ISSN 1816-6075 (Print), 1818-0523 (Online)
Journal of System and Management Sciences
Vol. 13 (2023) No. 6, pp.336 -346
DOI:10.33168/JSMS.2023.0620

Data Mining Techniques to Predict High Leadership Promotion of Civil Servants in Indonesia Based on Talent Pool Assessments

Dwi Pratiwi Kusumaningtyas, Nilo Legowo

Information Systems Management Department, Binus Graduate Program-Master of Information Systems Management, Bina Nusantara University, Jakarta, Indonesia, 11480

dwi.kusumaningtyas@binus.ac.id; nlegowo@binus.edu (Corresponding author)

Abstract. Civil servants play a crucial role in serving the government and the people, requiring them to possess both competent skills and organizational performance to fulfill the objectives of their respective organizations. To ensure the acquisition of high-quality civil servants with the right competencies and potential for promotion, a robust selection system is necessary. This paper focuses on utilizing data mining techniques to predict employee performance based on talent pool assessment results, thereby aiding in the identification of suitable candidates for target positions. The study employs predictive analytics to analyse the assessment data and predict the performance of employees in their designated roles. The data preprocessing stage involves the utilization of techniques such as one-hot encoding and feature selection to enhance the accuracy of the subsequent analysis. Several classification algorithms, including Naïve Bayes, Decision Tree, Random Forest, Support Vector Machine, and K-NN, are utilized to train models and attain the highest possible accuracy for the assessment system. The application of data mining techniques in the field of Human Resources, particularly in the government sector, offers valuable insights for predicting the suitability and qualification of candidates for specific positions. By leveraging these techniques, organizations can effectively identify individuals who possess the necessary competencies to thrive in their assigned roles. Overall, this paper emphasizes the importance of a robust selection process for civil servants and presents a practical approach utilizing data mining techniques to predict employee performance. The results obtained from this study can contribute to the development of more efficient and reliable assessment systems in the public sector, enabling organizations to make informed decisions when assigning individuals to key positions.

Keywords: talent pool, data mining, predictive analytic, classification

1. Introduction

Indonesia's State Civil Apparatuses consist of Civil Servants and Government Employees employed under employment agrrements based on Indonesian Law No. 5 of 2014. They are required to have the ability and competency stipulated by relevant laws and regulations, the principles of good, transparent, participatory, accountable, and fair government governance. An appointment of a civil servants in a particular position position must be based on competence, qualifications, and the requirements of the position through an open bidding or open auction following the Regulation of the Ministry of State Apparatus Utilization and Bureaucratic Reform No. 15 of 2019. The required conditions and qualifications are announced clearly, and the selection stage involves the head of ministry or government agency and a selection committee.

The selection process of high leadership officers is extremely vulnerable to corruption and nepotism that benefiting the interests of candidate's family or relations instead of the interests of society, nation, and state based on Indonesian Government Law No. 26 of 1985. Indonesia's Corruption Eradication Commission (KPK) stated that there are eight types of cases, including job auction bribes, especially for high leadership promotions that are not based on candidates' competence (Muhid, 2022). To avoid the corruption and nepotism loopholes in the selection process, a talent management system can be utilized to support the achievement of strategic golas of national development through quality public services by finding and preparing the best talents to fill key positions as future leaders, improving professionalism, competence, and talent performance, providing clarity and certainty of talent career, realizing an objective, planned, open, timely and accountable succession plan, ensuring the availability of talent supply to align the right employees with timely positions, as well as balancing employees' career development and agency needs. The aforementioned measures are in line with the Regulation of the Ministry of State Apparatus Utilization and Bureaucratic Reform No. 3 of 2020. A talent management system prepares successors or candidates who are qualified to occupy target positions based on the recommendation of talent pool assessment. Talent pool contributes to the overall talent management strategy in identifying and retaining essential talents (Jooss et al., 2021) and has the potential to contribute to organizations (Cappelli & Keller, 2014). Moreover, it is important that candidates are selected to occupy positions one level below the desired position.

The results of talent pool assessments are processed with the aid of information technology in an effort to reduce the error rate of examiners and expedite the procurement of test results used to select the most qualified candidate for the most suitable position at the appropriate time. However, this process is not easy as it depends on numerous variables, including human experience, knowledge, preferences, judgment, and the correlations among these variables. Simply having good experience in a particular position does not necessarily mean that one is suitable for that position. The individual should possess the right preferences and good judgment to be able to fit into a role, or they should have a relative amount of experience and the appropriate preferences. Therefore, the process of identifying existing talent within an organization poses a formidable obstacle for management and remains an ongoing concern.

Data mining has a positive impact on supporting management and policy development when utilized correctly. It combines statistics and machine learning using statistical methods to explore and build a model (Dahiya et al., 2021) and plays a crucial role in identifying data patterns that can assist organizations in making data-driven decisions based on talent pool data (Faqihi & Miah, 2023). It can be applied across various areas of human resource management, serving as a fundamental competency and competitive advantage for organizations, with promising prospects in its field. A current research trend involves combining data mining methods with optimization techniques to enhance the accuracy of classification and prediction (Yi & Yao, 2017), as well as to characterize and describe trends and patterns stored within data and information (McCue, 2007). In the public sector, data mining plays a vital role in optimizing organizational decision-making by extracting general trends from historical data (Wang et al., 2010).

However, in order to generate reliable recommendations based on talent pool assessment results and to accurately identify suitable candidates for the target position, it is necessary to assess the accuracy using data mining methods. This can be achieved by employing classification machine learning algorithms such as Decision Tree, Naïve Bayes, Random Forest, Support Vector Machine, and K-Nearest Neighbour. The objective of this research is to implement a prediction model that achieves the highest level of accuracy by utilizing some data pre-processing and machine learning approaches.

2. Related works

Researchers Qasem al Radeh and Eman Al Nagi (2012) made a preliminary attempt to use data mining concepts to build a classification model for predicting employee performance. They utilized the Decision Tree generated by the C4.5 algorithm to determine the effectiveness of performance prediction. This approach supported human resources directors and decision makers by evaluating employee's data to study the main attributes that may affect performance. The application of data mining concepts allowed for the development of a model to support the prediction of employee's performance (A & Al, 2012).

In 2017, Sepideh Hassankhani Dolatabadi and Farshid Keynia demonstrated the use of predictive analytics for employee churn or staff loss in organizations. They employed various algorithms, including Decision Tree, Naïve Bayes, SVM, and Neural Network. The results showed that SVM achieved the highest accuracy of 99.83% (Dolatabadi & Keynia, 2017).

Machine learning algorithms are useful in predicting job quitting probabilities. Sisodia, Vishwakarm, and Abinash (2017) used K-NN, SVM, Naïve Bayes, Decision Tree, and Random Forest algorithms to analyze accuracy, precision, recall, F-measure, specificity, false positive rate, and false negative rate. Random Forest outperformed all other classifiers, achieving an accuracy of 99%.

In 2019, Zarmina Jaffar, Waheed Noor, and Zartash Kanwal employed data mining techniques to predict employees who were likely to quit or leave an organization. They utilized J48, Naïve Bayes, and Logistic Regression algorithms. Their proposed information selection algorithm, based on conditional equivalence, achieved high efficiency, and utilized fewer characteristics in multiple data sets, resulting in higher classification accuracy. They suggested improving the algorithm by eliminating irrelevant and ineffective machine functions (Jaffar et al., 2019).

In 2021, Mingwei Xu and Cuang Li conducted research on the Data Mining Method of Enterprise Human Resources Management Based on the Simulated Annealing Algorithm. They applied the simulated annealing algorithm, using the Metropolis Algorithm, to generate a sequence of solutions for combinatorial optimization problems, ultimately finding the overall optimal solution (Xu & Li, 2021).

Meanwhile, Sahinbas (2022) studied predicting employee promotion using several machine learning approaches, such as SVM, ANN, and Random Forest. Some factors, including seniority, performance level, competencies, age, awards, training score, and organizational commitment, are taken into account when considering personnel for promotion. Random Forest achieved the highest performance with 98% accuracy, 96% precision, 1.0% recall, and 98% F1-score values using the ROS approach. Implementing this technique allows HR professionals and managers to predict the likelihood of promotion, enabling managers to identify the appropriate parameters for promoting individuals (Sahinbas, 2022).

3. Methodology

The methodology used for this research will be explained in several sub-sections. Section 3.1 will discuss the obtained dataset, including a description of each variable. In section 3.2, data preprocessing will be explained. This is intended to ensure that the model formed from the data yields good results. There are three processes involved in data preprocessing, including removing data points containing null values, vectorizing categorical data, and selecting the most relevant features to form a robust model that provides a high level of accuracy, while also considering other performance metrics such as precision, recall, and F1-score. Additionally, data modelling using various classification algorithms, such as Naive Bayes, Decision Tree, Random Forest, SVM, and KNN, will be explained in section 3.3. The evaluation process of each model will be discussed in section 3.5, which involves all the aforementioned performance metrics.

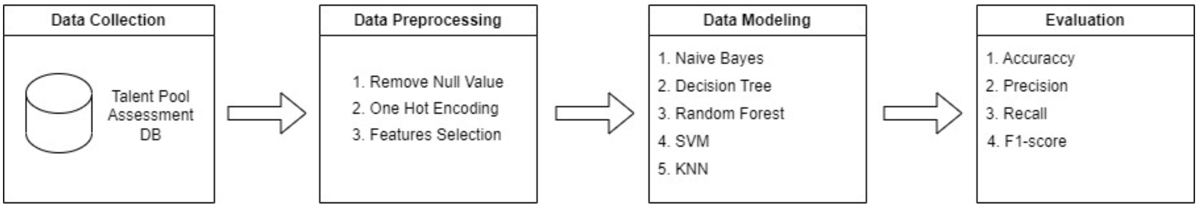


Fig. 1: Proposed methodology

3.1 Data Collection

This study employs the talent pool assessment database of government agencies, which includes 687 candidates who underwent talent pool assessment without any personal information attributes, such as name, identity number, birth date, job title, contact number, and address. The dataset consists of 24 features, including competencies, potentials, job levels, and targets. The assessment criteria and variables used in Table 1 are based on Regulation No. 26 of 2019 of the National Civil Agency of Indonesia and Regulation No. 3/2020 of the Minister of State Apparatus Empowerment and Bureaucratic Reform of the Republic of Indonesia, utilizing the computer-based test method.

Table 1: Dataset description

Category	Variables	Description	Data Type
Competencies	ING	Integrity	Numerical Data
	STW	Team Working	
	COM	Communication	
	RO	Result of Orientation	
	PUS	Public Service	
	SD	Self-Development	
	CM	Change Management	
	DM	Decision Maker	
	NAT	Nationality	
	CAT_TC	Category of Total Competencies	Categorial Data
Potentials	INL	Intellectual	Numerical Data
	INP	Interpersonal	
	SA	Self-Awareness	
	CST	Critical and Strategic Thinking	
	POS	Problem Solving	
	EQ	Emotional Quotient	
	GM	Growth Mindset	
	GR	Grit	
	CAT_TP	Category of Total Potential	Categorial Data
Job Levels	CAT_JLP	Job Level Position when Applied	
	CAT_TLP	Target Level Position	
	CAT_CLP	Current Level Position	

Category	Variables	Description	Data Type
	CAT_RTLP	Category of Result Target Level Position	
Target	CAT_REC	Recommendation	

3.2 Data Preprocessing

Data processing is conducted to ensure that the data to be analysed is clean, which involves removing empty values and renaming columns for further processing (Rajagukguk et al., 2023). The dataset is manually labelled, and the final count is presented in Table 2. For categorical data, vectorization is performed using one-hot encoding approach, allowing it to be trained using multiple classification algorithms.

Table 2: Total number of records of each class

Recommendation	Number of Records
Eligible	52
Fair Eligible	365
Less Eligible	270

To enhance the clarity of the study, the author appended the data samples prior to conducting feature selection and vectorization, as presented in Fig. 2. Having categorial data vectorized, the next process is to select significant attributes. The Chi-Square and ANOVA are employed as to perform attributes selection for categorial and numerical data, respectively.

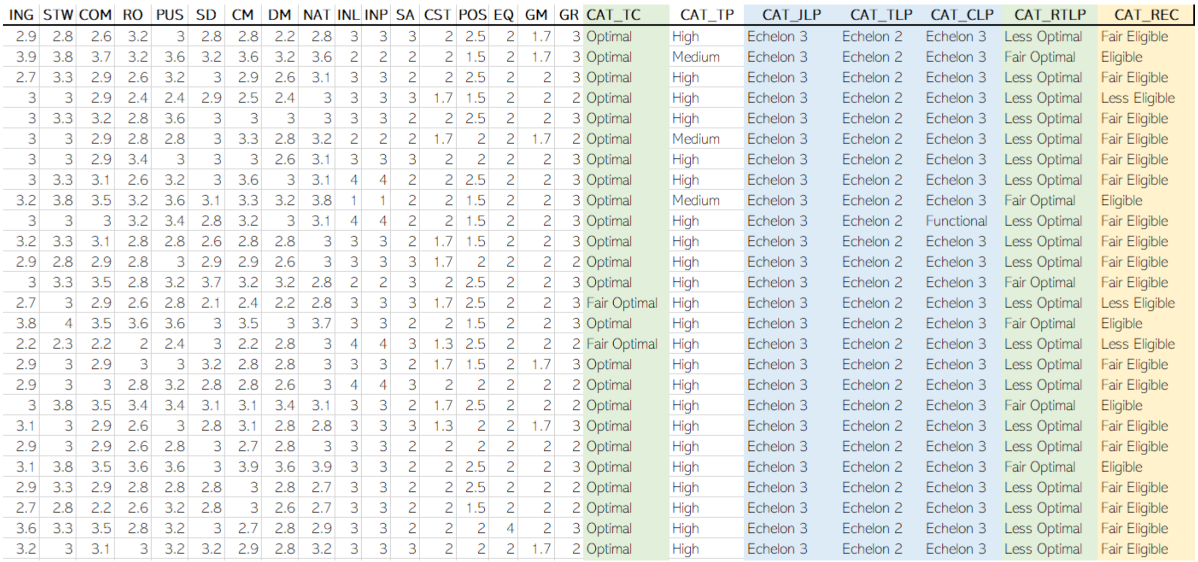
To determine the most relevant categorical features, one can use various statistical methods such as chi-square test or mutual information. These methods assess the relationship between the categorical feature and the target variable. The higher the statistical value, the more relevant the feature is considered to be. In the case of the chi-square test, it measures the independence between two categorical variables. It calculates the difference between the observed frequencies and the expected frequencies under the assumption of independence. The larger the chi-square statistic, the stronger the relationship between the feature and the target variable. By applying chi-square methods to the categorical features in the dataset, the author can obtain a ranking of their relevance.

In the given scenario where the target variable is categorical, ANOVA is the preferred choice. It allows ones to assess if there are significant differences in the mean values of the target variable across different categories or levels of the numerical features. By performing ANOVA, one can determine if the numerical features have a significant impact on the categorical target variable. Features with higher ANOVA scores demonstrate a better ability to separate or differentiate the different categories within the target variable. On the other hand, the author does not take into account Pearson correlation due to it is a measure of the linear relationship between two numerical variables. It quantifies the strength and direction of the linear association between the numerical features and the target. The results of attribute selection for both categorical and numerical data using chi-square and ANOVA, respectively can be seen in the following Table 3. The author prefers to use the attributes that have score number larger than 1 (the green rows on the table). On this research, the author will also show the difference between utilizing all the attributes and the less ones while generating model using several machine learning approaches. This will be discussed in section 4.

Table 3: Features selection using Chi-Square and ANOVA

	Attribute Name	**Score**
Chi-Square	CAT_RTLP_Fair Optimal	393.089
	CAT_TC_Fair Optimal	287.267
	CAT_TC_Optimal	227.784
	CAT_TC_Less Optimal	103.478

ANOVA	CAT_RTLP_Less Optimal	50.808
	CAT_TP_Medium	31.610
	CAT_RTLP_Optimal	24.423
	CAT_TP_High	16.117
	CAT_CLP_Functional	3.582
	CAT_CLP_Staff	0.906
	CAT_CLP_Echelon 4	0.722
	CAT_CLP_Echelon 3	0.516
	CAT_CLP_Echelon 2	0.367
	CAT_JLP_Echelon 3	0.000
	CAT_TLP_Echelon 2	0.000
	CM	395.514
	RO	365.601
	PUS	322.095
	DM	306.034
	COM	305.356
	STW	269.618
	SD	241.383
	ING	208.963
	NAT	155.397
	INL	29.320
	INP	29.320
	GM	4.358
	EQ	3.935
	POS	2.329
	GR	0.750
	CST	0.730
	SA	0.305



ING	STW	COM	RO	PUS	SD	CM	DM	NAT	INL	INP	SA	CST	POS	EQ	GM	GR	CAT_TC	CAT_TP	CAT_JLP	CAT_TLP	CAT_CLP	CAT_RTLP	CAT_REC
2.9	2.8	2.6	3.2	3	2.8	2.8	2.2	2.8	3	3	3	2	2.5	2	1.7	3	Optimal	High	Echelon 3	Echelon 2	Echelon 3	Less Optimal	Fair Eligible
3.9	3.8	3.7	3.2	3.6	3.2	3.6	3.2	3.6	2	2	2	2	1.5	2	1.7	3	Optimal	Medium	Echelon 3	Echelon 2	Echelon 3	Fair Optimal	Eligible
2.7	3.3	2.9	2.6	3.2	3	2.9	2.6	3.1	3	3	2	2	2.5	2	2	3	Optimal	High	Echelon 3	Echelon 2	Echelon 3	Less Optimal	Fair Eligible
3	3	2.9	2.4	2.4	2.9	2.5	2.4	3	3	3	3	1.7	1.5	2	2	2	Optimal	High	Echelon 3	Echelon 2	Echelon 3	Less Optimal	Less Eligible
3	3.3	3.2	2.8	3.6	3	3	3	3	3	3	2	2	2.5	2	2	3	Optimal	High	Echelon 3	Echelon 2	Echelon 3	Less Optimal	Fair Eligible
3	3	2.9	2.8	2.8	3	3.3	2.8	3.2	2	2	2	1.7	2	2	1.7	2	Optimal	Medium	Echelon 3	Echelon 2	Echelon 3	Less Optimal	Fair Eligible
3	3	2.9	3.4	3	3	3	2.6	3.1	3	3	3	2	2	2	2	3	Optimal	High	Echelon 3	Echelon 2	Echelon 3	Less Optimal	Fair Eligible
3	3.3	3.1	2.6	3.2	3	3.6	3	3.1	4	4	2	2	2.5	2	2	3	Optimal	High	Echelon 3	Echelon 2	Echelon 3	Less Optimal	Fair Eligible
3.2	3.8	3.5	3.2	3.6	3.1	3.3	3.2	3.8	1	1	2	2	1.5	2	2	3	Optimal	Medium	Echelon 3	Echelon 2	Echelon 3	Fair Optimal	Eligible
3	3	3	3.2	3.4	2.8	3.2	3	3.1	4	4	2	2	1.5	2	2	3	Optimal	High	Echelon 3	Echelon 2	Functional	Less Optimal	Fair Eligible
3.2	3.3	3.1	2.8	2.8	2.6	2.8	2.8	3	3	3	2	1.7	1.5	2	2	3	Optimal	High	Echelon 3	Echelon 2	Echelon 3	Less Optimal	Fair Eligible
2.9	2.8	2.9	2.8	3	2.9	2.9	2.6	3	3	3	3	1.7	2	2	2	2	Optimal	High	Echelon 3	Echelon 2	Echelon 3	Less Optimal	Fair Eligible
3	3.3	3.5	2.8	3.2	3.7	3.2	3.2	2.8	2	2	3	2	2.5	2	2	3	Optimal	High	Echelon 3	Echelon 2	Echelon 3	Fair Optimal	Fair Eligible
2.7	3	2.9	2.6	2.8	2.1	2.4	2.2	2.8	3	3	3	1.7	2.5	2	2	3	Fair Optimal	High	Echelon 3	Echelon 2	Echelon 3	Less Optimal	Less Eligible
3.8	4	3.5	3.6	3.6	3	3.5	3	3.7	3	3	2	2	1.5	2	2	3	Optimal	High	Echelon 3	Echelon 2	Echelon 3	Fair Optimal	Eligible
2.2	2.3	2.2	2	2.4	3	2.2	2.8	3	4	4	3	1.3	2.5	2	2	2	Fair Optimal	High	Echelon 3	Echelon 2	Echelon 3	Less Optimal	Less Eligible
2.9	3	2.9	3	3	3.2	2.8	2.8	3	3	3	2	1.7	1.5	2	1.7	3	Optimal	High	Echelon 3	Echelon 2	Echelon 3	Less Optimal	Fair Eligible
2.9	3	3	2.8	3.2	2.8	2.8	2.6	3	4	4	3	2	2	2	2	2	Optimal	High	Echelon 3	Echelon 2	Echelon 3	Less Optimal	Fair Eligible
3	3.8	3.5	3.4	3.4	3.1	3.1	3.4	3.1	3	3	2	1.7	2.5	2	2	2	Optimal	High	Echelon 3	Echelon 2	Echelon 3	Fair Optimal	Eligible
3.1	3	2.9	2.6	3	2.8	3.1	2.8	2.8	3	3	3	1.3	2	2	1.7	3	Optimal	High	Echelon 3	Echelon 2	Echelon 3	Less Optimal	Fair Eligible
2.9	3	2.9	2.6	2.8	3	2.7	2.8	3	3	3	2	2	2	2	2	2	Optimal	High	Echelon 3	Echelon 2	Echelon 3	Less Optimal	Fair Eligible
3.1	3.8	3.5	3.6	3.6	3	3.9	3.6	3.9	3	3	2	2	2.5	2	2	3	Optimal	High	Echelon 3	Echelon 2	Echelon 3	Fair Optimal	Eligible
2.9	3.3	2.9	2.8	2.8	2.8	3	2.8	2.7	3	3	2	2	2.5	2	2	2	Optimal	High	Echelon 3	Echelon 2	Echelon 3	Less Optimal	Fair Eligible
2.7	2.8	2.2	2.6	3.2	2.8	3	2.6	2.7	3	3	2	2	1.5	2	2	2	Optimal	High	Echelon 3	Echelon 2	Echelon 3	Less Optimal	Fair Eligible
3.6	3.3	3.5	2.8	3.2	3	2.7	2.8	2.9	3	3	2	2	2	4	2	3	Optimal	High	Echelon 3	Echelon 2	Echelon 3	Less Optimal	Fair Eligible
3.2	3	3.1	3	3.2	3.2	2.9	2.8	3.2	3	3	3	2	2	2	1.7	2	Optimal	High	Echelon 3	Echelon 2	Echelon 3	Less Optimal	Fair Eligible

Fig. 2: Data samples

3.3 Data Modelling

During the data modelling phase, the dataset was divided into two distinct groups: 20% for testing purposes and 80% for training the model. This allocation ensures a significant amount of data for the model to learn from, while also allowing for a reliable evaluation of its performance. The larger the amount of data available for training, the higher the potential for accurate predictions on unseen data. It is crucial to avoid using an insufficiently small training set, as it may hinder the learning capabilities of the machine learning algorithm (Nasir & Palanichamy, 2022).

In this study, the author opted for the linear as the SVM kernel as hyper parameter tuning shown that this kernel is best. Support Vector Machines (SVM) typically involve two hyperparameters: the C parameter, which governs the decision boundary margin, and the gamma parameter, which determines the impact of support vectors on the positioning of hyperplanes. As for Naïve Bayes, the author considers 'var_smoothing' as the parameter. The 'var_smoothing' is an additional hyperparameter present in the GaussianNB model. This hyperparameter is used to address potential numerical issues that may arise when estimating the variance of features that are extremely small or zero. By default, the value of 'var_smoothing' is set to 1e-9, which means a very small value will be added to the estimated feature variances. This helps maintain numerical stability and prevents division by zero when calculating the log-likelihood.

Meanwhile, performing classification using Decision Tree, some hyper-parameter that should be considered are criterion, max_depth, min_samples_split and min_samples_leaf. The criterion hyperparameter defines the quality measure used to evaluate the quality of a split at each node in the decision tree. It determines how the algorithm measures the impurity or the homogeneity of the target variable within a node. Max_depth limits the complexity and size of the tree, preventing it from overfitting the training data.

Tuning the hyperparameters n_neighbors, weights, and p in K-NN is crucial to find the optimal configuration that balances the model's ability to generalize and capture local patterns in the data. It involves finding the right number of neighbours, deciding the appropriate weighting scheme, and selecting the suitable distance metric for the specific problem at hand.

To fine-tune those hyperparameters above, grid search method is employed, and the findings will be discussed in section 4.

3.4 Model Evaluation

During the model evaluation phase, the dataset was divided into two distinct sections: the training data and the testing data, utilizing the hold-out validation method. Subsequently, both the training and testing data were prepared to assess the analysis outcomes based on their accuracy, precision, recall, and F1-score measurements.

4. Results and Findings

After going through the phase of removing data points with null values and vectorizing categorical data, a feature selection process was conducted for the existing dataset, and the results can be seen in Table 6. These selected features will be used for data modelling using various machine learning algorithms.

Table 4: Selected features for modelling data

BEFORE FEATURES SELECTION	AFTER FEATURES SELECTION
'CM', 'RO', 'PUS', 'DM', 'COM', 'STW', 'SD', 'ING', 'NAT', 'INL', 'INP', 'GM', 'EQ', 'POS', 'CAT_RTLP_Fair Optimal', 'CAT_TC_Fair Optimal', 'CAT_TC_Optimal', 'CAT_TC_Less Optimal', 'CAT_RTLP_Less Optimal', 'CAT_TP_Medium', 'CAT_RTLP_Optimal', 'CAT_TP_High', 'CAT_CLP_Functional'	'ING', 'STW', 'COM', 'RO', 'PUS', 'SD', 'CM', 'DM', 'NAT', 'CAT_TC', 'INL', 'INP', 'SA', 'CST', 'POS', 'EQ', 'GM', 'GR', 'CAT_TP', 'CAT_JLP', 'CAT_TLP', 'CAT_CLP', 'CAT_RTLP'

During the process of machine learning modelling, the author conducted hyperparameter tuning for each proposed algorithm using the GridSearch technique. This step is of utmost importance as hyperparameters are parameters of a machine learning model that are predefined and not learned from the data. They play a crucial role in determining the behaviour and performance of the model. Below are the best hyperparameter configurations for each algorithm (see Table. 5).

Table 5: Hyperparameter configurations for each machine learning algorithm

	Hyperparameter	Value
Naïve Bayes	var_smoothing	1e-09
Decision Tree	Criterion	entropy
	max_depth	5
	min_samples_leaf	2
	min_samples_split	2
Random Forest	max_depth	None
	min_samples_leaf	1
	min_samples_split	5
	n_estimators	100
SVM	C	1
	Gamma	0.1
	Kernel	linear
K-NN	n_neighbors	5
	P	1
	Weights	distance

By utilizing the aforementioned hyperparameter configurations, each trained model can be evaluated using multiple performance metrics. The results can be observed in Table 6. There is no significant difference among the models, as evidenced by several performance metrics showing the same values.

Table 6: The evaluation’s output for each machine learning algorithm

Algorithm	Accuracy	Precision	Recall	F1-Score

NB	0.942	0.953	0.942	0.945
DT	0.942	0.942	0.942	0.939
RF	0.942	0.942	0.942	0.940
SVM	0.942	0.942	0.942	0.940
K-NN	0.942	0.942	0.942	0.940

The author also conducted experiments on this dataset prior to the feature selection process. By utilizing the aforementioned best hyperparameter configuration, the evaluation results for each machine learning model can be observed in the following Table 7. All metrics indicate a value of 1.0, without exception.

Table 7: The evaluation's output for each ML algorithm without features selection

Algorithm	Accuracy	Precision	Recall	F1-Score
NB	1.0	1.0	1.0	1.0
DT	1.0	1.0	1.0	1.0
RF	1.0	1.0	1.0	1.0
SVM	1.0	1.0	1.0	1.0
K-NN	1.0	1.0	1.0	1.0

If the accuracy of a machine learning model is 1 (or 100%), it means that the model is predicting the correct class label for all instances in the dataset. When achieving a perfect accuracy score, it is typically a sign of some error, such as overfitting.

5. Conclusion

Promotions have a positive, significant, and advantageous impact on employee work performance within the human resources process. This study introduces prediction models for employee promotions utilizing various machine learning algorithms. The dataset is initially processed to produce valid outputs, such as features selection. The author use Chi-Square and ANOVA to select which attributes have significant influence on the accuracy. Similarly, the process of building machine learning models also involves hyperparameter tuning, resulting in robust models. The outputs show that all the machine learning algorithms proposed have score larger than 0.94 in all performance metrics. On the contrary, while the dataset is not being pre-processed, it poses an overfitting condition.

In terms of the talent pool assessment based on the algorithm above, there is a connected and engagement between the competency and the recommendation from assessment centre. By using features selection, it can be concluded that all the variables related to competencies exhibit a substantial influence on predicting the high leadership promotion of civil servants in Indonesia based on talent pool assessments. Conversely, variables such as Grit, Critical and Strategic Thinking, and Self-Awareness demonstrate a comparatively lower impact. For the future work, ones should consider applying SMOTE and ROS as to address imbalanced dataset.

References

A, Q., & Al, E. (2012). Using Data Mining Techniques to Build a Classification Model for Predicting Employees Performance. *International Journal of Advanced Computer Science and Applications*, *3*(2). https://doi.org/10.14569/ijacsa.2012.030225

Adhiselvam, A. (2021). Feature Selection Based Enhancement Of The Accuracy Of Classificationalgorithms. *Turkish Journal of Computer and Mathematics …*, *12*(10), 5621–5628. https://turcomat.org/index.php/turkbilmat/article/view/5373%0Ahttps://turcomat.org/index.php/turkbi lmat/article/download/5373/4480

Cappelli, P., & Keller, J. R. (2014). Talent Management: Conceptual Approaches and Practical Challenges. *Annual Review of Organizational Psychology and Organizational Behavior*, *1*(January), 305–331. https://doi.org/10.1146/annurev-orgpsych-031413-091314

Chauhan, R., & Kaur, H. (2013). Predictive analytics and data mining: A framework for optimizing decisions with R tool. In *Advances in Secure Computing, Internet Services, and Applications* (2015th ed.). https://doi.org/10.4018/978-1-4666-4940-8.ch004

Colleen McCue Ph.D. (2007). *Data Mining and Predictive Analysis* (Vol. 1, Issue 2). Butterworth-Heinemann, Elsevier Inc.

Dahiya, A., Gautam, N., & Gautam, P. K. (2021). Data mining methods and techniques for online customer review analysis: A literature review. *Journal of System and Management Sciences*, *11*(3), 1–26. https://doi.org/10.33168/JSMS.2021.0301

Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand ¨ Thirion, Olivier Grisel, M. B., Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, M. B., & Duchesnay, M. P. and E. (2011). *Scikit-learn: Machine Learning in Python*. *12*, 128–154. https://doi.org/10.4018/978-1-5225-9902-9.ch008

Jaffar, Z., Noor, W., & Kanwal, Z. (2019). Predictive Human Resource Analytics Using Data Mining Classification Techniques. *International Journal of Computer*, *32*(1), 9–20.

Jooss, S., Burbach, R., & Ruël, H. (2021). Examining talent pools as a core talent management practice in multinational corporations. *International Journal of Human Resource Management*, *32*(11), 2321–2352.https://doi.org/10.1080/09585192.2019.1579748

Kramer, O. (2013). Dimensionality Reduction with Unsupervised Nearest Neighbors. In *Intelligent Systems Reference Library* (Vol. 51). https://doi.org/10.1007/978-3-642-38652-7

Muhid, H.K. (2022, 08 15). Retrieved from https://nasional.tempo.co/read/1622996/lagi-lagi-kepala-daerah-tersandung-kasus-suap-jual-beli-jabatan-apa-alasannya

Nasir, A. A. B. M., & Palanichamy, N. (2022). Sentiment Analysis of Covid-19 Tweets by Supervised Machine Learning Models. *Journal of System and Management Sciences*, *12*(6), 50–69. https://doi.org/10.33168/JSMS.2022.0604

Potdar, K., S., T., & D., C. (2017). A Comparative Study of Categorical Variable Encoding Techniques for Neural Network Classifiers. *International Journal of Computer Applications*, *175*(4), 7–9. https://doi.org/10.5120/ijca2017915495

Santra, a. K., & Christy, C. J. (2012). Genetic Algorithm and Confusion Matrix for Document Clustering. *International Journal of Computer Science*, *9*(1), 322–328. http://ijcsi.org/papers/IJCSI-9-1-2-322-328.pdf

Sahinbas, K. (2022, April 15). Employee Promotion Prediction by using Machine Learning Algorithms for Imbalanced Dataset. 2022 2nd International Conference on Computing and Machine Intelligence (ICMI). https://doi.org/10.1109/icmi55296.2022.9873744

Sharda, R., Delen, D., & Turban, E. (2018). Business Intelligence, Analytics, and Data Science: A Managerial Perspective. In *Winning with Data*.

Suthaharan, S. (2003). Machine Learning Models and Algorithms for Big Data Classification - Thinking with Examples for Effective Learning. In *16th AIAA Computational Fluid Dynamics Conference*. https://doi.org/10.2514/6.2003-4110

Vercellis, C. (2009). Business Intelligence: Data Mining and Optimization for Decision Making. In *Business Intelligence: Data Mining and Optimization for Decision Making*. https://doi.org/10.1002/9780470753866

Wang, Z., Yan, R., & Chen, Q. (2010). *Data Mining in Nonprofit Organizations , Government Agencies , and Other Institutions*. *July*. https://doi.org/10.4018/jisss.2010070104

Xin-She Yang. (2019). Introduction to Algorithms for Data Mining and Machine Learning. In منشورات جامعة دمشق (Vol. 1999, Issue December).

Xu, M., & Li, C. (2021). Data Mining Method of Enterprise Human Resource Management Based on Simulated Annealing Algorithm. *Security and Communication Networks*, *2021*. https://doi.org/10.1155/2021/6342970

Yi, L., & Yao, W.-X. (2017). *Application of the Data Mining in the University Human Resource Management*. *91*(Edmi), 222–227. https://doi.org/10.2991/icwcsn-16.2017.111