Employee Attrition Predicting using Machine Learning

E .VIMAL RAJ[1], Y. GUNA SRI SAMHITHA[2], G. SAI[3], I.D.N.V. SABARISH[4]

Department of CSE, SRI VASAVI ENGINEERING COLLEGE, PEDATADEPALLI, India

- ***Abstract—*** *Predicting employee attrition can help organizations take the necessary steps to retain talent well within time. Employee attrition refers to an employee's voluntary or involuntary resignation from a workforce. We use, several classification models, namely Random Forest, AdaBoost, Support Vector Machine (SVM) and Multilayer Perceptron (MLP) have been trained and tested on the IBM HR Dataset. Oversampled data with PCA had the best performances on which Random Forest, AdaBoost, SVM, and MLP achieved accuracy and F1 score above 90%. Based on our analysis, attrition rates were higher in younger employees, doing overtime, having lower monthly incomes, and working for a shorter period.*

Keywords— *Machine learning; Employee attrition; Support vector machine; random forest; Feature ranking; Feature selection; Adaboost; multilayer perceptron*

1.INTRODUCTION:-

Employee attrition can be defined as the loss of employees due to any of the following reasons: personal reasons, low job satisfaction, low salary, and a bad business environment. Employee attrition can be categorized into two categories: voluntary and involuntary attrition. Involuntary attrition occurs when employees are terminated by their employer for different reasons, such as low employee performance or business requirements. In voluntary attrition, on the other hand, high-performing employees decide to leave the company of their own volition despite the company's attempt to retain them. Voluntary attrition can result from early retirement or job offers from other firms, for example. Although companies that realise the importance of their employees usually invest in their workforce by providing substantial training and a great working environment, they too suffer from voluntary attrition and the loss of talented employees. Another issue, hiring replacements, imposes high costs on the company, including the cost of interviewing, hiring and training.

Predicting employee's attrition at a company will help management act faster by enhancing their internal policies and strategies. Where talented employees with a risk of leaving can be offered several propositions, such as a salary increase or proper training, to reduce their likelihood of leaving. Using machine learning models can help companies predict employee's attrition. Using the historical data kept in human resources (HR) departments, analysts can build and train a machine learning model that can predict the employees who are leaving the company. Such models are trained to examine the correlation between the features of both active and terminated employees.

2.RELATED WORK: -

Voluntary employee turnover is one of any organization's greatest worries due to its impact. A company's capacity to replace brilliant employees, which can be difficult and time-consuming, is crucial to its success [1]. Researchers have investigated the reasons for voluntary employee attrition. The literature review indicates that a variety of factors can significantly affect the rate of employee attrition.

For instance, it was found by [2] and [3] that offering compensation significantly affects employee performance as well as turnover. When compensation is increased, the attrition rate falls. Money is not the main problem, as [1] found that other factors including job load, performance compensation, and a subpar career plan have increased the turnover rate in the retail industry.
Numerous studies have looked at the use of machine learning to predict employee behaviour. To predict employee performance, the authors of [4] used decision trees (ID3 C4.5) and the Naïve Bayes classifier. They found that age had no discernable effect, and that the most important component was job title. The authors of [5] investigated several data mining techniques to predict employee turnover using a dataset comprising 1575 records and 25 features.

The machine learning methods that they employed were naïve Bayes, logistic regression, support vector machines, decision trees, and random forests. The study's findings suggest employing an 84.12% accurate support vector machine (SVM). Several decision tree algorithms, such as C4.5, C5, REPTree, and classification and regression trees (CART), were investigated in [6]. With a dataset of 309 employee records out of 4326 records and a total of six attributes, the researchers evaluated and trained the decision trees. Consequently, when compared to other decision trees, the C5 decision tree yielded the best accuracy, at 74%. Additionally, their findings demonstrated the significance of employee pay and tenure in the dataset of the evaluated organization. Neural networks were employed by the authors in [7] to forecast the turnover rate for small-west manufacturing companies. As a result, they created the 10-fold cross validation approach in conjunction with the neural network simultaneous optimization algorithm (NNSOA), which accurately predicted the turnover rate by 94%. Furthermore, by utilizing a modified genetic algorithm, they were able to determine the most significant and pertinent "Tenure of employee on January 1." The online profiles of 6,909,746 employees were utilized in [8] to forecast staff turnover. Along with firm details, the individuals' profiles also contained information about their education and employment experiences. An SVM model might be trained and assessed by the researcher.

The average accuracy of the model forecast was just 55%, which is obviously not very good. In order to enhance the trained model, the researcher suggested including more personal information in the dataset, such as the age, gender, and workplace of the employees. [9] forecasted worker turnover for a multinational company with headquarters in the US. There were 33 features and 73,115 observations in the dataset. After analysing seven machine learning algorithms, the researchers concluded that XGBoost had the highest accuracy, with an area under the curve (AUC) of 0.88. It also performed better in terms of memory use than the other models. The author created a prediction model for Swedbank staff attrition in [10]. With 98.6% accuracy, a random forest model beat SVM and multi-layer perceptron (MLP) models in this investigation.

Earlier research that employed a variety of datasets and machine learning algorithms offered a range of accuracy metrics. It is so challenging to determine which model is the most appropriate to use. Additionally, the issue of class imbalance that appears in real-world attrition data was not addressed in earlier research. As a result, we investigated several approaches to address class imbalance, which greatly improved the training procedure.
This is how the rest of the paper is structured. The suggested techniques employed in this study are presented in Section 3. The experimental design and findings are presented in Section 4, and the study is concluded in Section 5.

3.PROPOSED METHODS:-

In this research, we have explored three main experiments to predict employee attrition. First, we have attempted to predict employee attrition using the original imbalanced dataset (data details presented in section IV). In the second experiment, we have introduced the adaptive synthetic sampling approach to solve the class imbalance problem. This approach involved oversampling the minority class which was in this case the "yes" class. The third experiment involved random under sampling of the data where we have randomly selected an equal subset of each class. Moreover, each experiment involved training and validating a set of machine learning classifiers to predict unseen dataset of employee attrition. All classifiers were validated using 5-fold cross validation. In addition, we have introduced feature selection method to minimize the trained model's complexity and enhance their performance. In each case, each classifier was trained and evaluated iteratively by increasing the number of features for each iteration. The proposed methodologies are presented below with further details.

- ***Classification:-***

To categorize unseen data, we employed a number of currently available machine learning classification models in this work. The classifiers employed in this study are described below.
A non-probabilistic supervised machine learning model for regression and classification is called a support vector machine (SVM). By dividing each class with a decision boundary—also referred to as a hyperplane—SVMs may be trained with assigned classes [11], [12].

Since it might be challenging to identify the decision boundary, some issues are regarded as nonlinear. But this may be resolved by utilizing a kernel function, sometimes referred to as a kernel trick. This function translates data points to a new, altered, high-dimensional space after returning the dot product of the two vectors. Furthermore, many kinds of kernel functions, including polynomial, Gaussian, and linear kernels, canbe employed [13], [14].
To producing regressions and classifications, one of the most effective supervised machine learning algorithms is random forest (RF). RF trains data using multiple decision trees [15]. After each tree casts a vote for a classification label for a particular dataset, the RF model determines which class received the most votes Since it might be challenging to identify the decision boundary, some issues are regarded as nonlinear. But this may be resolved by utilizing a kernel function, sometimes referred to as a kernel trick. This function translates data points to a new, altered, high-dimensional space after returning the dot product of the two vectors. Furthermore, many kinds of kernel functions, including polynomial, Gaussian, and linear kernels, can be employed [13], [14].

- ***Feature Selection:-***

Numerous characteristics may be present in real-world datasets. Certain characteristics may not be beneficial for training machine learning algorithms since they are seen as noise. Utilizing every feature will make the model more complicated, which will impact training time and model performance [21].
Every feature may be ranked and evaluated using a variety of techniques. The t-test technique is employed in this study to determine the mean and standard deviations of the binary class labels that were applied to the training data sets. The following represents the t-test formula [22].

$$t(x)=(\bar{y}_1(x)-\bar{y}_2(x))/\sqrt{(s_1^2(x)/n_1+s_2^2(x)/n_2)}$$

4. DATASET AND TOOLS:

The dataset we utilized in this study is available to the general public and may be acquired via IBM Watson Analytics1. IBM data scientists generated fake data for the dataset. The dataset includes 32 attributes and 1470 workers' HR-related data. Additionally, 237 former workers were from the "Yes" attrition group, whereas 1233 current employees were from the "No" attrition category. Two elements were eliminated from the research: "Standard hours," as every employee has the same standard hours, and "employee count," since it is a series of numbers (1, 2, 3..). Additionally, for processing purposes, all non-numerical variables were given numerical values, such as Sales=1, R&D=2, and Human Resources= 3. Additionally, the machine learning models in this study were trained and assessed using MATLAB R2017b.

5.EXPERIMENTS:-

We present the findings from the three primary experiments conducted on the dataset in this section. Multiple machine learning models were trained by each experiment. The models were assessed based on their F1 score, accuracy, precision, and recall. The subsections following go into further specifics.

- ***Performance Evaluation*:-**

The following metrics were used to assess each training model: accuracy, precision, recall, and F1 score.

: [17] [18] [19].

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FN+FP} \quad (2)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (3)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (4)$$

$$\text{F1 Score} = 2 * \frac{\text{Precesion} * \text{Recall}}{\text{Precesion}+\text{Recall}} \quad (5)$$

where TP are true positives values, TN are true negative, FP are false positives, and FN are the false negatives values.

- ***Imbalanced Data Experiments*:-**

This section predicts employee attrition using the original class-imbalanced dataset. In this research, SVM, random forest, and KNN classification models were evaluated. First, each classifier was evaluated based on using all features in the dataset. Next, classifiers were evaluated by ranking and selecting the important subset features only.

Table I compares the performance of several classification models. Training with linear SVM yielded 86.9% accuracy but a very low F1 score. This indicates that it is misclassifying most of the minority class 'Yes'. For further investigation, SVM was trained using different kernel types, such as quadratic, cubic and Gaussian. But, the F1 results were still low. The highest F1 score was 0.503, generated by quadratic SVM. The same applies to random forest and KNN. Although KNN was trained using different K values (1,3,5 and 10), the results were lower than those for SVM.

By using feature ranking it was possible to rank all features in the imbalanced dataset, where it first computed two-sample t-test and then returned an ordered index of the most important features which played a role in the training process. The top three of which were overtime, monthly income and job level. To further investigate, a linear SVM algorithm was trained and tested using only the top two features (overtime, monthly income), resulting in 83.9% accuracy. However, it did not classify any data point as 'Yes' for attrition; hence, F1 scored zero. Whereas when using SVM with several kernel types, the F1 scores slightlyincreased. Similar results were found using all three top features. Furthermore, training random forest with two features delivered a low F1 score. However, it

increased in accuracy when using the top three features (overtime, monthly income, job level), yet still had very low F1 scores. In addition, KNN was trained with the top two and three features. In both experiments, KNN showed zero F1 results when K =1 and 5. Also, KNN results were significantly low when K = 3. In this research, feature selection continued up to 12 features, but the results were insignificant. As a result, using feature selection with imbalanced data did not show any significant improvement on model performance.

TABLE I CLASSIFIER PERFORMANCE WITH IMBALANCED DATA

Model Type	Accuracy	Precision	Recall	F1 Score[a]
Linear SVM	0.869	0.814	0.240	0.371
Quadratic SVM	0.871	0.662	0.405	**0.503**
Cubic SVM	0.841	0.508	0.418	0.458
Gaussian SVM	0.865	0.788	0.219	0.343
Random Forest	0.856	0.75	0.164	0.269
KNN (K=1)	0.827	0.275	0.046	0.079
KNN (K=3)	0.8374	0.25	0.004	0.008

Bold values indicate highest F1 score

- ***Balancing Data Using Oversampling:***

Using a dataset that has been artificially balanced, this section forecasts employee attrition. To accomplish this phase, the dataset was scaled first, and then the ADASYN technique was applied. To oversample the minority class "Yes," additional synthetic data points were created. As a result, the total number of observations in the class "Yes" climbed to 1152, while the class "No" observations remained constant at 1233.

Table II presents a comparison of the performance of various classification models trained with all characteristics. It is evident that all predicting models performed much better overall when trained with balanced classes. Using linear SVM training improved the F1 score to 0.779. But when SVM was trained using quadratic, cubic, and Gaussian kernels, the F1 scores increased even more: the quadratic SVM produced 0.881 F1 ratings, the cubic SVM produced 0.927 F1 scores, and the Gaussian SVM produced 0.912 F1 values. This suggests that applying kernels to transfer data to higher dimensions aids in defining the ideal boundary and that the newly balanced dataset is nonlinearly separable.

Furthermore, random forest was used for training and assessment of the balanced dataset. In contrast to the unbalanced dataset, random forest produced F1 scores of 0.921. Moreover, many K values (one, three, five, and ten) were used to train KNN. When K = 1, KNN produced extremely high scores, which might be a sign of overfitting. In the meanwhile, KNN obtained F1 scores of 0.931 and.909 with K = 3 and K = 5, respectively. Ultimately, KNN produced outcomes with a lower F1 score of 0.88 when using K = 10.

The top features that were assisting in the training process were ranked using the feature ranking algorithm after the synthetic data points were generated. Consequently, it was discovered that the three most important characteristics were job level, total working year, and overtime. With an F1 score of 0.829, the random forest model produced the best results when compared to the other models, as seen in Table III. After being trained with the two characteristics, the remaining prediction models achieved relatively poor performance outcomes. While random forest achieved 0.806 with only three features, similar results were seen while training with only three features.

The 12 best characteristics were added as the tests went on. The top 12 aspects utilized in the training are listed in Table IV. Random forest only needed 12 subset features to get 0.909 F1 scores. Additionally, KNN was able to score as high as 0.882,.861, and 0.839 with K = 3, 5, and 10. Furthermore, the F1 scores of cubic and Gaussian SVMs exceeded 0.83.

TABLE II. CLASSIFIER PERFORMANCE WITH SYNTHETIC BALANCED DATA

Model Type	Accuracy	Pecision	Recall	F1 Score[b]
Linear SVM	0.782	0.763	0.795	0.779
Quadratic SVM	0.879	0.839	0.927	0.881
Cubic SVM	0.926	0.879	0.981	**0.927**
Gaussian SVM	0.912	0.885	0.941	**0.912**
Random Forest	0.926	0.950	0.893	**0.921**
KNN (K=1)	0.967	0.939	0.997	**0.967**
KNN (K=3)	0.929	0.877	0.992	**0.931**
KNN (K=5)	0.904	0.843	0.987	**0.909**
KNN (K=10)	0.872	0.804	0.970	0.880

TABLE III. CLASSIFIER PERFORMANCE WITH FEATURE SELECTION FOR SYNTHETIC BALANCED DATA

Model Type	No. Features	Accuracy	Precision	Recall	F1 Score[c]
Linear SVM	2	0.648	0.676	0.523	0.589
Cubic SVM	2	0.593	0.550	0.871	0.674
Gaussian SVM	2	0.722	0.755	0.628	0.686
Random Forest	2	0.852	0.935	0.745	**0.829**
KNN (K=1)	2	0.659	1	0.294	0.454
KNN (K=3)	2	0.537	1	0.045	0.087
KNN (K=5)	2	0.523	1	0.016	0.032
Linear SVM	3	0.649	0.676	0.523	0.590
Cubic SVM	3	0.562	0.530	0.823	0.645
Gaussian SVM	3	0.722	0.753	0.630	0.686
Random Forest	3	0.826	0.869	0.752	**0.806**
KNN (K=1)	3	66.4	1	0.303	0.466
KNN (K=3)	3	0.572	0.995	0.175	0.298
KNN (K=5)	3	0.553	1	0.137	0.242
Cubic Linear SVM	12	0.74	0.736	0.721	0.729
Cubic SVM	12	0.851	0.875	0.825	**0.850**
Quadratic SVM	12	0.801	0.796	0.791	0.794
Gaussian SVM	12	0.834	0.812	0.853	**0.832**
Random Forest	12	0.914	0.925	0.893	**0.909**
KNN (K=1)	12	0.641	1	0.256	0.407
KNN (K=3)	12	0.869	0.802	0.979	**0.882**

KNN (K=5)	12	0.844	0.771	0.976	**0.861**
KNN (K=10)	12	0.818	0.749	0.955	**0.839**

- **Balancing Data Using Under sampling**:-

To address class imbalance, we use manual under sampling of the dataset to estimate employee attrition in this part. To achieve this, an equal number of observations—237 for each class—were chosen at random. There were 474 total observations in the new dataset. The performance of various categorization models after they were trained with all characteristics is compared in Table V. SVM was used to get the highest F1 score possible; the quadratic SVM score was 0.74, and the linear and Gaussian SVM scores were 0.73. Additionally, random forest and cubic SVM both achieved 0.69 F1 ratings. Lastly, with K = 10, KNN produced poor results, up to 0.59. These findings suggest that employing manual under sampling might result in the loss of crucial data that could be involved in predicting attrition.

TABLE IV. CLASSIFIER PERFORMANCE FOR UNDER SAMPLED DATA

Model Type	Accuracy	Precision	Recall	F1 Score[d]
Linear SVM	0.745	0.754	0.725	**0.739**
Quadratic SVM	0.747	0.760	0.722	**0.740**
Cubic SVM	0.707	0.733	0.650	0.689
Gaussian SVM	0.751	0.779	0.700	**0.738**
Random Forest	0.717	0.756	0.641	0.694
KNN (K=1)	0.589	0.595	0.552	0.573
KNN (K=3)	0.573	0.572	0.586	0.579
KNN (K=5)	0.565	0.562	0.586	0.574
KNN (K=10)	0.588	0.584	0.611	0.597

Bold values indicate highest F1 score

In this part, feature ranking and selection were used even though the under sampling findings were modest. To rank the best features that were assisting in the training process, the feature ranking function was employed. The top three features were therefore

determined to be overtime, years under the present boss, and overall working years. The predictive models' performance with feature selection is displayed in Table VI. Results for KNN, random forest, and Gaussian SVM were all quite similar, ranging from 0.66 to 0.68. KNN rated 'Yes' on most observations. Additionally, it was shown that using all three characteristics during training produced quite similar outcomes.

TABLE V. CLASSIFIES PERFORMANCE WITH FEATURE SELECTION FOR UNDER SAMPLED DATA

Model Type	No. Faturs	Accuracy	Precision	Recall	F1 Score[e]
Linear SVM	2	0.652	0.698	0.536	0.606
Cubic SVM	2	0.631	0.661	0.536	0.592
Gaussian SVM	2	0.681	0.676	0.696	**0.686**
Random Forest	2	0.679	0.682	0.671	**0.677**
KNN (K=1)	2	0.523	0.511	0.995	**0.676**
KNN (K=3)	2	0.506	0.503	0.995	**0.668**
KNN (K=5)	2	0.5	0.5	1	**0.666**
Linear SVM	3	0.652	0.698	0.536	0.606
Cubic SVM	3	0.515	0.524	0.316	0.395
Gaussian SVM	3	0.67%	0.687	0.620	**0.652**
Random Forest	3	0.618	0.612	0.646	0.628
KNN (K=1)	3	0.5253	0.513	0.953	**0.667**
KNN (K=3)	3	0.5464	0.525	0.945	**0.675**
KNN (K=5)	3	0.5	0.527	0.919	**0.670**

Bold values indicate highest F1 score

6.CONCLUSION:

For businesses, a high staff churn rate is a serious issue. Retaining top performers is regarded as a significant setback for businesses, particularly those that make personnel investments. It can be challenging and expensive for the organization to find successors that perform at the same level, both in terms of money and time.

This study's primary goal was to forecast employee attrition using machine learning models by analysing their attributes. This will provide management of the organization with machine learning-backed indicators. This will thus enable management to act more quickly to lessen the possibility that brilliant people may go from their organization. Three experimental techniques were applied to the dataset in this study to create prediction models. Initially, a number of prediction models were trained on the initial unbalanced data, with quadratic SVM achieving the best results with 0.50 F1 scores. It was notable how much each model's performance improved: random forest (K = 3) had high F1 scores, ranging from 0.91 to 0.93. Additionally, extremely similar results were obtained when the top 12 features were used in the random forest feature selection, which produced F1 scores of 0.90 and 0.92 with only two features. The dataset was manually undersampled to achieve equal class sizes as the last strategy. Lower performance was the result of crucial information being lost. And yet SVMs managed to obtain more than 0.70 with all features and more than 0.60 with only two characteristics.

7.References:-

1)S. Kaur andR. Vijay," Work Fulfillment – A Major Calculate Behind squander or Maintenance in Retail Industry," Royal Diary of Intrigue Research,vol. 2,no. 8, 2016.

2)D.G. Gardner,L.V. Dyne andJ.L. Puncture," The products of pay position on affiliation-grounded tone- respect and execution a field consider," Diary of Word related and Organizational Psychology,vol. 77,no. 3,pp. 307- 322, 2004.

3)E. Moncarz,J. Zhao andC. Kay," An exploratory think about of US lodging parcels' organizational hones on hand improvement and maintenance," Universal Diary of Modern Neighborliness Management,vol. 21,no. 4,pp. 437- 458, 2009.

4)Q.A. Al- Radaideh andE.A. Nagi," Utilizing Information Mining ways to make a Bracket Show for Anticipating laborers Execution," Universal Diary of Progressed Computer Science and Applications,vol. 3,no. 2,p. 144 – 151, 2012.

5)G.K.P.V. Vijaya Saradhi," Hand churn vaticination," Master Frameworks with Applications,vol. 38,no. 3,pp. 1999- 2006, 2011.

6)D.A.B.A. Alao," assaying hand squander utilizing choice tree calculations," Computing, Data Frameworks, Advancement Informatics and Partnered Inquire about Journal,no. 4, 2013.

7)R.S. Sexton,S. McMurtrey,J.O. Michalopoulos, andA.M. Smith," Hand advancement a neural arrange result," Computers & Operations Research,vol. 32,no. 10,pp. 2635- 2651, 2005.

8)Z. Ö. KISAOˇGLU, Hand Advancement vaticination Utilizing Machine Learning Grounded styles(Proposal), Center EAST Specialized College, 2014.

9)R. Punnoose andP. Ajit," vaticination of Hand Improvement in Organizations utilizing Machine Learning Calculations," Worldwide.

10)M. Maisuradze, Prescient Examination On The outline Of Hand Improvement(Master's proposal), Tallinn Tallinn College of Innovation, 2017.

11)K.-B. Duan andS.S. Keerthi," Which is the smart multiclass SVM framework? An experimental think about," Universal production line on different classifier frameworks, 2005.

12)K.P. Bennett andC. Campbell," Back vector machines buildup or psalm?" Acm Sigkdd thinks about Newsletter,vol. 2,no. 2,pp. 1- 13, 2000.

13)S. Rogers andM. Girolami, A to begin with course in machine education, CRC Press, 2016.

14)N. Cristianini andB. Scholkopf," Bolster vector machines and part styles the modern era of learning machines," Ai Magazine,vol. 23,no. 3,p. 31, 2002.

15)T.K. Ho," Irregular choice timbers," in procedures of the third transnational conference on Record Examination and Acknowledgment, 1995.

16)L. Breiman," Irregular timbers," Machine literacy,vol. 45,no. 1,pp. 5- 32, 2001.

17)X. Zhu, Information Disclosure and Information Mining Challenges and Substances Challenges and Substances, Igi Worldwide, 2007.

18)D.M. Powers," Assessment from flawlessness, review and F- degree to ROC, informedness, markedness and relationship," Diary of Machine, 2011.

19) H. He andE.A. Garcia," Learning from Imbalanced Information," IEEE Bargains on Information and Information Engineering,vol. 21,no. 9,pp. 1263- 1284, 2009.

20)H. He,Y. Bai,E.A. Garcia, andS. Li," ADASYN Versatile Engineered Testing Approach for Imbalanced Learning," in IEEE Universal Joint Conference on Neural Systems, 2008.

21)I. Guyon andA. Elisseeff," An introduce to Variable and point Determination," Diary of Machine Learning Research,vol. 3,pp. 1157- 1182, 2003.

22)W. Zhu,X. Wang,Y. Ma,M. Rao,J. Glimm andJ.S. Kovach," Revelation of cancer-specific names in the midst of enormous mass unearthly information," Procedures of the National Institute of lores,vol. 100,no. 25,pp. 14666- 14671, 2003.