Machine Learning for Predicting Employee Attrition

Norsuhada Mansor[1], Nor Samsiah Sani[2]
Center for Artificial Intelligence Technology
Faculty of Information Science & Technology
Universiti Kebangsaan Malaysia
Bangi, Malaysia

Mohd Aliff[3]
Quality Engineering Research Cluster
Instrumentation and Control Engineering
Malaysian Institute of Industrial Technology
Universiti Kuala Lumpur, Johor, Malaysia

***Abstract*—Employee attrition has become a focus of researchers and human resources because of the effects of poor performance on organizations regardless of geography, industry, or size. In this context, the use of machine learning classification models to predict whether an employee is likely to quit could greatly increase the human resource department's ability to intervene on time and possibly provide a remedy to the situation to prevent attrition. This study is conducted with an objective to compare the performance machine learning techniques, namely, Decision Tree (DT) classifier, Support Vector Machines (SVM) classifier, and Artificial Neural Networks (ANN) classifier, and select the best model. These machine learning techniques are compared using the IBM Human Resource Analytic Employee Attrition and Performance dataset. Preprocessing steps for the dataset used in this comparative study include data exploration, data visualization, data cleaning and reduction, data transformation, discretization, and feature selection. In this study, parameter tuning and regularization techniques to overcome overfitting issues are applied for optimization purposes. The comparative study conducted on the three classifiers found that the optimized SVM model stood as the best model that can be used to predict employee attrition with the highest accuracy percentage of 88.87% as compared to the other classification models experimented with, followed by ANN and DT.**

Keywords—Artificial neural networks; decision tree; employee attrition; machine learning; support vector machines

I. Introduction

Machine learning is one of the artificial intelligence technologies that provide systems with the ability to automatically learn and improve from experience or gain human-like intelligence without explicit programming. In other words, machine learning focuses on developing computer programs that can access data and use it to learn for themselves [1]-[4]. Machine learning (ML) is one of the fastest-growing fields of research and has been developed and applied successfully to a wide range of real-world domains [5] – [9]. This study presents a comparative analysis of three machine learning algorithms, i.e., DT, Support Vector Machines (SVM), and Artificial Neural Networks (ANN), to predict employee attrition.

Employee attrition in an organization can mean the reduction of employees through normal means, such as retirement and resignation, clients due to old age, or retrenching them due to change in the target demographics of the organization. The high rate of employee attrition is a major issue in an organization as it greatly impacts them. When employees leave an organization, they carry with them invaluable tacit knowledge, which is often the source of competitive advantage for the organization [10]. Employee attrition causes the organization to bear the cost of business disruption, hiring and training new staff. On the other hand, higher retention means less hiring and training costs and more experienced workers to the company workforce over time. Organization nowadays has given a great business interest in understanding the drivers of staff attrition to reduce employee attrition. As a result, prediction on employee attrition and identifying the major contributing factors that lead to attrition becomes an important objective of an organization in order to enhance its human resource strategy [11].

The IBM Human Resource Analytic Employee Attrition and Performance dataset used in this paper is a publicly available dataset from Kaggle Dataset Repository. It was IBM's fictional dataset created by IBM data scientists. The dataset includes four (4) major components: employee satisfaction, income, seniority, and demographics data. The dataset contains several attributes influencing the predicted variable named 'Attrition' which signifies whether an employee left the company or not from 1,470 instances and 35 attributes. The identified class is labeled as 'Attrition' with 237 instances of 'Yes' and 1233 instances of 'No' having imbalanced data ratio of 1:5.

The purpose of this study is to conduct a comparative study to develop machine learning models, i.e., DT, SVM, and ANN, for predicting probable employee attrition and compare between the algorithms in terms of their accuracy and efficiencies.

II. Related Work

Human resources are considered an important aspect of an organization, and voluntary employee attrition has been identified as a key issue. Reference [10] in his study focused on identifying employee-related attributes to predict employee attrition using decision tree algorithms.

The classification has been identified as an important issue in the emerging field of data mining. Over the years, there have been several studies on classification algorithms. Data mining algorithms must be efficient and scalable for the effective extraction of information from huge amounts of data in many data repositories or dynamic data streams. The key criteria are efficiency, scalability, performance, optimization, and the ability to execute in real-time that drives the development of many new data mining algorithms [12]. Two

(2) important performance indicators for data mining algorithms are the accuracy of a classification and the time taken for training. These indicators are mainly useful for selecting the best algorithms for classification or prediction tasks in data mining [13].

A study conducted by [14] using the IBM HR Employee Attrition & Performance dataset indicated the imbalance in the retrieved data. The correlation plot and histogram visualization had been performed to indicate the correlation between the continuous variables in the model during the data exploration stage. Subsequently, the SMOTE (Synthetic Minority Oversampling Technique) was employed to balance the Attrition class.

The performance measurements observed in many literature reviews are mainly related to finding the best accuracy and speed to build a machine learning model. Table I briefly documents the literature review findings related to a comparative study on employee attrition using the machine learning classification algorithms:

TABLE I. RELATED WORK ON EMPLOYEE ATTRITION

No.	**Author**	**Objective of Study**	**Classification Techniques Studied**	**Recommendation of Classification Techniques by Author**
1.	Saradhi and Palshikar [15]	To predict employee churn	Naive Bayes, SVM, Logistic Regression, Decision Trees and Random Forests	SVM
2.	Alao and Adeyemo [10]	To analyze employee attrition using Decision Tree Algorithms	C4.5 Decision Tree, C5 Decision Tree, REPTree, CART (Simple Cart)	C5 Decision Tree
3.	Punnoose and Pankaj [16]	To predict employee turnover in organizations using machine learning algorithms	Logistic Regression, Naive Bayes, Random Forest, K-Nearest Neighbour (KNN), Linear Discriminant Analysis (LDA), SVM, Extreme Gradient Boosting (XGBoost)	Extreme Gradient XGBoost
4.	Alaskar, Crane and M. Alduailij [17]	To predict when workers will leave. It proposed a combination of five ML algorithms with three techniques for feature selection.	logistic regression, decision tree (DT), naïve Bayes, support vector machine (SVM) and AdaBoost	DT
5.	Mohbey [14]	To predict which customer or employee will leave their current company or organization	Naïve Bayes, SVM, decision tree, random forest, and logistic regression	DT
6.	Srivastava, D. K., & Nair, P. [18]	To analyze employee attrition using predictive techniques	ANN	ANN
7.	Frye et al. [19]	To present a model for predicting employee attrition	Logistic Regression, KNN, Random Forest	Logistic Regression
8.	Khera and Divya [20]	To predict employee turnover using machine learning techniques	SVM	SVM
9.	Ozdemir, Coskun, Gezer and Gungor [21]	To automatize the prediction of employee attrition utilizing data mining methods	Support Vector Machine (SVM), Random Forest, J48, LogitBoost, Multilayer Perceptron (MLP), K-Nearest Neighbors (KNN), Linear Discriminant Analysis (LDA), Naive Bayes, Bagging, AdaBoost, Logistic Regression	SVM
10.	Tharani and Raj [22]	To predict an employee's intention to leave the organization in the immediate future and identify the key features that influence the employee's intention to leave the organization	Logistic Regression and XG boost	XG boost

III. Methodology

A. Data Preprocessing

1) Data description: The initial step in carrying out this study is performing a data pre-preprocessing task. This study produces a data quality report to detect outliers and any unusual pattern about the dataset using statistical methods. Tables II and III show the data quality report of the dataset.

2) Detecting outliers: In addition to the above data quality report, forty-five (45) outliers were detected using the Interquartile Range filter based on the initial raw dataset, and the outliers were then checked. Those findings require further preprocessing, which are data cleaning, data reduction, and data transformation. There are also no missing values that are in existence, and the given data is complete.

3) Data visualization: An overview to understand each attribute pattern should be carried out and examined through data visualization. From the data visualization, we can see that a few attributes need to be examined to ensure accuracy during the model classification process. Fig. 1 shows the data visualization of each attribute in the dataset.

4) Data cleaning and reduction: The dataset is considered high dimensional as it consists of 35 attributes. Any irrelevant attributes that are not contributing to the objectives of this study should be removed. Based on the data quality report in Table III and data visualization in Fig. 1, ‘EmployeeCount,’ ‘StandardHours’ and ‘Over18’ features can be removed in view that the cardinality/distinction is ‘1’, which means it has the same values throughout the data. Other than that, ‘EmployeeNumber’ is found not useful for the modeling and prediction process and can be removed from the dataset. No spelling inconsistencies were detected as inconsistencies may cause problems in later merges or transformations. Further description of data cleaning and reduction is explained in Table IV.

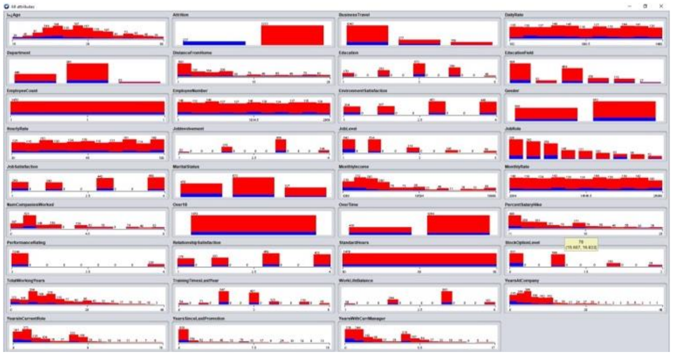

Fig. 1. Data Visualization.

TABLE II. The Data Quality Report (Continuous Attributes)

No	Feature Name	Count	% of Missing Value	Cardinality	Min	1st Qrt	Mean	Medi an	3rd Qrt	Max	Std. Dev
1	Age	1470	0	43	18.00	30.00	36.92	36.00	43.00	60.00	9.14
2	DailyRate	1470	0	886	102.00	465.00	802.49	802.00	1157.00	1499.00	403.50 9
3	DistanceFromHome	1470	0	29	1.00	2.00	9.19	7.00	14.00	29.00	8.11
4	Employee Count	1470	0	1	1.00	1.00	1.00	1.00	1.00	1.00	0.00
5	Employee Number	1470	0	1470	1.00	491.25	1024.86 5	1020. 5	1556.00	2068.00	602.02 4
6	Hourly Rate	1470	0	71	30.00	48.00	65.89	66.00	83.00	100.00	20.33
7	MonthlyIncome	1470	0	1349	1009.00	2911.00	6502.93	4919.00	8380.00	19999.00	4707.96
8	Monthly Rate	1470	0	1427	2094.00	8047.00	14313.1 03	14235.50	20462.00	26999.00	7117.79
9	NumCompaniesWorked	1470	0	10	0.00	1.00	2.69	2.00	4.00	9.00	2.50
10	PercentSal ryHike	1470	0	15	11.00	12.00	15.21	14.00	18.00	25.00	3.66
11	StandardH ours	1470	0	1	80.00	80.00	80.00	80.00	80.00	80.00	0.00
12	TotalWorkingYears	1470	0	40	0.00	6.00	11.28	10.00	15.00	40.00	7.78
13	TrainingTimesLastYear	1470	0	7	0.00	2.00	2.80	3.00	3.00	6.00	1.29
14	YearsAtCompany	1470	0	37	0.00	3.00	7.01	5.00	9.00	40.00	6.13
15	YearsInCurrentRole	1470	0	19	0.00	2.00	4.23	3.00	7.00	18.00	3.62
16	YearsSinceLastPromotion	1470	0	16	0.00	0.00	2.19	1.00	3.00	15.00	3.22
17	YearsWithCurrManager	1470	0	18	0.00	2.00	4.12	3.00	7.00	17.00	3.57

TABLE III. THE DATA QUALITY REPORT (CATEGORICAL ATTRIBUTES)

No.	Feature Name	Count	% of Missing Value	Card.	Mode	Mode Freq.	Mode %	2nd Mode	2nd Mode Freq.	2nd Mode %
1	Attrition	1470	0	2	No	1233	84	Yes	237	16
2	BusinessTravel	1470	0	3	Travel Rarely	1043	71	Travel Frequently	277	19
3	Department	1470	0	3	R & D	961	65	Sales	446	30
4	Education	1470	0	5	3	473	32	4	340	23
5	Education Field	1470	0	6	Life Science	606	41	Medical	464	32
6	Environment Satisfaction	1470	0	4	3	453	31	4	446	30
7	Gender	1470	0	2	Male	882	60	Female	588	40
8	Job Involvement	1470	0	4	3	868	59	2	375	25
9	Job Level	1470	0	5	1	543	37	2	534	36
10	Job Role	1470	0	9	Sales Exec	326	22	Research Scientist	292	20
11	Job Satisfaction	1470	0	4	4	459	31	3	442	30
12	Marital Status	1470	0	3	Married	673	46	Single	470	32
13	Over 18	1470	0	1	Y	1470	100	-	-	-
14	Over Time	1470	0	2	No	1054	72	Yes	416	28
15	Performance Rating	1470	0	2	3	1244	85	4	226	15
16	Relationship Satisfaction	1470	0	4	3	459	31	4	432	29
17	Stock Option Level	1470	0	4	0	631	43	1	596	41
18	Work Life Balance	1470	0	4	3	893	61	2	344	23

TABLE IV. DESCRIPTION OF ATTRIBUTES AND PRE-PROCESSING ACTION

No.	Feature Name	Type of Data	Type of Data	Data Description	Pre-processing action/Findings
1	Age	Continuous	Numeric	The age of individual employee	Min = 18, max = 60 Normalize, Discretize
2	Attrition	Categorical	Nominal	Employee leaving the company (Yes, No)	Set to class
3	BusinessTravel	Categorical	Nominal	Business travel frequency (No Travel, Travel Frequently, Travel Rarely)	Retain
4	DailyRate	Continuous	Numeric	Salary Level	Normalize, Discretize
5	Department	Nominal	Nominal	Employee department (HR, R&D, Sales)	Retain
6	DistanceFromHome	Continuous	Numeric	The distance from work to home	Min = 1, Max = 29 Normalize, Discretize
7	Education	Categorical	Numeric	Level of education attained (1 = ‘Below Collage’, 2 = ‘College’, 3 = ‘Bachelor’, 4 = ‘Master’, 5 = ‘Doctor’)	Change to Nominal
8	EducationField	Nominal	Nominal	Field of education (HR, Life Sciences, Marketing, Medical Sciences, Others, Technical)	Retain
9	EmployeeCount	Continuous	Numeric	Count of instance	Cardinality = 1 - To remove
10	EmployeeNumber	Continuous	Numeric	Employee ID	Cardinality = 1470 - To remove
11	EnvironmentSatisfaction	Categorical	Numeric	Employee satisfaction with the environment (1 = 'Low', 2 = 'Medium', 3 = 'High', 4 = 'Very High')	Change to Nominal
12	Gender	Categorical	Nominal	Female, Male)	Retain
13	HourlyRate	Continuous	Numeric	Hourly Salary	Normalize, Discretize
14	JobInvolvem ent	Categorical	Numeric	Job Involvement (1 = 'Low', 2 = 'Medium', 3 ='High', 4 = 'Very High')	Change to Nominal
15	JobLevel	Categorical	Numeric	Level Of Job (1 to 5)	Change to Nominal

16	JobRole	Categorical	Nominal	(1=Hc Rep, 2=Hr, 3=Lab Technician, 4=Manager, 5= Managing Director, 6=Reasearch Director, 7= Research Scientist, 8=Sales Executieve, 9= Sales Representative)	Retain
17	JobSatisfaction	Categorical	Numeric	Satisfaction with the job (1= 'Low', 2 = 'Medium', 3 ='High', 4 = 'Very High')	Change to Nominal
18	MaritalStatus	Categorical	Nominal	(1=Divorced, 2=Married, 3=Single)	Retain
19	MonthlyIncome	Continuous	Numeric	Monthly Salary	Min = 1 009 Max = 19 709 Normalize, Discretize
20	MonthlyRate	Continuous	Numeric	Monthy Rate	Normalize, Discretize
21	NumCompaniesWorked	Continuous	Numeric	No. Of Companies Worked At	Min = 0 Max = 9 Normalize, Discretize
22	Over18	Categorical	Nominal	(1=Yes, 2=No)	Cardinality = 1 To remove
23	OverTime	Categorical	Nominal	(1=No, 2=Yes)	Retain
24	PercentSalaryHike	Continuous	Numeric	Percentage Increase In Salary	Normalize, Discretize
25	PerformanceRating	Categorical	Numeric	Performance Rating	Min = 3, Max = 4 Change to Nominal
26	RelationshipSatisfaction	Categorical	Numeric	Relations Satisfaction (1 = 'Low', 2 = 'Medium', 3 = 'High', 4 = 'Very High')	Change to Nominal
27	StandardHours	Continuous	Numeric	Standard Hours	Cardinality = 1 - To remove
28	StockOptionLevel	Categorical	Numeric	Stock Options	Min = 0, Max = 3 Change to Nominal
29	TotalWorkin gYears	Continuous	Numeric	Total Years Worked	Normalize, Discretize
30	TrainingTimesLastYear	Continuous	Numeric	Hours Spent Training	Min = 0, Max = 6 Change to Nominal
31	WorkLifeBalance	Categorical	Numeric	Time Spent Between Work And Outside (1 'Bad' 2 'Good' 3 'Better' 4 'Best')	Change to Nominal
32	YearsAtCom pany	Continuous	Numeric	Total Number Of Years At The Company	Min = 0, Max = 40 Normalize, Discretize
33	YearsInCurrentRole	Continuous	Numeric	Years In Current Role	Min = 0, Max = 18 Normalize, Discretize
34	YearsSinceLastPromotion	Continuous	Numeric	Last Promotion	Min = 0, Max = 15 Normalize, Discretize
35	YearsWithCurrManager	Continuous	Numeric	Years Spent With Current Manager	Min = 0, Max = 17 Normalize, Discretize

5) Normalization and discretization: During the data transformation in the preprocessing stage, feature scaling or normalization is applied. Normalization is a method used to standardize the range of independent variables or features of data [23]. Applying feature scaling or normalization can avoid dependency on the choice of measurement units on attributes. This process made the range of features of data fall between 0 and 1. The data cleaning and reduction were performed, which include the discretization process and change of attribute type from numerical to nominal. Four (4) attributes were removed based on the findings above, leaving the remaining 30 attributes. No outliers were detected after the interquartile filter was regenerated.

6) Feature selection: The next preprocessing part in machine learning is feature selection, which involves selecting features in the data and removing irrelevant and redundant information as much as possible to reduce the dimensionality of the dataset. Feature selection is a process of data reduction that helps to improve accuracy, reduce overfitting, reduce training time and identify the fields that are most important and predictive for a given analysis. For this study, the top fifteen (15) out of 30 attributes had been selected based on several attribute selection methods that are Correlation Attribute, Gain Ratio Attribute, and Symmetrical Uncertainty Attributes as depicted in Table V:

Based on Table V, the selected fifteen (15) selected attributes that are used for the modeling phase are: Overtime, StockOptionLevel, JobLevel, MaritalStatus, YearsAt Company, MonthlyIncome, YearsWithCurrManager, TotalWorkingYears, BusinessTravel, Age, YearsInCurrent Role, JobRole, JobInvolvement, EnvironmenSatisfaction, and WorkLifeBalance.

7) Training dan test data: For this experiment, resample filter function is used, the data is divided into two sets of data, which are the training and testing data with a split ratio of 80:20 as per Table VI.

TABLE V. FEATURE SELECTION RESULT

Correlation Attribute		Gain Ratio Attribute		Symmetrical Uncertainty Attribute	
Rank	***Attributes***	***Rank***	***Attributes***	***Rank***	***Attributes***
0.24612	Overtime	0.0464	Overtime	0.0533	Overtime
0.1543	StockOption Level	0.0185	StockOption Level	0.0278	JobLevel
0.1373	JobLevel	0.0184	JobLevel	0.0266	StockOption Level
0.1172	MaritalStatus	0.0149	JobRole	0.0244	JobRole
0.1124	YearsAtCompany	0.0147	MonthlyIncome	0.0239	MonthlyIncome
0.0854	MonthlyIncome	0.0142	MaritalStatus	0.0200	TotalWorkingYears
0.0734	YearsWithCurrManager	0.0123	TotalWorkingYears	0.0200	MaritalStatus
0.0705	TotalWorkingYears	0.0121	YearsAtCompany	0.0187	YearsAtCompany
0.0644	BusinessTravel	0.0117	YearsWithCurrManager	0.0186	YearsWithCurrManager
0.05838	Age	0.0104	Age	0.0173	Age
0.0581	YearsInCurrentRole	0.0102	BusinessTravel	0.0158	YearsInCurrentRole
0.0577	JobRole	0.0099	YearsInCurrentRole	0.0131	BusinessTravel
0.0574	JobInvolvement	0.0083	JobInvolvement	0.0117	JobInvolvement
0.0549	EnvironmenSatisfaction	0.0051	EnvironmenSatisfaction	0.0077	EnvironmenSatisfaction
0.0485	WorkLifeBalance	0.0046	WorkLifeBalance	0.0064	WorkLifeBalance

TABLE VI. SPLIT OF DATA

Dataset	No of Instances
Training with k-fold cross-validation	1,176
Test	294
Total	**1,470**

8) Model validation technique: The k-fold cross-validation is applied to the training set in view of its simplicity. Generally, it results in a less biased or less optimistic estimate of the model trained as compared to the other methods, such as the simple train/test split. Apart from that, this method is chosen as compared to the other training methods in view of the limited data sample in this study. Hence, the cross-validation technique splits the data into *k* groups, and it enables the model to be trained and validated on different sets iteratively. Overfitting refers to a situation where a machine-learning model cannot generalize or match the unseen dataset well. A strong indication of machine learning overfitting is whether the testing or validation dataset error is greater than the training dataset. There are different ways to resolve overfitting; cross-validation is an effective preventive against overfitting. [24].

9) Imbalanced data: The data quality report indicated an imbalance in the class distribution, with 237 tuples predicted as 'Yes' and 1233 tuples predicted as 'No.' Data imbalance is a well-known issue in classification problems, where one class is frequently far more prevalent than the others. Class imbalance usually degrades the real performance of a classification algorithm by poorly predicting the minority class, which is often the center of attention for a classification problem. Imbalanced data requires techniques that can deal with unequal misclassification costs [25]. Hence, the SMOTE technique is applied to overcome the imbalance class at a 200% oversampling degree with five nearest neighborhoods on the training dataset. Using SMOTE, the minority class is over-sampled from 194 to 582 'Yes' instances by creating "synthetic" examples rather than by over-sampling with replacement as shown in Table VII.

B. Machine Learning Classification Algorithms

This section explains the three (3) algorithms that are used in this study:

1) Decision Tree (DT): DT is defined as a tree that classifies instances by sorting them based on feature values. The trees are made up of three fundamental segments: the root node, internal node, and leaf node as shown in Fig. 2. In a DT, each node represents a feature or attribute of the instance to be classified, each branch represents a test result, and leaf nodes represent class labels or class distribution. Classification of instances starts from the root node and is sorted based on their feature values. A sample of the decision tree, which is a flowchart like a tree structure, is as illustrated.

The basic algorithm for decision tree induction is a greedy algorithm that constructs decision trees in a top-down recursive divide-and-conquer manner [18]. C4.5 is an algorithm used to generate a decision tree based on information theory. C4.5 is known as J48 for Java. The classifiers, like filters, are organized in a hierarchy.

TABLE VII. NUMBER OF INSTANCES BEFORE AND AFTER SAMPLING (SMOTE)

Classification Model	No. of instances	Majority Class ("No" Attrition)	Minority Class ("Yes" Attrition)
Before Sampling	1176	982 (84%)	194 (16%)
After Sampling	1564	982 (63%)	582 (37%)

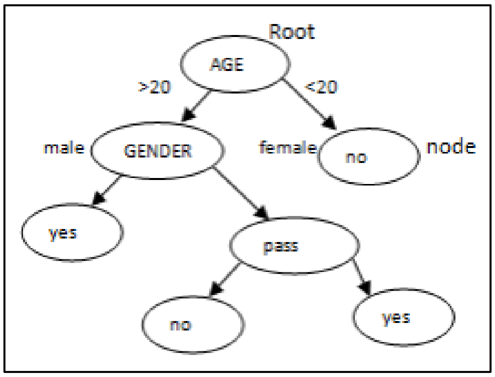

Fig. 2. Decision Tree.

The decision tree is induced by various algorithms. However, as it grows deeper, it happens that sometimes it generates unwanted and meaningless, and this is called overfitting. Therefore, pruning is needed to reduce the size of the tree that is too large and deep. The problem of noise and overfitting reduces the efficiency and accuracy of data [18]. There are various decision tree induction algorithms and various pruning parameters. In this study, pruning parameters such as the confidence factor and the number of objects (at the leaf node) were tuned to improve the DT classifier's performance.

2) Support Vector Machines (SVM): SVM is known as a popular supervised algorithm in machine learning. Also, based on literature, SVM is also commonly used for employee attrition dataset. SVM acts as a classifier that categorizes the data into different 'classes' or as a regression function to estimate the numerical value of the desired output based on a linear combination of features for both linear and non-linear data [27]; SVM is known as SMO.

In relation to his study, the SVM model which is based on the training dataset, will try to generalize the input data based on their features and make a prediction. SVM machine learning will then produce a model that predicts the test data's target values [27]. The basic idea of SVM is to separate classes with maximum margin created by hyperplanes.

The tuning parameter in SVM includes the kernel, regularization parameter (C parameter), and gamma. Polynomial and exponential kernels calculate separation lines in a higher dimension called kernel tricks [27].

3) Artificial Neural Networks (ANN): ANN is a machine learning technique that acquires knowledge through learning and is used to solve classification problems. The ANN can be organized in different topologies/architectures. There are different types of ANN architectures like feedforward and recurrent neural network. The most common neural network model is the Multilayer Perceptron (MLP), a non-linear predictive model that learns through training and is a feedforward network.

The objective in ANN in generic MLP is to find an unknown function f which relates the input vectors in X to the output vectors in *Y*,

$$Y = f(X) \tag{1}$$

Where X=[n ×k],Y=[n ×j].

n = number of training patterns.

k = the number of input nodes/variables.

j the number of output nodes/variables.

During the training of the dataset, the function *f* is optimized, where the network output for the input vectors in *X* is as close as possible to the target values in *Y*. Matrices *X* and *Y* represent the training data. The function f, for ANN architecture, is determined by the adjustable network weights. In ANN, the learning rate can be configured with a small positive value, often in the range between 0 and 1 [28].

C. Machine Learning Tasks Result

For this study, four (4) measures are used to compare the performance of the three (3) classifiers being studied i.e., J48, SVM, and ANN. Those four (4) common measures of the classifier are the accuracy rate, error rate, root mean square error (RMSE), receiver operating characteristic (ROC), and the time taken or speed to build a model. The prediction accuracy is defined as the percentage of correct prediction divided by the total number of predictions. The RMSE indicates an absolute measure of the fitness of the training dataset. A lower value of RMSE indicates a better fit. ROC tells how much the model is capable of distinguishing between classes. The time taken or the speed to build a model is another important consideration in choosing the best classifier model [4].

At the initial stage, the modeling task was carried out on the training dataset using the default parameter of each classifier, and SMOTE resampling technique was applied using 10-Fold cross-validation. Comparison of classifier performance is given in Table VIII.

As seen from the table, the following findings in the initial process of modeling were identified:

1) ANN had the highest accuracy result at 86.76% while SVM showed the lowest at 81.97%.

2) ANN showed the best RMSE with the lowest value of 0.3359.

3) ANN showed the best ROC at the highest value of 0.922.

4) J48 achieved the best time to build a model at 0.02 sec.

TABLE VIII. COMPARATIVE RESULT BETWEEN CLASSIFIERS USING 10-FOLD CROSS-VALIDATION ON DEFAULT PARAMETER ON THE TRAINING DATASET

Performance Measure	**J48**	**SVM**	**ANN**
Accuracy (%)	82.80	81.97	**86.76**
Error Rate (%)	17.20	18.03	**13.24**
RMSE	0.3756	0.4246	**0.3359**
ROC	0.853	0.808	**0.922**
Time taken to build model (second)	**0.02**	2.02	164.22

Machine learning algorithms can be optimized or configured in order to elicit different modeling behavior. Hence, in the next part, parameter tuning is conducted to optimize the model's current performance. The model will then be tested out with the unseen data after the parameter tuning is done on the model.

D. Parameter Tuning

Parameter tuning involves the process of optimizing the performance of a model, that is, to have the best result for each measurement. Parameter tuning is an important step in modeling as it is by no means the only way to improve performance.

1) J48: For the Decision Tree (J48) classifier, the value of the confidence factor and Minimum Number of Objects are tuned to achieve the best model and to avoid overfitting.

a) Confidence factor: The default confidence factor obtained above was run at 0.25. Table IX shows the results of confidence factor parameter tuning ranging from 0.1 to 1.0 run on the J48 model.

The confidence factor parameter is tuned in DT to test the effectiveness of post-pruning. Post-pruning is the process of evaluating the decision error that is the estimated percent of misclassifications, at each decision junction and propagating this error up the tree. Fig. 3 shows that the highest accuracy of 83.57% at 0.4 confidence factor and the accuracy of 82.61% remains constant starting at 0.6 confidence factor. Hence, the 0.4 confidence factor parameter is the optimal value for J48 classifier since increasing the confidence factor leads to lower accuracy.

b) Minimum number of objects: Also, parameter tuning is also conducted to get the optimal value for a minimum number of objects. For this study, the value of a minimum number of objects ranging from 0 to 30 is tuned at the confidence factor of 0.4. Table X shows the results for the minimum number of objects pruning parameter:

The minimum number of objects specifies the number of instances at the leaf node as a threshold value which means it specifies the minimum number of data separations per branch [26]. Fig. 4 shows that after the minimum number of objects of 1, the accuracy decreases when the minimum number of objects increases. The highest accuracy is at the parameter of 1 (minimum is 0 and cannot be a negative value) for the minimum number of objects with an accuracy of 84.40%. Hence, the minimum number of objects of 1 is the optimal number for the model.

2) SVM: The performance of the SVM classifier depends on the use of different kernel parameters in view that an appropriate kernel will provide a learning capability to SVM. For this experiment, as proposed in the literature, three (3) kernel functions were used for comparison in parameter tuning, which are the polynomial kernel, radial basis function (RBF) kernel, and Pearson VII kernel function (PUK) [29]-[31]. The regularization parameter (C) for these different kernels is tuned to improve the SVM model performance. The C determines how much penalty is given for misclassification. The result of the kernel with C tuning is indicated in Table XI as follows.

TABLE IX. CONFIDENCE FACTOR TUNING FOR DT

Confidence Factor	**Accuracy (%)**	**Error Rate (%)**
0.2	81.84	18.16
0.4	**83.57**	**16.43**
0.6	82.61	17.39
0.8	82.61	17.39
1.0	82.61	17.39

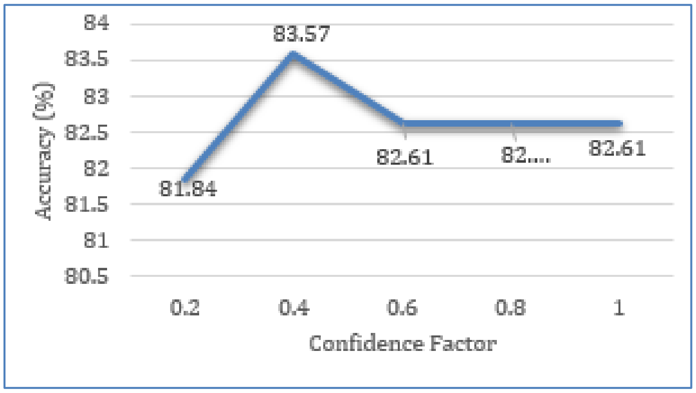

Fig. 3. Effect of Confidence Factor Tuning to Accuracy.

TABLE X. MINIMUM NUMBER OF OBJECTS TUNING FOR DT

Minimum Number of Objects	**Accuracy (%)**	**Error Rate (%)**
0	84.21	15.79
1	**84.40**	**15.60**
2 - default	83.57	16.43
5	83.38	16.62
10	79.80	20.20
15	78.71	21.29
20	77.69	22.31
25	77.11	22.89
30	76.15	23.85

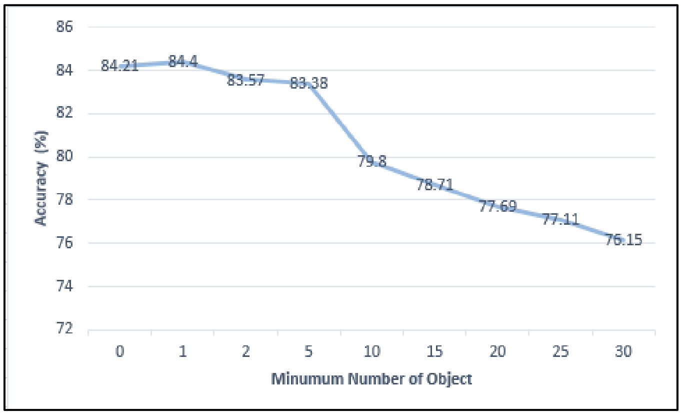

Fig. 4. Effect of MinNumObject Tuning to Accuracy.

TABLE XI. KERNEL AND REGULARIZATION PARAMETER (C) TUNING FOR SVM

Kernel	Regularization Parameter (C)	Accuracy (%)	Error Rate (%)	RMSE	ROC	Time taken to build model (s)
Polykernel	1	81.97	18.03	0.4246	0.808	2.19
	10	81.59	18.41	0.4291	0.806	6.57
Radial Basis Function (RBF)	1	82.23	17.77	0.4216	0.795	4.23
	10	85.23	14.77	0.3843	0.842	2.01
	100	86.51	13.49	0.3673	0.860	10.01
	200	85.55	14.45	0.3801	0.850	5.04
PUK	1	88.43	11.57	0.3402	0.847	3.68
	10	**88.87**	**11.13**	**0.3335**	**0.853**	**5.36**
	100	88.87	11.13	0.3335	0.853	5.35
	200	88.87	11.13	0.3335	0.853	5.67

The tuning result showed that the SVM model with PUK kernel produced the best fit with the highest accuracy of 88.87% and the lowest RMSE of 0.3335 compared to the other kernel when C is set to 10 using the PUK kernel. There is no change in the accuracy after the C value of 10; hence, the value is already optimized. This experiment also showed that the choices of kernel function gave an insightful effect on the performance of the SVM model for the employee attrition dataset after the parameter tuning.

3) ANN: In ANN, parameter tuning is performed by adjusting the learning rate. Table XII, Fig. 5 shows the performance result with parameter tuning on the learning rate.

TABLE XII. LEARNING RATE TUNING FOR ANN

Learning Rate	Accuracy (%)	Error Rate (%)	RMSE	ROC	Time taken to build model (s)
0.3	86.76	13.24	0.3359	0.922	84.27
0.4	**87.98**	**12.02**	**0.3274**	**0.925**	**86.41**
0.5	87.66	12.34	0.3329	0.924	90.86
0.6	87.08	12.92	0.3457	0.905	87.85

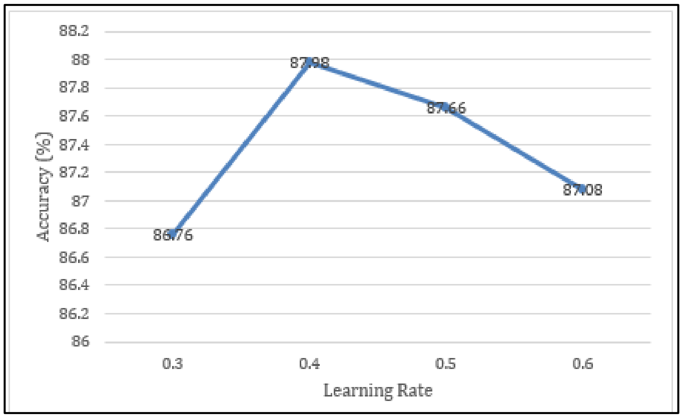

Fig. 5. Effect of Learning Rate Tuning to Accuracy.

The tuning result showed that ANN performed the best at a learning rate of 0.4 with an accuracy of 87.98%, and the time taken is 86.41sec as an optimal value. This algorithm was initially chosen in view of its capacity to detect all possible interactions between variables. However, even though this study used a small dataset with only 15 attributes after feature selection, ANN requires more time to create the model and requires more machine resources/capacity than the other machine learning algorithms. Moreover, the accuracy of 87.98% is still lower than the SVM. Hence, it is a less favorable option for this type of dataset.

E. Regularization

Regularization is basically a technique that was used to overcome the overfitting problem of a model. Overfitting refers to an occurrence where the model learns both the target function and noise during the training, which affects the performance of that model on the test/unseen data.

Regularization reduces the variance of the model without a substantial increase in its bias. In this study, few regularization techniques were performed to limit overfitting. As explained above, the tuning parameter is applied in each of the classifiers and is used as part of the regularization techniques to control the impact on bias and variance. As the value of parameter tuning rises, it reduces the coefficients' value, thus reducing the variance to avoid overfitting but not losing any important properties in the data. However, underfitting will occur when the model starts to lose important properties after a certain value, and this leads to the rising of bias in the model. Therefore, the value chosen during parameter tuning must be carefully selected [32].

Moreover, this study uses pruning to reduce the size of a decision tree to overcome overfitting. The SMOTE oversampling technique was applied to treat imbalanced minority classes in the dataset. Also, the use of the 10-fold cross-validation method, which is a resampling procedure, has given a coherent result and is used to overcome the overfitting issue in the dataset. Generally, regularization refers to a broad range of techniques for artificially forcing the machine learning model to be simpler and increase generalization chances.

IV. Results and Discussion

A. The Effect of Feature Selection on Classification Accuracies

The 10-fold cross-validation test option enables the accuracy improvement of 15 attributes in comparison to 30 attributes. The result is depicted in Table XIII.

Based on the table, the results indicated that the use of top 15 attributes through feature selection has very much reduced the time taken to build the model from 330.23sec to 28.01sec without affecting the accuracy much where there is only a slight change from 85.96% to 85.13%.

B. Comparative Result between Classifiers after Parameter Tuning and Regularization using 10-Fold Cross-Validation

Table XIV shows the result obtained after the parameter tuning and regularization are applied for each classifier. The result in the training dataset below represents the best result for each classifier after applying parameter tuning and regularization. The results were then be compared with the unseen/test data.

TABLE XIII. The Effect of Feature Selection

Classification Model	**Before feature Selection (30 attributes)**				**After Feature Selection (15 attributes)**			
	Accuracy (%)	**RMSE**	**ROC**	**Speed (sec)**	**Accuracy (%)**	**RMSE**	**ROC**	**Speed (sec)**
J48	84.48	0.3608	0.603	0.02	84.56	0.3619	0.602	0.01
SVM	87.00	0.3605	0.723	3.79	86.87	0.3623	0.659	1.77
MLP	86.39	0.3440	0.839	326.42	83.95	0.3737	0.780	82.25
Average	85.96	0.3551	0.722	330.23	85.13	0.3660	0.680	28.01

TABLE XIV. Performance Comparison between DT, SVM and ANN Classifiers

Classifier/ Results	**Dataset**	**Accuracy (%)**	**Error Rate (%)**	**RMSE**	**ROC**
DT – J48	Training	84.40	15.60	0.3704	0.850
	Test	80.95	19.05	0.4038	0.633
SVM	**Training**	**88.87**	**11.13**	**0.3335**	**0.853**
	Test	**87.76**	**12.25**	**0.3499**	**0.990**
ANN	Training	87.98	12.02	0.3274	0.925
	Test	85.03	14.97	0.3571	0.88

From the result in Table XI, SVM is revealed to be the best model that separates the class that can later be used to decide the class of a new set of data in predicting attrition. SVM ranks first at an accuracy rate of 88.87% (with parameter tuning at C=10 under the PUK kernel) while closely followed by ANN at 87.38%. DT showed the lowest accuracy rate of 84.40%. The performance measure result of the test dataset also showed a close result as compared to the training data and does not exceed the training result. It is proved that the model is not overfitted, and it is useful for predicting attrition for the new unseen dataset.

As data preprocessing may affect the outcomes of the final model be interpreted, hence a tremendous effort is emplaced during the preprocessing stage for this study as it took a considerable amount of processing time. Several challenges and critical constraints faced in this study include the limited size of the dataset, imbalanced class, and high dimensional dataset. Hence, data preprocessing is an important stage to ensure only relevant features are selected for the training set.

The crucial part during the modeling stage is the parameter tuning conducted for each algorithm as different parameters require a different setting. In this study, this fact is proven when the initial accuracy for SVM was the lowest with no parameter tuning applied. However, SVM showed the highest accuracy after the parameter tuning due to its capacity to handle high-dimensional data with the use of different kernel functions. Also, the regularization technique is applied throughout the experiment to overcome the issue of overfitting during the modeling phase.

V. Conclusion

The comparative study on IBM Human Resource Analytic Employee Attrition and Performance was conducted to evaluate the classification models, i.e., J48, SVM, and ANN. SVM model stood at the best accuracy, RMSE, and Speed value after parameter tuning and regularization. Each of the three (3) classifiers used in this study has advantages and limitations; thus, evaluation is required to determine its suitability to solve the problem in relation to the dataset being studied.

This paper is mainly focusing on the comparative study of the machine learning model to predict whether an employee would leave the company or not given an employee attrition dataset. Hence, future work may look into identifying the key features that lead to employee attrition. Apart from that, the use of the hyperparameter tuning approaches like grid search or random search can further be deliberated to find the best combination of parameters to enhance the model to ensure its efficiency and scalability.

Acknowledgment

The authors would like to thank Universiti Kebangsaan Malaysia (UKM) and Ministry of Education, Malaysia (MOE) under the FRGS/1/2018/ICT02/UKM/02/6) for funding and supporting this research.

References

[1] S. Das, A. Dey, A. Pal and N. Roy, "Applications of artificial intelligence in machine learning: Review and prospect," Int. J. Comp. Appl., vol. 115, pp. 31–41, January 2015.

[2] A. Abu, R. Hamdan, R. and N.S. Sani, "Ensemble Learning for Multidimensional Poverty Classification," Sains Malaysiana," vol. 49(2), pp.447-459 2020.

[3] Nor Samsiah Sani, Abdul Hadi Abd Rahman, Afzan Adam, Israa Shlash and Mohd Aliff, "Ensemble Learning for Rainfall Prediction" International Journal of Advanced Computer Science and Applications, vol. 11(11), pp. 153-162, 2020.

[4] Nor Samsiah Sani, Ahmad Fikri Mohamed Nafuri, Zulaiha Ali Othman, Mohd Zakree Ahmad Nazri and Khairul Nadiyah Mohamad, "Drop-Out Prediction in Higher Education Among B40 Students" International Journal of Advanced Computer Science and Applications, 11(11), pp. 550-559, 2020.

[5] A. B. Abdulkareem, N. S. Sani, S. Sahran, Z. A. A. Alyessari, A. Adam et al., "Predicting covid-19 based on environmental factors with machine learning," Intelligent Automation & Soft Computing, vol. 28 (2), pp. 305–320, 2021.

[6] N. S. Sani, I. I. S. Shamsuddin, S. Sahran, A. H. A. Rahman, and E. N. Muzaffar, "Redefining selection of features and classification algorithms for room occupancy detection," International Journal on Advanced Science, Engineering and Information Technology, vol. 8, pp. 1486-1493, 2018.

[7] S. Shabudin, N. S. Sani, K. A. Z. Ariffin and M. Aliff, "Feature Selection for Phishing Website Classification," International Journal of Advanced Computer Science and Applications, vol. 11(4), pp. 587-595, 2020.

[8] R. Hamdan, A. Abu and N. S. Sani. "Does Artificial Intelligence Prevail in Poverty Measurement?." In Journal of Physics: Conference Series, vol. 1529(4), pp. 042082. IOP Publishing, 2020.

[9] Z. A. Othman, A. A. Bakar, N. S. Sani, and J. Sallim, "Household Overspending Model Amongst B40, M40 and T20 using Classification Algorithm," International Journal of Advanced Computer Science and Applications, vo1. 11(7), pp. 392-399, 2019.

[10] D. Alao and A. B. Adeyemo, "Analyzing employee attrition using decision tree algorithms," Comput., Inf. Syst., Dev. Inform. Res. J., vol. 4, pp. 17–28, March 2013.

[11] J. Rohan, A. Shahid, S. Saud, and J. Ramirez, "IBM HR analytics employee attrition & performance," January 2018 [Online] http://inseaddataanalytics.github.io/INSEADAnalytics/groupprojects/January2018FBL/IBM_Attrition_VSS.html#business_problem.

[12] H. Jiawei and M. Kamber, Data Mining: Concepts and Techniques, Burlington, MA: Morgan Kaufmann, 2001.

[13] S. O. Akinola and O. J. Oyabugbe, "Accuracies and training times of data mining classification algorithms: An empirical comparative study," J. Softw. Eng. Appl., vol. 8, pp. 470–477, September 2015.

[14] K.K Mohbey, "Employee's Attrition Prediction Using Machine Learning Approaches," In Machine Learning and Deep Learning in Real-Time Applications, pp. 121-128, 2020.

[15] M. E. Kara, S. Ü. O. Fırat and A. Ghadge, "A data mining-based framework for supply chain risk management," Computers & Industrial Engineering, vol. 139, pp. 1-12, 2020.

[16] R. Punnoose and A. Pankaj, "Prediction of employee turnover in organizations using machine learning algorithms: A case for extreme gradient boosting," Int. J. Adv. Res. Artif. Intel., vol. 5, pp. 22–26, October 2016.

[17] L. Alaskar, M. Crane and M. Alduailij, "Employee Turnover Prediction Using Machine Learning," In International Conference on Computing, pp. 301-316, 2020.

[18] D. K. Srivastava and P. & Nair, "Employee attrition analysis using predictive techniques," in Int. Conf. Inform. Commun. Technol. for Intell. Syst., pp. 293–300, March 2017.

[19] A. Frye, C. Boomhower, M. Smith, L. Vitovsky, and S. Fabricant, "Employee attrition: What makes an employee quit?. SMU Data Sci. Rev., vol. 1, pp. 1–29, 2018.

[20] S. N. Khera and Divya, "Predictive modelling of employee turnover in Indian IT industry using machine learning techniques," Vis. J. Bus. Perspect., vol. 23, pp. 12–21, March 2018.

[21] F. Ozdemir, M. Coskun, C. Gezer and V.C Gungor, "Assessing Employee Attrition Using Classifications Algorithms," In Proceedings of the 2020 the 4th International Conference on Information System and Data Mining, pp. 118-122, May 2020.

[22] S.M. Tharani and S.V. Raj, "Predicting employee turnover intention in IT&ITeS industry using machine learning algorithms," In 2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC), pp. 508-513, 2020.

[23] S. Srivastava, "Weka: A tool for data preprocessing, classification, ensemble, clustering and association rule mining," Int. J. Comput. Appl., vol. 88, pp. 26–29, February 2014.

[24] T. Shah, "About train, validation and test sets in machine learning," Towards Data Science, 6 Dec. 2017.

[25] A. Singh and A. Purohit, "A survey on methods for solving data imbalance problem for classification," Int. J. Comput. Appl., vol. 127, pp. 37–41, October 2015.

[26] N. Patel and S. Upadhyay, "Study of various decision tree pruning methods with their empirical comparison in WEKA," Int. J. Comput. Appl., vol. 60, pp. 20–25, December 2012.

[27] J. Cervantes, F. Garcia-Lamont, L. Rodriguez-Mazahua and A. Lopez, "A comprehensive survey on support vector machine classification: Applications, challenges and trends," Neurocomputing, vol. 408, pp. 189-215, 2020.

[28] D. S. Jat, P. Dhaka and A. Limbo, "Applications of statistical techniques and artificial neural networks: A review," Journal of Statistics and Management Systems, vol. 21(4), pp. 639-645, 2018.

[29] D. Tien Bui, B. Pradhan, O. Lofman and I. Revhaug, "Landslide susceptibility assessment in Vietnam using support vector machines, decision tree, and Naive Bayes Models," Math. Probl. Eng., pp. 1–26, July 2012.

[30] K. A. Abakar and C. Yu, "Performance of SVM based on PUK kernel in comparison to SVM based on RBF kernel in prediction of yarn tenacity," Indian J. Fibre Text. Res., vol. 39, pp. 55–59, 2014.

[31] M. Abdul Rahman, N.S. Sani, R. Hamdan, Z. Ali Othman and A. Abu Bakar, "A clustering approach to identify multidimensional poverty indicators for the bottom 40 percent group," Plos one, vol. 16(8), pp. e0255312, 2021.

[32] Y. Li, C. Wei and T. Ma, "Towards explaining the regularization effect of initial large learning rate in training neural networks," In Advances in Neural Information Processing Systems, pp. 11674-11685, 2019.