



Deep Learning Based Employee Attrition Prediction

Kerem Gurler^{1(✉)}, Burcu Kuleli Pak¹, and Vehbi Cagri Gungor²

¹ adesso Turkey, Istanbul, Turkey

{kerem.gurler,burcu.kuleli}@adesso.com.tr

² Abdullah Gul University, Kayseri, Turkey

cagri.gungor@agu.edu.tr

Abstract. Employee attrition is a critical issue for the business sectors as leaving employees cause various types of difficulties for the company. Some studies exist on examining the reasons for this phenomenon and predicting it with Machine Learning algorithms. In this paper, the causes for employee attrition is explored in three datasets, one of them being our own novel dataset and others obtained from Kaggle. Employee attrition was predicted with multiple Machine Learning and Deep Learning algorithms with feature selection and hyperparameter optimization and their performances are evaluated with multiple metrics. Deep Learning methods showed superior performances in all of the datasets we explored. SMOTE Tomek Links were utilized to oversample minority classes and effectively tackle the problem of class imbalance. Best performing methods were Deep Random Forest on HR Dataset from Kaggle and Neural Network for IBM and Adesso datasets with F1 scores of 0.972, 0.642 and 0.853, respectively.

Keywords: Machine learning · Deep learning · Employee attrition · Data imbalance

1 Introduction

Employees unexpectedly leaving their company is a crucial problem in many business sectors today. There may be various reasons affecting people's decision to leave their company such as working overtime for long periods of time, finding another job that pays higher wage, etc. This poses a major problem for companies because employees leaving can cause projects to be interrupted or slowed down, therefore harming the company. Even if companies can quickly replace the workers, adaptation time for new employees will potentially decrease the overall work efficiency. Recently, companies have started to use statistical methods to prevent employee attrition. They also use predictive machine learning models to determine which employees might leave.

In this paper, we worked on three different datasets to analyze the reasons of employee attrition. These datasets are IBM Human Resources (HR) Dataset, another anonymous HR dataset from Kaggle and finally our own dataset collected in Adesso Turkey HR department. Data is obtained in an anonymized way not to violate employee privacy. We also built a predictive model with machine learning (ML) methods. Evaluated ML methods are Support Vector Machines (SVM), Linear Discriminant Analysis (LDA), Logistic Regression (LR), Random Forest (RF), K Nearest Neighbours (KNN), Naive Bayes (NB), AdaBoost (AB), XG Boost (XGB), Deep Random Forest [1] (DRF) and Artificial Neural Networks (ANN). In addition, feature importance scores were calculated with permutation importances using a random forest classifier. Features with negative or near zero scores were dropped for training. Hyperparameter optimization with Bayesian Search and cross validation was done to optimize classification performance.

It is important to note that the Kaggle and IBM HR datasets were very imbalanced. Attrition data in general is very likely to be imbalanced as people that left the company will increase over time or in the case of a fast growing company, active workers may increase rapidly thus, resulting in large number of negative attrition values. In order to handle imbalance SMOTE and Tomek Links were utilized [2,3]. Novel contributions that we provide are as follows:

- A novel dataset is obtained at Adesso company and employee attrition analysis and prediction are conducted on this dataset.
- Three different datasets are used for employee attrition analysis and Deep Learning methods showed the best performance for all of them. Therefore, deep learning approach can be utilized by other companies to effectively deal with problem of employee attrition prediction.
- Comparative performance results showed that Neural Network method performed better compared to existing studies conducted based on IBM dataset.

This paper is organized as follows. Related work on employee attrition studies are analyzed in Sect. 2. Datasets that are used are explained in Sect. 3. In Sect. 4, algorithms used in the paper and their results on all datasets are discussed. Finally in Sect. 5, the paper is concluded.

2 Related Work

Yadav et al. [4] worked on the Kaggle HR Dataset listed above. They used Recursive Feature Elimination with Cross Validation for feature selection. This approach uses various subsets of features to determine the best set of features. They applied LR, SVM, RF, Decision Tree (DT) and AdaBoost (AB). Best performing method was Random Forest with feature selection for both accuracy and F1 score metrics.

Another study was conducted on IBM HR dataset and a dataset from a bank by Zhao et al. [5] They achieved highest accuracy with LR, precision with

LDA, Recall and F1 score with a Neural Network (NN) and AUC with Gradient Boosting on IBM data with 1500 samples. On 1000 sample bank dataset, RF showed the best precision while XGB was best in the rest of the metrics. Some of the top feature importance scores for the bank data obtained with XGB are shown in Table 1.

Table 1. Feature Importances of Bank Data [5]

Feature	Importance Score
Last Pay Raise	214
Job Tenure	77
Age	63
Compensation	40
Specialized Area	36
Department	31
Education Background	31

Qutub et al. [6] applied DTs, RF, LR, GB, AB and Stochastic Gradient Descent (SGD) and also pairwise ensembles of some of these methods and found that Logistic Regression alone performed best. Another study conducted by Ozdemir. et al. [11] also identified Logistic Regression as the best method with accuracy of 0.871. In Table 2, some previous studies conducted on IBM data were listed.

Table 2. Comparison of best method of studies on IBM Dataset

Authors	Year	Method	Accuracy	Precision	Recall	F1 Score	AUC
Yigit et al. [7]	2017	Support Vector Machines	0.897	0.98	0.37	0.53	
Boomhower et al. [8]	2018	Logistic Regression	74.59	-	-	-	-
Qutub et al. [6]	2021	Logistic Regression	0.88	0.74	0.46	0.57	0.8593
Fallucchi et al. [9]	2020	Logistic Regression	0.875	0.663	0.337	0.445	-
Proposed Method	2023	ANN	0.885	0.66	0.625	0.642	0.831

Existing studies give insights on causes of employee attrition and predict employee attrition with various models. However, deep learning methods are usually not used or were not very effective. In this study, we utilize two deep learning methods DRF and ANN and show that they outperformed other machine learning models.

3 Dataset

HR Analytics dataset obtained from Kaggle has 15000 samples with 3571 of them leaving the company. IBM HR Analytics is a synthetic dataset which has 35 features, 1470 samples and 237 of them have positive attrition value. Lastly, our own dataset collected in Adesso Turkey has 1087 samples and 18 features with 569 positive attrition. Permutation importances were used with base classifiers on three different datasets and results can be seen in Fig. 1, 2 and 3.

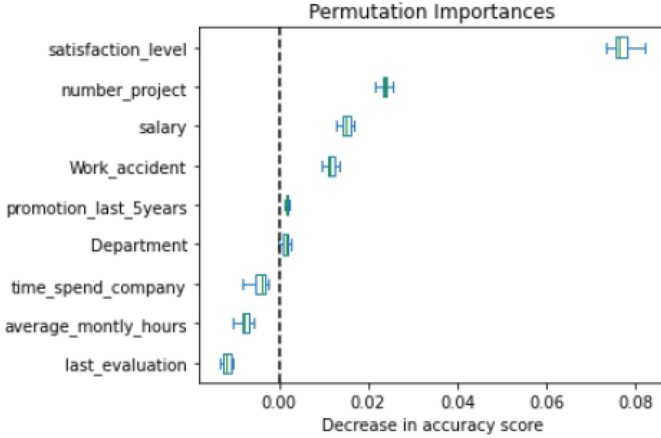


Fig. 1. Feature Importances on Kaggle HR Analytics Dataset

Overall, satisfaction level of employees, number of projects assigned, salary and whether they had a work accident is observed as a most important attrition factor for the first dataset. On IBM dataset, working overtime was the most important followed by income, distance from home, age, years at company and so on. City of residence in Adesso data had the highest score. This is caused by the fact that most of the employees working outside İstanbul left the company. We also see total experience, Adesso experience, graduated university, age and marital status have considerable impacts on employee attrition.

We can see that similar features are important across the datasets. For instance, income and satisfaction levels in first two datasets have high scores. We also observe evaluation scores of employees, gender and department had little to no effect on attrition across our datasets. Both IBM and Adesso datasets show that attrition is highly dependent on how long the employee was working in the company and also in their overall career. Employees with lower experience and age tended to leave more than the others in Adesso, however the opposite case was seen in IBM. Distance from home in IBM and city of residence in Adesso are similar features and they are both important.

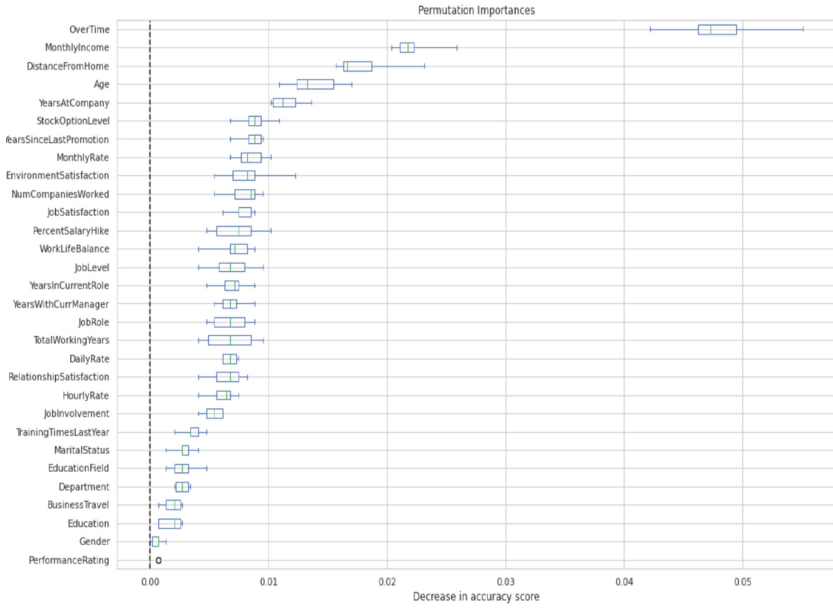


Fig. 2. Feature Importances on IBM HR Dataset

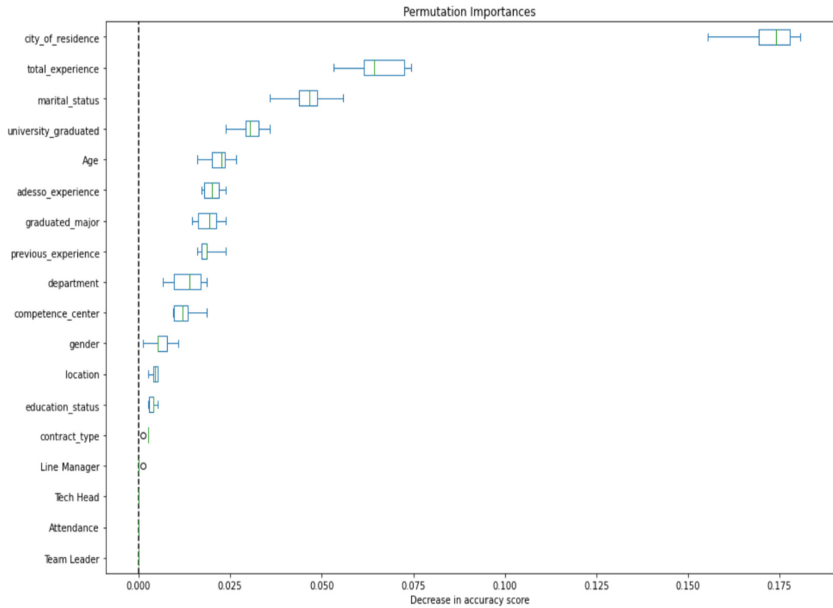


Fig. 3. Feature Importances on Adesso HR Dataset

4 Methodology

4.1 Handling Data Imbalance

In both IBM and other Kaggle HR datasets, there is a considerable data imbalance that needs to be handled. Oversampling and undersampling are two main approaches to data imbalance and for these datasets, oversampling is more compatible because the sample count is not very high. There are multiple ways of oversampling a minority class. Chawla et al. [3] proposed SMOTE method for this problem which works by creating synthetic samples instead of repeating the same examples multiple times. For every minority class sample, k amount of neighbours are selected and new samples are created in their direction.

Another method SMOTE with Tomek Links is proposed by Batista et al. [10] which utilizes smote for oversampling the minority class, but also uses Tomek links on over sampled data for cleaning the data and preventing overfitting. We applied SMOTE, Random Oversampling and SMOTE with Tomek Link on our data and observed great improvements to the performance with SMOTE Tomek Links.

4.2 Methods

In this study, we also applied various traditional machine learning and statistical methods and also two deep learning approaches, namely Deep Random Forests and a feed forward neural network on our three datasets.

Datasets were split as 70% train, 15% validation and 15% test sets. Different balancing strategies were tested on train dataset and feature selection was performed after determining the best balancing strategy. Hyperparameter tuning was performed with validation and train set and final model classification performance evaluation was completed. Figure 4 shows the overall workflow diagram. Metrics utilized for evaluation are described in the equations below.

$$Precision = \frac{TruePositive}{TruePositive + FalsePositive} \quad (1)$$

$$Recall = \frac{TruePositive}{TruePositive + FalseNegative} \quad (2)$$

$$F1Score = 2 * \frac{precision * recall}{precision + recall} \quad (3)$$

$$Specificity = \frac{TrueNegative}{TrueNegative + FalsePositive} \quad (4)$$

$$AUC = \int_0^1 sensitivity(Specificity^{-1}(x)) d(x) \quad (5)$$

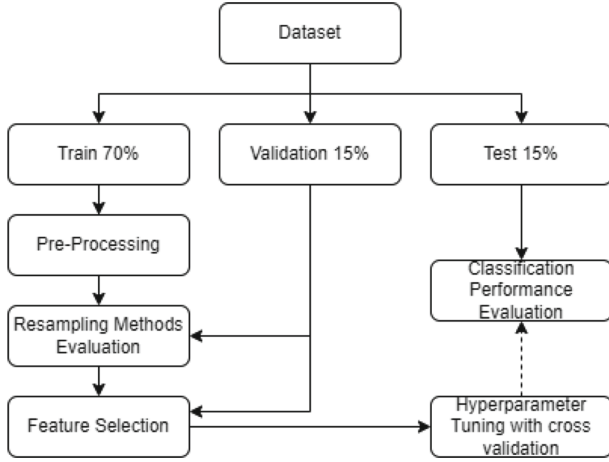


Fig. 4. Proposed Method Diagram

4.3 Hyperparameter Optimization

Hyperparameter optimization was applied with Bayesian search on ML models with search spaces that are commonly used for each algorithm. Optimization process was applied manually on Deep Random Forest and Neural Network models. For Neural Networks; various layer sizes, number of layers, activation functions, loss functions, optimizers, learning rates, weight initializations and regularizations were tested.

The final architecture used for the first dataset is 3 hidden layers with 64 neurons with relu activation, output layer with sigmoid, l2 regularization on layers with alpha $1e-3$, node dropout of $5e-2$ and uniform weight initialization in hidden layers, Xavier on output layer. Training was done with Adam optimizer with learning rate of $1e-2$, batch size of 1024 and early stopping patience of 30.

For IBM, network of 4 layers of 128 neurons with l2 regularization with alpha $1e-3$, dropout of $1e-1$ with tanh activation in hidden layers and sigmoid at output and Xavier uniform initialization for weights was trained with Adam optimizer, mini batches of 256, learning rate of $1e-2$ and early stopping patience of 30.

For Adesso data, 3 layers of 256 neurons with tanh activation and output neuron with sigmoid were used with Xavier uniform initialization, l2 regularization with alpha of $1e-1$ and dropout of $1e-1$. Adam optimizer with full batch, learning rate of $1e-3$ and early stopping patience of 50 was used at training.

Binary cross entropy was used as a loss function for all datasets. For imbalanced datasets, binary cross entropy with weights based on class ratios was tested and although showing improvement when oversampling is not used, SMOTE Tomek link oversampling with normal binary cross entropy loss performed better.

4.4 Results

Experiment results on all datasets are shown in Tables 3, 4 and 5. Deep Random Forest showed the best scores in all metrics in Kaggle HR Dataset which is followed by XGB, RF and ANN by their F1 score. F1 is a critical metric in both IBM and Kaggle HR datasets because of their highly imbalanced distribution.

On IBM data, ANN showed the best accuracy and F1 score among all methods. SVM and KNN classifiers had the best precision and recall, respectively.

ANN performed the best in terms of F1 score, precision and accuracy. Deep RF showed slightly higher AUC score and slightly lower F1 score. NB performed the worst among all three datasets with a high margin.

Table 3. Results on Kaggle HR Dataset

Method	Accuracy	Precision	Recall	F1 Score	AUC	Specificity
RF	0.984	0.984	0.947	0.965	0.99	0.995
SVC	0.935	0.88	0.84	0.862	0.953	0.964
LDA	0.763	0.502	0.804	0.619	0.827	0.75
KNN	0.92	0.88	0.834	0.834	0.959	0.948
Bagging	0.935	0.875	0.843	0.862	0.92	0.964
AdaBoost	0.956	0.929	0.882	0.905	0.98	0.978
XGBoost	0.986	0.988	0.953	0.97	0.991	0.996
LR	0.771	0.514	0.81	0.629	0.828	0.759
NB	0.49	0.3	0.85	0.445	0.795	0.375
DRF	0.987	0.99	0.955	0.972	0.99	0.997
ANN	0.965	0.961	0.892	0.925	0.983	0.988

Table 4. Results on IBM HR Dataset

Method	Accuracy	Precision	Recall	F1 Score	AUC	Specificity
RF	0.864	0.647	0.392	0.488	0.783	0.957
SVC	0.894	0.812	0.464	0.59	0.849	0.978
LDA	0.882	0.9	0.321	0.473	0.824	0.992
KNN	0.62	0.284	0.857	0.426	0.738	0.573
Bagging	0.882	0.807	0.375	0.512	0.757	0.982
AdaBoost	0.858	0.571	0.571	0.571	0.813	0.915
XGBoost	0.852	0.55	0.589	0.568	0.801	0.904
LR	0.876	0.652	0.535	0.588	0.833	0.943
NB	0.676	0.275	0.589	0.375	0.696	0.693
DRF	0.87	0.636	0.5	0.56	0.8	0.943
ANN	0.885	0.66	0.625	0.642	0.831	0.936

Table 5. Results on Adesso HR Dataset

Method	Accuracy	Precision	Recall	F1 Score	AUC	Specificity
RF	0.802	0.802	0.821	0.811	0.87	0.782
SVC	0.697	0.676	0.797	0.732	0.84	0.589
LDA	0.771	0.783	0.773	0.778	0.833	0.769
KNN	0.802	0.77	0.88	0.822	0.862	0.717
Bagging	0.641	0.603	0.904	0.723	0.801	0.358
AdaBoost	0.777	0.76	0.833	0.795	0.837	0.717
XGBoost	0.796	0.793	0.821	0.807	0.864	0.769
LR	0.746	0.752	0.642	0.734	0.819	0.884
NB	0.759	0.857	0.589	0.375	0.696	0.693
DRF	0.839	0.822	0.88	0.85	0.928	0.794
ANN	0.845	0.839	0.869	0.853	0.912	0.82

Random oversampling, SMOTE and SMOTE with Tomek Links are applied for oversampling and balancing the class distributions on datasets except Adesso, which is already balanced. The results are compared with Logistic Regression as a base method in Tables 6 and 7. On IBM dataset, SMOTE with Tomek Links showed the best performance. Pure SMOTE was the worst performing one in terms of F1 score. On Kaggle HR dataset, three oversampling methods showed similar performance, but SMOTE methods were slightly better than Random Oversampling. Accuracy metric tends to be inflated in imbalanced datasets, because model can learn to mostly predict majority class. This is the reason Specificity and therefore, accuracy is higher with no oversampling in IBM dataset.

Table 6. Tests of Resampling Methods on IBM

Method	Accuracy	Precision	Recall	F1 Score	AUC	Specificity
No Oversampling	0.873	0.600	0.375	0.462	0.758	0.957
Random	0.745	0.300	0.562	0.391	0.747	0.777
SMOTE	0.841	0.444	0.375	0.407	0.723	0.920
SMOTE + Tomek Link	0.873	0.577	0.469	0.517	0.804	0.941

Table 7. Tests of Resampling Methods on Kaggle HR

Method	Accuracy	Precision	Recall	F1 Score	AUC	Specificity
No Oversampling	0.779	0.560	0.313	0.401	0.813	0.924
Random	0.759	0.495	0.798	0.611	0.822	0.747
SMOTE	0.761	0.498	0.807	0.616	0.823	0.747
SMOTE + Tomek Link	0.761	0.498	0.807	0.616	0.822	0.747

In addition, feature importances were calculated with permutation importance scores. Effect of feature selection was measured on a logistic regression base model and features were dropped iteratively from lowest to highest score until performance drops. For IBM, Performance Rating and Gender, for Kaggle data last evaluation and for Adesso data team leader, attendance, tech head, line manager and contract type features were dropped.

On IBM dataset, feature selection resulted in slight improvement for all five metrics with considerable around 6% increase in F1, recall and precision. For other datasets a notable improvement is not observed. Results are shown in Tables 8, 9 and 10.

Table 8. Feature Selection on IBM

Method	Accuracy	Precision	Recall	F1 Score	AUC	Specificity
Without Feature Selection	0.873	0.577	0.469	0.517	0.804	0.941
With Feature Selection	0.886	0.630	0.531	0.576	0.811	0.947

Table 9. Feature Selection on Kaggle HR

Method	Accuracy	Precision	Recall	F1 Score	AUC	Specificity
Without Feature Selection	0.761	0.498	0.807	0.616	0.822	0.747
With Feature Selection	0.763	0.501	0.792	0.613	0.822	0.754

Table 10. Feature Selection on Adesso

Method	Accuracy	Precision	Recall	F1 Score	AUC	Specificity
Without Feature Selection	0.778	0.736	0.848	0.788	0.827	0.711
With Feature Selection	0.778	0.736	0.848	0.788	0.831	0.711

5 Conclusion

In this paper, employee attrition was predicted with multiple Machine Learning and Deep Learning algorithms with feature selection and hyperparameter optimization and their performances are evaluated with multiple metrics. ANN in IBM dataset and Adesso HR dataset, and Deep RF in Kaggle HR dataset showed the best overall performance considering all metrics. In the first two datasets, positive attrition samples are minority, therefore specificity and accuracy values tend to be high. Precision and recall are more important for our use case and F1 score which gives as harmonic average of the two metric evaluates the overall performance of our models.

Multiple other studies on IBM dataset were also analyzed and the proposed method performed better than existing methods the literature. Our experiments

suggest that deep learning methods are promising for the problem of predicting employee attrition.

Balancing the data with first oversampling the minority class with SMOTE and undersampling it with Tomek Links was very effective for our datasets and improved training performance considerably. SMOTE with Tomek Links performed better than no oversampling, random oversampling and SMOTE in IBM dataset. SMOTE Tomek Links and SMOTE only showed the best performance on Kaggle HR followed by random oversampling and no oversampling.

We also observed that similar features across different datasets showed similar permutation importance ranks. This shows us which factors should or should not be considered for employee attrition problem. Features, such as income, working overtime, experience and age, are observed to be important factors for employee attrition, whereas performance evaluation and gender were not critical features in multiple datasets.

In the future, Adesso HR dataset can be expanded with salary level or employee satisfaction values, since they proved to be a strong predictor for employee attrition in other two datasets. Experiments also demonstrate that Deep Learning approaches show very promising results for predicting employee attrition and can be studied further in the future with new models, architectures and approaches.

References

1. Zhou, Z.-H., Feng, J.: Deep Forest: towards an alternative to deep neural networks. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence 2017, pp. 3553–3559. <https://doi.org/10.24963/ijcai.2017/497>
2. Tomek, I.: Two modifications of CNN. *IEEE Trans. Syst. Man Cybern.* **6**, 769–772 (1976)
3. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **16**, 321–357 (2002)
4. Yadav, S., Jain, A., Singh, D.: Early prediction of employee attrition using data mining techniques. In: 2018 IEEE 8th International Advance Computing Conference (IACC) (2018). <https://doi.org/10.1109/iadcc.2018.8692137>
5. Zhao, Y., Hryniewicki, M.K., Cheng, F., Fu, B., Zhu, X.: Employee turnover prediction with machine learning: a reliable approach. In: Intelligent Systems and Applications, pp. 737–758 (2018). https://doi.org/10.1007/978-3-030-01057-7_56
6. Qutub, A., Al-Mehmadi, A., Al-Hssan, M., Aljohani, R., Alghamdi, H.: Prediction of employee attrition using machine learning and ensemble methods. *Int. J. Mach. Learn. Comput.* **11**, 110–114 (2021). <https://doi.org/10.18178/ijmlc.2021.11.2.1022>
7. Yiğit, I.O., Shourabizadeh, H.: An approach for predicting employee churn by using data mining. In: International Artificial Intelligence and Data Processing Symposium (IDAP), Malatya, Turkey, 16–17 September 2017. IEEE (2017). <https://doi.org/10.1109/IDAP.2017.8090324>
8. Frye, A., Boomhower, C., Smith, M., Vitovsky, L., Fabricant, S.: Employee attrition: what makes an employee quit? *SMU Data Sci. Rev.* **1** (2018). Article 9. <https://scholar.smu.edu/cgi/viewcontent.cgi?article=1010&context=datasciencereview>

9. Fallucchi, F., Coladangelo, M., Giuliano, R., William De Luca, E.: Predicting employee attrition using machine learning techniques. *Computers* **9**(4), 86 (2020). <https://doi.org/10.3390/computers9040086>
10. Batista, G., Prati, R., Monard, M.-C.: A study of the behavior of several methods for balancing machine learning training data. *SIGKDD Exp.* **6**, 20–29 (2004). <https://doi.org/10.1145/1007730.1007735>
11. Ozdemir, F., Coskun, M., Gezer, C., Gungor, V.: Assessing employee attrition using classifications algorithms, pp. 118–122 (2020). <https://doi.org/10.1145/3404663.3404681>