Malaysian Journal of Computing, 8 (1): 1264-1286, 2023
Copyright © UiTM Press
eISSN: 2600-8238

PREDICTION OF EMPLOYEE PROMOTION USING HYBRID SAMPLING METHOD WITH MACHINE LEARNING ARCHITECTURE

Shahidan bin Shafie[1], Soek Peng Ooi[2], Khai Wah Khaw[3*]

[1,2,3]School of Management*
Universiti Sains Malaysia, 11800 Minden, Pulau Pinang.
[1]shahidan@usm.my, [2]soekpeng915@student.usm.my, [3*]khaiwah@usm.my

ABSTRACT

Employee promotion plays an important role in an organization. It aids to inspire employees to grow and develop their skills, thus increase employee loyalty and reduce the turnover rate. This study predicts employee job promotion based on employee promotion data by using a hybrid sampling method with machine learning. The purpose of this study is to accelerate the promotion process and share the important features that might be determined when promoting an employee. In this study, there are eight machine learning algorithms have been used, such as Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbors, Support Vector Machine, Naïve Bayes, Adaptive Boosting Classifier, and Extreme Gradient Boost. The purpose of using eight machine learning algorithms is to find out the most suitable model to predict employee promotion. Additionally, hybrid sampling methods like Synthetic Minority Oversampling Technique combined with Edited Nearest Neighbor (SMOTE+ENN) and Synthetic Minority Oversampling Technique combined with Tomek Link (SMOTE+Tomek) were adopted. These two techniques are to cure the imbalanced dataset. For the importance of feature selection, the Recursive Feature Elimination method with Random Forest Classifier model (RFE-RFC), Explained Variance Ratio method with Principal Component Analysis (EVR-PCA), and the Rank Feature Importance method with Extra Classifier Tree model (RFI-ECT) is applied. The first 5, 8, and 12 features are selected based on the RFI-ECT to train the machine learning algorithms. As a result, the model is evaluated by precision, recall, and F1-score. In conclusion, the top five rank feature importance methods with the Extra Classifier Tree model are region, department, previous year rating, KPIs met and above 80%, and award won. The results suggest that SMOTE+ENN and Extreme Gradient Boost with eight features have the highest-performing model in this study.

Keywords*: Employee Promotion Prediction, Hybrid Sampling, Imbalanced Data*

Machine Learning,

Received for review: 23-11-2022; Accepted: 23-03-2023; Published: 10-04-2023
DOI: 10.24191/mjoc.v8i1.18456

1. Introduction

Employee promotion means the ascension of an employee to higher ranks like a salary increase, higher status, more benefits will receive, and job responsibilities will become heavy. Employees are most motivated by this duty because it is the greatest honour for their loyalty and dedication to the organization (Jyoti, 2022). Promoting an employee will take a long time and need to collect data or feedback and analyse the data. It will increase the workload of the human resource (HR) team.

Human Resource Analytics (HRA) is a variety of tools, technologies, and methods for acquiring, saving, retrieving, and interpreting data to assist business users in making better choices to reduce the HR team's workload (Bandi *et al.*, 2021; Jain & Bhushan, 2020). For example, the HR team used HRA to estimate the requirement of human resources in recruitment, training, development, retention, promotion, transfer, performance appraisal, retirement and others. (Kakulapati *et al.*, 2020). The goal is to increase the quality of people-related decisions so that individuals and organizations can perform better (Jomthanachai *et al.*, 2022).

Research works on machine learning have received great attention in the past decade such as Aimran *et al.* (2022), Pisal *et al.* (2022), Malik *et al.* (2022), to name a few. Machine learning is training and testing past data and envisioning a future outcome. Machine learning is categorized into two types, (1) supervised learning – dealing with classification data; and (2) unsupervised learning – classifying the cluster data (Punnoose & Ajit, 2016). It is an intelligent algorithm that helps tackle some problems by increasing efficiency, decreasing the cost and workload of data analysis; increasing effectiveness by enhancing data quality when deciding the future (Garg *et al.*, 2021). For example, International Business Machines Corporation (IBM) fills the job gaps by using intelligent algorithms to tailor applicants suitable for particular positions (Castellanous, 2019). Furthermore, Club Med custom-made rewards for each of the employees by drilling into data and analytics to identify employee contribution and performance in the workplace (Bolton *et al.*, 2019)

Promotion studies can help workers obtain development opportunities and assist enterprises in selecting and retaining talents. Human Resource Management (HRM) has always been a research hotspot. However, formal studies rely on gathered data through questionnaires and interviews, there will be limitations in sample size, and subjective considerations are likely to impact the outcome (Aleem & Bowra, 2020; Hetland *et al.*, 2018). In the era of big data, machine learning has progressively become more prevalent in human resource management. Despite some accomplishments in utilizing big data analytic tools in HRM, relatively little research has applied machine learning to promotion attributes, and further exploration of employees' promotion is necessary (Garg *et al.*, 2021; Zhu, 2021).

In a previous study, a professional development dataset has been used to predict the employees' promotion. The features in this dataset include department, region, education, gender, recruitment channel, no. of training, age, previous year rating, length of service, KPIs, awards, and average training score. Thus, the selected machine learning techniques applied in the study are Decision Tree, Random Forest, and Support Vector Machine. The model is validated by evaluating it with three distinct training and testing groups, and there are 80:20, 70:30, and 60:40 ratios. In addition, SMOTE is adopted to deal with data imbalances. In conclusion, validation of Random Forest with SMOTE in 80:20 proportions achieve good accuracy (96.32%) (Keawwiset *et al.*, 2021).

Another research work is conducted to identify the employees most likely to get promoted by using 38,312 samples from the training dataset to train the model and 16,496 samples from the test dataset to test the model. The features used to predict employees' promotion include division, foreign school, geographical zone, working experiences, and education. The researchers replaced the missing values with mode values for each feature and used the resampling technique to solve the imbalanced response feature. Thus, Gradient Boost (GB), Random Forest (RF), Catboost, and Extreme Gradient Boosting (XGBoost) are the machine learning techniques used in this prediction. In a nutshell, Catboost and XGBoost scored the highest (93%), followed by RF (88%) and GB (84%) (Ibrahim *et al.*, 2020).

The previous study showed the prediction of employees' promotion based on the geographical position (area, administrative division, particular region) and structural position (department, level, size of company). This dataset contains 17,704 samples and is split into training and testing models. Thus, three supervised machine learning algorithms like Random Forest, Logistic Regression, and AdaBoost, are adopted. In split training data and test data, cross-validation (cv=5) is used to prevent random factors. Besides, the synthetic minority over-sampling technique (SMOTE) is applied to deal with the imbalanced dataset. After that, to determine the best classifier, the researchers used a grid search to adjust hyper-parameters. In the paper, the champion model is the Random Forests Classifier, which returns the accuracy and AUC at 85.6%, recall at 88.9%, and precision at 83.4% (Liu *et al.*, 2019).

In another study, the prediction of 77,218 employee promotion information in Chinese state-owned enterprises based on personal primary information data (birth date, gender, degree of education, hometown, and nationality) and position information data (work department, department level, position type and level, personnel nature, start time of current position). The features are built based on unique values, mode, highest or lowest value, count the number of different values, and calculate the difference between two dates. It uses min-max normalization to deal with numerical features and One-Hot encoding to transform discrete elements, dividing the dataset into a train and test set with 80:20 ratios. Thus, six classification algorithms such as Support Vector Machine (SVM), Random Forest (RF), K-Nearest Neighbor, Decision Tree, Logistic Regression (LR), and AdaBoost are used. The performance in the cross-validation accuracy score, RF (0.902), is the best, followed by SVM (0.894), LR (0.892), and Adaboost (0.892). The working years, positions and department level were ranked as the top three features in this research (Long *et al.*, 2018).

Another study is to solve the process of employee promotion based on ability (dependability, job knowledge, interpersonal relationship, performance under pressure, creativity) and loyalty (attendance, quantity of work, accuracy, courtesy, and housekeeping). In this work, the scoring rules are sorted from excellent (5) to the worst (1) and analyzed by using the Mamdani Fuzzy method. In conclusion, it had an accuracy of 91.4% in testing data with an average speed of 1.24 seconds (Zulfikar *et al.*, 2018).

The regression analysis is used to describe the variance of an independent variable over the dependent variable, such as predicting employee happiness, employee loyalty, and employee performance (Saxena *et al.*, 2021). The company uses a performance rating scale to measure the quality of an employee. There are nineteen features used to calculate an employee, such as intelligence, ability, responsibility, morals, etc. According to the forecast, if an employee's performance is exceptional, they will be promoted. However, if an employee's performance is good, medium, or poor, they will need to acquire some training and identify their weak areas to improve (Sarker *et al.*, 2018).

This study aims to use machine learning to accelerate the whole promotion process. Specifically, this study focuses on predicting employee promotion using hybrid sampling methods with machine learning. This study also captures the important and relevant features which affect employees getting a promotion. Three main objectives need to be addressed in this study, which is (1) to identify machine learning algorithms that are suitable for users to predict employees' promotion, (2) to identify performance metrics suitable for evaluating the models' performance and lastly, (3) to compare the models' performance and determine the champion model

Employees' promotion prediction can be marked as a binary classification because it classifies data points into one of two classes: "Promoted" or "Non-Promoted." Since there is no one of the best algorithms can be applied to any dataset, a set of algorithms will be experimented with to evaluate the most suitable algorithm for the dataset in this study. For this study, eight supervised machine learning algorithms were tested with: (1) Logistic Regression Classifier (LRC); (2) Decision Tree Classifier (DTC); (3) Random Forest

Classifier (RFC); (4) K-Nearest Neighbors (KNN); (5) Support Vector Machine (SVM); (6) Naïve Bayes (GNB); (7) Adaptive Boosting Classifier (ABC); (8) Extreme Gradient Boost (XGB). For the details of supervised machine learning algorithms will explained at section 2 Methodology.

1.1 Hybridization Sampling

An imbalanced dataset refers to one of the classes in a binary category that is lower than another one (Lin *et al.*, 2021). This issue is of utmost importance since it affects numerous fields with considerable environmental, vital, or commercial significance and has been demonstrated in some instances to significantly impede the performance achievable by conventional learning methods (Malik *et al.*, 2022).

A standard classification algorithm frequently misclassifies the patterns of the minority class when used directly to imbalanced data because of their bias towards the dominant class. It will assume that all classes will experience equal misclassification costs to another problem with typical classification algorithms; nonetheless, minority classes are frequently linked with greater misclassification costs. Misclassification of minority patterns in this situation could have devastating effects. Because of this, the problem of class imbalance must be properly addressed when constructing efficient categorization systems. There are divided into three categories: (1) data-level solutions, (2) algorithmic solutions, and (3) ensemble learning-based solutions to overcome imbalanced datasets. Amongst these three solutions, data-level solutions are the most popular solution that will be applied. Due to it being user-friendly, easy to understand, viability, and great proficiency. Under-sampling and over-sampling are the two prevalent methods, and both are efficient in various problem circumstances (Devi *et al.*, 2020).

For datasets with a smaller ratio of class imbalance, the under-sampling methods are good to consider. The under-sampling method is to redistribute the training dataset and weights by fewer majority instances. This covers a variety of techniques, including random and non-random under-sampling. However, employing this method can result in the loss of some important data. Additionally, data under-sampling produces a within–class distribution if the majority class is made up of several classes.

High-class imbalance scenarios can be successfully handled by oversampling approaches. The oversampling method is creating synthetic instances and adding them to the minority class, these strategies aim to redistribute the training data. In order to achieve the appropriate class ratios, various techniques are used, such as reproducing the minority class and adding some produced synthetic samples. These methods have the advantage over under-sampling methods in that all training instances are preserved. However, the over-sampling strategy drowns the retrieved real-world cases in synthetic ones and distorts the classification results when there is a substantial imbalance ratio between classes.

The basic concept for hybrid sampling is to remove some samples from the majority class (negative examples) and progressively replace them with new positive examples. A hybrid sampling includes two parts: an under-sampling method and an over-sampling method. In this study, we present a hybrid sampling strategy that under-sampling the majority class using the Edited Nearest Neighbor (ENN) and Tomek Link and over-sampling the minority instance using the Synthetic Minority Oversampling Technique (SMOTE) (Gazzah et al, 2015)

The Edited Nearest Neighbor (ENN) and Tomek Link are under-sampling methods. The Edited Nearest Neighbor (ENN) can reduce the majority class by applying the KNN approach and detecting and deleting noisy examples (Guan *et al.*, 2021; Jeon & Lim, 2020). Tomek Link is to discover all instances in the majority class and helps the classifier make better borderline judgments (Sawangarreerak & Thanathamathee, 2020).

Furthermore, the over-sampling method - the Synthetic Minority Oversampling Technique (SMOTE) generates some positive class samples and mollifies the imbalance of network traffic data. According to Xu *et al.*'s (2020) empirical results, SMOTE is a perfect match for ENN. As a result, imbalanced data can be transformed into balanced data, and skewed distribution also can be corrected (Jiang *et al.*, 2020; Lu *et al.*, 2016).

2. Methodology

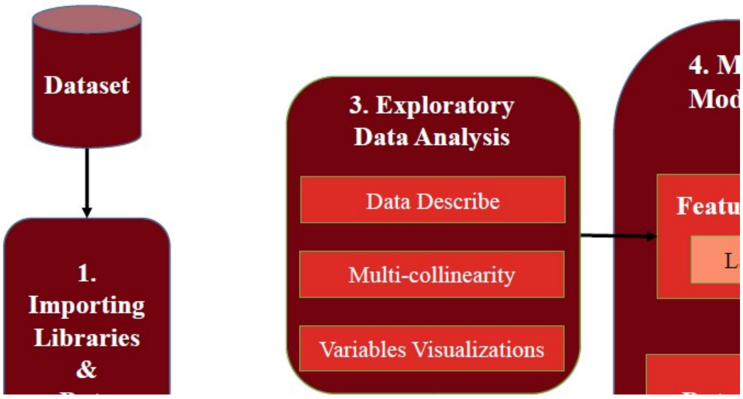


Figure 1. Methodological Framework.

2.1 Data Retrieval

"HR Analytics: Employee Promotion Data" was obtained from Kaggle.com. A multinational company's Data Scientist creates the dataset. The dataset contains 54,808 instances with 14 attributes. In addition, this dataset includes numeric and categorical data types in columns. The study aims to identify which machine learning technique can produce high accuracy by using different hybridization sampling methods and different features. The dataset consists of 14 attributes; 13 are input attributes, and 1 is a target attribute referred to as "is_promoted" with binary labels (0=No, 1=Yes). The description of features is shown in Table 1.

Table 1. Attributes Description in the Employee Promotion Dataset.

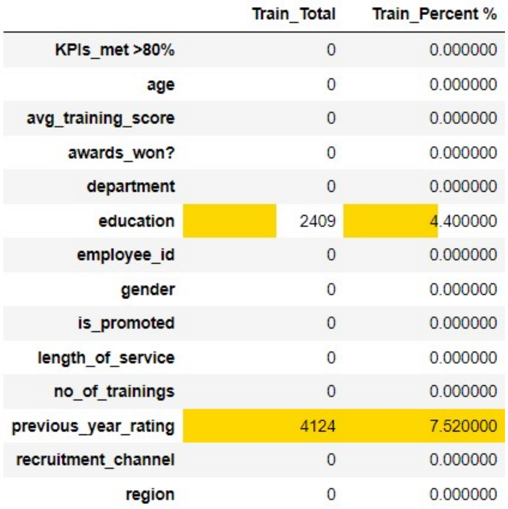
	Attribute	Description
1	employee_id	Unique ID for the employee
2	department	Department of employee
3	Region	Region of employment (unordered)
4	Education	Education level
5	Gender	Gender of employee
6	recruitment_channel	Channel of recruitment for employee
7	no_of_trainings	no training was completed in the previous year on soft skills, technical skills, etc.
8	Age	Age of employee
9	previous_year_rating	Employee rating for the previous year
10	length_of_service	Length of service in years
11	KPIs_met >80%	if % of KPIs (key performance indicators) > 80% then 1 else 0
12	awards_won?	if awards were won during the previous year, then 1 else 0
13	avg_training_score	The average score in current training evaluations
14	is_promoted (target)	Recommended for promotion

2.2 Data Cleaning

Data cleaning includes checking missing values, duplicating data, inspecting useless features, and converting object variables into categorical variables.

2.2.1 Check Missing Value

Missing values have a big impact on how well a classification model works (Zahin *et al.*, 2018). So, checking missing values is an important process before analyzing the dataset. The function to check the missing value is isnull().sum(). Figure 2 shows that the "education" features and "previous_year_rating" features have missing values. To solve the present missing values, replace them with mode values based on each feature.



	Train_Total	Train_Percent %
KPIs_met >80%	0	0.000000
age	0	0.000000
avg_training_score	0	0.000000
awards_won?	0	0.000000
department	0	0.000000
education	2409	4.400000
employee_id	0	0.000000
gender	0	0.000000
is_promoted	0	0.000000
length_of_service	0	0.000000
no_of_trainings	0	0.000000
previous_year_rating	4124	7.520000
recruitment_channel	0	0.000000
region	0	0.000000

Figure 2. Missing Values Count and Percentage.

2.2.2 Check Duplicate Data

Besides that, use the function duplicated().sum() to check duplicate data. In this dataset, there is no duplicated data.

2.2.3 Inspect Useless Features

Next is to inspect the useless features and drop them. For example, this dataset found that "employee_id" has only one unique value for each observation and did not impact or change anything in the dataset. So, the function drop() from Pandas was used to remove this column.

2.2.4 Convert Object Variables to Categorical Variables

The dataset has five object types which are "department", "region", "education", "gender", and "recruitment_channel". The object type in the dataset has been changed to the category type to have greater memory use and become faster. The first memory usage was 5.9+MB, and after changing the data types, the memory usage only left 3.6MB.

2.2.5 Exploratory Data Analysis (EDA)

The purpose of exploratory data analysis (EDA) is to find out why people get promoted and what factors they are commonly looking for as potentially significant to get a promotion. Graphical techniques were used to accomplish this. It allowed researchers to investigate the elements that influence employees to determine whether or not the factors are the primary causes of promotion. In this research, we have applied descriptive statistics to calculate, describe, and summarize datasets. Besides, a correlation matrix has been applied to find out the correlation between each variable. The results of the analysis will be shown in section 3. Results and Discussions.

2.4 Machine Learning Modeling

Machine learning modelling includes feature engineering, data preprocessing, data modelling, and model evaluation.

2.4.1 Feature Engineering

2.4.1.1 Label Encoder

The scikit-learn package for machine learning algorithms only allows numerical input and output. The columns of "department", "region", "education", "gender", "recruitment_channel", "no_of_trainings", "length_of_service", "avg_training_score", and "age" contain textual or categorical data. Therefore, these textual or categorical data must be encoded as integer values before the training and testing of the machine learning models. The LabelEncoder package from the Scikit-learn preprocessing library encoded the categorical features. All the categorical values were transformed to a value between 0 and n-1, with n being the total number of classes.

2.4.2 Data Preprocessing

2.4.2.1 Split Data

Before building the machine learning model, the dataset needs to be split into train and test sets using the train_test_split() function. The entire employee promotion dataset is separated based on the rule of thumb, the 70:30 ratio (train: test) (Ayoubi *et al.*, 2018). 70% of train data will be used to fit the machine learning model, and 30% of test data will be used to evaluate the models.

2.4.2.2 Feature Scaling

The standardscaler() function from the scikit-learn preprocessing package is applied to scaling X_train and X_test in the feature scaling section. The purpose of using standardscaler() is to standardize features by eliminating the mean and scaling to unit variance (Pedregosa *et al.*, 2011).

2.4.2.3 Handling Imbalanced Data

In this dataset, the target variable is imbalanced. 91.48% of employees have "No" and only 8.52% of employees have "Yes." To deal with such imbalanced data, hybrid sampling SMOTE+ENN and SMOTE+Tomek were used in the dataset.

2.4.2.4 Feature Importance

Feature selection methods with libraries and parameters used are library "pandas" and "seaborn." We use the Recursive Feature Elimination method with Random Forest Classifier model (RFE-RFC), Explained Variance Ratio method with Principal Component Analysis (EVR-PCA), and the Rank Feature Importance method with Extra Classifier Tree model (RFI-ECT) to determine the feature importance.

2.5 Data Modeling

The machine learning algorithms used in this study are LogisticRegressionClassifier (LRC), DecisionTreeClassifier (DTC), RandomForestClassifier (RFC), KNeighborsClassifier (KNN), SVC (SVM), GaussianNaïveBayes (GNB), AdaBoostClassifier (ABC) and XGBClassifier (XGB). The training data is fed into each machine learning model and created as a model fit.

2.5.1 Logistic Regression Classifier (LRC)

Cox (1958) proposed Logistic Regression as a typical classification approach employing linear discriminants. It is an essential tool for modelling because the response variable logistic regression is a powerful modelling technique. Moreover, it is used to forecast the analysis of a project (Jaffar *et al.*, 2019).

2.5.2 Decision Tree Classifier (DTC)

Morgan and Sonquist (1963) published a Decision Tree in 1963. It is a type of classification analysis utilized to create a tree model with a gain ratio. Essentially, it identifies decision factors that might be discounted without increasing cost, and it simply requires a dataset with correctly predicted variables. According to Saxena *et al.* (2021), this is the best model to train modules by Human Resource teams. The Decision Tree algorithm's benefit is that it will return good accuracy and robustness of the classifier. However, the drawback of Decision Tree algorithms is overfitting because the subtree may be repeated several times (Zhou *et al.*, 2021). This successive model effectively and cohesively connects a sequence of actual tests. A numeric attribute is matched to a threshold value and tested one by one (Charbuty & Abdulazeez, 2021).

2.5.3 Random Forest Classifier (RFC)

The Random Forest algorithm is famous in the ensemble learning technique because each branch is split from a subset (choose best predictors) and chosen randomly. However, it primarily uses bootstrap aggregation or bagging for tree learning. In bagging (bootstrap + aggregating), each model uses a bootstrapped data set, which is aggregated to predict the model. In the end, a simple majority vote is used to make a prediction (Punnoose & Ajit, 2016).

2.5.4 K-Nearest Neighbors (KNN)

Cover and Hart (1967) proposed the K-Nearest Neighbors algorithm in 1968. KNN is a lazy learner algorithm. It gathers training examples and defers model construction until the classification test is delivered (Mulak and Talhar, 2013). KNN is a simple and successful approach because it is a non-parametric technique frequently employed in various domains (Yuan *et al.*, 2021). KNN categorizes data items based on their closest neighbours. It is well worth considering (Punnoose and Ajit, 2016).

2.5.5 Support Vector Machine (SVM)

Cortes and Vapnik (1995) suggested a Support Vector Machine in 1995. SVM can help to solve massive classification data. It is like a discriminative classifier because separating the new sample data into a suitable group is good. The process for SVM is to design a hyperplane. After that, divide the categorical data into two classes. Then, it will maximize the geometric distance between the nearest data points (Zhao *et al.*, 2018).

2.5.6 Naïve Bayes (GNB)

One of the supervised learning techniques with no connections between each attribute is called Naïve Bayes (Varmedja *et al.*, 2019). The theorem of Bayes is the foundation of Naïve Bayes. It is a popular classification strategy that divides examples into categories based on the likelihood of events occurring (Sisodia *et al.*, 2017). It is the most attractive classification algorithm because it is simple, has computing efficiency, and has excellent performance for real-world issues. In addition, Naïve Bayes analyzes data considerably more rapidly and accurately (Jaffar *et al.*, 2019).

2.5.7 Adaptive Boosting Classifier (ABC)

AdaBoost is also called 'Adaptive Boosting,' which Freund and Schapire (1999) proposed. This model can adapt to the problem by merging numerous "weak classifiers" into one "strong classifier." However, this model is not easy to over-fit compared with other machine learning methods. The classifier used in the AdaBoost process may be weak (the classification error rate is high). Nevertheless, we must note that the random classification model will be higher than the classification error rate (Tsai & Hung, 2021).

2.5.8 Extreme Gradient Boosting Classifier (XGB)

One of the boosting algorithm members is Extreme Gradient Boosting, introduced by Chen and Guestrin (2016). Multiple regression trees are used to integrate XGB, such that the predicted value of the tree group is as close as possible to the actual value. The goal for XGB is to reduce the risk from the structural (Zhang & Lu, 2021). XGB's strength is its speed, which is faster than other standard machine learning algorithms since it can efficiently analyse massive volumes of data in parallel (Chen & Fan, 2021).

2.6 Model Evaluation

Traditionally, the accuracy measure is used to determine the performance of the predicted model, but because of the imbalanced data used in this research, this measure will not be efficient owing to the overwhelming majority class (Malik *et al.*, 2022). Consequently, different criteria are needed to evaluate the model's performance. The common evaluation metrics in classification cases other than accuracy measures are recall, precision and F1-score (Nandipati *et al.*, 2020). A recall is the proportion of real employee promotion predicted correctly by the model as successful cases. On the other hand, precision is the proportion of predicted observations such as the successful cases of employee promotion predicted by the model that is accurate (Cruz and Wishart, 2006). Performance measurements such as the F1 measure give equal consideration to precision and recall. Moreover, the misclassification rate or error rate will be used which determines the percentage of misclassified observations by the model (Al Khaldy & Kambhampati, 2018). The description of the evaluation metrics is mentioned in Table 2(a) and Table 2(b).

Table 2(a). Model Evaluation.

<table>
<tr><th>Evaluation Metrics</th><th>Explanation</th></tr>
<tr><td>Confusion Matrix</td><td>It is a primary means to evaluate classification problems' errors (Malik <i>et al.</i>, 2022). When the TP and TN classes are higher, FP and FN classes are lower than the algorithms that have done an excellent job.
<table>
<tr><td></td><td></td><td colspan="2">Predicted</td></tr>
<tr><td></td><td><b>Confusion Matrix</b></td><td><b>Negative</b></td><td><b>Positive</b></td></tr>
<tr><td rowspan="2">Actual</td><td><b>Negative</b></td><td>True Negative (TN)</td><td>False Positive (FP)</td></tr>
<tr><td><b>Positive</b></td><td>False Negative (FN)</td><td>True Positive (TP)</td></tr>
</table>
*Binary Classifier: 0-Negative; 1-Positive
TN: Predict the employee non-promoted, and the model also predicts non-promoted employees.
FN: Predict incorrectly. When the employee is promoted, but the model thinks the employee non-promoted
TP: Predict the employee promoted, and the model also predicts promotion.
FP: Prediction incorrectly. When the employee is non-promoted, the model thinks the employee is promoted (Lanier, 2020).</td></tr>
</table>

Evaluation Metrics	Explanation	Formula
Support	The number of actual occurrences of the class in the provided dataset.	$Support_0 = TN + FP$ (1) $Support_1 = FN + TP$ (2) $Support = Support_1 + Support_0$ (3)
Precision	It can also be called positive predictive value. It explains the number of expected cases that happened (Das, 2015; Tatbul *et al.*, 2018).	$Precision_0 = \frac{TN}{TN + FN}\ x\ \frac{Support_0}{Support}$ (4) $Precision_1 = \frac{TP}{TP + FP}\ x\ \frac{Support_1}{Support}$ (5) $Precision = Precision_0 + Precision_1$ (6)
Recall	It can also be called sensitivity or true positive rate (TPR). It explains how many of the actual positive cases with our algorithm were able to accurately predict (Das, 2015; Tatbul *et al.*, 2018).	$Recall_0 = \frac{TN}{TN + FP}\ x\ \frac{Support_0}{Support}$ (7) $Recall_1 = \frac{TP}{TP + FN}\ x\ \frac{Support_1}{Support}$ (8) $Recall = Recall_0 + Recall_1$ (9)

Table 2(b). Model Evaluation.

Evaluation Metrics	Explanation	Formula
F1-score (Harmonic mean)	The F1 score is a metric that combines precision and recalls to assess the accuracy of anomaly predictions (Tatbul *et al.*, 2018). The worst value is 0, and the best deal is 1 (Chicco and Jurman, 2020).	$F1_0 = \left(2\left(\frac{Precision_0 + Recall_0}{Precision_0 + Recall_0}\right)\right)$ *(10)* $F1_1 = \left(2\left(\frac{Precision_1 + Recall_1}{Precision_1 + Recall_1}\right)\right)$ *(11)* $F1\,score = F1_0 + F1_1$ *(12)*

3. Results and Discussions

This section shows descriptive statistics, multi-collinearity, imbalanced dataset and Hybrid Sampling dataset performance, and a comparison of imbalanced and hybridization sampling.

3.1 Descriptive Statistics

Descriptive statistics are a set of methods for calculating, describing, and summarizing datasets in a logical, comprehensible, and time-effective manner (Vetter, 2017). The descriptive analysis was done by using the describe(include="all") function from Pandas. Table 3 shows the statistical properties of each numerical feature, such as mean, standard deviation, and interquartile values. In addition, to the categorical variables, this study provides an overall sense of unique, top, and frequency values for each categorical feature.

Table 3. Descriptive Profile of Variable.

Numerical Variable					Categorical Variable			
Variable	**Mean**	**Std.**	**Min**	**Max**	**Variable**	**Unique**	**Top**	**Frequency**
No of trainings	1.25	0.61	1	10	Department	9	Sales & Marketing	16840
Age	34.80	7.66	20	60				
Previous year rating	3.30	1.21	1	5	Region	34	region_2	12343
Length of service	5.87	4.27	1	37	Education	3	Bachelor's	39078
					Gender	2	Male	38496
					Recruitment channel	3	Other	30446
Avg. training score	63.39	13.37	39	99				

3.2 Preview of Multi-collinearity

Next, will look at how variables are related to each other. Various methods/visualizations can be used for this, such as scatter plots, correlation matrices, variance inflation factors, and others (Cheruku, 2019). In this study, a correlation matrix was applied. A correlation matrix is a matrix that shows how two variables in a dataset are statistically related. There are three types of relationships: positive, negative, and none. A positive correlation indicates that the two variables rise and fall in sync.

Moreover, negative correlation means when two variables move in opposite directions, i.e., two variables rise/fall in sync (Oladunni & Sharma, 2016). When creating models, highly correlated variables are avoided because they can skew the output and make "noise" or inaccuracy in the model (Anderson, 2019). For example, based on the correlation matrix Figure 3, "age" and "length_of_service" have moderate correlations, and other features have weak correlations with each other.

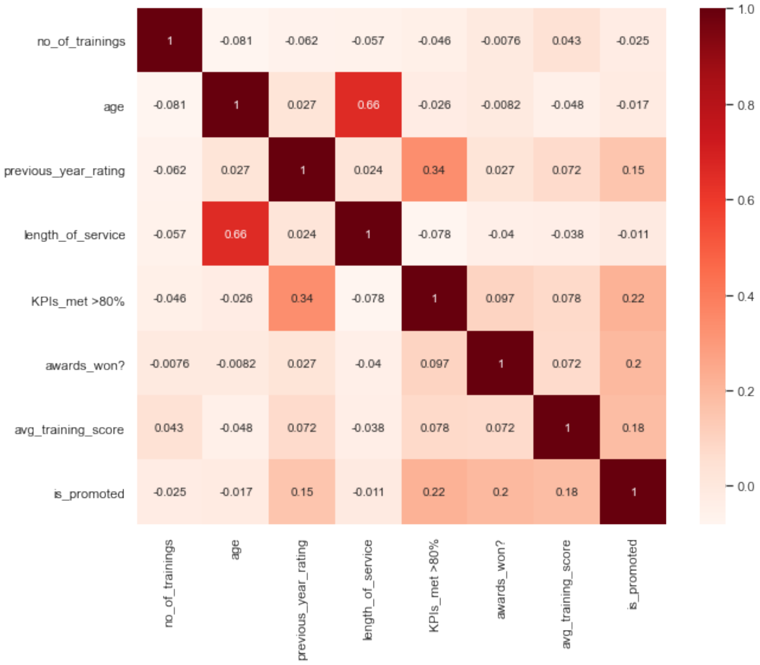


Figure 3. Correlation Matrix of Variables.

A selection of relevant features is recognized as effective in improving model performance. From Figure 3, correlation analysis revealed that the dataset lacks strongly associated features. As a result, no attributes can be removed as highly associated features. So, the feature-selected approaches have been applied to help evaluate the essential features. The different feature selection approaches revealed different sequences for feature selection. Table 4 shows the results for the Recursive Feature Elimination-Random Forest Classifier (RFE-RFC) used to select suitable features and remove the weakest variable within a dataset, Principle Component Analysis (PCA) used to reduce the number of features within a dataset, and Rank Feature Importance method with Extra Tree Classifier (RFI-ECT) used to compute the rank of each variable in a dataset.

Table 4. The Sequence for Features Selected from Different Feature Selection Methods.

FS method	Sequenced for features selected of FS method
RFE-RFC	department, avg_training_score, KPIs_met >80%, region, previous_year_rating, awards_won?, age, recruitment_channel, gender, education, length_of_service, no_of_trainings.
PCA	department, region, education, gender, recruitment_channel, no_of_trainings, age, previous_year_rating, length_of_service, KPIs_met >80%, awards_won?, avg_training_score
RFI-ECT	Region, department, previous_year_rating, avg_training_score, KPI_met >80%, awards_won?, age, recruitment_channel, gender, length_of_service, education, no_of_trainings

In the Rank Feature Importance method, the feature has a higher score, which means that the feature is more important or related to the output variable. This score aids in the selection of the most important aspects for model construction and the elimination of the less important ones. This situation frequently leads to increased accuracy and prevents overfitting by discarding irrelevant features (Abubaker *et al.*, 2020; Sharaff & Gupta, 2019). An extra-tree classifier was chosen due to its explicit meaning, simple features, and ease of conversion to "if-then" rules. Meanwhile, selected the extra-tree method because it implements a meta estimator that fits several randomized decision trees (a.k.a. extra-trees) on various sub-samples of the dataset. So, we opted for the "RFI-ECT" method to compute the features. Table 5 showed the RFI-ETC for the top 5 and 8 features.

Table 5. Two Feature-selected Datasets Selected from RFI-ETC.

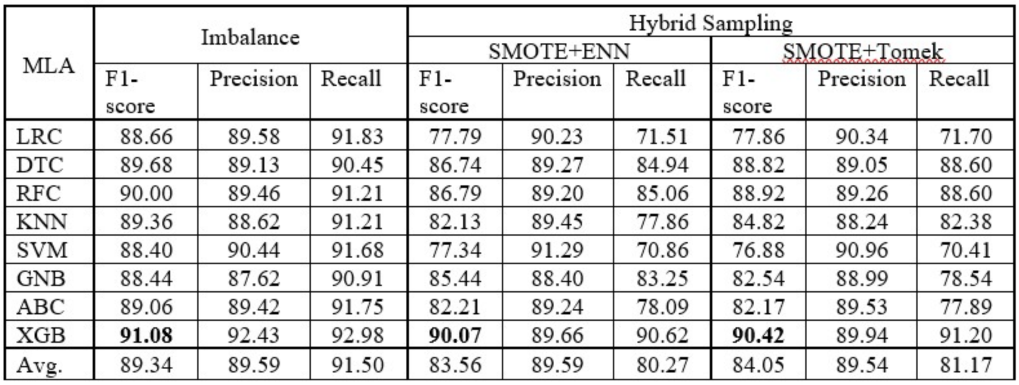
No. Feature	**Feature selected attributes**
8	region, department, previous_year_rating, avg_training_score, KPIs_met >80%, awards_won, age, recruitment_channel
5	region, department, previous_year_rating, avg_training_score, KPIs_met >80%

3.3 Imbalanced dataset and Hybrid Sampling dataset performance with 12 features/ 8 features / 5 features

This study aims to learn how classifiers performed with hybrid sampling methods of the employee promotion datasets. To the best of the author's knowledge, the employee promotion dataset is normally analyzed using machine learning with SMOTE to handle imbalanced data. As a result, no comparative studies between the imbalanced dataset utilized a hybrid sampling method with eight machine learning approaches. The F1 score has been used to determine the performance of each classification algorithm.

In Table 6 show the performance comparisons of the imbalanced dataset and hybrid dataset with 12 features. The imbalanced dataset shows the highest F1 score (91.08%), precision (92.43%), and recall (92.98%) in XGB, followed by RFC with a 90% of F1 score. Next, the SMOTE+ENN hybrid sampling dataset shows the highest F1 score with 90.07% in XGB, followed by RFC with an 86.79% F1 score. Finally, the SMOTE+Tomek hybrid sampling dataset shows the highest F1-score (90.42%) and recall (91.20%) in XGB. Based on the table, the imbalance dataset had the highest overall average in F1-score, and recall, followed by SMOTE+Tomek and SMOTE+ENN.

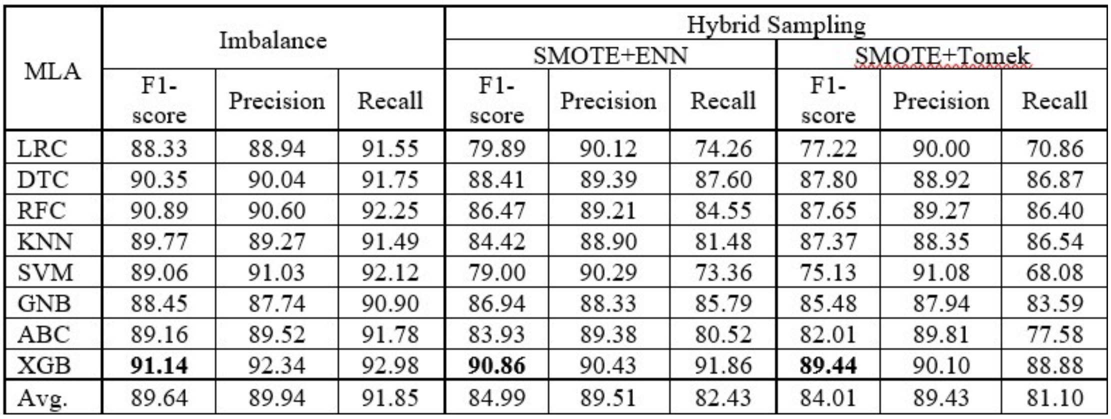
Table 6. The Performance Comparisons of the Imbalanced Dataset and Hybrid Sampling Dataset with 12 Features.



MLA	Imbalance			Hybrid Sampling					
				SMOTE+ENN			SMOTE+Tomek		
	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall
LRC	88.66	89.58	91.83	77.79	90.23	71.51	77.86	90.34	71.70
DTC	89.68	89.13	90.45	86.74	89.27	84.94	88.82	89.05	88.60
RFC	90.00	89.46	91.21	86.79	89.20	85.06	88.92	89.26	88.60
KNN	89.36	88.62	91.21	82.13	89.45	77.86	84.82	88.24	82.38
SVM	88.40	90.44	91.68	77.34	91.29	70.86	76.88	90.96	70.41
GNB	88.44	87.62	90.91	85.44	88.40	83.25	82.54	88.99	78.54
ABC	89.06	89.42	91.75	82.21	89.24	78.09	82.17	89.53	77.89
XGB	**91.08**	92.43	92.98	**90.07**	89.66	90.62	**90.42**	89.94	91.20
Avg.	89.34	89.59	91.50	83.56	89.59	80.27	84.05	89.54	81.17

Next is the eight selected features were taken into consideration to build a model and the result showed in Table 7. First, the imbalanced dataset shows the highest F1 score (91.14%), precision (92.34%), and recall (92.98%) in XGB, followed by RFC with a 90.89% of F1 score. Next, the SMOTE+ENN hybrid sampling dataset shows the highest F1-score with 90.86% in XGB, followed by DTC with an 88.41% F1-score. Finally, the SMOTE+Tomek hybrid sampling dataset shows the highest F1-score (89.44%) and recall (88.88%) in XGB. Based on the table, the imbalance dataset had the highest overall average in F1-score, precision, and recall, followed by SMOTE+ENN and SMOTE+Tomek.

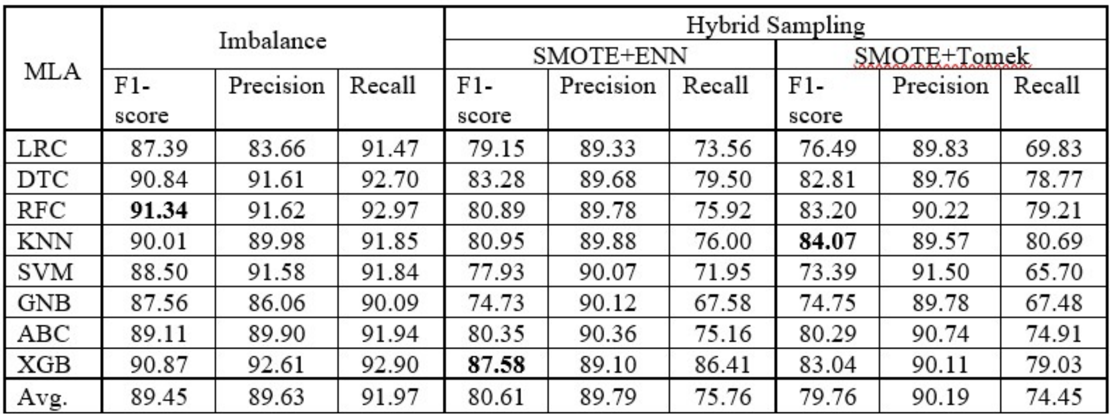
Table 7. The Performance Comparisons of the Imbalanced Dataset and Hybrid Sampling Dataset with 8 Features.



MLA	Imbalance			Hybrid Sampling					
				SMOTE+ENN			SMOTE+Tomek		
	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall
LRC	88.33	88.94	91.55	79.89	90.12	74.26	77.22	90.00	70.86
DTC	90.35	90.04	91.75	88.41	89.39	87.60	87.80	88.92	86.87
RFC	90.89	90.60	92.25	86.47	89.21	84.55	87.65	89.27	86.40
KNN	89.77	89.27	91.49	84.42	88.90	81.48	87.37	88.35	86.54
SVM	89.06	91.03	92.12	79.00	90.29	73.36	75.13	91.08	68.08
GNB	88.45	87.74	90.90	86.94	88.33	85.79	85.48	87.94	83.59
ABC	89.16	89.52	91.78	83.93	89.38	80.52	82.01	89.81	77.58
XGB	**91.14**	92.34	92.98	**90.86**	90.43	91.86	**89.44**	90.10	88.88
Avg.	89.64	89.94	91.85	84.99	89.51	82.43	84.01	89.43	81.10

Lastly, the five selected features were taken into consideration to build a model. From Table 8, the imbalanced dataset shows the highest F1 score (91.34%) and recall (92.97%) in RFC, followed by XGB with a 90.87% F1 score. Next, the SMOTE+ENN hybrid sampling dataset shows the highest F1-score with 87.58% in XGB, followed by DTC with an 83.28% F1-score. Finally, the SMOTE+Tomek hybrid sampling dataset shows the highest F1-score (84.07%) and recall (80.69%) in KNN. Based on the table, the imbalance dataset had the highest overall average in F1-score, and recall, followed by SMOTE+ENN and SMOTE+Tomek.

Table 8. The Performance Comparisons of the Imbalanced Dataset and Hybrid Sampling Dataset with 5 Features.



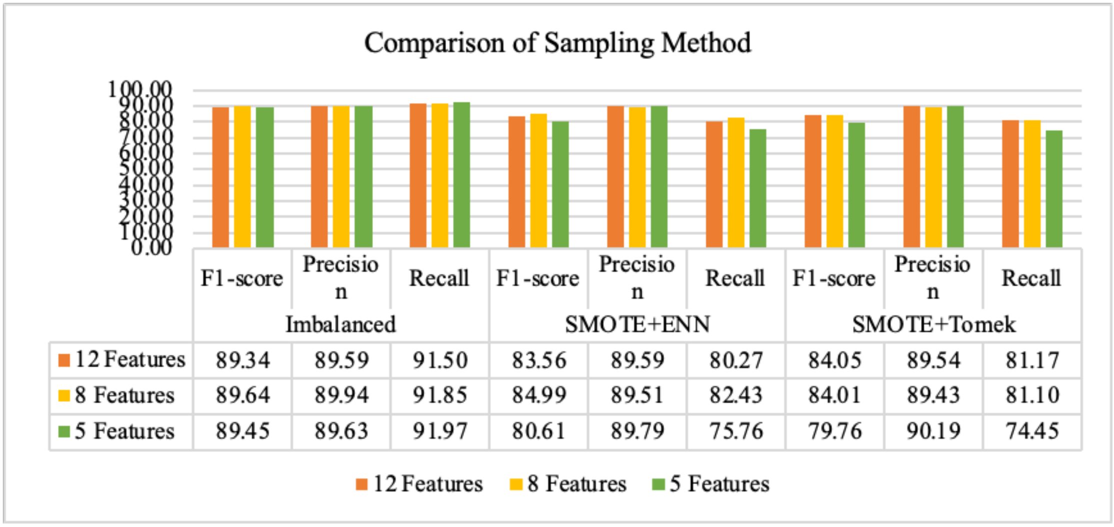
MLA	Imbalance			Hybrid Sampling					
				SMOTE+ENN			SMOTE+Tomek		
	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall
LRC	87.39	83.66	91.47	79.15	89.33	73.56	76.49	89.83	69.83
DTC	90.84	91.61	92.70	83.28	89.68	79.50	82.81	89.76	78.77
RFC	**91.34**	91.62	92.97	80.89	89.78	75.92	83.20	90.22	79.21
KNN	90.01	89.98	91.85	80.95	89.88	76.00	**84.07**	89.57	80.69
SVM	88.50	91.58	91.84	77.93	90.07	71.95	73.39	91.50	65.70
GNB	87.56	86.06	90.09	74.73	90.12	67.58	74.75	89.78	67.48
ABC	89.11	89.90	91.94	80.35	90.36	75.16	80.29	90.74	74.91
XGB	90.87	92.61	92.90	**87.58**	89.10	86.41	83.04	90.11	79.03
Avg.	89.45	89.63	91.97	80.61	89.79	75.76	79.76	90.19	74.45

3.4 Comparison of imbalanced and hybridization sampling in 3 datasets

In comparing machine learning algorithms between an imbalanced dataset and a hybrid sampling dataset, the approximate rank order based on F1-score is XGB, DTC, RFC, and KNN. Besides that, within the comparison of 3 datasets, the 12, 8, and 5 selected features show the overall average F1-score in the three different sampling methods is within the range of 79.76-89.64% (refer to Figure 5). In the three datasets (12, 8, and 5 features), imbalanced data have a higher F1-score and recall than hybrid sampling. When comparing the overall average performance with three datasets (12, 8, and 5 features), the eight features' dataset has a good F1-score in imbalance data and hybrid sampling data.

In comparing performance between an imbalanced dataset and a hybrid sampling dataset, the imbalanced dataset has the highest overall average performance of F1-score, and recall to 3 datasets (12, 8, and 5 features) show a range of 89.34 – 91.97%. On the other hand, the SMOTE+ENN hybrid sampling dataset's overall average performance of F1-score, precision, and recall to 3 datasets (12, 8, and 5 features) show a range of 75.76 – 89.79%; the SMOTE+Tomek hybrid sampling dataset has the lowest overall average performance in recall to 8 and 5 features' dataset with a content of 75.76% and 74.45% respectively (refer to Figure 3.2).

From the comparison, XGB is a good model for handling imbalanced datasets because it usually returns a higher F1 score, precision, and recall. Among the three tables (Table 3.4, Table 3.5, Table 3.6), the overall average performance shows that the imbalanced dataset's F1 score is higher than the hybrid sampling dataset. XGB shows the highest F1-score (90.86%) for the SMOTE+ENN hybrid sampling dataset with eight features; the highest F1-score (90.42%) by XGB for the SMOTE+Tomek hybrid sampling dataset with 12 features. After resampling the imbalanced dataset, XGB has achieved a higher F1-score than other machine learning algorithms compared with the seven different machine learning algorithms. Therefore, it can be concluded that XGB is the champion model in this study because it has a higher F1 score with balanced data.



	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall
	Imbalanced			SMOTE+ENN			SMOTE+Tomek		
12 Features	89.34	89.59	91.50	83.56	89.59	80.27	84.05	89.54	81.17
8 Features	89.64	89.94	91.85	84.99	89.51	82.43	84.01	89.43	81.10
5 Features	89.45	89.63	91.97	80.61	89.79	75.76	79.76	90.19	74.45

Figure 4. The Average F1-Score, Precision, and Recall Comparison with Three Datasets in the Imbalanced Dataset and Hybrid Sampling Dataset.

3.5 Empirical Implications

In general, all the eight classifiers experimented with within this study can predict employee promotion as they yield satisfactory F1-score, precision scores, and recall scores. However, the three datasets (12, 8, and 5 features) do not significantly impact F1-score, precision, and recall scores. This study showed that the F1-score for the imbalanced dataset is higher than the hybrid sampling dataset (SMOTE+ENN, SMOTE+Tomek). The predictive model that developed imbalanced data using a conventional machine learning algorithm does not consider the class distribution proportion or balance of classes and might be biased or inaccurate to the performance (Peng *et al.*, 2019). So, the imbalanced dataset's version looks better than the hybrid sampling dataset.

In a hybrid sampling dataset, most of the time, SMOTE+ENN has the higher F1-score, precision score, and recall score compared with SMOTE+Tomek. The Tomek Links technique was less effective in predicting low precision and recall values than the other oversampling technique (Sawangarreerak and Thanathamathee, 2020). Xu *et al.* (2020) also stated that SMOTE matches ENN perfectly. Based on the studies of Keawwiset *et al.* (2021); Liu *et al.* (2019) and Long *et al.* (2018) (Table 4.6), the Random Forest always has a high accuracy or F1-score with a range of 85.60 – 96.32%. Keawwiset *et al.*'s (2021) study had better accuracy after resampling the imbalanced dataset using the oversampling method (SMOTE). Compared with our study, which uses a hybrid sampling method, the performance of our dataset is a bit lower than the Keawwiset *et al.* (2021). This is because over-sampling, such as making exact copies of existing examples, will cause overfitting and return a high result (Weiss *et al.*, 2007).

This study's champion model is XGB, an ensemble tree technique that uses the gradient descent architecture to boost weak learners. XGB is based on systems optimization (e.g., parallelized tree building, depth-first tree pruning, and cache awareness) and algorithmic enhancements (e.g., regularization, sparsity awareness, and cross-validation) to improve the results (Morde, 2019).

3.6 Potential Implications

The findings of this study present some theoretical implications for academia and industry professionals and some practical implications for the company under investigation for machine learning. Furthermore, this research points to applying machine learning algorithms for employee promotion. Before, some studies successfully predicted employee promotion using machine learning, but no studies have applied a hybrid sampling method with machine learning to predict employee promotion.

The contribution of this study is to give an introductory guideline to apply machine learning, especially SMOTE+ENN with XGB, to predict employee promotion. As a result, future users or researchers can reduce many resources such as money, intention, and time to analyze different algorithms and create the most suitable framework that will lead to the best performance. Hence, practitioners can use the findings of this study to save time and money when choosing the best algorithms that should be used.

The following contribution to this research is to expedite the HR team throughout the promotion cycle. The HR team can save much time to choose the right candidate when applying machine learning. Besides that, the prediction of employee promotion may contribute to improving employee performance. Machine learning can analyze the factors that affect employees getting a promotion. After that, the company can base on the element to train their employee. As a result, it can increase employee productivity, the chance for the employee to get a promotion will be higher, and the turnover rate will reduce in the company.

3.7 Limitations

The first limitation of this study is the imbalanced dataset. The majority class has 91.48%, while the minority class only has 8.52%. This indicates that the minority class is more challenging to predict because learning the features of the samples from the minority class is more difficult for the models. It is also possible that the classification result is skewed toward the majority class. Therefore, a balanced dataset applied to training would probably accelerate more dependable and powerful results. However, in a real-world dataset, having an imbalanced dataset is a common occurrence, as the number of employees who will be promoted is always significantly smaller than the number of employees who will not be promoted. Therefore, the hybrid sampling method has been applied and used the performance metrics suitable for imbalanced data to reduce bias.

Besides, this study lacks statistical generalizability. So, there is no way to describe the relationship between each attribute and the target variable in detail. Despite that, this constraint does not invalidate the study's conclusion, as the goal stated in this study is to develop the best-performing model for predicting employee promotion. Hence, this research is exploratory rather than confirming.

Finally, this study was conducted using the stated dataset, and the results are confined to the information included inside it. Therefore, the results cannot draw broad conclusions about various datasets in various contexts or domains.

4 Conclusion

The goal of this study is to predict employee promotions. First, a literature review was undertaken to investigate a related study and identify a set of machine learning algorithms and a hybrid sampling strategy appropriate for this employee promotion prediction problem. Then, using a 70:30 train to test ratio, each of the eight classifiers (LRC, DTC, RFC, KNN, SVM, GNB, ABC, and XGB) was trained with the dataset tested using the three performance metrics (F1-score, precision, and recall).

Based on the rank feature importance method with the Extra Classifier Tree model, the top five feature-selected attributes are "region", "department", "previous_year_rating", "KPIs_met and_above_80%", and "award_won". The results suggest that SMOTE+ENN and XGB with eight features have the highest-performing model in this study when all three performance measures are considered. As a result, it can be concluded that XGB is the best algorithm for solving the prediction problem, whereas SMOTE+ENN is the best for dealing with imbalanced datasets.

In this study, the goal of forecasting employee promotion is to assist the HR team in expediting the promotion process. In conclusion, the organization can use SMOTE+ENN with XGB to predict whether or not an employee will be promoted. This allows them to identify the factor of non-promoted and allows non-promoted employees to enhance their skills and gain a promotion.

Future research can iterate this study using the same dataset but different methodologies to better deal with the limits given by the imbalanced dataset. For example, resampling the data to obtain balanced classes by using an algorithm approach (e.g., cost-sensitive learning), modifying existing machine learning algorithms (e.g., bagging, stacking techniques), applying hyper-parameter tuning (e.g., random search, grid search), or using other performance evaluation metrics to evaluate the model (e.g., ROC AUC, Log-loss). The idea is to get more accurate and reliable predictions.

Second, the scope of this study could be expanded by integrating other features that are relevant but not available in the current dataset to increase model accuracy. Next is to identify other independent variables that significantly impact promoting an employee. This can be accomplished using statistical analysis, which entails formulating multiple hypotheses and testing them to see their statistical significance.

Acknowledgement

Special thanks are extended to the SOM Research Enhancement Lab 2022, School of Management, Universiti Sains Malaysia.

Funding

The authors received no specific funding for this work.

Author Contribution

All authors have involved themselves equally in every section when writing this paper.

Conflict of Interest

The authors have no conflicts of interest to declare.

References

Abubaker, H., Ali, A., Shamsuddin, S. M., & Hassan, S. (2020). Exploring permissions in android applications using ensemble-based extra tree feature selection. *Indonesian Journal of Electrical Engineering and Computer Science*, 19(1), 543–552.

Aimran, N., Rambli, A., Afthanorhan, A., Mahmud, A., Sapri, A., & Aireen, A. (2022). Prediction of Malaysian women divorce using machine learning techniques. *Malaysian Journal of Computing*, 7(2), 1067-1081.

Aleem, M., & Bowra, Z. A. (2020). Role of training & development on employee retention and organizational commitment in the banking sector of Pakistan. *Review of Economics and Development Studies*, 6(3), 639–650.

Al Khaldy, M., & Kambhampati, C. (2018). Resampling imbalanced class and the effectiveness of feature selection methods for heart failure dataset. *International Robotics & Automation Journal*, 4, 37-45.

Anderson, C. (2019). Hot or not? Heatmaps and correlation matrices. *A post at Medium available at https://medium.com/@connor.anderson_42477/hot-or-not-heatmaps-and-correlation-matrix-plots-940088fa2806*

Ayoubi, S., Limam, N., Salahuddin, M. A., Shahriar, N., Boutaba, R., Estrada-solano, F., & Caicedo, O. M. (2018). Machine learning for cognitive network management. *IEEE Communications Magazine*, 158–165.

Bandi, G. N. S., Rao, T. S., & Ali, S. S. (2021). Data Analytics Applications for Human Resource Management. 2021 *International Conference on Computer Communication and Informatics,* 2021, 31–34.

Bolton, R., Dongrie, V., Saran, C., Ferrier, S., Mukherjee, R., Soderstrom, J., Brisson, S., & Adams, N. (2019). The future of HR 2019: In the know or in the no. *A post at KPMG available at https://assets.kpmg/content/dam/kpmg/pl/pdf/2019/05/pl-Raport-KPMG-The-future-of-HR-2019-In-the-Know-or-in-the-No.pdf*

Castellanous, S. (2019). HR departments turn to AI-enabled recruiting in race for talent. *A post at The Wall Street Journal available at https://www.wsj.com/articles/hr-departments-turn-to-ai-enabled-recruiting-in-race-for-talent-11552600459*

Charbuty, B., & Abdulazeez, A. (2021). Classification based on Decision Tree algorithm for Mmachine learning. *Journal of Applied Science and Technology Trends*, 2(01), 20–28.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *International Conference on Knowledge Discovery and Data Mining*, 785–794.

Chen, Z., & Fan, W. (2021). A freeway travel time prediction method based on an xgboost model. *Sustainability (Switzerland)*, 13(15).

Cheruku, S. K. (2019). What is multicollinearity and how affects model performance in machine learning? *A post a LinkedIn available at https://www.linkedin.com/pulse/what-multicollinearity-how-affects-model-performance-machine-cheruku/.*

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21(6), 1–13.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *IEEE Expert-Intelligent Systems and Their Applications*, 20, 273–297.

Cover, T. & Hart, P. (1967). Nearest neighbour pattern classification, *IEEE Transactions on Information Theory*, 13(1), 21-27.

Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society*, 20(2), 215–242.

Cruz, J. A., & Wishart, D. S. (2006). Applications of machine learning in cancer prediction and prognosis. *Cancer Inform*, 2, 59-77.

Das, T. K. (2015). A customer classification prediction model based on machine learning techniques. *IEEE*, 321–326.

Devi, D., Biswas, S. K., & Purkayastha, B. (2020). A review on solution to class imbalance problem: undersampling approaches. *International Conference on Computational Performance Evaluation*, 626-631.

Freund, Y. & Schapire, R. E. (1999). A short introduction to boosting. *Journal of Japanese Society for Artificial Intelligence*, 14(5), 771-780.

Garg, S., Sinha, S., Kar, A. K., & Mani, M. (2021). A review of machine learning applications in human resource management. *International Journal of Productivity and Performance Management*.

Gazzah, S., Hechkel, A., & Amara, N. E. B. (2015). A hybrid sampling method for imbalanced data. *International Multi-Conference on Systems, Signals & Devices*, 1-6.

Guan, H., Zhang, Y., Xian, M., Cheng, H. D., & Tang, X. (2021). SMOTE-WENN: Solving class imbalance and small sample problems by oversampling and distance scaling. *Applied Intelligence*, 51(3), 1394–1409.

Hetland, J., Hetland, H., Bakker, A. B., & Demerouti, E. (2018). Daily transformational leadership and employee job crafting: The role of promotion focus. *European Management Journal*, 36(6), 746–756.

Ibrahim, A., Muhammed, M. M., Sowole, S. O., Raheem, R., & Rabiat, O. (2020). Performance of catboost classifier and other machine learning methods. 1–14.

Jaffar, Z., Noor, W., & Kanwal, Z. (2019). Predictive human resource analytics using data mining classification techniques. *International Journal of Computer*, 32(1), 9–20.

Jain, N., & Bhushan, M. (2020). Transforming human resource perspective through HR analytics. In Management Dynamics in Digitalization Era.

Jeon, Y. S., & Lim, D. J. (2020). PSU: Particle Stacking Undersampling Method for Highly Imbalanced Big Data. *IEEE Access*, 8, 131920–131927.

Jiang, K., Wang, W., Wang, A., & Wu, H. (2020). Network intrusion detection combined hybrid sampling with deep hierarchical network. *IEEE Access*, 8(3), 32464–32476.

Jomthanachai, S., Wong, W. P. & Khaw, K. W. (2022). An application of machine learning regression to festure selection: A study of logistics performance and economic attribute, *Neural Computing and Applications*, 34, 15781-15805.

Jyoti, P. B. (2022). Employee promotion: The types, benefits, & whom to promote. *A post at Vantage Circle available at https://blog.vantagecircle.com/employee-promotion/*

Kakulapati, V., Chaitanya, K. K., Chaitanya, K. V. G., & Akshay, P. (2020). Predictive analytics of HR - A machine learning approach. *Journal of Statistics and Management Systems*, 23(6), 959–969.

Keawwiset, T., Temdee, P., & Yooyativong, T. (2021). Employee classification for personalized professional training using machine learning techniques and SMOTE. *The 6th International Conference on Digital Arts, Media and Technology*, 376–379.

Lanier, S. T. (2020). Choosing performance metrics. *A post at Towards Data Science available at https://towardsdatascience.com/choosing-performance-metrics-61b40819eae1*

Lin, M., Zhu, X., Hua, T., Tang, X., Tu, G., & Chen, X. (2021). Detection of ionospheric scintillation based on xgboost model improved by smote-enn technique. *Remote Sensing*, 13(13), 1–22.

Liu, J., Wang, T., Li, J., Huang, J., Yao, F., & He, R. (2019). A data-driven analysis of employee promotion: The role of the position of organization. *IEEE International Conference on Systems, Man and Cybernetics (SMC)*, 4056–4062.

Long, Y., Liu, J., Fang, M., Wang, T., & Jiang, W. (2018). Prediction of employee promotion based on personal basic features and post features. *ACM International Conference Proceeding Series on* (pg. 5–10).

Lu, Y., Cheung, Y., & Tang, Y. Y. (2016). Hybrid sampling with bagging for class imbalance learning. *Springer International Publishing Switzerland 2016*, 14–26.

Malik, E. F., Khaw, K. W., Belaton, B., Wong, W. P., & Chew, X. Y. (2022). Credit card fraud detection using a new hybrid machine learning architecture. *Mathematics*, 10, 1480.

Morde, V. (2019). XGBoost algorithm: Long may she reign! *A post at Towards Data Science available at https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d.*

Morgan, J. N., & Sonquist, J. A. (1963). Problems in the analysis of survey data, and a proposal. *Journal of the American Statistical Association,* 58(302), 415–434.

Mulak, P., & Talhar, N. (2013). Analysis of Distance Measures Using K-Nearest Neighbor Algorithm on KDD Dataset. *International Journal of Science and Research*, 4, 2319–7064.

Nandipati, S., Chew, X. Y., & Khaw, K. W. (2020). Hepatitis C virus (HCV) prediction by machine learning techniques. *Applications of Modelling and Simulation*, 4, 89-100.

Oladunni, T., & Sharma, S. (2016). Hedonic housing theory – A machine learning investigation. *15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 522–527.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12, 2825–2830.

Peng, Z., Yan, F., & Li, X. (2019). Comparison of the different sampling techniques for imbalanced classification problems in machine learning. *11th International Conference on Measuring Technology and Mechatronics Automation, ICMTMA*, 431–434.

Pisal, N. S., Abdul-Rahman, S., Hanafiah, M., & Kamarudin, S. I. (2022). Prediction of life expectancy for Asian population using machine learning algorithms. *Malaysian Journal of Computing*, 7(2), 1150-1161.

Punnoose, R., & Ajit, P. (2016). Prediction of employee turnover in organizations using machine learning algorithms. *International Journal of Advanced Research in Artificial Intelligence*, 5(9), 22–26.

Sarker, A., Shamim, S. M., Shahiduz, M., Rahman, Z. M., Shahiduz Zama, M., & Rahman, M. (2018). Employee's performance analysis and prediction using k-means clustering & Decision Tree algorithm. *International Research Journal Software & Data Engineering Global Journal of Computer Science and Technology*, 18(1), 7.

Sawangarreerak, S., & Thanathamathee, P. (2020). Random Forest with sampling techniques for handling imbalanced prediction of university student depression. *Information*, 11(11), 1–13.

Saxena, M., Bagga, T., & Gupta, S. (2021). Fearless path for human resource personnel's through analytics: a study of recent tools and techniques of human resource analytics and its implication. *International Journal of Information Technology (Singapore)*, 13(4), 1649–1657.

Sharaff, A., & Gupta, H. (2019). Extra-Tree Classifier with Metaheuristics Approach for Email Classification. *Advances in Computer Communication and Computational Sciences*, 189–197.

Sisodia, D. S., Vishwakarma, S., & Pujahari, A. (2017). Evaluation of machine learning models for employee churn prediction. *Proceedings of the International Conference on Inventive Computing and Informatics, 2017 ICICI Conference on* (pg. 1016–1020).

Tatbul, N., Lee, T. J., Zdonik, S., Alam, M., & Gottschlich, J. (2018). Precision and recall for time series. *32nd Conference on Neural Information Processing Systems*, 1–11.

Tsai, J. K., & Hung, C. H. (2021). Improving adaboost classifier to predict enterprise performance after covid-19. *Mathematics*, 9(18), 1–10.

Varmedja, D., Karanovic, M., Sladojevic, S., Arsenovic, M., & Anderla, A. (2019). Credit card fraud detection - Machine learning methods. *18th International Symposium INFOTEH-JAHORINA*, 1–5.

Vetter, T. R. (2017). Descriptive statistics : Reporting the answers to the 5 basic questions of who, what, why, when, where, and a sixth, so what? *Anesthesia & Analgesia*, 125(5), 1797–1802.

Weiss, G., McCarthy, K., & Zabar, B. (2007). Cost-sensitive learning vs. sampling: Which is best for handling unbalanced classes with unequal error costs? 1–7.

Xu, Z., Shen, D., Nie, T., & Kou, Y. (2020). A hybrid sampling algorithm combining M-SMOTE and ENN based on Random Forest for medical imbalanced data. *Journal of Biomedical Informatics*, 107, 1–11.

Yuan, B. W., Luo, X. G., Zhang, Z. L., Yu, Y., Huo, H. W., Johannes, T., & Zou, X. D. (2021). A novel density-based adaptive k nearest neighbor method for dealing with overlapping problem in imbalanced datasets. *Neural Computing and Applications*, 33(9), 4457–4481.

Zahin, S. A., Ahmed, C. F., & Alam, T. (2018). An effective method for classification with missing values. *Applied Intelligence*, 48(10), 3209–3230.

Zhang, Y., & Lu, S. (2021). Multi-model fusion method and its application in prediction of stock index movements. *ACM International Conference Proceeding Series on* (pg. 58–64).

Zhao, Y., Hryniewicki, M. K., Cheng, F., Fu, B., & Zhu, X. (2018). Employee turnover prediction with machine learning: A reliable approach. *Advances in Intelligent Systems and Computing*, 869, 737–758.

Zhou, H. F., Zhang, J. W., Zhou, Y. Q., Guo, X. J., & Ma, Y. M. (2021). A feature selection algorithm of Decision Tree based on feature weight. *Expert Systems with Applications*, 164(July 2020), 113842.

Zhu, H. (2021). Research on human resource recommendation algorithm based on machine learning. Scientific Programming.

Zulfikar, W. B., Jumadi, Prasetyo, P. K., & Ramdhani, M. A. (2018). Implementation of mamdani fuzzy method in employee promotion system. *IOP Conference Series: Materials Science and Engineering*, 288(1), 1–5.