Employee Turnover Prediction Using Machine Learning

Lama Alaskar[1(✉)], Martin Crane[2(✉)], and Mai Alduailij[3(✉)]

[1] College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia
lama.m.alaskar@gmail.com
[2] School of Computing, Dublin City University, Dublin, Ireland
martin.crane@dcu.ie
[3] College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia
MAAlduailij@pnu.edu.sa

Abstract. High employee turnover is a common problem that can affect organizational performance and growth. The ability to predict employee turnover would be an invaluable tool for any organization seeking to retain employees and predict their future behavior. This study employed machine learning (ML) algorithms to predict whether employees would leave a company. It presented a comparative performance combination of five ML algorithms and three Feature Selection techniques. In this experiment, the best predictors were identified using the SelectKBest, Recursive Feature Elimination (RFE) and Random Forest (RF) model. Different ML algorithms were trained, which included logistic regression, decision tree (DT), naïve Bayes, support vector machine (SVM) and AdaBoost with optimal hyperparameters. In the last phase of the experiment, the predictive models' performance was evaluated using several critical metrics. The empirical results have demonstrated that two predictive models performed better: DT with SelectKBest and the SVM-polynomial kernel using RF.

Keywords: Employee turnover · Machine learning · Support vector machine · Feature selection · Decision tree

1 Introduction

One of the most valuable assets of any organization is its employees. The key to the success of any organization is attracting and retaining talent. Human resources (HR) play an essential role in driving organizational performance and helping develop a company, and those resources have a substantial influence on company performance. Further, one of the most common issues affecting organizational performance and growth is the phenomenon of employee turnover, particularly when it may be unexpected, or when a departing employee held an important position within the organization. In any company it is essential to retain skilled employees, as they are considered a crucial element of HR management (HRM).

A. Alfaries et al. (Eds.): ICC 2019, CCIS 1097, pp. 301–316, 2019.
https://doi.org/10.1007/978-3-030-36365-9_25

Significant growth in the volume of HR data has created a desire to extract meaningful information from those data. Ideally, this large volume of data would be used to identify behavioural patterns, and those patterns would be understood to create the knowledge needed to help decision-makers improve policies that both increase employee performance and maintain high levels of employee satisfaction.

Employee turnover may be one of the biggest challenges facing many companies today. Use of machine-learning (ML) techniques, like those discussed in this study, to develop predictive models can potentially help a company identify staff that are likely considering leaving. Knowing the most important factors that lead to turnover can help provide a crucial element to the HRM system. Talent retention plans and strategies can then be used to address this issue in an effective and timely manner to enhance employee performance and job satisfaction, and in turn, minimize departures.

This study proposed several supervised ML algorithms in order to find the one most suitable for the HR domain. In addition, the study identified the predictors (or features) that contributed most to employee turnover, which could be used to help decision-makers both manage new staff and analyse the departure of existing employees to improve overall employee retention rates. The study also employed a survival analysis technique to determine time-to-turnover within an organization.

The feature selection (FS) technique was applied on proposed models that offered the best prediction results. This resulted in an optimal feature subset that greatly affected the prediction results.

A combination of five ML models and three FS techniques were created, and the experiment was conducted using the train-validate-test and 10-fold cross-validation methods. The proposed models that were developed for predicting employee turnover were a naïve Bayes classifier, a decision tree (DT) classifier, logistic regression (LR), a support vector machine (SVM) and an AdaBoost ensemble classifier. The popular FS techniques were the f_classif function using the SelectKBest method, Recursive Feature Elimination (RFE), and the Random Forest (RF) model.

The performance of the predictive models was studied using the critical metrics for accuracy, sensitivity, precision, specificity, F1-score, misclassification rate, and AUC value.

Experimental comparisons were made to address several key research questions (RQs). However, the results of this study also demonstrated the value of prediction outcomes to decision-makers in their efforts to successfully manage human capital.

This study aims to improve employee retention plans for organizations. Toward that end, it sought to answer the following the RQs:

- RQ1: Can a prediction model determine whether an employee will leave the company (thus, contribute to turnover) or not?
- RQ2: What are the key factors or best predictors that significantly contribute to employee turnover?

The rest of this paper is organized into five sections. Review of the literature is examined in Sect. 2, and an overview of the proposed methods used in the experiment is presented in Sect. 3. Sections 4 and 5 show research methodology and results, respectively. Section 6 is the conclusions of this paper.

2 Literature Review

In recent years, the employee-departure phenomenon has gained considerable attention within the data mining (DM) domain. It is becoming increasingly necessary for organizations to attempt to predict employee turnover and to determine the factors most effective in helping retain employees going forward. In this section, the DM techniques and experiment methodologies are discussed.

In [1], researchers concluded that proper integration between HRM and data mining would improve the recruitment and decision-making processes, which in turn would significantly reduce turnover. Moreover, the researchers showed that the probabilistic neural network (PNN), support vector machine (SVM), and k-nearest neighbours (KNN) models were sensitive to parameters. Conversely, they found that the naïve Bayes classifier is both the most user-friendly and demonstrated high performance in the classification problem. The random forest model outperformed other classifiers with 91% accuracy, followed by the naïve Bayes classifier at 88.8%, while SVM was found to be the least accurate [1]. In [2], researchers employed four prominent models for churn prediction, including decision tree (DT), KNN, artificial neural network (ANN), and SVM models. The authors found that the ANN model slightly outperformed the SVM model. Meanwhile, the proposed hybrid model, which combined four classifiers, was found to be more accurate, at 95% in precision and recall measures than the other four models. In other work, the unsupervised DM method employed by [3] for predicting turnover among technology experts, along with clustering analysis, involved a self-organizing map (SOM) and a hybrid artificial neural network. These data were obtained from surveys of Taiwanese company employees. Three models were used, k-mean, SOM with a backpropagation network algorithm (BPN), and BPN. The SOM–BPN model had the highest accuracy at 92%.

As data mining has a role in selecting candidates in order to limit turnover, empirical research has shown practical benefits from using data-mining processes to create useful rules for companies [4]. This empirical research has shown the viability of using decision trees to create useful rules for HRM within the semiconductor industry.

Another study surveyed ML techniques and compared the models used in predicting customer-churn behaviour [5]. The results indicated that ML algorithms, such as SVM could be extended to build accurate predictive models for employee turnover [5]. The comparison revealed that SVM outperformed other classifiers that have higher accuracy score. In [6], two proposed methods, DT and ANN, were examined to both manage and prevent turnover for a Taiwanese telecommunications company. The resulting research revealed that ANN performed better in predicting employee churn [6].

In a practical experiment [7], data extracted from the first two months of an employee contract was used to predict departure in the third month. In this study, an understanding of whether call centre employees would continue through to the third month, thereby earning permanent contracts, was of importance to the human resource manager. The results of various predictive models were shown using 10-fold cross-validations, where the naïve Bayes algorithm slightly outperformed other techniques, with an accuracy of 85%.

A further overview of the recent literature on turnover prediction is provided in [8]. The literature on churn analysis has employed intelligent data analysis in addressing the turnover issue through construction of predictive models (such as SVM).

Another study demonstrated the effectiveness of logistic regression to predict voluntary employee turnover [9]. The data in this research were collected by a private company in Belgium specializing in human resources. The finding was that the AUC of logistic regression was around 74%. The authors concluded that the results of this study helped the HR manager prepare interviews with the at-risk group to prevent and reduce turnover in the company.

Despite other efforts in turnover prediction, no research examined different feature selection techniques. Hence, this paper has attempted to improve performance predictive models. This research has employed various techniques to select important features, and it has compared those techniques with different models. In addition, optimized models by hyperparameters such as grid search are discussed.

3 Methods

3.1 Logistic Regression

Logistic regression (LR) is a kind of classification approach that measures a relationship between the target variable in dichotomies (such as no turnover/turnover) and one or more explanatory variables. The LR calculation is shown in Eq. 1 [9].

$$y = e^{\left(b_0 + \sum_{i=1}^{\rho} b_i, x_i\right)} / 1 + e^{\left(b_0 + \sum_{i=1}^{\rho} b_i, x_i\right)} \quad (1)$$

Where b_0 is the intercept and b_i is the regression coefficient for the input value of predictors x_i.

3.2 Decision Tree

One powerful technique commonly used in classification and as a predictor for decision making is a decision tree (DT). DT is a top-to-bottom approach that recursively splits data into smaller subsets, based on the attribute value of some test [10]. The partition is complete when the subsets include instances with pure values (homogenous) [11]. The type of DT algorithm that applied in this study was a classification tree (e.g., classification and regression tree (CART)).

3.3 Naïve Bayes

A Naïve Bayes (NB) is a probabilistic model based on the Bayes Theorem [12]. It assumes that the value of a particular attribute in a class is independent of the presence of another attribute [7]. The Bayes rule computes conditional probabilities of each feature for a given class and class probabilities.

The algorithm predicts by estimating the probability for a particular class by considering the maximum *a posteriori* probability according to Eq. 2 [13], where c is a class and x_j are features of vector x.

$$y = \arg \underset{C=\{0,1\}}{\text{MAX}}\ p(y) \prod_{j=1}^{n} p(x_j|y) \tag{2}$$

3.4 Support Vector Machine

Support vector machine (SVM) algorithm [2] attempts to distinctly separate data points of different classes by finding the optimal hyperplane in a high dimensional space.

The hyperplanes H1 and H2 that separate instances (data points) of the two classes are parallel, where the margin is the distance between them. The instances on the planes are called the support vectors, as illustrated in Fig. 1.

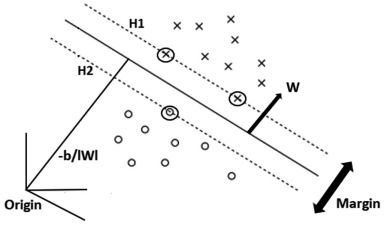

Fig. 1. SVM model [14].

The main benefit of using the SVM is its ability to use a kernel function, which can classify nonlinear data by mapping a space to a higher dimension, as given by Eq. 3 [2].

$$y = \text{sign}\left(\sum_{i=1}^{n} \alpha_i y_i k(x, x_i) + b\right) \tag{3}$$

Where $k(x,x_i)$ is the kernel function; α_i and b are free parameters. The kernel applied in this experiment was polynomial as it was considered the best method for tuning in this type of classification problem [2].

3.5 AdaBoost

AdaBoost is an abbreviation of Adaptive Boosting. AdaBoost is an ensemble method that aims build a strong learning algorithm from multiple, weak classifiers [12].

Initially, the same weight is set for all data points, and after each iteration the weights of the incorrectly classified observations are increased. The focus is on iteratively improving incorrectly classified instances (hard cases) with updated weights in the training set. The resultant model from this process combines the most effective classifiers and creates a stronger classifier, which has high accuracy on the training set and eventually improves prediction power [12].

4 Research Methodology

4.1 CRISP-DM Approach

This study follows the Cross-Industry Standard Process for DM (CRISP-DM) methodology, comprised of six phases, as illustrated in Fig. 2.

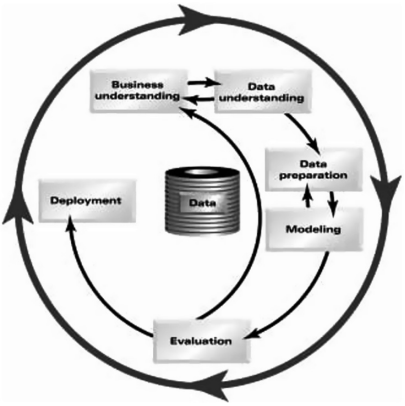

Fig. 2. Flowchart illustrating the CRISP-DM process [14].

In researching this topic, it is understood that this method is considered the *de facto* standard for planning successful DM projects [14]. As such, this methodology was chosen for use in this work.

4.2 Business Understanding

The first step in the CRISP approach is business understanding. This study aims to address the costly issue of employee turnover faced by many organisations. Employee turnover is a significant issue for several reasons, such as the financial losses and the time required to replace, hire and train new employees. This study aims, in part, to predict whether an employee will leave a company. Moreover, the work uncovers key factors, or predictors, that significantly contribute to employee turnover, which can help with developing retention planning strategies within HR management.

Predicting employee departure helps decision makers develop measures to identify areas for improvement, which is a crucial part of an organisational management system.

4.3 Data Understanding

The dataset examined in this paper was collected by Kaggle [15]. The dataset contains 14,999 data points and ten attributes. Class label is the 'Left' attribute; whether the employee departs the company or not. The ten attributes and their descriptions are provided in Table 1.

Table 1. Attribute description

Attributes	Description
Satisfaction level	Satisfaction level of employee; ranges from 0 to 1
Average monthly hours	Average numbers of hours worked by the employee in a month
Last evaluation	Evaluated performance of the employee by the company; ranges from 0 to 1
Number projects	The number of projects assigned to the worker
Promotion last 5 years	Whether the employee has had a promotion during the last five years
Time spent company	The number of years the staff member has worked at the company
Department	Employee's department/division
Work accident	Whether the employee has had an accident or not at work
Salary	Salary levels: low, medium or high relative to employees the company
Left	Whether the employee has left the job or not

Exploratory Data Analysis (EDA). EDA was employed for the goal of exploring potentially meaningful patterns determine the reasons for turnover, and the relationships between variables. This was done using graphical techniques. It enabled to examine the factors affecting staff so as to identify either the main causes of departure or any pre-existing symptoms of a problem. Another objective was defining the factors that would be useful for distinguishing between employees who leave their organisations and those who do not [6].

The dataset showed that the percentage of employees who remained in their jobs was much higher than those who left, at 76% and 24%, respectively. This indicated the class imbalance problem, which was addressed with the synthetic minority oversampling technique (SMOTE) in the data-preparation step.

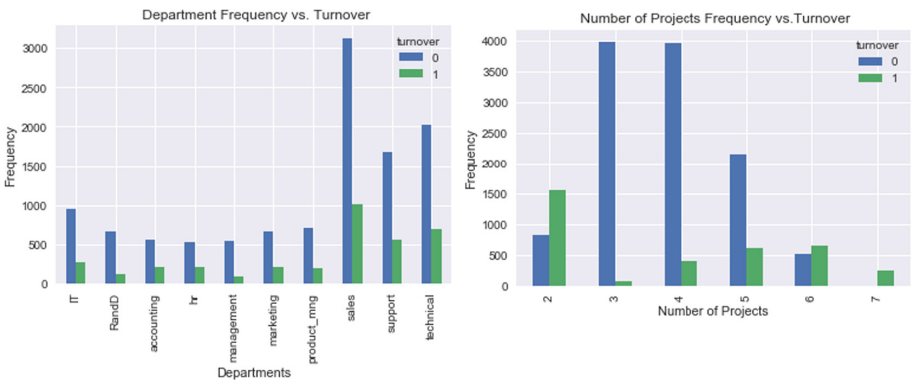

Fig. 3. Exploratory data analysis plots for the Department (a) and Number of Projects (b) features

Examining the graphs of the different variables, Fig. 3(b) shows that the employees with two, six and seven projects were more likely to leave the organisation, while the employees with three, four and five projects tended to stay with the company. A plot of various company departments is shown in Fig. 3(a), which clearly illustrates that the departments with the highest employee turnover were sales, technical, and support. In contrast, the management and research and development (R&D) departments experienced less turnover.

Survival Analysis. Survival analysis is a statistical technique used for analysis of time to event variables. It indicates the time until the occurrence of an event of interest; in this case, the event of interest is employee turnover [16]. The Kaplan–Meier (KM) approach [17] is one of the most important quantitative measures in survival analysis for estimating the survival curve. The survival function, *S(t)*, estimates the probability of survival at a certain time, *t*. It is defined as [17]:

$$S(t) = \prod_{j:t_j \leq t} \frac{n_j - d_j}{n_j} \tag{4}$$

Where n_j represents individuals at risk just before time t_j, and d_j is the number of turnover events at time t_j.

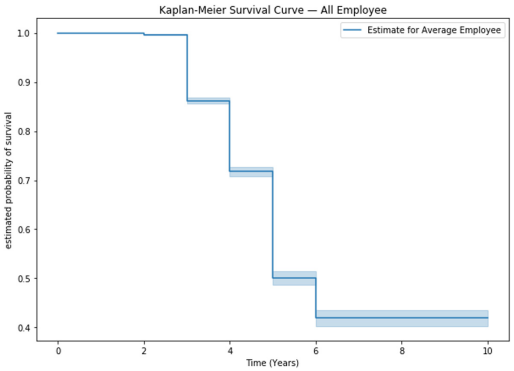

Fig. 4. Kaplan–Meier total survival curve for the first 10 years

As shown in Fig. 4, survival probability is 1.0 for the first three years, then drops to approximately 0.85, whereas 50% of the population, the employees, survive at five years.

The KM curve in Fig. 5 shows survival curves by salary level, labelled “low”, “medium”, and “high”. The graph illustrates that survival across all salary categories becomes vulnerable after six years. It is also clear that after two years, the survival probability of a lower-salaried employee exceeds that for employees in the other categories, where the probability of a low-salary employee surviving to five years exceeds 85%; for high-salary staff, the survival rate is about 78%, while for medium-salary staff,

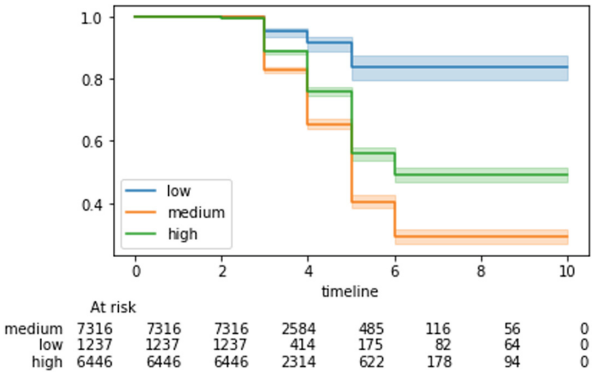

Fig. 5. Survival curve for different groups of employees based on salary level

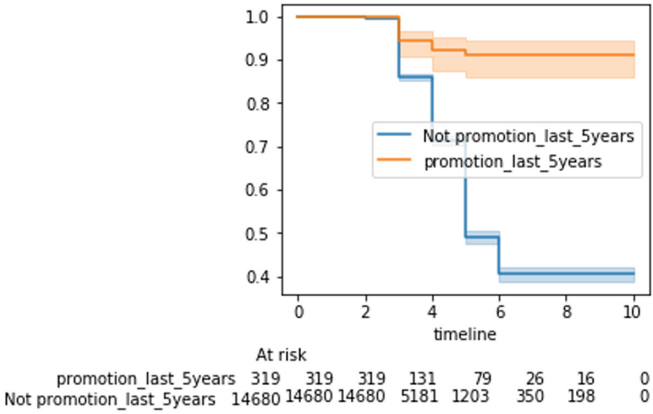

Fig. 6. Survival curve based on promotions variable for different groups of employees

the rate is approximately 60%. The curve in Fig. 6 illustrates the survival proportion of the group that had received promotions in the last 5 years, which was much higher than that of the no-promotion group. It can be observed that the promotion factor significantly extended the time until the employee's departure.

4.4 Data Preparation

Data preparation is an important step that is executed before modelling in order to maximize the chances of achieving accurate results. This experiment started with the application of a pre-processing phase on the HR dataset, which included checking for missing values, identifying outliers, feature correlation, data transformation and normalization. The employee dataset was considered imbalanced, as only 24% of the employees in the dataset left the company. The SMOTE approach was applied to address the class imbalance problem and was used only in the training set [18].

The experiments were executed on a personal computer (PC) containing an Intel Core i7-8550U 1.80 GHz CPU and 8.00 GB of RAM. A Jupyter Notebook was used as the primary working environment, with Python version 3.6 and its libraries such as matplotlib, lifeline, pandas and Scikit-learn.

Feature Selection. Feature selection (FS) technique is an essential process in ML, where the set of all possible features is reduced to those that are expected to contribute most to the prediction output [19]. The FS technique includes selecting a features subset by ignoring features that are irrelevant or redundant from the original feature set. FS therefore plays a significant role in building accurate models and identifying the critical features (best predictors) that lead to turnover. This study contributed to finding effective predictors in the HR dataset through application of various FS techniques; these techniques were the SelectKBest method using f_classif function based on analysis of variance (ANOVA) F-Test, the RFE-based logistic regression model and the RF technique based on training the RF algorithm. The five most important features obtained by the RF technique were ranked, as shown in Fig. 7. Table 2 defines the features that were selected by each of the three FS techniques.

Table 2. Best predictors obtained by various FS techniques

FS techniques	Features selected of FS techniques
SelectKBest (f_classif)	Satisfaction level, average monthly hours, work accident, salary promotion last 5 years, and time with company
RFE	Satisfaction level, time with company, last evaluation, promotion last 5 years, number of projects, salary, department R&D, department HR, work accident, and department management
RF model	Time with company, satisfaction level, average monthly hours, number of projects, and last evaluation

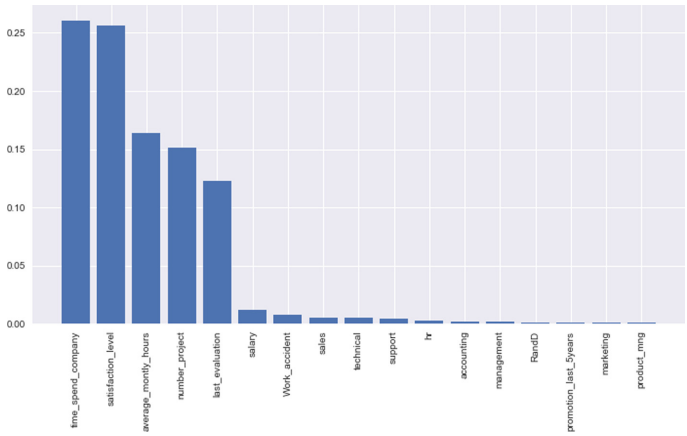

Fig. 7. Features importance as ranked by the RF model

4.5 Modelling

In this phase, a set of candidate models has been prepared to find the best performance. The literature suggests that no single classifier achieves the best results in all contexts, therefore it is necessary to investigate which classifiers are more suited to the specific data being analysed. Several classifiers can be applied to predict employee attrition, such as the SVM, naïve Bayes, Logistic Regression, DT and AdaBoost ensemble classifiers.

Moreover, each model can be applied to three different feature sets, as established by the different FS techniques to determine the best combination of model and feature set. In this work, each candidate model was tuned with hyperparameters using the GridSearchCV function. Grid Search [20] is an approach to find the optimal parameters to models that trained. The values of the hyperparameters obtained via Grid Search were: {'C': 0.5, 'gamma': 1.0, 'kernel': 'poly'} for SVM, and {'criterion': 'entropy', 'max_depth': 8, 'max_features': 0.5} for DT. AdaBoost used {'learning_rate': 0.1, 'n_estimators': 500}.{'C': 1.0, 'fit_intercept': True, 'penalty': 'l1'} was used for the LR model.

4.6 Evaluation

To determine the performance of each model, the classifier was evaluated by use of the appropriate metrics. In the experiment, the performance of the predictive models was evaluated using several critical metrics that included accuracy, sensitivity, precision, specificity, F1-score, misclassification rate, and AUC value.

5 Results

5.1 The Train-Validate-Test Split

In this experiment, the dataset was divided into three subsets: training, validating and testing. The training set was used to train the classifier, while the validating set was applied to optimize the model. Finally, the testing set was used to evaluate the ability of a classifier to predict given new, unseen data [21]. Generalization on testing data indicates how well the models perform to new data. The generalization refers to the ability of the ML model that trained to perform well on new unseen data (e.g., testing data). The results in Table 3 show the performance of each of the five models paired with the different FS techniques.

Overall, SVM with RF (feature importance) achieved the best performance by most metrics as compared with the other models, with an accuracy of 0.97. Decision Tree-RFE followed in performance, with an accuracy score of 0.967, and the DT-RF model was next, with an accuracy of 0.96. The AdaBoost-RF classifier followed, with an accuracy score of 0.95. The LR and naïve Bayes classifiers were the weakest performers by all measures.

SVM-RF performed better than the other models by most metrics, including accuracy, precision, specificity, F1-score, misclassification rate and AUC value. The only metric by which SVM-RF was not superior was the sensitivity score, where the

Table 3. Algorithm performance results using various feature selection techniques

Classifier	FS methods	Accuracy	Precision	Sensitivity	Specificity	F1	Mis-classification	AUC
LR	SelectKBest	0.75	0.702	0.785	0.739	0.71	0.25	0.76
	RFE	0.751	0.687	0.687	0.771	0.698	0.25	0.73
	RF	0.737	0.687	0.753	0.732	0.694	0.26	0.74
NB	SelectKBest	0.737	0.688	0.758	0.73	0.695	0.26	0.74
	REF	0.711	0.674	0.771	0.692	0.674	0.29	0.73
	RF	0.816	0.754	0.49	0.919	0.722	0.18	0.70
SVM	SelectKBest	0.942	0.916	0.90	0.955	0.922	0.6	0.93
	REF	0.959	0.945	0.911	0.974	0.944	0.4	0.94
	RF	**0.971**	**0.964**	0.922	**0.986**	**0.961**	**0.3**	**0.95**
DT	SelectKBest	0.954	0.932	**0.931**	0.962	0.939	0.5	0.95
	REF	0.967	0.959	0.916	0.983	0.954	0.3	0.95
	RF	0.964	0.949	0.927	0.975	0.95	0.4	0.95
AdaBoost	SelectKBest	0.948	0.936	0.87	0.973	0.928	0.5	0.92
	REF	0.944	0.932	0.853	0.973	0.922	0.6	0.91
	RF	0.951	0.941	0.872	0.977	0.932	0.5	0.92

DT model using SelectKBest feature selection performed better, with a score of 0.93 versus 0.92.

5.2 10-Fold Cross-Validation

10-fold cross-validation was applied to the dataset to attempt to determine if the model results would improve. The dataset was divided into ten equal subsets [7]. Nine subsets were used to train the model, validating the model with the remaining fold (one of the subsets). Then, the average of all 10 trials were computed for a final estimation of performance [7]. 10-fold cross-validation was used this research due to its powerful method to prevent overfitting by splitting the data into two parts (train and test), applying the cross-validation only to the training set and leaving out the testing set to evaluate the model's ability to predict the unseen dataset.

The experimental results using cross-validation identified the two best models, as illustrated in Table 4. The SVM-RF model achieved the highest accuracy (0.97), precision (0.94), specificity (0.98), F1-score (0.93) and lowest misclassification rate (0.3). DT using SelectKBest performed best in sensitivity (0.92) and AUC value (0.96). In general, model performance was little changed between the train-validate-test and cross-validation methods. However, the logistic regression classifier did produce better results using cross-validation; the accuracy of LR-RF increased from 73% to 83%. Also, misclassification of LR improved from 25% to 17% across all feature sets.

Table 4. Algorithm results using 10-fold cross-validation and FS methods

Classifier	FS methods	Accuracy	Precision	Sensitivity	Specificity	F1	Mis-classification	AUC
LR	SelectKBest	0.830	0.596	0.886	0.812	0.713	0.17	0.75
	RFE	0.830	0.596	0.886	0.812	0.713	0.17	0.75
	RF	0.833	0.593	0.940	0.797	0.730	0.17	0.77
NB	SelectKBest	0.803	0.567	0.739	0.824	0.642	0.20	0.74
	REF	0.794	0.541	0.878	0.768	0.670	0.21	0.73
	RF	0.749	0.458	0.816	0.892	0.356	0.25	0.71
SVM	SelectKBest	0.938	0.853	0.891	0.952	0.872	0.6	0.94
	REF	0.951	0.897	0.899	0.968	0.898	0.5	0.95
	RF	**0.966**	**0.944**	0.913	**0.983**	**0.928**	**0.3**	0.95
DT	SelectKBest	0.948	0.870	**0.918**	0.957	0.894	0.5	**0.96**
	REF	0.962	0.930	0.909	0.979	0.920	0.4	0.95
	RF	0.947	0.873	0.913	0.958	0.892	0.5	0.96
AdaBoost	SelectKBest	0.945	0.902	0.862	0.971	0.882	0.6	0.93
	REF	0.944	0.912	0.848	0.974	0.879	0.6	0.92
	RF	0.954	0.918	0.888	0.975	0.903	0.5	0.93

5.3 Decision Tree Result

Figure 8 illustrates the constructed DT model, built with the five most important features, as defined through RF feature selection. This representation can facilitate the decision-making process by organizing features in order of condition [10]. The condition represents a subset of values for a given attribute in order to classify the chance of turnover as a series of if-then-else decision rules. According to the results of the DT classifier, the following rules can be extracted:

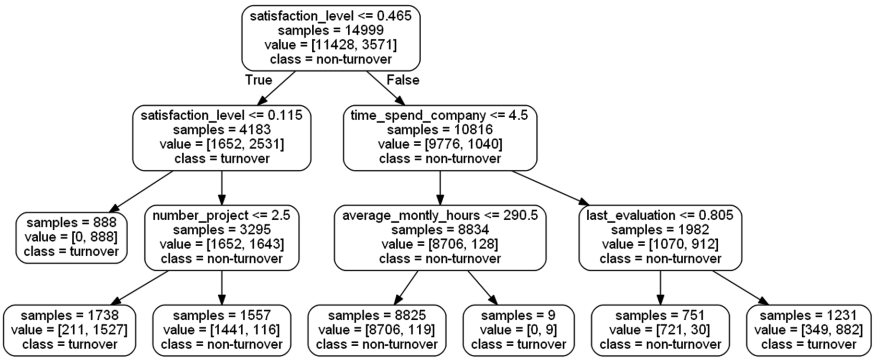

Fig. 8. Decision tree result

Rule 1: If (satisfaction_level <= 0.115), THEN Class = Turnover
Rule 2: If (satisfaction_level <= 0.465 and satisfaction_level > 0.115 and number_project <= 2.5), THEN Class = Turnover
Rule 3: If (satisfaction_level <= 0.465 and satisfaction_level > 0.115 and number_project > 2.5), THEN Class = Non-Turnover
Rule 4: If (satisfaction_level > 0.465 and time_spend_company <= 4.5 and average_montly_hours <= 290.5), THEN Class = Non-Turnover
Rule 5: If (satisfaction_level > 0.465 and time_spend_company <= 4.5 and average_montly_hours > 290.5), THEN Class = Turnover
Rule 6: If (satisfaction_level > 0.465 and time_spend_company > 4.5 and last_evaluation <= 0.805), THEN Class = Non-Turnover
Rule 7: If (satisfaction_level > 0.465 and time_spend_company > 4.5 and last_evaluation > 0.805), THEN Class = Turnover.

6 Conclusion

This study proposed supervised ML models, which aimed to predict whether an employee would leave a company or not. The experiment was conducted with a combination of five predictive models (SVM, DT, NB, LR and the Adaboost classifier) using three different FS techniques (SelectKBest, RFE, and RF), while considering both train-validate-test and 10-fold cross-validation methods. Several important evaluation metrics were examined to analyse the performance of the supervised machine learning techniques. These metrics were accuracy, sensitivity, precision, specificity, F1-score, misclassification rate and AUC value.

The feature selection approaches were implemented within the HR dataset to identify the best predictors that lead to turnover. Finding the best predictors would also help an employer explore how to retain employees with high satisfaction and performance.

In the context of employee turnover prediction, sensitivity is one of the most critical metrics, as correctly identifying the potential staff that may leave the company is critical in the decision-making procedure. In this case, the decision maker would be able to provide a better solution and implement some form of retention planning as early as possible. Survival analysis was used to estimate patterns of time-to-the-event (i.e., the decision to leave), which was useful in measuring the length of time employees remained before finally leaving. This feature also allowed for the creation of a summary of survival data and made it possible to create comparisons across different situations. Within the context of this study and the dataset, satisfaction level, number of projects, last evaluation, time spent with company and average monthly hours were the most significant factors, as identified by the RF technique. This technique performed the best with the predictive models, followed by the SelectKBest and the RFE FS techniques.

The experimental results identified two models as most able to predict whether employee turnover: SVM using the RF algorithm and DT with SelectKBest method. It was found that when using ten-fold cross-validation, the SVM-RF algorithm performed

best (accuracy = 0.97, precision = 0.94, specificity = 0.98, F1-score = 0.93, misclassification error = 0.3). While the DT considering SelectKBest performed better than other classifiers (sensitivity = 0.92 and AUC = 0.96).

On the other hand, when using the train-validate-test approach, the SVM-RF model provided very good indicators for predicting turnover. It performed best compared to other models in term of accuracy, precision, specificity, F1-score, misclassification rate and AUC value. The DT using SelectKBest method performed significantly better in sensitivity (i.e., the prediction of true turnover cases) with 0.93.

In future research, the best models identified in this study will be applied to various datasets (such as those containing the multi-class problem) in order to extend the performance evaluation in predicting employee turnover. Deep learning could be used to improve the results of the predictive models. Also, the genetic algorithm could be examined for feature selection in order to increase predictive model performance.

References

1. Sikaroudi, E., et al.: A data mining approach to employee turnover prediction (case study: Arak automotive parts manufacturing). J. Ind. Syst. Eng. **8**(4), 106–121 (2015)
2. Keramati, A., et al.: Improved churn prediction in telecommunication industry using data mining techniques. Appl. Soft Comput. **24**, 994–1012 (2014)
3. Fan, C.-Y., et al.: Using hybrid data mining and machine learning clustering analysis to predict the turnover rate for technology professionals. Expert Syst. Appl. **39**(10), 8844–8851 (2012)
4. Chien, C.-F., Chen, L.-F.: Data mining to improve personnel selection and enhance human capital: a case study in high-technology industry. Expert Syst. Appl. **34**(1), 280–290 (2008)
5. Saradhi, V.V., Palshikar, G.K.: Employee churn prediction. Expert Syst. Appl. **38**(3), 1999–2006 (2011)
6. Hung, S.-Y., Yen, D.C., Wang, H.-Y.: Applying data mining to telecom churn management. Expert Syst. Appl. **31**(3), 515–524 (2006)
7. Valle, M.A., Ruz, G.A.: Turnover prediction in a call center: behavioral evidence of loss aversion using random forest and Naïve Bayes algorithms. Appl. Artif. Intell. **29**(9), 923–942 (2015)
8. García, D.L., Nebot, À., Vellido, A.: Intelligent data analysis approaches to churn as a business problem: a survey. Knowl. Inf. Syst. **51**(3), 719–774 (2017)
9. Rombaut, E., Guerry, M.-A.: Predicting voluntary turnover through human resources database analysis. Manag. Res. Rev. **41**(1), 96–112 (2018)
10. Lima, E., Mues, C., Baesens, B.: Domain knowledge integration in data mining using decision tables: case studies in churn prediction. J. Oper. Res. Soc. **60**(8), 1096–1106 (2017)
11. De Caigny, A., Coussement, K., De Bock, K.W.: A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. Eur. J. Oper. Res. **269**(2), 760–772 (2018)
12. Vafeiadis, T., et al.: A comparison of machine learning techniques for customer churn prediction. Simul. Model. Pract. Theory **55**, 1–9 (2015)
13. Valle, M.A., Varas, S., Ruz, G.A.: Job performance prediction in a call center using a Naive Bayes classifier. Expert Syst. Appl. **39**(11), 9939–9945 (2012)
14. Shearer, C.: The CRISP-DM model: the new blueprint for data mining. J. Data Warehous. **5**(4), 13–22 (2000)

15. Kaggle. HR Analytics (2017). https://www.kaggle.com/colara/hr-analytics
16. Sainani, K.L.: Introduction to Survival Analysis. PM R **8**(6), 580–585 (2016)
17. Kartsonaki, C.: Survival analysis. Diagn. Histopathol. **22**(7), 263–270 (2016)
18. Amin, A., et al.: Comparing oversampling techniques to handle the class imbalance problem: a customer churn prediction case study. IEEE Access **4**, 7940–7957 (2016)
19. Jain, D., Singh, V.: Feature selection and classification systems for chronic disease prediction: a review. Egypt. Inform. J. **19**(3), 179–189 (2018)
20. Gao, X., Hou, J.: An improved SVM integrated GS-PCA fault diagnosis approach of Tennessee Eastman process. Neurocomputing **174**, 906–911 (2016)
21. Moosavi, M., Soltani, N.: Prediction of the specific volume of polymeric systems using the artificial neural network-group contribution method. Fluid Phase Equilib. **356**, 176–184 (2013)