and non-graphical approach on univariate or multivariate level [14]. Univariate analysis focuses on understanding a single variable at a time, while multivariate analysis explores the relationships between two or more variables. To gain a comprehensive understanding of the factors influencing data scientist salaries, this study utilizes a multivariate approach to EDA.

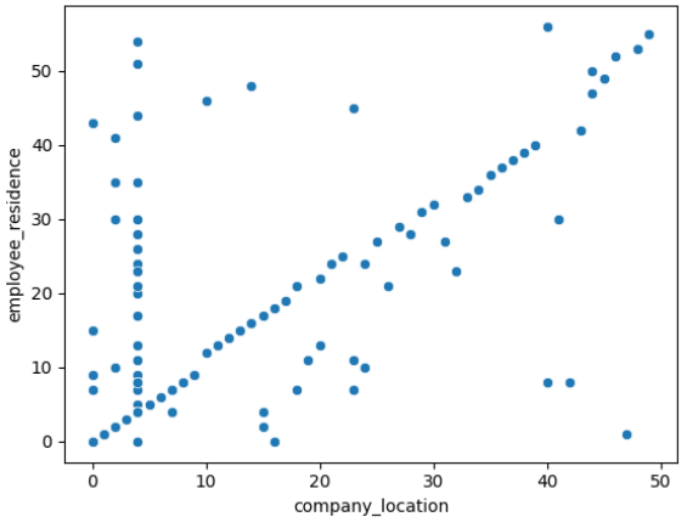

Fig. 3. Average salary of individuals categorized by different employment types

Figure 3 shows the visualisation of company location and employee residence. The correlation value of 0.9 between 'company_location' and 'employee_residence' indicates a strong positive correlation between these two variables. This implies that there is a tendency for the location of the company where the job is based to be closely related to the geographical location where the employee resides. In other words, employees tend to live in or near the same geographic area where their company is located.

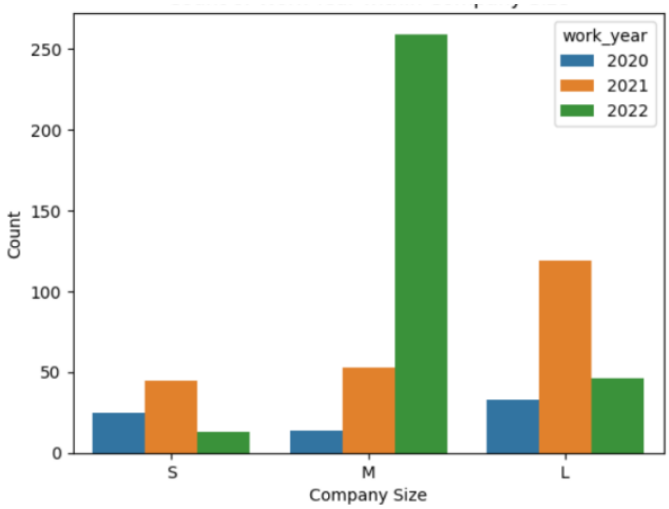

Fig. 4. Count of work year within company size

Figure 4 distribution of company size and work year. In 2020, the data indicates a relatively lower proportion of individuals employed in Medium-sized companies, with small and large companies potentially dominating the employment landscape. However, by 2022, there is a notable shift, revealing a significant increase in employment within Medium-sized companies, surpassing other size categories. Interestingly, the employment distribution in 2021 appears to favour Large-sized companies, indicating fluctuations in company size preferences across the observed years.

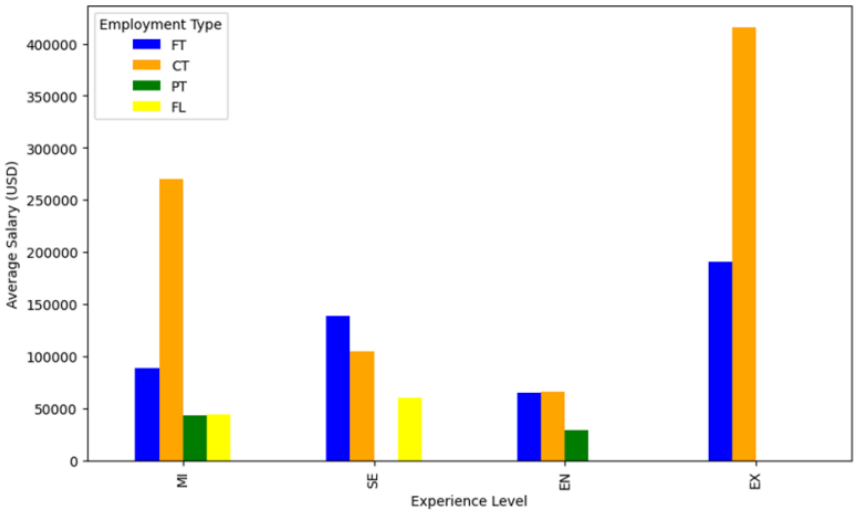

Fig. 5. Average salary of individuals categorized by different employment types

Figure 5 illustrates the average salary of individuals categorized by different employment types and experience levels. The vertical axis denotes salary ranges from 0 to 400,000 USD, while the horizontal axis displays the various employment types, including Full-time (FT), Contract (CT), Part-time (PT), and Freelance (FL).

The highest salaries are observed at the executive level, particularly among those employed on a contractual basis. Notably, executive roles do not typically offer part-time or freelance employment arrangements. Freelance positions predominantly align with senior-level roles, while the majority of senior-level roles are full-time employment. Intermediate or mid-level positions tend to be contracted, with approximately one-third held as full-time roles. For entry-level positions, there is a relatively equal distribution across part-time, contract, and full-time employment arrangements.

Table 2 shows the mean, minimum, maximum values of remote_ratio and salary_usd in the dataset. The data highlights prevalence of remote work, with an average ratio of 70.92%. This indicates that a significant portion of work is conducted remotely, with a substantial number of individuals engaged in 100% remote work arrangements. On the other hand, the average salary in USD is approximately $107,168.86, with a standard deviation of $58,555.52. The salary range extends from a minimum of $2,859 to a maximum of $276,000.

Table 2
Mean, minimum and maximum values of remote_ratio and salary_usd

	remote_ratio	salary_usd
Mean	70.68	107168.86
Min	0.00	2859.00
Max	1000.00	276000.00

The exploratory data analysis revealed an interesting finding: a seemingly linear relationship between company location and employee residence. This suggests that data scientists tend to be located geographically close to their employers. Additionally, the data indicates a positive trend in the number of data science professionals, with a significant increase observed in 2022 compared to the previous two years. This growth highlights the increasing demand for data science expertise. Furthermore, the analysis suggests a shift in work preferences, with a growing number of data scientists favoring remote work arrangements over physical or hybrid models. This trend could be

attributed to the flexibility and work-life balance offered by remote work. Finally, a noteworthy observation pertains to salary distribution within the data science field. The data revealed a significant gap between minimum and maximum salaries, indicating a high degree of variability in compensation. This suggests that factors such as experience level, type of employment, and company size can significantly impact earning potential in data science.

2.4 Data Modelling

Before proceeding with modelling, the dataset was divided into training and testing sets using an 80/20 split. The decision to select 80% of the data for training was influenced by research conducted by Nguyen *et al.,* [15], indicating that increasing the size of the training set from 30% to 80% could lead to improved testing performance.

K-Means clustering technique was employed in this study. It is a technique that aims to divide a dataset into distinct groups (clusters) based on shared characteristics, ensuring data with similar traits are grouped together while different data points are placed into separate clusters [16]. K-Means is a non-hierarchical algorithm that operates by iteratively assigning data points to clusters based on their proximity to the specified centre points [16]. To determine the appropriate number of clusters, the elbow method was employed, revealing that k=4 is the optimal choice for cluster count. Figure 6 shows the elbow method for optimal k.

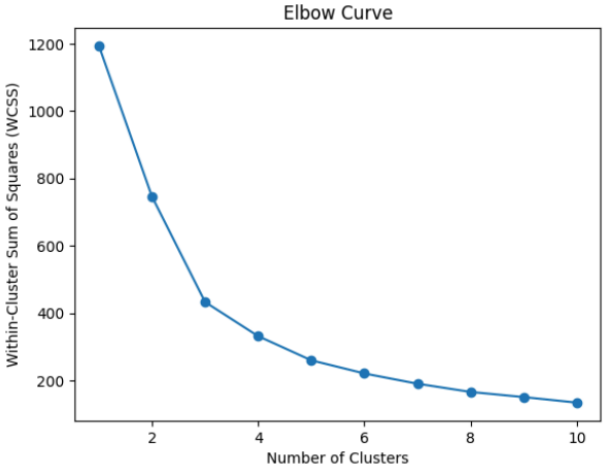

Fig. 6. Elbow method for optimal k

In the K-Means algorithm, the starting point for each cluster group is the centroid, which is determined by calculating the closest distance to the initial cluster center point, serving as the central point for the formation of each group or cluster [17]. The steps to compute K-Means are as follows [18]:

i. Specify the number of groups (K) for the K-Means clustering method.
ii. Randomly select K data points and assign them to individual groups based on data point counts.
iii. Calculate the cluster centroids.
iv. Iterate steps 1-3 until optimal centroids are found, ensuring minimal variance within groups:
   a) Compute the total squared distances between data points and centroids.
   b) Assign each data point to the nearest centroid.
   c) Average all data points within each cluster.

3. Results and Discussion

3.1 Correlation Analysis

Figure 7 shows the correlation between all the attributes in the dataset.

Fig. 7. Correlation analysis between all the variables

There is a strong positive correlation of 0.9 between employee residence and company location, indicating that employees tend to live close to where their company is located. Health benefits can be gained by living close to the workplace, as stress from long commutes is reduced, time is saved for exercise or hobbies, and improved sleep is facilitated by the opportunity to wake up later [19].

Additionally, there are moderately positive correlations between salary and company location (0.46) and salary and residence (0.48). From the analysis, companies with employees in multiple locations often implement regional pay differentials or adjust pay rates based on geographical factors [20]. Besides, geographic pay differential is extra pay given to employees to address differences in labour costs and living expenses across various locations [21]. Some companies use factors like the cost of living, which includes expenses for goods and services, to calculate these differentials.

Salary and experience level also show a moderate relationship with 0.41. Employees with over five years of work experience typically earn an average of 36% more than those with less than five years of experience [22]. These correlations suggest that salaries tend to be influenced by the geographical location of the company, with higher salaries potentially being offered in certain locations. Besides, employees with higher salaries are more likely to reside in specific areas and salaries tend to increase with experience level.

3.2 Clustering Analysis

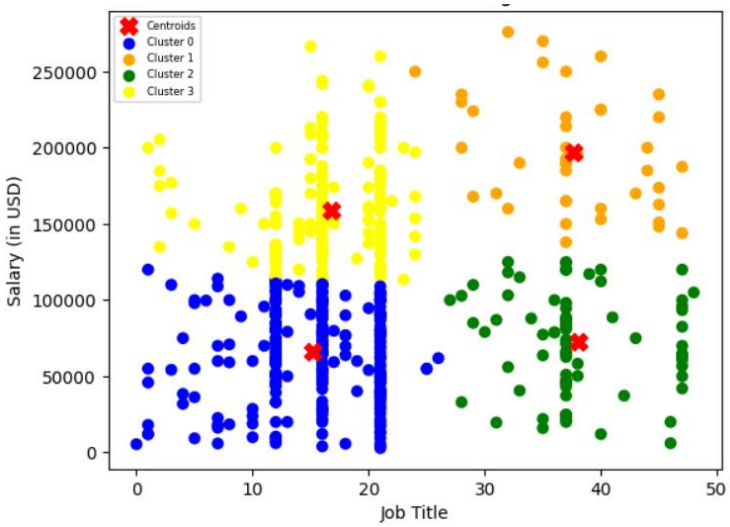

Fig. 8. K-mean Clustering Result

Based on the K-means clustering plot with k=4 (as shown in Figure 8), the following insights can be derived:

i. Cluster 0: This cluster comprises popular roles in data science characterized by salaries ranging from low to moderate. These roles likely experience high demand but offer salaries within a relatively lower range compared to other clusters.
ii. Cluster 1: Representing less popular roles in data science, this cluster features salaries ranging from moderate to high. These roles may be specialized or niche positions that command higher compensation despite being in lower demand.
iii. Cluster 2: This cluster encompasses low-profile yet popular roles in data science, offering salaries ranging from low to moderate. Although not as widely recognized as roles in Cluster 0, they still hold significance within the industry.
iv. Cluster 3: Popular roles in data science with salaries ranging from moderate to high are depicted in this cluster. These roles experience high demand and offer higher compensation compared to Cluster 0, highlighting their importance and desirability within the field.

4. Conclusion

This study employed EDA, correlation analysis and K-means clustering to gain insights into data scientist salaries. The EDA and correlation revealed a positive correlation between company location and employee residence, suggesting data scientists tend to live near their employers. Additionally, a positive trend in data science professionals was observed, highlighting the increasing demand for this expertise. Furthermore, the analysis points towards a shift in work preference favouring remote work arrangements. The K-means clustering identified four distinct clusters of data science roles, categorized by popularity and salary range. These clusters provide valuable insights into the data science landscape, highlighting the varying salary potential associated with different roles and specializations. Overall, the findings offer valuable guidance for aspiring and experienced data scientists alike, allowing them to make informed career decisions based on their desired location, work style, and salary expectations.

Acknowledgement
This research was not funded by any grant.

References
[1] Quan, Tee Zhen, and Mafas Raheem. "Salary Prediction in Data Science Field Using Specialized Skills and Job Benefits–A Literature." *Journal of Applied Technology and Innovation* 6, no. 3 (2022): 70-74.
[2] Zhang, Junyu, and Jinyong Cheng. "Study of Employment Salary Forecast using KNN Algorithm." In *2019 International Conference on Modeling, Simulation and Big Data Analysis (MSBDA 2019)*, pp. 166-170. Atlantis Press, 2019.. https://doi.org/10.2991/msbda-19.2019.26
[3] Chen, Jingyi, Shuming Mao, and Qixuan Yuan. "Salary prediction using random forest with fundamental features." In *Third International Conference on Electronics and Communication; Network and Computer Technology (ECNCT 2021)*, vol. 12167, pp. 491-498. SPIE, 2022. https://doi.org/10.1117/12.2628520
[4] Saeed, Ashty Kamal Mohamed, Pavel Younus Abdullah, and Avin Tariq Tahir. "Salary Prediction for Computer Engineering Positions in India." *Journal of Applied Science and Technology Trends* 4, no. 01 (2023): 13-18. https://doi.org/10.38094/jastt401140
[5] Eckhardt, Christina M., Sophia J. Madjarova, Riley J. Williams, Mattheu Ollivier, Jón Karlsson, Ayoosh Pareek, and Benedict U. Nwachukwu. "Unsupervised machine learning methods and emerging applications in healthcare." *Knee Surgery, Sports Traumatology, Arthroscopy* 31, no. 2 (2023): 376-381. https://doi.org/10.1007/s00167-022-07233-7
[6] Chen, Yewang, Xiaoliang Hu, Wentao Fan, Lianlian Shen, Zheng Zhang, Xin Liu, Jixiang Du, Haibo Li, Yi Chen, and Hailin Li. "Fast density peak clustering for large scale data based on kNN." *Knowledge-Based Systems* 187 (2020): 104824. https://doi.org/10.1016/j.knosys.2019.06.032
[7] Montaño, Vicente E., and Glenndon C. Sobrejuanite. "CLUSTER AND CAREERS: EXPLORING SALARY STRUCTURES AND FACULTY RECRUITMENT IN ACADEMIC PROGRAMS." *European Journal of Education Studies* 11, no. 2 (2024). https://doi.org/10.46827/ejes.v11i2.5187
[8] Harahap, Leliana, Sartika Dewi Purba, Sutrisno Situmorang, Jonas Franky R. Panggabean, and Kamson Sirait. "Analysis Of Salary Of Permanent Employees And Contract Employees On The Medicom Campus Using The K–Means Algorithm." *Journal of Intelligent Decision Support System (IDSS)* 6, no. 4 (2023): 231-240.
[9] Fan, Cheng, Meiling Chen, Xinghua Wang, Jiayuan Wang, and Bufu Huang. "A review on data preprocessing techniques toward efficient and reliable knowledge discovery from building operational data." *Frontiers in energy research* 9 (2021): 652801. https://doi.org/10.3389/fenrg.2021.652801
[10] Bhatia, R. Data Science Job Salaries. Kaggle. (2022) [Internet].
[11] Domański, Paweł D. "Study on statistical outlier detection and labelling." *International Journal of Automation and Computing* 17, no. 6 (2020): 788-811.. https://doi.org/10.1007/s11633-020-1243-2
[12] Chen, Shuxiao, and Jacob Bien. "Valid inference corrected for outlier removal." *Journal of Computational and Graphical Statistics* 29, no. 2 (2020): 323-334. https://doi.org/10.1080/10618600.2019.1660180
[13] Joshi, ASHISH P., and BIRAJ V. Patel. "Data preprocessing: the techniques for preparing clean and quality data for data analytics process." *Orient. J. Comput. Sci. Technol* 13, no. 2-3 (2020): 78-81. https://doi.org/10.13005/OJCST13.0203.03
[14] Indrakumari, R., T. Poongodi, and Soumya Ranjan Jena. "Heart disease prediction using exploratory data analysis." *Procedia Computer Science* 173 (2020): 130-139. https://doi.org/10.1016/j.procs.2020.06.017
[15] Nguyen, Quang Hung, Hai-Bang Ly, Lanh Si Ho, Nadhir Al-Ansari, Hiep Van Le, Van Quan Tran, Indra Prakash, and Binh Thai Pham. "Influence of data splitting on performance of machine learning models in prediction of shear strength of soil." *Mathematical Problems in Engineering* 2021 (2021): 1-15. https://doi.org/10.1155/2021/4832864
[16] Sari, Indah Purnama, Al-Khowarizmi Al-Khowarizmi, and Ismail Hanif Batubara. "Cluster Analysis Using K-Means Algorithm and Fuzzy C-Means Clustering For Grouping Students' Abilities In Online Learning Process." *Journal of Computer Science, Information Technology and Telecommunication Engineering* 2, no. 1 (2021): 139-144. https://doi.org/10.30596/jcositte.v2i1.6504
[17] Mughnyanti, M., S. Efendi, and M. Zarlis. "Analysis of determining centroid clustering x-means algorithm with davies-bouldin index evaluation." In *IOP Conference Series: Materials Science and Engineering*, vol. 725, no. 1, p. 012128. IOP Publishing, 2020. https://doi.org/10.1088/1757-899X/725/1/012128
[18] Naeem, Samreen, Aqib Ali, Sania Anam, and Muhammad Munawar Ahmed. "An unsupervised machine learning algorithms: Comprehensive review." *International Journal of Computing and Digital Systems* (2023). https://doi.org/10.12785/ijcds/130172
[19] Vertex. How good is it to Have Your Workplace Close to Home? [Internet].
[20] Pierson, L. Salary Differences by Geographic Location - What to Know Linkedin. (2022) [Internet].

[21] Culpepper & Associates. Geographic Pay Difference. SHRM. (2009) [Internet]
[22] Verma, A. 4 Criteria for What Is Considered a Good Salary (Plus Tips) (2023) [Internet].