The Indonesian Journal of Computer Science
www.ijcs.net
Volume 13, Issue 4, August 2024
https://doi.org/10.33022/ijcs.v13i4.4120

Prediction of Employee Reassignment Using Supervised Machine Learning Algorithms and Rule-Based Experts: A Case Study in Educational Institution

Bayu Suciono Romdhoni[1], Denny[2]
bayu_bsr@hotmail.com[1], denny@cs.ui.ac.id[2]
[1,2] Faculty of Computer Science, University of Indonesia

Article Information

Received : 11 Jun 2024
Reviewed: 26 Jul 2024
Accepted : 1 Aug 2024

Keywords

Placement, Supervised Machine Learning, Random Forest

Abstract

Education plays a crucial role in shaping the future of a nation. To maintain the quality of education, effective human resource management is essential in educational institutions. This study addresses the challenges of employee's placement under the Educational Institution. According data from December 2021 to May 2024, only 2,452 out of 41,722 employees were reassignment, which is significantly below the target set by regulation. This study evaluates several supervised machine learning algorithms, including Gaussian Naive Bayes, Decision Tree, Support Vector Machine, and Random Forest. Random Forest emerges as the most suitable algorithm due to its superior accuracy, precision, recall, and F1 Score. Following the evaluation of the chosen algorithm, the deployment phase includes comprehensive data preprocessing steps, such as handling missing values, data normalization, and categorical feature encoding. This system integrates with Google API for geospatial data, ensuring accurate and efficient decision-making.

A. Introduction

Education is characterized as an intentional and well-organized initiative aimed at cultivating an environment and process conducive to learning, with the aim of nurturing inner strength, self-regulation, moral growth, cognitive abilities, moral principles, and the proficiencies necessary for individuals, communities, the nation, and the state.. To maintain the quality of education, educational institutions undoubtedly conduct effective human resource management as one of many approachs.

From the total number of employees in Educational Units, from December 2021 to May 2024, the number of employee'ss reassignment that occurred was only 2,452 out of a total of 41,722, or about 5.8% of the population. This is certainly not in line with regulation which states that the reassignment is carried out once every 4 years. This issue arises because the Institution stiill manually conducts the mapping process for reassignment. This leads to the desired is not being achieved and, of course, falls far short of expectations.

Using technology that combines the principles of data science and have knowledged from data is technique of data mining [1]. Data mining is divided into two types: unsupervised learning and supervised learning. The approach in which machine learning models learn patterns from labeled or targeted data is called supervised learning [1]. These models are supervised during the learning process and are given guidance on what to predict. Meanwhile, unsupervised learning is an approach in which machine learning models attempt to find patterns and structures in data without specific guidance or labels [1].

By leveraging data mining techniques, the Educational Institution can conduct deeper and more optimal analyses of employee's reassignment, considering various variables that are standard in the process. This would allow for the goal of equalizing human resources within the Educational Institution to be achieved.

There are several studies which use Machine Learning to approach this issue. The first study have problem addressed is the need for a support tool to assist HR recruiters in selecting and placing candidates for specific positions [2]. The study's limitation is that it employs the Machine Learning (ML) VOBN (Variable-Order Bayesian Network) algorithm was used to research that aims to create a support tool for HR with hybrid technique to facilitate recruitment and candidate placement processes [2]. The second study have objective to make a tool utilizing a machine learning model that can precisely predict a student's likelihood of obtaining a job with a company, taking into account factors like age, gender, academic background, internships, CGPA, hostel accommodation, and historical data. However, the research is confined to using the Decision Tree Classifier and Naïve Bayes algorithms. Ultimately, the study aims to confirm the efficacy of machine learning in accurately predicting student placements [3]. The third study is aiming to address the need for accurate and early forecasting to implement suitable strategies ensuring every student receives placement, using a tool to predict placements based on various variables, within the constraint of utilizing 7 algorithms such as Ensemble Voting Classifier, SVM, LR, RF, DT, and Gaussian NB, and k-NN Classification [4]. The last study is addressing issues such as the influence of salary on student placement by course specialization and gender, placement status, and

identification, utilizing logical computational and algorithm of machine learning approaches for real-time student employment decision making[5].

Furthermore, by leveraging technology such as Google API to obtain distance data and combining it with employee status data, age ranges, and other variables to be used, the reassignment decision-making process can become more accurate, efficient, and avoid previous data irrelevance issues, ultimately achieving the desired equalization.

B. Literature Review

This section covers the knowledge and theoretical framework applied in the research.

1. Data

Information that is being collected, recorded, and subsequently analyzed by qualitative and quantitave technique is the meaning of data. [1]. Data includes facts, numbers, letters, symbols, images, sounds, or measurements gathered from diverse sources such as scientific research, government records, commercial transactions, social media, and other platforms. [6]. Data can also be represented based on the material structure and various properties of physical objects [7]. Insights gained from data analysis underpin strategies and innovations. The presence of data also extends to the landscape of digital social media, where user interactions and content creation contribute to its ceaseless growth. However, the significance of data lies in its transformative potential because once collected and utilized, data becomes information for progress and innovation. To perform data analysis, several steps are required such as reading data, exploring data, identifying data, analyzing data, and visualizing data [8]. Through the science of data analysis, unstructured raw data is transformed into highly valuable information. This information underpins strategic decision-making in various industries and domains as it helps businesses understand market trends and consumer preferences, shaping their strategies for growth and sustainability. Similarly, in the realm of scientific research, data analysis assists in obtaining new information that enables researchers to uncover the mysteries of new phenomena.

2. Data Mining

Data mining entails extracting insights by identifying patterns, knowledge, or hidden insights within large, complex, and unstructured datasets [1]. Data mining primarily aims to uncover valuable relationships, trends, or patterns that can improve decision-making. The quality of the data greatly impacts the effectiveness of the data mining outcomes [9]. The data mining process involves various computational and statistical techniques to unravel data, identify patterns or information not directly visible, and transform it into actionable knowledge [1].

3. Supervised Learning

Supervised learning is an approach where the model's labels are predetermined based on selected features and targets [1]. The goal of supervised learning is to explore unseen in data's structures, such as patterns, groups, or correlations that may not be directly visible. Learning the relationship between input and output

samples according to the given model is also the objective of Supervised Learning [9].

In a business context, supervised learning can assist organizations in identifying groups of customers with similar behaviors, categorizing similar products or services, or recognizing anomalies in data that may indicate business problems or opportunities. By implementing supervised learning, organizations can make more informed decisions because they receive information based on the algorithms used for consideration, enabling them to execute more effective business strategies. Some algorithms used in supervised learning include the following.

4. Gaussian Naive Bayes

A widely used technique in text mining for sentiment analysis is the Naive Bayes algorithm. [10], which is a straightforward classifier relying on probability and statistics through Bayes' Theorem [11]. This method is theoretically robust concerning data coherence and classification computation. Twitter uses classification techniques which are based on the Naive Bayes algorithm such as Unigram, Multinomial and Maximum Entropy of Naive Bayes. [12][13]. The ability of Naive Bayes Classification to generate strong hypotheses from a given condition is one of its main features to be considered. The calculation of group probabilities in Naive Bayes is carried out using the Bayesian algorithm approach using equations.

$$P(X|Y) = \frac{P(x|y)P(Y)}{P(X)}$$

In the equation, Y denotes a specific category, X represents data without a defined category, and P(Y|X) signifies the probability of a hypothesis given a condition, probabilities of P(Y) and P(X|Y) are derived previously from a category based on the given hypothesis condition, and P(X) is the probability obtained from Y. Efficiently extracting text data involves extracting essential information from diverse records. Initially, unstructured text is observed in content data, then structured and stored in a database, demonstrating the strategy of opinion mining using the Naive Bayes method on Twitter. Data retrieval is conducted using specific keywords within a defined timeframe. The process of sentiment assessment labeling occurs after data retrieval. Subsequently, systematic preprocessing and data set transformation take place, involving the cleaning process to reduce noise and eliminate unnecessary words like 'I' and 'and.'

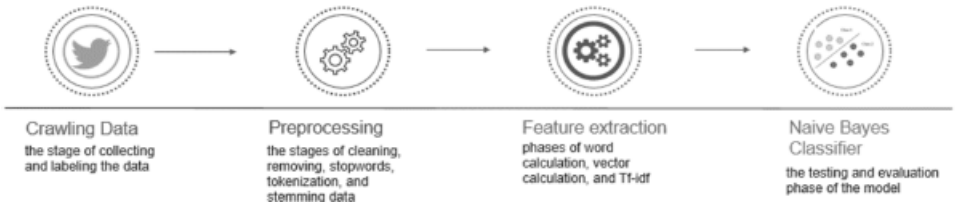

Figure 1. Naïve Bayes Research Stages

The process of tokenization in the Naive Bayes algorithm defines words by segmenting sentences into space-separated phrases and punctuation [10]. The last preprocessing step involves transforming affixes into base words. The third phase of opinion mining involves extraction, aiming to streamline Naive Bayes

Classification. During this phase, a model is generated to showcase the effectiveness of Naive Bayes Classification accuracy.

5. Decision Tree

An algorithm used in machine learning which are classification is Decision Tree, characterized by a tree-based approach where each path, originating from the root node and leading to the Boolean outcome at the leaf node, is established [14]. This is a hierarchical representation of relations that contain information between nodes and connections. Nodes and trees form a tree structure in each decision tree. Features in categorized data are represented by each node, and the values that each node can take are defined as a tree [15].

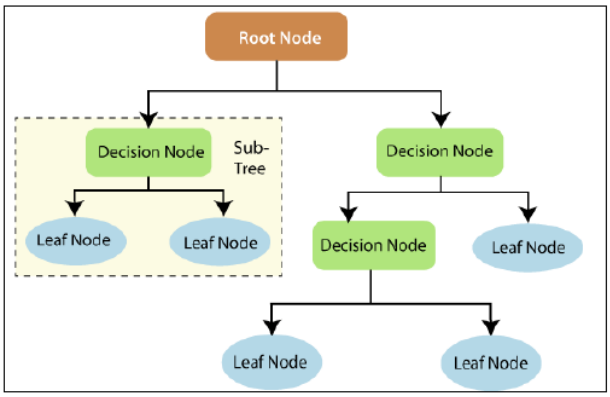

Figure 2. Decision Tree Process

The Decision Tree method finds widespread application across diverse domains such as pattern recognition, image processing, and machine learning. Some advantages of the Decision Tree algorithm include simplicity, usefulness for data exploration, less necessity for data cleaning processes, no constraints on data type, and utilization of non-parametric methods [16].

6. Random Forest

Random forest has become one of the most powerful learning algorithms. The development of this algorithm is beneficial as long as they can access its implementation [17].

```
for i ← 1 to B do
    Draw a bootstrap sample of size N from the training data;
    while node size != minimum node size do
        randomly select a subset of m predictor variables from total p;
        for j ← 1 to m do
            if jth predictor optimizes splitting criterion then
                split internal node into two child nodes;
                break;
            end
        end
    end
end
return the ensemble tree of all B subtrees generated in the outer for loop;
```

Figure 3. Machine Learning Algorithm (RF)

The category of ensemble learning technique is belong to RF, specifically based on trees, where predictions are averaged across multiple individual trees [18]. Every tree is built using bootstrap samples instead of the initial dataset, employing

a method referred to as bootstrap aggregating or bagging, which aids in reducing overfitting. While each individual tree in a random forest is easily interpretable, the interpretability diminishes when combining many trees. However, this trade-off often leads to superior predictive performance. Compared to decision trees, random forests tend to provide more accurate estimates of error rates, with the inaccuracy consistently converging linear with many increase of trees. In classification tasks, random forests are highly favored due to their intuitive decision-making process and excellent results [19].

7. Support Vector Machine

Vapnik introduced the algorithm that applied for regression task and classification model called Support Vector Machine (SVM) which are a kernel-based model [19]. Compared to other supervised learning methods, SVM has demonstrated superiority, making it a robust solution for practical binary classification problems. Its widespread adoption is attributed to its strong theoretical foundation and impressive generalization capabilities. [20].

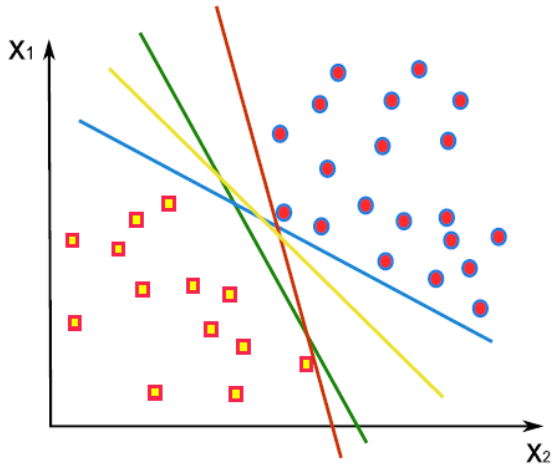

Figure 4. Support Vector Machine Separation Hyperlanes

The SVM algorithm derives the decision function directly from the training data. This technique minimizes classification errors during training and enhances the algorithm's ability to generalize. SVM is particularly effective when handling small input data sets, earning it recognition as an efficient classification algorithm for high-dimensional spaces [19]. Additionally, SVM provides notable benefits by selecting a subset of support vectors during the training process, which often represent only a small portion of the original dataset and play a crucial role in specific classification tasks.

C. Research Method

This section contains the results of research and discussion of the developed system design.

1. CRISP-DM

CRISP-DM or called Cross-Industry Standard Process for Data Mining is a commonly utilized methodology for data science, incorporating diverse methods or

frameworks. [1]. It has emerged as an industry standard extensively employed across diverse data analysis and data mining endeavors. CRISP-DM comprises six primary stages that span the entire data mining cycle [21], which encompass business comprehension, data comprehension, data preprocessing, data modeling and assessment, and deployment. Each stage entails distinct objectives and tasks essential for achieving success in the data mining process.

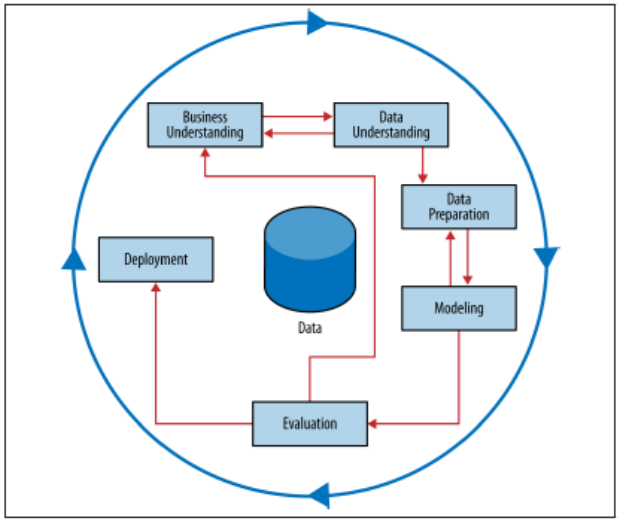

Figure 5. CRISP-DM Data Mining Process

Business Understanding: During this phase, the emphasis is on grasping the business issue that needs resolution. Typically, this is achieved by communicating with business stakeholders to understand the objectives, challenges, and opportunities. The outcome of this stage is a clear statement of business goals.

Data Understanding: After understanding the business obstacle, it is also necessary to knows about the data to be used. This involves data collection and evaluation to ensure that the data aligns with the objectives. In this stage, it is also important to identify issues such as missing values or outliers.

Data Preparation: The collected data needs to be prepared for analysis. Techniques such as data cleaning, data transformation, and data merging from various sources may be used if necessary. The goal of this stage is to produce a dataset ready for modeling.

Data Modeling and Evaluation: This is the stage where the actual data mining models are built. A range of machine learning algorithmic approaches, including regression, clustering, random forest, or other methods, may be employed, depending on the nature of the problem and the data accessible. Subsequently, these models undergo assessment to gauge their effectiveness and precision, a process that entails testing them on new data to prevent overfitting. If the models do not meet expectations, revision or further development is needed.

Deployment: The final stage is the deployment of the validated models into real data (production). This means using the evaluated model on real data so that it can be used to make decisions or provide recommendations.

One of the advantages of CRISP-DM is its common sense, cyclical nature, adaptability, right start, and flexibility [22]. This approach is adaptable to diverse data mining endeavors and applicable across various sectors. It enables project

teams to identify business problems, collect, process, and analyze data, and implement solutions in a more structured and efficient manner. CRISP-DM also encourages collaboration among various stakeholders in a project, including data analysts, data scientists, and business stakeholders.

D. Result and Discussion

The study developed an automatic tool for educational institutions to predict employee's reassignment. Below are the steps involved:

1. Data Understanding

   The data consist several tables of data with the following details.

Table 1. Employee's Attributes

Atribute's Name	*Size*
ptk_id	Unique id of employee
name	Full name of employee
gender	Gender
birthdate	Birthdate
employee_status	Status of employee's job
office_id	Office's id
office_name	Office's name
office_address	Office's status
office_stage	Office's stage
office_district_address	Office's district address
office_occupation	Occupation's name in each office
employee_category	Employee's category status
office_region	Office's region address
employee_address	Employee's address
employee_district_address	Employee's district address
employee_province_address	Employee's province address
employee_age_group	Additional column to group range of employee's age
employee_status_group	Additional column to grouping employee's status
office_occupation_group	Additional column to grouping office's occupation
employee_district_group	Additional column to grouping district of employee's address

Table 2. Distance's Attributes

Atribute's Name	*Size*
origin	Origin address
destination	Destination address
mode	Google map's mode (driving)
distance	Distance between origin and destination
duration	Duration between origin and destination

2. Data Preprocessing

   Based on the previously obtained data, the next process is data preprocessing. In this method, all data will be combined into one dataset for further modeling using the specified algorithm and evaluation. The processes include:

   a. Adjusting for missing values.
   b. Correcting data errors.

c. Flagging employee's data who's already have reassignment.
d. Merging the result from step 3 with distance data.

3. Data Modelling and Evaluation

Feature selection is performed based on the interview results are age range, gender, employee's occupation, distance, and reassignment's flagging. Theholdout technique is used to determine the division of data training and data testing.

From the results obtained in the initial trial employing the Gaussian Naive Bayes algorithm, the following observations were made..

- Accuracy: 0.7908 (79.08%)
- Precision: 0.6254 (62.54%)
- Recall: 0.7908 (79.08%)
- F1 Score: 0.6985 (69.85%)

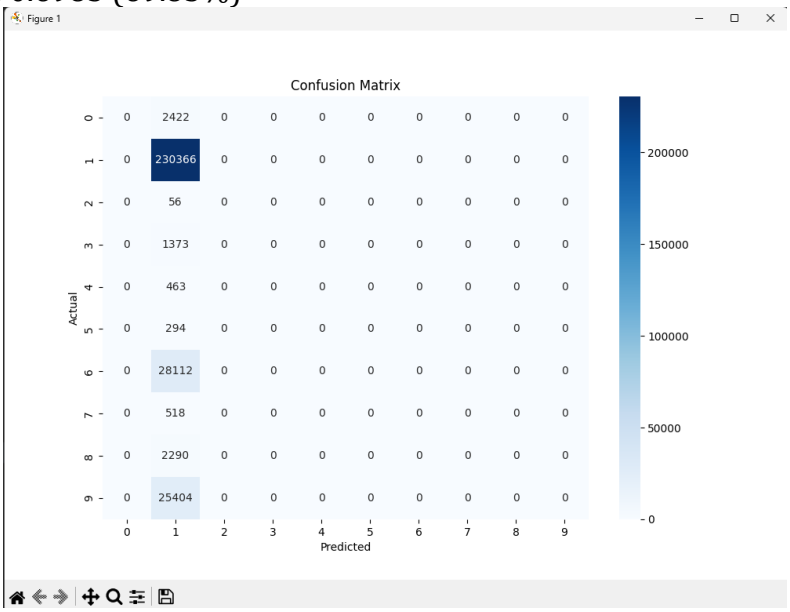

Figure 6. Gaussian Naive Bayes's Confusion Matrix

The test outcomes indicated a notable accuracy level with the Gaussian Naive Bayes algorithm, albeit its performance remained average due to difficulties in predicting minority classes. Enhanced performance can be achieved with this algorithm when addressing class imbalances.

In the subsequent experiment, utilizing the Decision Tree algorithm, the following outcomes were observed.

- Accuracy: 0.7928 (79.28%)
- Precision: 0.7421 (74.21%)
- Recall: 0.7928 (79.28%)
- F1 Score: 0.7579 (75.79%)

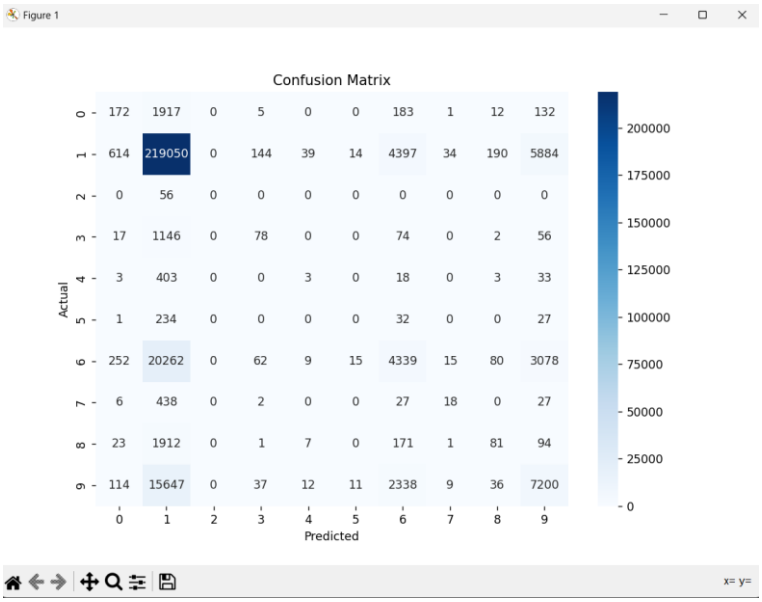

Figure 7. Decision Tree's Confusion Matrix

The testing results revealed class imbalance, which affected the performance of the Decision Tree algorithm. Techniques that can be used to achieve better testing results include:

a. Utilizing ensemble techniques, which enhance the model's capacity to address minority class considerations.
b. Modifying the classification threshold to attain a more balance between True Positive Rate (TPR) and False Positive Rate (FPR) for challenging categories..

The outcomes of the third experiment, employing the Random Forest algorithm, yielded the following results.

- Accuracy: 0.7916 (79.16%)
- Precision: 0.7452 (74.52%)
- Recall: 0.7916 (79.16%)
- F1 Score: 0.7609 (76.09%)

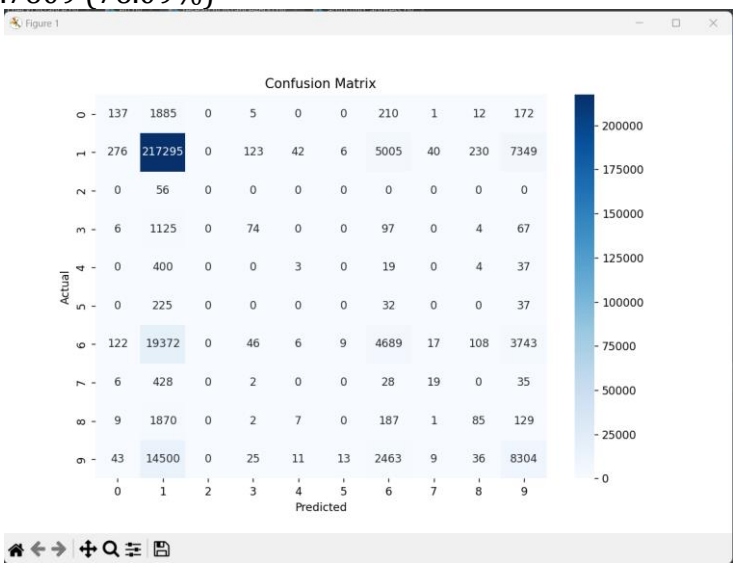

Figure 8. Random Forest's Confusion Matrix

Overall, the testing results using the Random Forest algorithm yielded high accuracy. The performance of this algorithm varies across each class. However, techniques are still needed to address class imbalance to achieve better testing results.

The last experiment based on Support Vector Machine (SVM) algorithm yielded the following testing results.

- Accuracy: 0.7908 (79.08%)
- Precision: 0.6254 (62.54%)
- Recall: 0.7908 (79.08%)
- F1 Score: 0.6985 (69.85%)

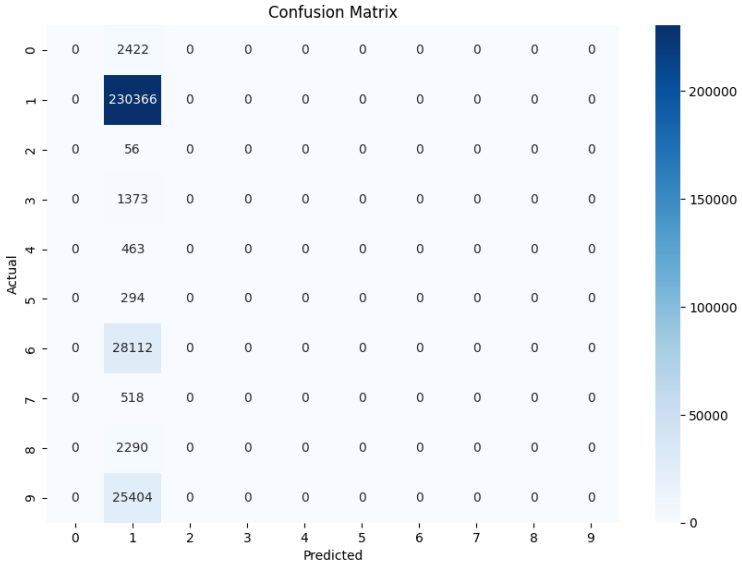

Figure 9. Support Vector Machine (SVM)'s Confusion Matrix

4. Deployment

From several testing results of models using Supervised Machine Learning algorithms, differences were observed as shown in Table 3.

Table 3. Experiment Results

Algorithm	Accuracy	*Precision*	*Recall*	*F1 Score*
GNB	0.7908258	0.6254055	0.7908258	0.6984549
DT	0.7927998	0.7420621	0.7927998	0.7578700
RF	0.7916497	0.7452436	0.7916497	0.7609008
SVM	0.7908258	0.6256055	0.7908258	0.6984549

Various machine learning algorithms underwent comparison, with Random Forest emerging as the preferred choice for this system. The algorithms evaluated included DT, SVM, and GNB, among others. RF demonstrated superior performance across various metrics such as Accuracy, Precision, Recall, and F1 Score compared to its counterparts. Specifically, it achieved an accuracy of 0.7916, precision of 0.7452, recall of 0.7916, and F1 Score of 0.7609, surpassing both Support Vector Machine and Gaussian Naive Bayes, which had an accuracy of 0.7908 and an F1 Score of 0.6985, and outperforming Decision Tree. Random Forest's ensemble

approach, which combines multiple decision trees, was the primary reason for its selection, as it provides more accurate and stable predictions while mitigating overfitting issues often observed in individual decision trees.. Additionally, Random Forest is adept at handling datasets with numerous features and coping effectively with missing values. Furthermore, it offers insights into feature importance, contributing to a better understanding of the factors influencing the employee mapping process.

E. Conclusion

In this study, an exploration of the use of technology was conducted to increase the number of employee's reassignment. To resolve this obstacle, data mining techniques will be used, particularly the Random Forest algorithm, is proposed. Through comparative trials, this particular algorithm consistently outperformed others such as GNB, DT, and SVM. Upon analyzing the results, Random Forest emerged as the most suitable algorithm for predicting employee reassignment and analyzing their mapping. With an accuracy rate of 0.7916, precision of 0.7452, recall of 0.7916, and an F1 Score of 0.7609, Random Forest demonstrated superior performance compared to alternative algorithms, largely attributed to its ensemble nature, which combines the results of many decision trees, reducing overfitting and being capable of handling datasets with high features and missing values.

The deployment process of this system involves several important technical steps. First, adequate technological infrastructure must be prepared, including the use of computers with high enough specifications. After that, preprocessing of the data is conducted to ensure the proper formatting of employee data for input into the machine learning model, which includes addressing missing values, normalizing data, and encoding categorical features. Next, hyperparameter tuning for Random Forest is performed to find the best combination that provides the most optimal results.

Moving forward, future research could explore the application of class imbalance techniques to address any imbalance in employee's reassignment. Class imbalance techniques such as oversampling or undersampling could be applied to guarantee that the model have maintains balance and accuracy in forecasting employee's reassignment across different educational levels.

By incorporating class imbalance techniques, future studies can further enhance the effectiveness and reliability of the proposed system in optimizing employee's reassignment and addressing the challenges faced by Educational Institution.

F. Acknowledgment

We extend our heartfelt appreciation to the University of Indonesia for furnishing the necessary resources and support for this research endeavor. We also wish to express our gratitude to all individuals and organizations who played a part in this research endeavor, whether through direct involvement or indirect support. Your collaboration and assistance have been indispensable in the accomplishment of this project.

G. References

[1] Provost & Fawcett, "Data science-what you need to know about analytic-thinking and decision-making," *J Chem Inf Model*, vol. 53, no. 9, pp. 1689–1699, 2013.

[2] D. Pessach, G. Singer, D. Avrahami, H. Chalutz 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, no. April, p. 113290, 2020, doi: 10.1016/j.dss.2020.113290.

[3] S. Chavhan, O. Joshi, S. Deshpande, A. Rambhad, P. Wanjari, and S. Tiwari, "Machine Learning Based Placement Prediction - A Comparative Study," *Proceedings of the 5th International Conference on Inventive Research in Computing Applications, ICIRCA 2023*, no. Icirca, pp. 343–350, 2023, doi: 10.1109/ICIRCA57980.2023.10220636.

[4] P. Jain, S. Khare, and M. K. Gourisaria, "A Data Mining Solution to Predict Campus Placement," *2021 IEEE 4th International Conference on Computing, Power and Communication Technologies, GUCON 2021*, pp. 1–7, 2021, doi: 10.1109/GUCON50781.2021.9573551.

[5] D. Kumar, C. Verma, P. K. Singh, M. S. Raboaca, R. A. Felseghi, and K. Z. Ghafoor, "Computational statistics and machine learning techniques for effective decision making on student's employment for real-time," *Mathematics*, vol. 9, no. 11, 2021, doi: 10.3390/math9111166.

[6] B. Chaudhary, "Importance of Data (A Term Paper)", doi: 10.13140/RG.2.2.23837.69602.

[7] R. van Koningsbruggen, H. Waldschütz, and E. Hornecker, "What is Data?-Exploring the Meaning of Data in Data Physicalisation Teaching," p. 21, 2022, doi: 10.1145/3490149.

[8] K. Yuditskiy, I. Bezdvornykh, A. Kazantseva, A. Kanapin, and A. Samsonova, "BSXplorer: analytical framework for exploratory analysis of BS-seq data," *BMC Bioinformatics*, vol. 25, no. 1, Dec. 2024, doi: 10.1186/s12859-024-05722-9.

[9] A. Dogan and D. Birant, "Machine learning and data mining in manufacturing," *Expert Systems with Applications*, vol. 166. Elsevier Ltd, Mar. 15, 2021. doi: 10.1016/j.eswa.2020.114060.

[10] Pristiyono, M. Ritonga, M. A. Al Ihsan, A. Anjar, and F. H. Rambe, "Sentiment analysis of COVID-19 vaccine in Indonesia using Naïve Bayes Algorithm," *IOP Conf Ser Mater Sci Eng*, vol. 1088, no. 1, p. 012045, Feb. 2021, doi: 10.1088/1757-899x/1088/1/012045.

[11] T. Setiadi, F. Noviyanto, H. Hardianto, A. Tarmuji, A. Fadlil, and M. Wibowo, "Implementation Of Naïve Bayes Method In Food Crops Planting Recommendation," *INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH*, vol. 9, p. 2, 2020, [Online]. Available: www.ijstr.org

[12] Yuyun, Nurul Hidayah, and Supriadi Sahibu, "Algoritma Multinomial Naïve Bayes Untuk Klasifikasi Sentimen Pemerintah Terhadap Penanganan Covid-19 Menggunakan Data Twitter," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 5, no. 4, pp. 820–826, Aug. 2021, doi: 10.29207/resti.v5i4.3146.

[13] I. Esa Tiffani, "Optimization of Naïve Bayes Classifier By Implemented Unigram, Bigram, Trigram for Sentiment Analysis of Hotel Review."
[14] L. Zhao, S. Lee, and S. P. Jeong, "Decision tree application to classification problems with boosting algorithm," *Electronics (Switzerland)*, vol. 10, no. 16, Aug. 2021, doi: 10.3390/electronics10161903.
[15] B. Charbuty and A. Abdulazeez, "Classification Based on Decision Tree Algorithm for Machine Learning," *Journal of Applied Science and Technology Trends*, vol. 2, no. 01, pp. 20–28, Mar. 2021, doi: 10.38094/jastt20165.
[16] M. Bansal, A. Goyal, and A. Choudhary, "A comparative analysis of K-Nearest Neighbor, Genetic, Support Vector Machine, Decision Tree, and Long Short Term Memory algorithms in machine learning," *Decision Analytics Journal*, vol. 3, p. 100071, Jun. 2022, doi: 10.1016/j.dajour.2022.100071.
[17] M. Schonlau and R. Y. Zou, "The random forest algorithm for statistical learning," *Stata Journal*, vol. 20, no. 1, pp. 3–29, Mar. 2020, doi: 10.1177/1536867X20909688.
[18] M. Savargiv, B. Masoumi, and M. R. Keyvanpour, "A new random forest algorithm based on learning automata," *Comput Intell Neurosci*, vol. 2021, 2021, doi: 10.1155/2021/5572781.
[19] M. Sheykhmousa, M. Mahdianpari, H. Ghanbari, F. Mohammadimanesh, P. Ghamisi, and S. Homayouni, "Support Vector Machine Versus Random Forest for Remote Sensing Image Classification: A Meta-Analysis and Systematic Review," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13. Institute of Electrical and Electronics Engineers Inc., pp. 6308–6325, 2020. doi: 10.1109/JSTARS.2020.3026724.
[20] J. Cervantes, F. Garcia-Lamont, L. Rodríguez-Mazahua, and A. Lopez, "A comprehensive survey on support vector machine classification: Applications, challenges and trends," *Neurocomputing*, vol. 408, pp. 189–215, Sep. 2020, doi: 10.1016/j.neucom.2019.10.118.
[21] C. Schröer, F. Kruse, and J. M. Gómez, "A systematic literature review on applying CRISP-DM process model," in *Procedia Computer Science*, Elsevier B.V., 2021, pp. 526–534. doi: 10.1016/j.procs.2021.01.199.
[22] J. S. Saltz, "CRISP-DM for Data Science: Strengths, Weaknesses and Potential Next Steps," in *Proceedings - 2021 IEEE International Conference on Big Data, Big Data 2021*, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 2337–2344. doi: 10.1109/BigData52589.2021.9671634.