Jurnal Teknik Informatika (JUTIF)
Vol. 4, No. 3, June 2023, pp. 657-667

DOI: https://doi.org/10.52436/1.jutif.2023.4.3.1084
p-ISSN: 2723-3863
e-ISSN: 2723-3871

DECISION SUPPORT SYSTEM FOR PREDICTING EMPLOYEE LEAVE USING THE LIGHT GRADIENT BOOSTING MACHINE (LIGHTGBM) AND K-MEANS ALGORITHM

Vasthu Imaniar Ivanoti[*1], Megananda Hervita P.[2], Gandung Triyono[3], Dyah Puji Utami[4]

[1,2,3]Master of Computer Science, Faculty of Information Technology, Universitas Budi Luhur, Indonesia
[4]Faculty of Engineering, Architecture & Information Technology, The University of Queensland, Australia
Email: [1]vasthu.ivanoti@gmail.com, [2]2111602153@student.budiluhur.ac.id, [3]gandung.triyono@budiluhur.ac.id, [4]deeutami000@gmail.com

(Article received: May 04, 2023; Revision: May 27, 2023; published: June 26, 2023)

Abstract

Nowadays, decision support systems have gained wide popularity not only in private companies but also in government sectors. These systems play a crucial role in assisting leaders during the decision-making process. The effective functioning of the government heavily relies on employee performance, which requires discipline in carrying out their duties and responsibilities. Employee discipline is closely linked to their attendance, including leave-taking. Therefore, analyzing employee leave data can reveal trends and interrelationships, providing leaders with valuable information and insights for determining employee leave policies. To address this issue, data mining applications such as the Light Gradient Boosting Machine (LightGBM) regression prediction model can be utilized. This model takes into account factors like gender, age, and the starting year of leave to predict the number of employees who take annual leave simultaneously with holidays. Additionally, clustering algorithms like K-Means can be employed to group reasons for leave into clusters, identifying common leave patterns among employees. In this study, employee leave application data from January 2018 to July 2022 was collected from the Leave module within the HRIS (Human Resource Information System) application. The research outcomes encompass a dashboard visualization presenting descriptive analysis and modeling using LightGBM. The modeling results yielded reasonably accurate predictions, as evidenced by model testing that showed a difference of only 1 employee. Additionally, K-Means clustering formed 4 clusters of leave reasons, with the majority being family-related, illness, childcare, and elderly care. The dashboard can be used by management as a consideration for approving employee leaves, ensuring well-planned leave scheduling for the following year and minimizing disruption to work execution in each department.

Keywords: *decision support system, employee leave, k-means clustering, light gradient boosting machine*

1. INTRODUCTION

The progress of an organization is greatly influenced by its human resources, which are considered its primary assets [1].Factors such as heavy workloads, high work pressure, and repetitive daily tasks can contribute to increased stress levels and boredom among employees, ultimately leading to a decline in performance [2]. To maintain optimal employee performance, it is necessary for employees to take leave, allowing them to rest and recover both physically and psychologically for a specific period of time. Leave is a fundamental right for every employee [3]. However, when a significant number of employees take leave simultaneously, particularly around religious holidays, it can have a significant impact on work operations, especially for service-oriented organizations. Therefore, effective management of employee leave becomes crucial in ensuring the smooth continuity of business processes within the organization.

The Ministry of Finance has an existing system in place for managing employee leave, specifically the Leave module within the HRIS (Human Resource Information System) Application. Within this module, employees provide details regarding the start and end dates of their leave, reasons for their leave, as well as contact information such as addresses and phone numbers for communication during their absence. The decision to grant employee leave is influenced by two main factors, the chosen dates for leave and the reasons provided. Approving leave dates requires careful consideration of the organization's work plan to minimize any significant impact on operations due to employee absence. Furthermore, the reasons stated for leave help determine the level of urgency in granting employee requests. Since employees submit diverse reasons for their leaves, it becomes necessary to classify frequently cited leave reasons. These classifications can then be incorporated as parameters within the Leave module of the HRIS Application.

In order to enhance the efficiency of the leave approval process, it is necessary to implement a system that utilizes data mining techniques to analyze information regarding trends in employee leave-taking, frequent leave-takers, and the annual percentage of employee leaves. By doing so, valuable insights can be generated, serving as useful references or recommendations for leaders when approving employee leave requests in alignment with organizational requirements.

Past studies have explored various decision support systems for determining employee leave. For instance, [4]developed a system using the Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) method. [5] focused on the Army's financial education center and utilized the Simple Additive Weighting (SAW) method with attendance criteria. Similarly, [6] also employed the SAW method for employee leave applications. Another study by [7] combined the Evaluation based on Distance from Average Solution (EDAS) approach with the Pivot Pairwise Relative Criteria Importance Assessment (PIPRECIA) method to build an HRIS application with decision support system features.

In terms of data mining approaches, [2] utilized the description method to present information on the most frequently taken leave categories by employees. On the other hand, [8] employed clustering techniques, specifically the Affinity Propagation algorithm, to analyze disease patterns based on sick leave data and employee health. Other research on employee leave primarily focuses on developing information systems or leave management applications, as exemplified by the following studies. [9] developed a web-based employee information system due to the manual recording of employee data using an iterative methodology. [10] built a web-based employee leave application website using WhatsApp blast using a waterfall methodology because to streamline the manual leave application process and facilitate obtaining supervisor signatures. [11]–[14] also developed a web-based leave application information system due to the manual recording of employee data using a waterfall methodology. On the other hand, [15] developed a Java-based leave application information system using a waterfall methodology.

As described earlier, there is a scarcity of research specifically focusing on decision support systems for determining employee leave, particularly those leveraging data mining methods. Therefore, we conduct research in this area, specifically employing classification and clustering techniques within data mining. The LGBM method was chosen due to its relatively high accuracy, especially when dealing with complex and unstructured data. On the other hand, K-Means was selected for its simplicity and ease of implementation, as it is a commonly used and straightforward approach for clustering.

2. RESEARCH METHOD

2.1. Research Stages

The research stages conducted in this study are illustrated in the provided diagram.

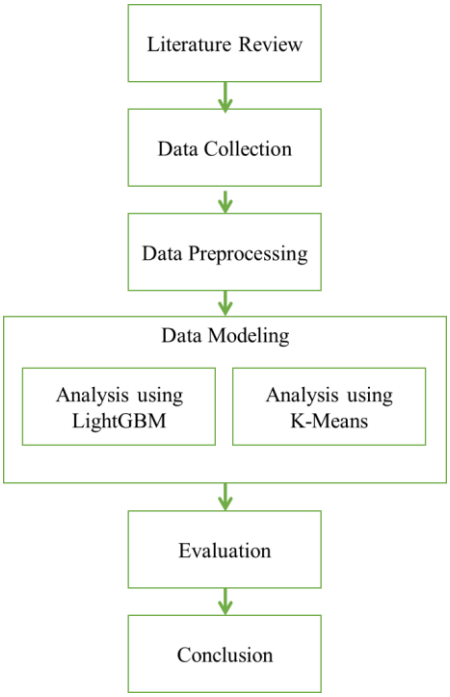

Figure 1. Research stages

Figure 1 depicts the sequential stages undertaken in the research. The process commences with a comprehensive review of relevant literature from journals or libraries pertaining to the research topic. Subsequently, data collection is performed, which serves as the basis for analysis. The dataset utilized is obtained from the HRIS application database and comprises a solitary table containing 414,809 rows and 36 columns. Following data collection, the data undergoes preprocessing or preparation utilizing various data cleansing techniques. To obtain the best prediction model or algorithm, changes were made to the data type of the 'LeaveStartYear' column. It was initially categorical and converted to numerical. Additionally, the 'EmployeeID' and 'NumberofDaysApproved' columns were removed during the prediction process. Furthermore, parallel prediction analysis was conducted using the Light Gradient Boosting Machine (LightGBM) regression method and K-Means clustering on the prepared data. These two methods do not mutually influence each other but are expected to complement each other in providing input or recommendations to the management. The LightGBM method is used to predict the number of employees who will apply for leave based on certain criteria. Meanwhile, the K-Means method aims to cluster the reasons for employee leave and determine the average difference in days between the submission date and the start date of leave. The

resulting model from the analysis stage is evaluated to determine its efficacy. The final stage encompasses drawing conclusions based on the research findings.

2.2. Data Mining

Data mining involves extracting valuable patterns from extensive datasets, resulting in knowledge that can be categorized into two main types, namely predictive and descriptive [16]. The Cross-Industry Standard Process Model for Data Mining (CRISP-DM) is a widely used methodology in data mining, comprising six stages: (1) Business Understanding, which focuses on comprehending the organization's business processes; (2) Data Understanding, which entails understanding the data; (3) Data Preparation; (4) Modeling, where algorithms are utilized as needed; (5) Evaluation, which assesses the outcomes of the modeling; and (6) Deployment [17].

Data mining encompasses various techniques, including classification and clustering. Classification involves grouping data into classes or labels based on predetermined criteria, considering variables from existing data groups [18]. For classification, the dataset utilized should possess a target label or attribute [19]. On the other hand, clustering involves grouping data based on similarities or shared characteristics [20]. The main distinction between clustering and classification lies in the absence of a target variable used to group data in the clustering process [21].

2.3. Light Gradient Boosting Machine (LightGBM)

Light Gradient Boosting Machine (LightGBM) is an alternative version of Gradient Boosting that employs the leaf-wise algorithm to construct Decision Trees in a vertical manner. In the leaf-wise algorithm, branches are continuously added until they cannot be further subdivided and reach the maximum depth. LightGBM incorporates the Gradient-based One-Side Sampling (GOSS) algorithm to reduce the volume of data by focusing on instances with significant gradients while disregarding those with small gradients. However, this approach introduces bias towards samples with larger gradients and alters the original data distribution. To address this, random sampling is conducted for data with small gradients. Furthermore, LightGBM implements Exclusive Feature Bundling (EFB) to address sparsity in the dataset. This technique combines certain features, resulting in a reduced number of features while still preserving the crucial ones [22].

Assuming a raw dataset with $N = \{1, 2, ..., n\}$ samples, the LGBM model consists of $T = \{1, 2, ..., t\}$ trees. The final prediction after t iterations is obtained by summing the prediction from the first (1 - t) trees and the tth tree. The iteration process is outlined as follows [23], [24]:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_i(x_i) \quad (1)$$

For the tth iteration, the predicted value of the i-th sample, denoted as $\hat{y}_i^{(t)}$. While, $\hat{y}_i^{(t-1)}$ denotes the previously generated tree model and $f_i(x_i)$ denotes the newly built model. Thus, according to equation (1), each new prediction is derived by considering the residual and the previous prediction.

The training process is further described in equation (2). Additionally, a regularization term is introduced to reduce the complexity of the model and enhance its applicability to other datasets, as illustrated by equation (3).

$$\begin{cases} \hat{y}_i^{(0)} = 0 \\ \hat{y}_i^{(i)} = f_1(x_i) = \hat{y}_i^{(1)} + f_1(x_i) \\ \hat{y}_i^{(2)} = f_1(x_i) + f_2(x_i) = \hat{y}_i^{(1)} + f_2(x_i) \end{cases} \quad (2)$$

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)\right) + \sum_{t=1}^{T} \Omega(f_t) \quad \Omega(f) = \gamma T + \frac{1}{2}\lambda \|\omega\|^2 \quad (3)$$

Where y_i is the actual value, and $y_i^{(t)}$ is the predicted value. $\sum l$ represents the sum of losses between each group y_i and $y_i^{(t)}$, while $\Omega(f_t)$ is the regularization term. T denotes the number of leaves, and ω represents the leaf weight. Coefficients λ and γ are included, with default values set as $\gamma = 0$ and $\lambda = 1$.

2.4. K-Means

The K-Means algorithm is a data mining technique used to group or cluster data based on their proximity, conditions, criteria, or characteristics. It aims to ensure that data within the same cluster have the shortest distance and share similar conditions, criteria, or characteristics [25]. The steps involved in clustering using the K-Means algorithm are as follows:

a. Determine the desired number of clusters (k).
b. Randomly initialize k centroids using the formula (4):

$$d(i,k) = \sqrt{\sum_i^m (C_i j - C_k j)^2} \quad (4)$$

c. Calculate the distance between each data point and the centroids from the previous step using the Euclidean Distance equation (5) as follows:

$$min \sum_k^i - \alpha_{ik} - = \sqrt{\sum_i^m (C_i j - C_k j)^2} \quad (5)$$

d. Group each data point based on its closest distance to its centroid using the following equation (6):

$$C_{kj} = \frac{\sum_k^i x_{ij}}{p} \quad (6)$$

e. Determine the new centroid positions (k).
f. Repeat steps c and d if the new centroid positions are not the same as the previous centroids.

To determine the optimal number of clusters, the elbow method is employed, which visually examines the consistency of the best number of clusters by comparing the differences in the sum of squared errors (SSE) for each cluster. The graph typically forms a distinct angle when the most significant difference occurs, indicating the optimal number of clusters [26]–[28] The SSE is calculated using the equation (7):

$$SSE = \sum_{k-1}^{k} \sum_{xi \epsilon Sk} \|Xi - Ck\|^2 \quad (7)$$

Here, Xi represents the attribute value of the i-th data point, and Ck is the attribute value of the centroid of cluster k. The process starts with calculating for k=2 and iteratively increments the value of k by 1. The number of clusters and associated error are computed for each value of k. The error typically decreases significantly at certain values of k, after which it reaches a stable point.

To evaluate the quality of the resulting clusters, the silhouette value analysis method is used [29][30]. The silhouette value is calculated by considering the intra-cluster distance (a), and the average nearest-cluster distance for each data point. The silhouette value can be interpreted into three categories:

a. A value close to +1 or positive indicates that the data point is correctly assigned to its cluster.
b. A value close to 0 suggests that the data point may belong to another cluster as well.
c. A value close to -1 or negative implies that the data point is in the wrong cluster.

3. RESULT AND DISCUSSION

3.1. Data Collection

The data collection process commences by conducting interviews with employees responsible for managing leave data using the Leave module in the HRIS application. These interviews aim to understand the workflow of employee leave management. Additionally, discussions with application developers are held to comprehend the data structure and obtain necessary permissions to access relevant databases. Once access is granted, data is extracted from the production area (data source) to the staging area using Microsoft Visual Studio. The extracted data is then exported into CSV format for subsequent processing. The dataset utilized in this study comprises leave application data for all employees of the Ministry of Finance, including approved, unapproved, and pending applications from January 2018 to July 2022.

	IDPermohonan	IDPegawai	IDRefJenis	KodeOrganisasi	IDRefJenisOrganisasi	JenisOrganisasi	Es1	Es2	Es3	NamaKota	...	TanggalPersetujuan2	TanggalPer
0	50889	114660	1	3.503040e+09	1	Kantor Pusat	Sekretariat Jenderal	Biro Hukum	Bagian Hukum Pajak dan Kepabeanan	Kota Adm. Jakarta Pusat	...	2018-01-02 07:49:19.077	201 07:4
1	50890	106455	1	3.503111e+09	1	Kantor Pusat	Sekretariat Jenderal	Pusat Pembinaan Profesi Keuangan	Bidang Pemeriksaan Profesi Akuntansi	Kota Adm. Jakarta Pusat	...	2018-01-03 08:58:40.427	201 08:5
2	50891	72567	1	3.503060e+09	1	Kantor Pusat	Sekretariat Jenderal	Biro Sumber Daya Manusia	Bagian Manajemen Informasi Sumber Daya Manusia	Kota Adm. Jakarta Pusat	...	2018-01-02 11:01:03.013	201 11:0
3	50892	119641	1	3.508011e+09	1	Kantor Pusat	Direktorat Jenderal Kekayaan Negara	Sekretariat Direktorat Jenderal	Bagian Umum	Kota Adm. Jakarta Pusat	...	2018-01-02 08:26:38.743	201 08:2
...	...	...	...	...	...	...	...	...	...	...	...	...	
414804	527461	94906	1	3.509100e+05	1	Kantor Pusat	Direktorat Jenderal Perimbangan Keuangan	Direktorat Dana Transfer Khusus	NaN	Kota Adm. Jakarta Pusat	...	NaN	
414805	527462	135711	1	3.509090e+09	1	Kantor Pusat	Direktorat Jenderal Perimbangan Keuangan	Direktorat Dana Transfer Umum	Subdirektorat Dana Desa	Kota Adm. Jakarta Pusat	...	NaN	
414806	527463	153712	1	3.504010e+09	1	Kantor Pusat	Direktorat Jenderal Anggaran	Sekretariat Direktorat Jenderal	Bagian Umum	Kota Adm. Jakarta Pusat	...	NaN	
414807	527464	83020	1	3.504080e+09	1	Kantor Pusat	Direktorat Jenderal Anggaran	Direktorat Harmonisasi Peraturan Penganggaran	Subdirektorat Harmonisasi Peraturan Jaminan So...	Kota Adm. Jakarta Pusat	...	NaN	
414808	527465	70734	1	3.508160e+09	1	Kantor Pusat	Direktorat Jenderal Kekayaan Negara	Kantor Wilayah Direktorat Jenderal Kekayaan Ne...	Bidang Piutang Negara	BANDUNG	...	NaN	

414809 rows × 36 columns

Figure 2. Inisial dataset for employee leave application

The acquired dataset consists of a single table containing 414,809 rows and 36 columns, as depicted in Figure 2. It encompasses various information such as employee details, organizational units, destination cities for leave, leave dates, and reasons for leave. Details in the initial dataset which include column names, count, and data types are shown in Figures 3.

```
Data columns (total 36 columns):
 #   Column                    Non-Null Count   Dtype
---  ------                    --------------   -----
 0   IDPermohonan              414809 non-null  int64
 1   IDPegawai                 414809 non-null  int64
 2   IDRefJenis                414809 non-null  int64
 3   KodeOrganisasi            414809 non-null  float64
 4   IDRefJenisOrganisasi      414809 non-null  int64
 5   JenisOrganisasi           414809 non-null  object
 6   Es1                       414809 non-null  object
 7   Es2                       406772 non-null  object
 8   Es3                       392146 non-null  object
 9   NamaKota                  358715 non-null  object
 10  NamaProvinsi              358715 non-null  object
 11  Nama                      414809 non-null  object
 12  IDRefJenisKelamin         414809 non-null  object
 13  TanggalLahir              414809 non-null  datetime64[ns]
 14  Alamat                    414787 non-null  object
 15  TanggalPengajuan          414809 non-null  datetime64[ns]
 16  TanggalMulai              414809 non-null  datetime64[ns]
 17  TanggalSelesai            414809 non-null  datetime64[ns]
 18  JumlahHari                414809 non-null  int64
 19  AlamatCuti                414131 non-null  object
 20  Keperluan                 410838 non-null  object
 21  Uraian                    414809 non-null  object
 22  JumlahHariDisetujui       412782 non-null  float64
 23  IDRefStatusPersetujuan1   412776 non-null  float64
 24  TanggalPersetujuan1       412776 non-null  object
 25  IDRefStatusPersetujuan2   401312 non-null  float64
 26  TanggalPersetujuan2       401311 non-null  object
 27  TanggalPenetapan          398585 non-null  object
 28  AlasanPenolakan           8522 non-null    object
 29  AlasanPenundaanPenetapan  3758 non-null    object
 30  IsLuarNegeri              414809 non-null  int64
 31  Kuota2021                 414809 non-null  int64
 32  Kuota2022                 414809 non-null  int64
 33  Kuota2020                 414809 non-null  int64
 34  Kuota2019                 414809 non-null  int64
 35  Kuota2018                 414809 non-null  int64
dtypes: datetime64[ns](4), float64(4), int64(11), object(17)
```

Figure 3. Initial dataset details

3.2. Data Preparation

To enable data analysis, it is essential to preprocess the data by making modifications such as removing or rectifying characters, adding or eliminating columns, and altering the data format to enhance comprehensibility. The steps taken in data preprocessing include:

a. Case Folding: This entails converting the text's letter format to lowercase. Non-alphanumeric characters like punctuation marks and spaces are treated as delimiters.
b. Tokenization: This technique involves dividing the text into smaller units such as words, sentences, or bi-grams, while retaining the punctuation symbols.
c. Stopword Removal: This process involves filtering out common words and retaining significant and unique words from the tokenized results.

Some adjustments were made, as seen in Figure 3, where certain columns such as 'ApprovalDate1', 'ApprovalDate2', and 'DeterminationDate' displayed as object data type instead of the expected datetime data type. Therefore, it was necessary to change their data type accordingly. Additionally, there were some columns with missing values, such as 'ReasonsforRejection' and 'ReasonsforPostponement', where the number of non-null values was less than the total number of records. This aligns with the number of leave requests that were rejected, resulting in those columns being empty for approved leaves. To aid in the analysis, several columns were added, including ‘LeaveStartYear’, ‘Age’, ’EidDay’, ’ChristmasDay’, ‘SeclusionDay', ‘DifferenceStartsLeave', dan ‘LeaveWithHolidays’. Furthermore, an examination of outlier distribution was conducted, and outliers were removed to generate a robust model.

	IDPermohonan	IDPegawai	IDRefJenis	KodeOrganisasi	IDRefJenisOrganisasi	JenisOrganisasi	Es1	Es2	Es3	NamaKota	...	TanggalPersetujuan2	TanggalPer
0	50889	114660	1	3.503040e+09	1	Kantor Pusat	Sekretariat Jenderal	Biro Hukum	Bagian Hukum Pajak dan Kepabeanan	Kota Adm. Jakarta Pusat	...	2018-01-02 07:49:19.077	201 07:4
1	50890	106455	1	3.503111e+09	1	Kantor Pusat	Sekretariat Jenderal	Pusat Pembinaan Profesi Keuangan	Bidang Pemeriksaan Profesi Akuntansi	Kota Adm. Jakarta Pusat	...	2018-01-03 08:58:40.427	201 08:5
2	50891	72567	1	3.503060e+09	1	Kantor Pusat	Sekretariat Jenderal	Biro Sumber Daya Manusia	Bagian Manajemen Informasi Sumber Daya Manusia	Kota Adm. Jakarta Pusat	...	2018-01-02 11:01:03.013	201 11:0
3	50892	119641	1	3.508011e+09	1	Kantor Pusat	Direktorat Jenderal Kekayaan Negara	Sekretariat Direktorat Jenderal	Bagian Umum	Kota Adm. Jakarta Pusat	...	2018-01-02 08:26:38.743	201 08:2
...	...	...	...	...	...	...	...	...	...	...	...	...	...
414804	527461	94906	1	3.509100e+05	1	Kantor Pusat	Direktorat Jenderal Perimbangan Keuangan	Direktorat Dana Transfer Khusus	NaN	Kota Adm. Jakarta Pusat	...	NaN	
414805	527462	135711	1	3.509090e+09	1	Kantor Pusat	Direktorat Jenderal Perimbangan Keuangan	Direktorat Dana Transfer Umum	Subdirektorat Dana Desa	Kota Adm. Jakarta Pusat	...	NaN	
414806	527463	153712	1	3.504010e+09	1	Kantor Pusat	Direktorat Jenderal Anggaran	Sekretariat Direktorat Jenderal	Bagian Umum	Kota Adm. Jakarta Pusat	...	NaN	
414807	527464	83020	1	3.504080e+09	1	Kantor Pusat	Direktorat Jenderal Anggaran	Direktorat Harmonisasi Peraturan Penganggaran	Subdirektorat Harmonisasi Peraturan Jaminan So...	Kota Adm. Jakarta Pusat	...	NaN	
414808	527465	70734	1	3.508160e+09	1	Kantor Pusat	Direktorat Jenderal Kekayaan Negara	Kantor Wilayah Direktorat Jenderal Kekayaan Ne...	Bidang Piutang Negara	BANDUNG	...	NaN	

414809 rows × 46 columns

Figure 4. Dataset of preprocessed data

After adjusting the data types, organizing the data (e.g., annual leave, leave approval, leave coinciding with holidays), and eliminating outliers, the resulting dataset containing 414,809 rows and 46 columns, as depicted in Figure 4, is ready for further analysis.

3.3. Data Modeling

3.3.1. Descriptive Analysis

To identify the number of employees who take the most leave together with holidays or joint leave, categorized by Echelon I and gender, a descriptive analysis was carried out on the prepared dataset. The analysis uses the 'Es1' column and the 'Number of Employees' column which represents the count of 'Employee ID' grouped by 'IDRefGender', 'Age', 'Es1', and 'LeaveStartYear' for approved leave from the preprocessed dataset. The dataset used to perform descriptive analysis is presented in Figure 4.

	JenisKelamin	Usia	Es1	TahunMulaiCuti	JumlahPegawai
0	P	19	Direktorat Jenderal Perbendaharaan	2020	2
1	P	20	Direktorat Jenderal Kekayaan Negara	2019	4
2	P	20	Direktorat Jenderal Kekayaan Negara	2020	1
3	P	20	Direktorat Jenderal Perbendaharaan	2020	4
4	P	20	Direktorat Jenderal Perbendaharaan	2021	1
...	...	...	...	...	...
1830	W	58	Direktorat Jenderal Perbendaharaan	2021	14
1831	W	58	Direktorat Jenderal Perbendaharaan	2022	13
1832	W	58	Inspektorat Jenderal	2019	2
1833	W	58	Inspektorat Jenderal	2022	1
1834	W	63	Badan Pendidikan dan Pelatihan Keuangan	2020	1

1796 rows × 5 columns

Figure 5. Dataset for descriptive analysis

Based on the dataset in Figure 4, a descriptive analysis was visualized to showcase the number of employees who took leave in conjunction with holidays based on Echelon I.

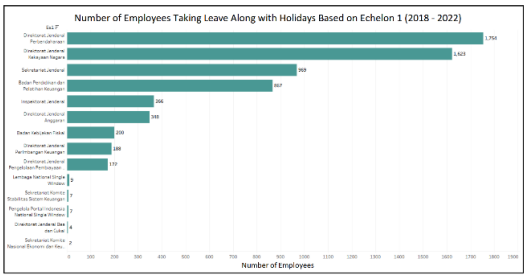

Figure 6. Descriptive analysis based on Echelon I Unit

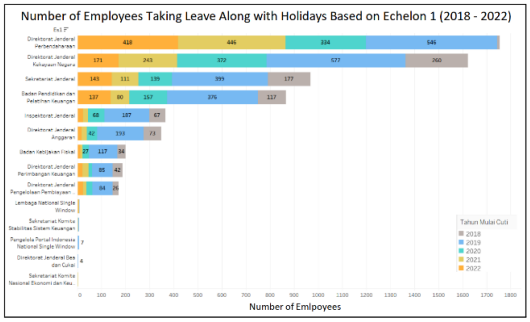

Figure 7. Descriptive analysis based on Echelon I Unit per year

As shown in Figure 5 and Figure 6, the Directorate General of Treasury holds the highest ranks in both total period and yearly comparison, with 1.754 employees. Furthermore, a descriptive analysis was conducted to describe the number of employees who took leave in conjunction with holidays based on gender within each Echelon I Unit.

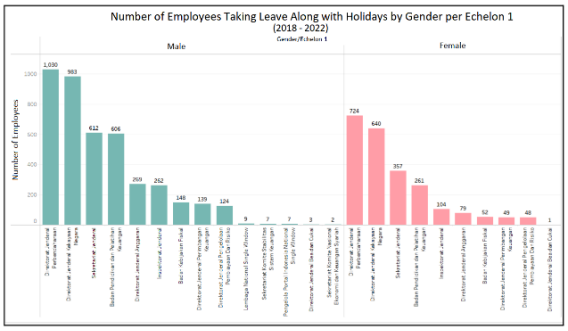

Figure 8. Descriptive analysis based on gender per Echelon I Unit

According to Figure 7 above, it is evident that, on average, male employees took more annual leave in conjunction with holidays (Nyepi, Lebaran, and Christmas) than female employees during the period from January 2018 to July 2022.

3.3.2. Predictive Analysis Using LightGBM Method

Predicting employees who take annual leave along with holidays or collective leave is achieved through modeling with the LightGBM algorithm. The predictor variables used include gender, age, echelon 1, and the year in which the leave commenced. Initially, a comparison of various regression methods is performed using the PyCaret library.

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
lightgbm	Light Gradient Boosting Machine	1.7903	6.7201	2.5818	0.5077	0.4911	0.7753	0.0380
rf	Random Forest Regressor	1.7384	7.2104	2.6726	0.4702	0.4647	0.7097	0.3230
et	Extra Trees Regressor	1.7322	7.4685	2.7177	0.4542	0.4722	0.6689	0.2830
gbr	Gradient Boosting Regressor	1.9394	7.9508	2.8028	0.4263	0.5053	0.8496	0.0510
knn	K Neighbors Regressor	2.2915	10.1183	3.1762	0.2537	0.5936	1.1191	0.0460
ridge	Ridge Regression	2.2602	10.9688	3.2890	0.2130	0.5741	0.9974	0.0130
lr	Linear Regression	2.2621	10.9709	3.2894	0.2127	0.5748	0.9986	0.0120
br	Bayesian Ridge	2.2626	10.9807	3.2908	0.2122	0.5737	1.0001	0.0120
ada	AdaBoost Regressor	2.5271	11.0990	3.3236	0.1838	0.6373	1.3231	0.0260
huber	Huber Regressor	2.1830	12.2188	3.4675	0.1278	0.5584	0.7686	0.0300
dt	Decision Tree Regressor	2.0875	11.7477	3.4129	0.1149	0.5688	0.7566	0.0170
omp	Orthogonal Matching Pursuit	2.5423	12.8987	3.5725	0.0685	0.6415	1.1912	0.0130
en	Elastic Net	2.7098	13.9519	3.7160	-0.0064	0.6718	1.2881	0.0130
llar	Lasso Least Angle Regression	2.7149	13.9566	3.7167	-0.0068	0.6724	1.2910	0.0130
dummy	Dummy Regressor	2.7149	13.9566	3.7167	-0.0068	0.6724	1.2910	0.0130
lasso	Lasso Regression	2.7128	13.9585	3.7169	-0.0069	0.6722	1.2898	0.0130
par	Passive Aggressive Regressor	2.5469	17.8321	4.1877	-0.2760	0.7399	0.6953	0.0170
lar	Least Angle Regression	640.7582	490438983.1287	7160.9949	-37985456.6502	0.7764	639.2304	0.0150

Figure 9. Comparison of several regression methods

Based on the findings presented in Figure 8, it can be deduced that the LightGBM model exhibits the highest R2 value of 0.5077, establishing it as the most effective regression model. An R2 value ranging from 0.50 to 0.75 suggests a moderate correlation between the independent variable and the dependent variable [31].

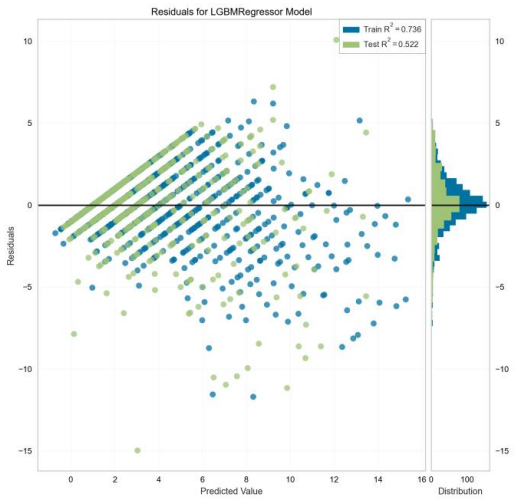

Figure 10. The residual graph of the LightGBM regression modeling results

Figure 9 presents the residual graph illustrating the results of the LightGBM regression modeling. The R2 value for the training data is 0.736, while it decreases to 0.522 for the testing data. This implies that 52.2% of the leave data variables can be accounted for by variations in independent variables such as age, gender, echelon 1, and the year of starting leave. The remaining portion can be attributed to other factors outside the scope of this study.

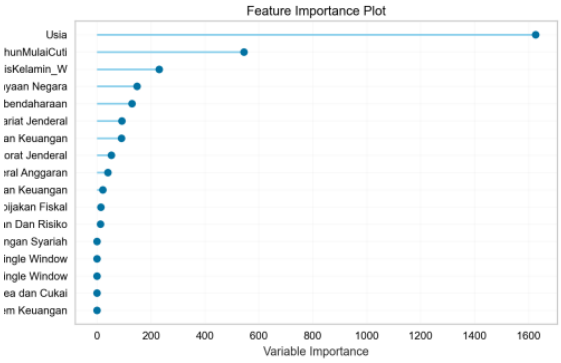

Figure 11. Important features in prediction

Regarding important features in prediction, Figure 10 highlights age as the most influential feature in the best model. Subsequently, two prediction tests were conducted using the regression modeling results, focusing on 10 rows from the dataset used in the descriptive analysis.

	JenisKelamin	Usia	Es1	TahunMulaiCuti	JumlahPegawai
40	P	23	Inspektorat Jenderal	2020	2
41	P	23	Sekretariat Jenderal	2019	9
42	P	23	Sekretariat Jenderal	2020	2
43	P	24	Badan Pendidikan dan Pelatihan Keuangan	2020	2
44	P	24	Direktorat Jenderal Anggaran	2020	1
45	P	24	Direktorat Jenderal Kekayaan Negara	2018	3
46	P	24	Direktorat Jenderal Kekayaan Negara	2019	10
47	P	24	Direktorat Jenderal Kekayaan Negara	2020	9
48	P	24	Direktorat Jenderal Kekayaan Negara	2021	3
49	P	24	Direktorat Jenderal Pengelolaan Pembiayaan Dan...	2018	1
50	P	24	Direktorat Jenderal Pengelolaan Pembiayaan Dan...	2019	3

Figure 12. Leave dataset

```
# Mengambil baris ke 46 dalam data untuk diprediksi

test_prediksi_df = prediksi_df.loc[46:46].copy()
test_prediksi_df
```

	JenisKelamin	Usia	Es1	TahunMulaiCuti	JumlahPegawai
46	P	24	Direktorat Jenderal Kekayaan Negara	2019	10

Figure 13. Rows 46 for prediction test

Figure 11 is the original dataset that was also used in the previous descriptive analysis. Figure 12, on the other hand, illustrates the first prediction test involving row number 46 of the original dataset. In this case, the employee who took leave is 24 years old and commenced the leave in 2019.

```
# Melakukan prediksi tahun mulai cuti 2020 dan usia 24

test_prediksi_df['Usia'] = 24
test_prediksi_df['TahunMulaiCuti'] = 2020
```

```
result = predict_model(model_lightgbm, test_prediksi_df)
```

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	Light Gradient Boosting Machine	3.4731	12.0621	3.4731	nan	0.3794	0.3473

```
result
```

	JenisKelamin	Usia	Es1	TahunMulaiCuti	JumlahPegawai	Label
46	P	24	Direktorat Jenderal Kekayaan Negara	2020	10	6.526941

Figure 14. First prediction test

Figure 13 illustrates the results of the first prediction test. The objective of this test is to predict whether an employee who is 24 years old and started their leave in 2020 will take annual leave along with joint leave. The prediction result in Figure 13 indicates that there are 10 employees who meet these criteria. By comparing this prediction result to the original dataset in Figure 11, it can be observed that the number of employees with these characteristics, specifically at Echelon 1 of the Directorate General of State Assets and starting leave in 2020, is close to row 47. In that row, the original dataset shows 9 employees, whereas the prediction result suggests 10 employees. Hence, it can be concluded that the first prediction result is relatively close to the original data.

```
test_prediksi_df['Usia'] = 24
test_prediksi_df['TahunMulaiCuti'] = 2021
```

```
result2 = predict_model(model_lightgbm, test_prediksi_df)
```

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	Light Gradient Boosting Machine	6.3143	39.8708	6.3143	nan	0.8534	0.6314

```
result2
```

	JenisKelamin	Usia	Es1	TahunMulaiCuti	JumlahPegawai	Label
46	P	24	Direktorat Jenderal Kekayaan Negara	2021	4	3.685671

Figure 15. Second prediction test

Similarly, Figure 14 represents the second prediction test, which also employs the same row as the first test, namely row 46 from the original dataset. However, in this case, the prediction is made for a 24-year-old employee starting leave in 2021. The predicted label indicates that there are 4 employees who would take annual leave along with joint leave

based on these criteria. Comparing this with the original dataset in Figure 11, where the number of employees meeting these criteria is found at Echelon 1 of the Directorate General of State Assets and starting leave in 2021, it closely aligns with row 48. The original dataset displays 3 employees in that row, while the prediction result suggests 4 employees. Therefore, the prediction results are reasonably close to the original data.

3.3.3. Predictive Analysis Using K-Means Method

Clustering prediction using the K-Means method was performed on the leave dataset with rejected status, as the process run by PyCaret took too long and couldn't display prediction results when using the dataset with both approved and rejected status. Thus, initial modeling was conducted on the subset of the dataset with rejected status.

	IDPegawai	KodeOrganisasi		Est	JumlahHari	IDRefJenisKalamin	Usia	Keperluan_string_fdist	SelisihPegajuanCutiMenujuMulaiCuti
131	94938	3512020402	Badan Kebijakan Fiskal	2	P	39	keluarga	-1
145	115315	3508050403	Direktorat Jenderal Kekayaan Negara	1	P	33	keluarga	-1
149	138489	3505331208	Direktorat Jenderal Pajak	1	P	24	keluarga	-1
197	129330	3504070305	Direktorat Jenderal Anggaran	2	P	26	nan	-1
217	74460	3508190706	Direktorat Jenderal Kekayaan Negara	1	P	43	keluarga	-1

Figure 16. Leave dataset filter results

Figure 15 presents a subset of rows from the dataset used for clustering with the K-Means algorithm. The evaluation of the data clustering quality was carried out using the silhouette coefficient, considering different combinations of input cluster values.

```
kmeans = create_model('kmeans')
```

	Silhouette	Calinski-Harabasz	Davies-Bouldin	Homogeneity	Rand Index	Completeness
0	0.2278	4663.8523	1.2156	0	0	0

Figure 17. Silhouette value of clustering with K-Means algorithm

The silhouette value obtained for the K-Means clustering, as shown in Figure 16, is 0.2278. This value indicates a relatively weak cluster structure, prompting the need for further experimentation using the elbow method to determine the optimal number of clusters (k).

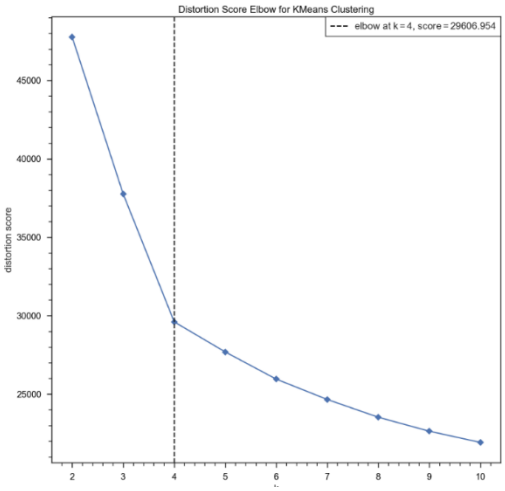

Figure 18. Elbow graph

Figure 17 presents the elbow graph, where the x-axis represents the number of clusters (k), and the y-axis displays the sum of squared errors (SSE) as an error measure. The graph clearly demonstrates a significant decrease in error when k changes from 2 to 3 and from 3 to 4, followed by a slower decrease until k=10. The "elbow" shape in the graph occurs at k=4, indicating that it is the best value of k for this study. The modeling results using the K-Means algorithm with k=4 are presented in Figure 18 to Figure 21.

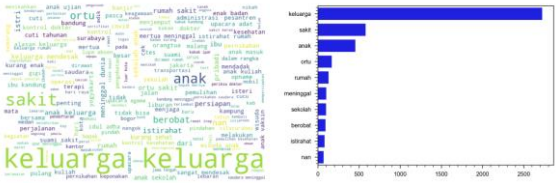

Figure 19. Cluster 0

Figure 18 showcases cluster 0, which reveals that rejected leave applications in this cluster have an average difference of 3 days between the application date and the start date of leave. The average age of employees applying for leave is 50 years old, and the average duration of leave applied for is 2 days. The three most common reasons for leave in cluster 0 are family needs, illness, and children.

Figure 20. Cluster 1

Cluster 1, shown in Figure 19, is the cluster with the highest frequency of leave requests. Within this cluster, rejected leave applications have an average difference of 2 days between the application date and the start date of leave. The average age of employees applying for leave is 31 years old, and the average duration of leave applied for is 1.5 days. The top three reasons for leave in cluster 1 are family needs, illness, and children.

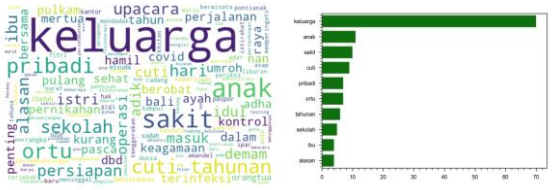

Figure 21. Cluster 2

Figure 20 illustrates cluster 2, where rejected leave applications exhibit an average difference of 5 days between the application day and the start day of leave. The average age of employees applying for

leave in this cluster is 34 years, and the average duration of leave applied for is 2 days. The primary reasons for leave within cluster 2 are family needs, illness, and children.

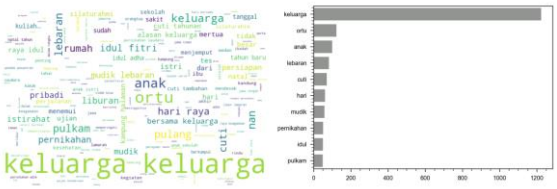

Figure 22. Cluster 3

In Figure 21, cluster 3 is presented, representing rejected leave applications with an average difference of 18 days between the application day and the start day of leave. The average age of employees applying for leave in this cluster is 35 years, and the average duration of leave applied for is 3 days. The top three reasons for leave within cluster 3 are family needs, parents, and children.

Based on the 4 clusters formed, the majority of reasons for leave are family-related, illness, childcare, and eldercare, albeit in different orders and quantities. Furthermore, there is a significant difference in the average age of employees requesting leave between cluster 0 and the other three clusters, with cluster 0 having an average age of 50 years, while the other three clusters fall within the range of 31-35 years. Based on the average number of days between the leave application date and the start of leave, it can be observed that employees in cluster 3 have better leave planning compared to those in cluster 1.

3.4. Evaluation

The performance of the model developed in the previous stages is assessed to determine its effectiveness. The LightGBM regression model, which achieved an R2 value of 0.5077, emerged as the best prediction method. Through two prediction tests, the prediction results are closely aligned with the original data. Hence, the LightGBM regression model, incorporating gender, age, echelon 1, and year of leave start as independent variables, proves capable of accurately predicting the number of employees taking annual leave alongside holidays. Nonetheless, incorporating additional independent variables is recommended to enhance the accuracy of the predictions.

Regarding the K-Means clustering approach, the silhouette value of 0.2278 indicates that the data points are correctly assigned to their respective clusters. To determine the optimal number of clusters (k) for improved accuracy, the Elbow method was employed. Based on the Elbow method graph, the value of k = 4 was selected since it exhibits an "elbow" shape in the graph, indicating a significant decrease in error up to that point.

4. DISCUSSION

Leave is a crucial benefit provided to employees to facilitate their physical and mental well-being. However, when a significant number of employees request leave at the same period, it becomes crucial to effectively manage the leave systems to maintain the continuity of the organization's business processes.

Previous research in the field of leave management decision support systems, as mentioned in the introduction, has explored various methods such as TOPSIS, SAW, EDAS, and PIPRECIA. However, limited research has been conducted in the domain of data mining specifically related to leave management, with only descriptive and affinity propagation methods being employed.

This study adopts the CRISP-DM methodology, encompassing stages ranging from a comprehensive literature review to evaluation, utilizing data obtained from the Leave module in the HRIS application. The modeling techniques employed involve regression prediction using LightBGM to forecast the number of employees likely to take annual leave alongside joint leave, and K-Means clustering to categorize the reasons for leave. The initial dataset of 414,809 data rows and 46 columns was then cleansed and analyzed using python language. The modeling results derived from the analysis are thoroughly evaluated, leading to the formulation of conclusive findings.

From the analysis results using LGBM modeling and K-Means clustering, predictions were obtained for leave-taking and clustering of common reasons employees use for taking leave. The LGBM prediction results achieved accuracy in determining employees who take leave simultaneously. Meanwhile, the K-Means method resulted in 4 clusters of leave reasons.

5. CONCLUSION

Predictive analysis was conducted using regression and clustering methods. Among the regression algorithms compared, the LightGBM algorithm demonstrated the best performance with an R2 value of 0.5077. The predictions generated by this model closely matched the original data in two separate tests. Therefore, the LightGBM regression model, utilizing independent variables such as gender, age, echelon 1, and leave year, can effectively predict the number of employees taking annual leave along with holidays. This provides valuable insights for decision-makers when granting or approving employee leave.

Meanwhile, the K-Means clustering algorithm was used to group the reasons for employee leave, offering recommendations for setting leave reason parameters in the HRIS application's Leave module. Additionally, it provided insights into the average difference in days of leave requests that are likely to be rejected. Analysis revealed that most rejected leave applications had no difference in days between

the application and the start of leave, indicating a high likelihood of rejection. The clustered reasons for leave included family, illness, children, and parents, which can inform the selection of leave reason parameters in the HRIS application's Leave module.

For future research, it is recommended to analyze leave data encompassing both approved and rejected statuses to improve prediction accuracy. Comparisons using combinations of additional variables would allow for a better understanding of the characteristics of different clusters. Furthermore, analyzing leave data with more comprehensive holiday and joint leave dates could lead to more accurate predictions regarding the number of employees taking annual leave alongside specific holidays or joint leave periods. Exploring the incorporation of other independent variables is also highly advised in order to obtain more precise prediction outcomes.

REFERENCES

[1] I. Mardeli and Y. Yansahrita, "Pengaruh Kedisiplinan Terhadap Prestasi Kerja Pegawai Pada Kantor Kecamatan Belitang Madang Raya Oku Timur," *J. Aktual STIE Trisna Negara*, vol. 17, no. 1, pp. 41–52, 2019, doi: 10.47232/aktual.v17i1.32.

[2] K. Ummi, Daifiria, and Irwansyah, "Penerapan Metode Deskripsi Pada Sistem Monitoring Data Cuti Pegawai Di Dinas Pemerintahan," *CSRID J.*, vol. 13, no. 3A, pp. 190–200, 2021.

[3] *Peraturan Badan Kepegawaian Negara Republik Indonesia Nomor 24 Tahun 2017 tentang Tata Cara Pemberian Cuti Pegawai Negeri Sipil*. Jakarta: Badan Kepegawaian Negara, 2017.

[4] M. I. Dwipayani, A. S. Honggowibowo, and D. Nugraheny, "Sistem Pendukung Keputusan Penentuan Cuti Pegawai Menggunakan Metode Technique For Order Preference By Similarity To Ideal Solution (TOPSIS)," *Compiler*, vol. 1, no. 1, pp. 47–61, 2012, doi: http://dx.doi.org/10.28989/compiler.v1i1.5.

[5] A. F. A. Natsir and M. K. Najiyah, Ina S.Kom., "Sistem Pendukung Keputusan Cuti Pegawai Di PUSDIKKU TNI AD dengan Menggunakan Metode SAW Berbasis Web," *eProsiding Tek. Inform.*, vol. 2, no. 1, pp. 21–27, 2021.

[6] M. K. Zuhri, L. Yulianti, and R. Supardi, "The Implementation of SAW Method in Applying Employees Leave At The Regional Secretariat Office Of Bengkulu," *J. Komput. Indones.*, vol. 1, no. 1, pp. 43–48, 2022, [Online]. Available: https://jurnal-unived.com/index.php/JK/article/view/12

[7] C. S. Suriady, L. P. Dewi, and A. Setiawan, "Aplikasi Human Resource Information System dengan Fitur Sistem Pendukung Keputusan (Studi Kasus PT. Industri Kreatif Digital)," *J. Infra*, vol. 9, no. 2, pp. 314–319, 2021, [Online]. Available: http://publication.petra.ac.id/index.php/teknik-informatika/article/view/11466%0Ahttps://publication.petra.ac.id/index.php/teknik-informatika/article/download/11466/10076

[8] T. P. Putri and Febriani, "Clustering Data Cuti Sakit Menggunakan Algoritma Affinity Propagation (Studi Kasus: Perusahaan Telekomunikasi Di Jakarta)," *J. Ilm. Teknol. dan Rekayasa*, vol. 27, no. 1, pp. 69–84, 2022, doi: http://dx.doi.org/10.35760/tr.2022.v27i1.5823.

[9] Y. Hermawan, A. Mulyana, and N. F. Rizky I., "Rancang Bangun Sistem Informasi Kepegawaian Berbasis Web di STIE Kesatuan," *JAS-PT (Jurnal Anal. Sist. Pendidik. Tinggi Indones.*, vol. 3, no. 2, pp. 153–160, 2019, doi: 10.36339/jaspt.v3i2.420.

[10] D. A. Firmansah, R. S. Rohman, and Y. Farlina, "Aplikasi Website Pengajuan Cuti Karyawan Rumah Sakit Islam Assyifa Sukabumi Berbasis Whatsapp Blast," *J. Teknol. dan Inf.*, vol. 10, no. 2, pp. 129–143, 2020, doi: 10.34010/jati.v10i2.2854.

[11] R. Taufiq, A. A. Permana, and M. A. Marfino, "Rancang Bangun Sistem Informasi Pengajuan Cuti Berbasis Web Pada Pt. Tribuana Gasindo," *JIKA (Jurnal Inform. Univ. Muhammadiyah Tangerang*, vol. 6, no. 1, pp. 99–104, 2022, doi: 10.31000/jika.v6i1.5472.

[12] M. Badrul and L. N. Janah, "Penerapan Metode Waterfall dalam Sistem Informasi Cuti Kepegawaian Madrasah Istiqlal," *J. Sains Komput. Inform. (J-SAKTI*, vol. 6, no. 1, pp. 282–293, 2022, doi: 10.30645/j-sakti.v6i1.444.

[13] M. A. Rizaldi and A. Primajaya, "Rancang Bangun Sistem Informasi Pengajuan Cuti Pegawai Berbasis Web pada PT. Mitratiga Perkasa Abadi," *J. Ilm. Wahana Pendidik.*, vol. 8, no. 16, pp. 381–388, 2022, doi: 10.5281/zenodo.7067874.

[14] A. Rachman and Effiyaldi, "Sistem Informasi Cuti Pegawai Berbasis Web Pada Universitas Jambi," *Manaj. Sist. Inf.*, vol. 8, no. 1, pp. 55–66, 2023, doi: https://doi.org/10.33998/jurnalmsi.2023.8.1.763.

[15] C. A. Gemawaty, "Perancangan Sistem

Pengajuan Cuti Kerja PT DCM Fatih Medika Menggunakan Netbeans," *Innov. J. Soc. Sci. Res.*, vol. 2, no. 2, pp. 95–101, 2022, doi: https://doi.org/10.31004/innovative.v2i2.80.

[16] J. Yang *et al.*, "Brief introduction of medical database and data mining technology in big data era," *J. Evid. Based. Med.*, vol. 13, pp. 57–69, 2020, doi: 10.1111/jebm.12373.

[17] F. Martinez-Plumed *et al.*, "CRISP-DM Twenty Years Later : From Data Mining Processes to Data Science Trajectories," *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 8, pp. 3048–3061, 2019, doi: 10.1109/TKDE.2019.2962680.

[18] D. A. Nasution, H. H. Khotimah, and N. Chamidah, "Perbandingan Normalisasi Data untuk Klasifikasi Wine Menggunakan Algoritma K-NN," *J. Comput. Eng. Syst. Sci.*, vol. 4, no. 1, pp. 78–82, 2019, doi: 10.24114/cess.v4i1.11458.

[19] F. M. Hana, "Klasifikasi Penderita Penyakit Diabetes Menggunakan Algoritma Decision Tree C4.5," *J. SISKOM-KB (Sistem Komput. dan Kecerdasan Buatan)*, vol. 4, no. 1, pp. 32–39, 2020, doi: 10.47970/siskom-kb.v4i1.173.

[20] A. A. Aldino, D. Darwis, A. T. Prastowo, and C. Sujana, "Implementation of K-Means Algorithm for Clustering Corn Planting Feasibility Area in South Lampung Regency," *J. Phys. Conf. Ser.*, vol. 1751, no. 1, 2021, doi: 10.1088/1742-6596/1751/1/012038.

[21] E. Buulolo, *Data Mining Untuk Perguruan Tinggi*. Deepublish, 2020. [Online]. Available: https://books.google.co.id/books?id=-K%5C_SDwAAQBAJ

[22] M. K. Pangkasidhi, H. N. Palit, and A. Gunawan, "Analisis Sentimen Mahasiswa di Surabaya Terhadap Pelayanan Vaksinasi COVID-19 Menggunakan Beberapa Classifier," *J. Infra*, vol. 10, no. 2, pp. 21–27, 2022.

[23] Y. Zhang, C. Zhu, and Q. Wang, "LightGBM-based model for metro passenger volume forecasting," *IET Intell. Transp. Syst.*, vol. 14, no. 13, pp. 1815–1823, 2020, doi: 10.1049/iet-its.2020.0396.

[24] P. Septiana Rizky, R. Haiban Hirzi, and U. Hidayaturrohman, "Perbandingan Metode LightGBM dan XGBoost dalam Menangani Data dengan Kelas Tidak Seimbang," *J Stat.*, vol. 15, no. 2, pp. 228–236, 2022, doi: https://doi.org/10.36456/jstat.vol15.no2.a5548.

[25] D. N. P. Sari and Y. Sukestiyarno, "Analisis Cluster dengan Metode K-Means pada Persebaran Kasus Covid-19 Berdasarkan Provinsi di Indonesia," *Prism. Pros. Semin. Nas. Mat.*, vol. 4, pp. 602–610, 2021, [Online]. Available: https://journal.unnes.ac.id/sju/index.php/prisma/

[26] E. Umargono, J. E. Suseno, and V. G. S. K., "K-Means Clustering Optimization using the Elbow Method and Early Centroid Determination Based-on Mean and Median," *Proc. 2nd Int. Semin. Sci. Technol. (ISSTEC 2019)*, vol. 1, pp. 234–240, 2019, doi: 10.5220/0009908402340240.

[27] R. Y. Sari, H. Oktavianto, and H. W. Sulistyo, "Algoritma K-Means dengan Metode Elbow untuk Mengelompokkan Kabupaten/Kota di Jawa Tengah Berdasarkan Komponen Pembentuk Indeks Pembangunan Manusia," *J. Smart Teknol.*, vol. 3, no. 2, pp. 104–108, 2022, [Online]. Available: http://jurnal.unmuhjember.ac.id/index.php/JST/article/view/6928

[28] V. A. Permadi, S. P. Tahalea, and R. P. Agusdin, "K-Means and Elbow Method for Cluster Analysis of Elementary School Data," *Prog. Pendidik.*, vol. 4, no. 1, pp. 50–57, 2023, doi: 10.29303/prospek.v4i1.328.

[29] K. R. Shahapure and C. Nicholas, "Cluster quality analysis using silhouette score," *Proc. - 2020 IEEE 7th Int. Conf. Data Sci. Adv. Anal.*, pp. 747–748, 2020, doi: 10.1109/DSAA49011.2020.00096.

[30] R. Hidayati, A. Zubair, A. H. Pratama, and L. Indana, "Analisis Silhouette Coefficient pada 6 Perhitungan Jarak K-Means Clustering," *Techno.COM*, vol. 20, no. 2, pp. 186–197, 2021, doi: 10.33633/tc.v20i2.4556.

[31] C. Dr. Meiryani, S.E., Ak., M.M., M.Ak., "Memahami R Square (Koefisien Determinasi) Dalam Penelitian Ilmiah," 2021. https://accounting.binus.ac.id/2021/08/12/memahami-r-square-koefisien-determinasi-dalam-penelitian-ilmiah/ (accessed May 26, 2023).