Juni Khyat
(UGC Care Group I Listed Journal)
ISSN: 2278-4632
Vol-10 Issue-1 No. 1 January 2020

Feature Selection Methods for Improving Classification Accuracy – A Comparative Study

S. Chitra [#1], Dr.P.Srivaramangai [*2]
[#1]*Research Scholar,* [*2] *Associate Professor*
[#1, *2]*Department of Computer Science, Marudupandiyar College (Affiliated to Bharathidasan University), Thanjavur - 613 403, Tamilnadu,India*
[1]chitrasathish1979@gmail.com

***Abstract*— In any organization's talent management is becoming an increasingly crucial method of approaching HR functions. Talent management can be defined as an outcome to ensure the right person in the right job. Human capital is the most effective resource to hiring the highly qualified personnel for improving the world economy and also for developing company's management. Turnover of employee considers as one of the major issues that every company faces. Especially, if the employee has advance skills at his/her working field, then the company faces great loss during that period. To find out the most dominant reasons of employee attrition, we approach by determining features and using machine learning algorithms where features have been processed and reduced beforehand. In this paper, four different feature selection methods are used to find the relevant features of the HR datasets to improve the classification accuracy on the Employee Attrition of the company. The Machine Learning classifiers like Random Forest, K- Nearest Neighbor, Gradient Boosting Tree, Neural Network and Naïve Bayes algorithms used to evaluate the performance of the feature selection methods.**

***Keywords*— Machine Learning, Feature Selection, Random Forest, Classifier, K-Nearest Neighbor, Gradient Boosting Tree, Neural Network, Naïve Bayes**

I. INTRODUCTION

Employee turnover means the ratio of leaving and total employee within a period of time [1]. In modern day, employee turnover has considered as very common event [2]. Lack of satisfaction, heavy work load, workplace environment, poor performance, less salary, etc. is some of the trigger points that lead to employee attrition. Turnover of employees obviously a major issue for any reputed company as they suffer since the skilled employee leaves [1]. A company's reputation depends on employee attrition also [2]. Therefore, it is a major concern for any Human Resource Management of a company to identify the key facts behind the employee's turnover to retain the reputation and prosperity.

Employee Attrition is one of the major problems faced by any organization. In this age of cut-throat competition there are many factors which lead to dissatisfaction in employee. long working hours, peer pressure, job location, job role, travelling time, office space, amenities in the office, perks and many more reasons could be a factor for employee attrition. It is very important for the HR department to understand employee satisfaction level. Sometimes the employee many not have any problem in the company but others many offer a better profile with better pay package. So, the employee may be willing to leave. Retaining one employee needs a lot of insight in many areas. In this research we try to find out important factors that lead to employee attrition [3]. The results of our model can be used by HR department to plan a strategy before the employee sends his resignation.

The systematic application of analytical methods on human resources (HR) related (big) data is referred to as HR analytics or people analytics [4]. Typical problems in HR analytics are the estimation of churn rates, the identification of knowledge and skill in an organization or the prediction of success on a job. HR analytics, as opposed to the simple use of key performance indicators, is a growing field of interest because of the rapid growth of volume, velocity and variety of HR data, driven by the digitalization of work processes. Personnel files used to be in steel lockers in the past, they are now stored in company systems, along with data from hiring processes, employee satisfaction surveys, emails, and process data [5].

II. IMPORTANCE OF FEATURE SELECTION

ISSN: 2278-4632
Vol-10 Issue-1 No. 1 January 2020

The abundance of data in contemporary datasets demands development of clever algorithms for discovering important information. Data models are constructed depending on the data mining tasks, but usually in the areas of classification, regression and clustering. Often, pre-processing of the datasets takes place for two main reasons: 1) reduction of the size of the dataset in order to achieve more efficient analysis, and 2) adaptation of the dataset to best suit the selected analysis method. The former reason is more important nowadays because of the plethora of developed analysis methods that are at the researcher's disposal, while the size of an average dataset keeps growing both in respect to the number of features and samples [6][7].

Dataset size reduction can be performed in one of the two ways: feature set reduction or sample set reduction. The problem is important, because a high number of features in a dataset, comparable to or higher than the number of samples, leads to model overfitting, which in turn leads to poor results on the validation datasets. Additionally, constructing models from datasets with many features is more computationally demanding [8]. All of these leads researchers to propose many methods for feature set reduction. The reduction is performed through the processes of feature extraction (transformation) and feature selection. Feature extraction methods such as Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA) and Multidimensional Scaling work by transforming the original features into a new feature set constructed from the original one based on their combinations, with the aim of discovering more meaningful information in the new set [9]. The new feature set can then be easily reduced by taking into consideration characteristics such as dataset variance coverage. Feature selection, on the other hand, is a process of taking a small subset of features from the original feature set without transformation (thus preserving the interpretation) and validating it with respect to the analysis goal. The selection process can be achieved in a number of ways depending on the goal, the resources at hand, and the desired level of optimization. Feature set reduction is based on the terms of feature relevance and redundancy with respect to goal. More specifically, a feature is usually categorized as: 1) strongly relevant, 2) weakly relevant, but not redundant, 3) irrelevant, and 4) redundant. A strongly relevant feature is always necessary for an optimal feature subset; it cannot be removed without affecting the original conditional target distribution [10]. Weakly relevant feature may not always be necessary for an optimal subset, this may depend on certain conditions. Irrelevant features are not necessary to include at all. Redundant features are those that are weakly relevant but can be completely replaced with a set of other features such that the target distribution is not disturbed (the set of other features is called Markov blanket of a feature). Redundancy is thus always inspected in multivariate case (when examining feature subset), whereas relevance is established for individual features. The aim of feature selection is to maximize relevance and minimize redundancy. It usually includes finding a feature subset consisting of only relevant features. In order to ensure that the optimal feature subset with respect to goal concept has been found, feature selection method has to evaluate a total of 2m - 1 subsets, where m is the total number of features in the dataset (an empty feature subset is excluded).

III. RELATED WORKS

Xue, Bing, et al [11] presented a comprehensive survey of the state-of-the-art work on evolutionary computation for feature selection, which identifies the contributions of the different algorithms. A Variety of methods have been applied to solve feature selection problems, where evolutionary computation techniques have recently gained much attention and shown more success. However, there are no comprehensive guidelines on the strengths and weakness of alternative approaches. This leads to a disjointed and fragmented field with ultimately lost opportunities for improving performance and successful applications.

Win, Thee Zin, and Nang Saing Moon Kham [12] presented Feature selection, a data preprocessing technique, is effective and efficient to enhance data mining, data analytics and machine learning. Most

feature selection algorithms have been trying to eliminate irrelevant features. However, removing only irrelevant features is not enough to get the best insight and patterns. Not only irrelevant features but also redundant features can degrade learning performance. Feature selection methods which can eliminate both irrelevant and redundant features are demanding in high dimensional data analytics. To solve this problem, information gain measured feature selection is presented in this work.

Moran, Michal, and Goren Gordon [13] addressed the challenge of continues change in data structures by implementing concepts from the field of intrinsically motivated computational learning, also known as Artificial Curiosity (AC). The authors presented a novel method of intrinsically motivated learning, based on the curiosity loop, to learn the data structures in large and varied datasets. An autonomous agent learns to select the subset of relevant features in the data, i.e., feature selection to be used later for model construction.

Chiew, Kang Leng, et al [14] proposed a new feature selection framework for machine learning based phishing detection system, called the Hybrid Ensemble Feature Selection (HEFS). In the first phase of HEFS, a novel Cumulative Distribution Function gradient (CDF-g) algorithm is exploited to produce the priority feature subsets, which are then fed into a data perturbation ensemble to yield secondary feature subsets. The second phase derives a set of baseline feature from the secondary feature subsets by using a function perturbation ensemble.

Huang, Changqin, et al [15] conducted a deep analysis, and find that simply extracting the features based on the score calculated by a metric may not always be the best strategy as it may turn many documents into zero length, which make them not suitable for training. Then model the feature selection process as a multiple objectives optimization problem to gain the best number of selected features rationally and automatically.

Singh, Ajeet, and Anurag Jain [16] focused on credit cards fraud detection at application level using feature selection methods. The authors used J48 Decision Tree, Ada boost, Random Forest, Naïve Bayes and PART machine learning techniques for detection of financial frauds of a credit card and the performance of these techniques are compared.

Tsamardinos, Ioannis, et al [17] presented the Parallel, Forward-Backward with Pruning (PFBP) algorithm for feature selection (FS) for Big Data of high dimensionality. PFBP partitions the data matrix both in terms of rows as well as columns. By employing the concepts of p-values of conditional independence tests and meta-analysis techniques, PFBP relies only on computations local to a partition while minimizing communication costs, thus massively parallelizing computations.

IV. FEATURE SELECTION AND CLASSIFICATION METHODS

A. Chi-Square Feature Selection Method

The main reason to use the Chi-Square algorithm is to find the highest valued features from the test chi2 statistics. Chi-Square algorithm is basically on X^2 statistics. The Chi2 algorithm works in two phases. In phase 1 it calculates the X^2 values for each pair of intervals. Then it combines the pair of intervals with the smallest X^2 values until all the pairs have X^2 values gone beyond the sigLevel determined parameters (in phase 1 Chi2 algorithm starts with significant level). Until the inconsistency rate has gone beyond the discretized data, the phase 1 continues its operation. On the other hand, phase 2 is more refined process of phase 1 where it begins with SigLevel0 (in phase 1) and each attribute is associated with sigLevel[i]. In addition to this it also merges the attribute and checks the consistency. Now, if the inconsistency rate has not been surpassed, then sigLevel[i] is decremented for each attribute's (i) next phase of merging. This stops when all attributes are merged [8].

ISSN: 2278-4632
Vol-10 Issue-1 No. 1 January 2020

The mathematical equation of Chi2 algorithm is the following [8]

$$X^2 = \sum_{i=1}^{2} \sum_{j=1}^{k} \frac{(A_{ij} - E_{ij})}{E_{ij}}$$

Here k is the number of classes, the number of patterns is given by A_{ij}, and the expected frequency is denoted by E_{ij}.

B. Information Gain Feature Selection Method

Entropy is commonly used in the information theory measure, which characterizes the purity of an arbitrary collection of examples [7][8]. It is in the foundation of Gain Ratio, Information Gain and Similarity Uncertainity (SU). The entropy measure is considered a measure of the system's unpredictability. The entropy of Y is

$$H(Y) = \sum_{y \in Y} p(y) \log_2(p(y)) \quad (3.1)$$

where *p(y)* is the marginal probability density function for the random variable *Y*. If the observed values of *Y* in the training data set *S* are partitioned according to the values of a second feature *X*, and the entropy of *Y* with respect to the partitions induced by *X* is less than the entropy of *Y* prior to partitioning, then there is a relationship between features *Y* and *X*. The entropy of *Y* after observing *X* is then:

$$H(Y|X) = \sum_{x \in X} p(x) \sum_{y \in Y} p(y|x) \log_2(p(y|x)) \quad (3.2)$$

where *p(y |x)* is the conditional probability of *y* given *x*.

Given the entropy is a criterion of impurity in a training set *S*, we can define a measure reflecting additional information about *Y* provided by *X* that represents the amount by which the entropy of *Y* decreases. This measure is known as IG. It is given by

$$IG = H(Y) - H(Y|X) = H(X) - H(X|Y) \quad (3.3)$$

IG [9] is a symmetrical measure and it is given by equation (3.3). The information gained about *Y* after observing *X* is equal to the information gained about *X* after observing *Y*. A weakness of the IG criterion is that it is biased in favor of features with more values even when they are not more informative.

C. Gain Ratio Feature Selection Method

The Gain Ratio [8] is the non-symmetrical measure that is introduced to compensate for the bias of the Information Gain (IG) [7]. GR is given by

$$GR = \frac{Information\ Gain\ (IG)}{H(X)} \quad (3.4)$$

Information Gain (IG) is a symmetrical measure.

$$IG = H(Y) - H(Y|X) = H(X) - H(X|Y) \quad (3.5)$$

The information gained about Y after observing X is equal to the information gained about X after observing Y. A weakness of the IG criterion is that it is biased in favor of features with more values even when they are not more informative.

D. Random Forest

Random Forest is a very popular and highly accurate learning algorithm for high-dimensional and ill-posed classification and regression tasks, based on model aggregation idea. The key idea behind the random forests framework is to grow a large number of unbiased decision trees from the random samples of the training data with replacement, where each tree votes for a class and the forest choose the classification having the most votes over all the trees in the forest [18].

ISSN: 2278-4632
Vol-10 Issue-1 No. 1 January 2020

One of the key advantages of random forests is that random forests can measure importance score of each feature to learn the impact of each feature regarding the prediction of the classes. However, for high dimensional problem, the number of features may be huge that makes the manual investigation of the feature importance scores and selection of the most relevant features for classification very challenging.

E. K- Nearest Neighbor Classification

K-nearest neighbors is a non-parametric algorithm used for classification and regression problems [19]. For classification problems, the idea is to identify the K data points in the training data that are closest to the new instance and classify this new instance by a majority vote of its K neighbors. In practice, the popular distance measures include the Euclidean distance, the Manhattan distance as well as the Minkowski distance. For regression problems, the idea is to calculate the new instance value by taking the average of its K neighbors. KNN could work well with a small number of features, but it struggles when the feature dimensions increase drastically.

F. Gradient Boosting Tree Classification

Gradient boosting trees is an ensemble machine learning method proposed in 2001 by Friedman for regression and classification purposes. The difference between RF and GBT is the gradient boosted tree models learn sequentially. In GBT, a series of trees are built and each tree attempts to correct the mistakes of the previous tree in the series. Trees are added sequentially until no further enhancement can be achieved. Making predictions in GBT is fast and memory-efficient; boosting could be viewed as a form of ‘1 regularization to reduce overfitting [20]. However, unlike highly interpretable single DT, GBT is harder to visualize and interpret.

G. Neural Network Classification

Neural networks, also known as multi-layer perceptron, are designed to simulate the operations of the human nervous system. The simplest form of a neural network is a single perceptron. Essential elements for a perceptron are input values, associated weights, bias, activation functions and a computed output. Commonly used activation functions include the sigmoid, hyperbolic tangent (Tanh) and rectified linear units (ReLU). A neural network may contain more than one layer between input and output to handle complex problems. This sophisticated structure of neural networks makes it a universal approximation tool which could model any smooth function to any desired level of accuracy, given enough hidden units [21]. One can extend the model to become deep with more advantages, in what is commonly referred to as deep learning. Due to the rapid development of hardware and the continuous exploration of backpropagation techniques, neural networks are currently the most heavily researched topic in machine learning.

H. Naïve Bayes Classification

Naïve Bayes is a probabilistic approach that uses Bayes Theorem. The Bayes Theorem describes the occurrence probability of an event based on the prior knowledge of related features. The other important characteristic of Naïve Bayes is the conditional independence assumption of its features. This assumption indicates that the presence of a feature would not influence any other features. Naïve Bayes classifiers first learn joint probability distribution of their inputs by utilizing the conditional independence assumption. Then, for a given input, the methods produce an output by computing the maximum posterior probability with Bayes Theorem [22].

ISSN: 2278-4632
Vol-10 Issue-1 No. 1 January 2020

V. RESULT AND DISCUSSION

The three Employee Attrition datasets are considered from the famous Kaggle Repository [23][24][25]. The first dataset is composed of 35 features, second dataset (Dataset_2) is composed of 35 features, and the third dataset (Dataset_3) contains 24 features. The performance metrics like Accuracy, True Positive Rate, False Positive Rate, Precision, Miss Rate, and Specificity are considered for evaluating the features selection methods. Table 1 depicts the description of the HR dataset.

TABLE 1 DEPICTS THE FEATURES IN THE CONSIDERED HR ANALYTICS EMPLOYEE ATTRITION DATASETS

Sl.No	Dataset_1	Dataset_2	Dataset_3
1	Age	Employee_name	Age
2	Attrition	Employee_ID	Attrition
3	Business Travel	Married_ID	Business Travel
4	Daily Rate	Marital Status ID	Department
5	Department	Gender ID	Distance from home
6	Distance From Home	Emp Status ID	Education
7	Education	Dept ID	Education Field
8	Education Field	Performance Score ID	Employee Count
9	Employee Count	From Diversity Job Fair ID	Employee Number
10	Employee Number	Pay Rate	Gender
11	Environment Satisfaction	Attrition	Job Level
12	Gender	Position ID	Job Role
13	Hourly Rate	Position	Marital Status
14	Job Involvement	State	Monthly Income
15	Job Level	Zip	Number of companies worked
16	Job Role	DateofBirth	Over18
17	Job Satisfaction	Sex	Percentage salary Hike
18	Martial Status	Marital Description	Standard Hours
19	Monthly Income	Citizen Description	Stock Option Hours
20	Monthly Rate	Hispanic Latino	Total working years
21	Number of Companies worked	Race Description	TrainingTimes last year
22	Over18	Date of Hire	Years at company
23	Overtime	Date of Termination	Years since last promotion
24	Percentage Salary Hike	Termination Reason	Year with current manager
25	Performance Rating	Employment Status	
26	Relationship Satisfaction	Department	
27	Standard Hours	Manager Name	
28	Stock Option Level	Manager ID	
29	Total Working Years	Recruitment Source	
30	Training time last year	Performance Score	
31	Work life balance	Engagement Survey	
32	Years at Company	Employee Satisfaction	
33	Years in current role	Special project count	
34	Years since last promotion	Last performance review date	
35	Years with current manager	Dayslatelast30	

I. Number of Features obtained

Table 2 depicts the features obtained by the Chi-Square Feature Selection algorithm, Random Forest (Feature of Importance), Information Gain, Gain Ratio for Dataset_1. Table 3 gives the features obtained by the Chi-Square Feature Selection algorithm, Random Forest (Feature of Importance), Information Gain, Gain Ratio for Dataset_2. Table 4 gives the features obtained by the Chi-Square Feature Selection algorithm, Random Forest (Feature of Importance) Information Gain, Gain Ratio for Dataset_3.

TABLE 2: FEATURES OBTAINED BY CHI-SQUARE, INFORMATION GAIN, GAIN RATIO AND RF (FEATURE OF IMPORTANCE) FEATURE SELECTION METHODS FOR THE DATASET_1

Sl.No	Feature Selection Methods			
	Chi-Square Algorithm	Information Gain	Gain Ratio	Random Forest – Features of Importance

1	Age	Age	Age	Age
2	Attrition	Attrition	Attrition	Business Travel
3	Business Travel	Business Travel	Business Travel	Daily Rate
4	Daily Rate	Daily Rate	Daily Rate	Department
5	Department	Department	Department	Distance from Home
6	Distance from Home	Distance from Home	Distance from Home	Education
7	Education	Education	Education	Education Field
8	Education Field	Education Field	Education Field	Employee Count
9	Environment Satisfaction	Employee Count	Employee Count	Environment Satisfaction
10	Gender	Employee Number	Employee Number	Gender
11	Hourly Rate	Environment Satisfaction	Environment Satisfaction	Hourly Rate
12	Job Involvement	Gender	Gender	Job Involvement
13	Job Level	Hourly Rate	Hourly Rate	Job Level
14	Job Role	Job Involvement	Job Involvement	Job Role
15	Job Satisfaction	Job Level	Job Level	Job Satisfaction
16	Marital Status	Job Role	Job Role	Monthly Income
17	Monthly Income	Job Satisfaction	Job Satisfaction	Monthly Rate
18	Monthly Rate	Monthly Income	Marital Status	Number of Companies worked
19	Number of Companies worked	Monthly Rate	Monthly Income	Over18
20	Over18	Number of Companies worked	Monthly Rate	Overtime
21	Overtime	Percentage Salary Hike	Number of Companies worked	Percentage Salary Hike
22	Percentage Salary Hike	Performance Rating	Over18	Performance Rating
23	Performance Rating	Relationship Satisfaction	Overtime	Relationship Satisfaction
24	Standard Hours	Standard Hours	Percentage Salary Hike	Standard Hours
25	Stock Option Level	Stock Option Level	Performance Rating	Total Working Years
26	Total Working Years	Total Working Years	Standard Hours	Years at Company
27	Training time last year	Training time last year	Stock Option Level	Years in current role
28	Work life balance	Work life balance	Total Working Years	Years since last promotion
29	Years in current role	Years at Company	Training time last year	Years with current manager
30	Years with current manager	Years in current role	Work life balance	
31		Years since last promotion	Years at Company	
32		Years with current manager	Years in current role	
33			Years since last promotion	
34			Years with current manager	

TABLE 3: FEATURES OBTAINED BY CHI-SQUARE, INFORMATION GAIN, GAIN RATIO AND RF (FEATURE OF IMPORTANCE) FEATURE SELECTION METHODS FOR THE DATASET_2

Sl. No	**Feature Selection Methods**			
	Chi-Square Algorithm	**Information Gain**	**Gain Ratio**	**Random Forest – Features of Importance**

1	Emp Status ID	Employee_name	Employee_name	Employee_ID
2	Dept ID	Employee_ID	Marital Status ID	Emp Status ID
3	Performance Score ID	Married_ID	Emp Status ID	Dept ID
4	Pay Rate	Marital Status ID	Dept ID	Performance Score ID
5	Position	Gender ID	Performance Score ID	From Diversity Job Fair ID
6	State	Emp Status ID	From Diversity Job Fair ID	Pay Rate
7	Zip	Dept ID	Pay Rate	Position ID
8	DateofBirth	Performance Score ID	Attrition	Position
9	Sex	From Diversity Job Fair ID	Position ID	State
10	Marital Description	Pay Rate	Position	DateofBirth
11	Citizen Description	Attrition	State	Sex
12	Hispanic Latino	Position ID	DateofBirth	Hispanic Latino
13	Race Description	Position	Sex	Race Description
14	Date of Hire	State	Marital Description	Date of Hire
15	Date of Termination	Zip	Citizen Description	Date of Termination
16	Termination Reason	DateofBirth	Hispanic Latino	Termination Reason
17	Employment Status	Sex	Race Description	Employment Status
18	Department	Marital Description	Date of Hire	Department
19	Manager Name	Citizen Description	Date of Termination	Manager Name
20	Manager ID	Hispanic Latino	Termination Reason	Manager ID
21	Recruitment Source	Race Description	Employment Status	Recruitment Source
22	Performance Score	Date of Hire	Department	Performance Score
23	Engagement Survey	Date of Termination	Manager Name	Engagement Survey
24	Employee Satisfaction	Termination Reason	Manager ID	Employee Satisfaction
25	Special project count	Employment Status	Recruitment Source	Special project count
26	Last performance review date	Department	Performance Score	Last performance review date
27	Dayslatelast30	Manager Name	Engagement Survey	Dayslatelast30
28		Manager ID	Employee Satisfaction	
29		Recruitment Source	Special project count	
30		Performance Score	Last performance review date	
31		Special project count	Dayslatelast30	
32		Last performance review date		
33		Dayslatelast30		

TABLE 4: FEATURES OBTAINED BY CHI-SQUARE, INFORMATION GAIN, GAIN RATIO AND RF (FEATURE OF IMPORTANCE) FEATURE SELECTION METHODS FOR THE DATASET_3

Sl.No	**Feature Selection Techniques**			
	Chi-Square	**Information Gain**	**Gain Ratio**	**Random Forest – Features of Importance**
1	Age	Age	Age	Age
2	Business Travel	Business Travel	Business Travel	Business Travel
3	Department	Department	Department	Department
4	Distance from home	Distance from home	Distance from home	Distance from home
5	Education	Education	Education	Education
6	Education Field	Education Field	Education Field	Education Field
7	Gender	Employee Count	Employee Count	Gender
8	Job Level	Employee Number	Employee Number	Job Level
9	Job Role	Gender	Gender	Job Role
10	Marital Status	Job Level	Job Level	Monthly Income
11	Monthly Income	Job Role	Job Role	Number of companies worked
12	Number of	Marital Status	Marital Status	Percentage salary Hike

	companies worked			
13	Over18	Monthly Income	Monthly Income	Standard Hours
14	Percentage salary Hike	Number of companies worked	Number of companies worked	Stock Option Hours
15	Standard Hours	Over18	Over18	Total working years
16	Stock Option Hours	Percentage salary Hike	Percentage salary Hike	TrainingTimes last year
17	Total working years	Stock Option Hours	Stock Option Hours	Years at company
18	Training Times last year	Total working years	Total working years	Years since last promotion
19	Years at company	TrainingTimes last year	TrainingTimes last year	Year with current manager
20	Years since last promotion	Years at company	Years at company	
21	Year with current manager	Years since last promotion	Years since last promotion	
22		Year with current manager	Year with current manager	

Table 5 depicts the number of features obtained by the various feature selections for three datasets. From the table 5, it is clear that the Random Forest (Feature of Importance) gives a smaller number of features for the given three datasets than the other feature selection methods.

TABLE 5: NUMBER OF FEATURES OBTAINED IN THE GIVEN DATASETS USING CHI-SQUARE, INFORMATION GAIN, GAIN RATIO AND RF-FEATURE OF IMPORTANCE

Feature Selection Techniques	Number of Features obtained in Datasets		
	Dataset_1	**Dataset_2**	**Dataset_3**
Original Dataset	35	35	24
Chi-Square	30	**27**	21
Information Gain	32	33	22
Gain Ratio	34	31	22
RF- Feature of Importance	**29**	**27**	**19**

J. Performance Analysis

The machine learning classifiers like Random Forest, K- Nearest Neighbor, Gradient Boosting Tree, Neural Network and Naïve Bayes algorithms are considered to evaluate the performance of the feature selection methods for improving the classification accuracy of the models. The above-mentioned performance metrics are considered in this research work.

Result obtained for Dataset_1

Table 6a depicts the performance analysis of the feature selection methods for the given dataset_1 using Random Forest (RF), K Nearest Neighbor (K-NN), Gradient Boosting Tree (GBT), Neural Network (NN) and Naïve Bayes (NB) classifiers.

TABLE 6A: PERFORMANCE ANALYSIS OF THE ORIGINAL DATASET_1 USING RF, KNN, GBT, NN AND NB CLASSIFIERS

Performance Metrics	Classification Techniques				
	RF	**KNN**	**GBT**	**NN**	**NB**
Accuracy (in %)	43.099	46.44	**48.32**	42.98	41.65
TPR (in %)	52.61	52.94	**52.80**	51.78	50.26
FPR (in %)	67.17	61.08	**56.83**	68.89	69.04
Precision (in %)	45.81	49.01	**51.72**	44.54	43.66
Miss Rate (in %)	47.39	47.06	**47.2**	48.96	49.17

Specificity (in %)	32.83	38.92	**43.17**	31.74	30.25

Table 6b gives the performance analysis of the CS Processed Dataset_1 using Random Forest (RF), K Nearest Neighbor (KNN), Gradient Boosting Tree (GBT), Neural Network and Naïve Bayes classifiers.

TABLE 6B: PERFORMANCE ANALYSIS OF THE CS PROCESSED DATASET_1 USING RF, KNN, GBT, NN AND NB CLASSIFIERS

Performance Metrics	Classification Techniques				
	RF	**KNN**	**GBT**	**NN**	**NB**
Accuracy (in %)	69.63	69.97	**70.84**	68.45	67.91
TPR (in %)	76.07	74.59	**71.35**	73.63	72.21
FPR (in %)	35.62	34.77	**29.73**	36.47	37.15
Precision (in %)	68.79	68.81	**73.60**	67.89	66.34
Miss Rate (in %)	**23.93**	25.41	28.65	29.54	29.32
Specificity (in %)	64.38	65.23	**70.27**	63.08	62.87

Table 6c gives the performance analysis of the Information Gain Processed Dataset_1 using Random Forest (RF), K Nearest Neighbor (KNN), Gradient Boosting Tree (GBT), Neural Network and Naïve Bayes classifiers.

TABLE 6C: PERFORMANCE ANALYSIS OF THE INFORMATION GAIN PROCESSED DATASET_1 USING RF, KNN, GBT, NN AND NB CLASSIFIERS

Performance Metrics	Classification Techniques				
	RF	**KNN**	**GBT**	**NN**	**NB**
Accuracy (in %)	66.54	66.86	**68.75**	63.34	62.82
TPR (in %)	69.18	67.68	**65.45**	65.72	64.32
FPR (in %)	46.53	45.66	**40.82**	47.82	48.26
Precision (in %)	59.68	59.72	**62.51**	56.78	55.43
Miss Rate (in %)	**32.82**	36.52	39.76	40.63	40.43
Specificity (in %)	53.49	54.32	**59.38**	52.19	51.98

Table 6d gives the performance analysis of the Gain Ratio Processed Dataset_1 using Random Forest (RF), K Nearest Neighbor (KNN), Gradient Boosting Tree (GBT), Neural Network and Naïve Bayes classifiers.

TABLE 6D: PERFORMANCE ANALYSIS OF THE GAIN RATIO PROCESSED DATASET_1 USING RF, KNN, GBT, NN AND NB CLASSIFIERS

Performance Metrics	Classification Techniques				
	RF	**KNN**	**GBT**	**NN**	**NB**
Accuracy (in %)	65.46	65.77	**67.64**	62.23	61.73
TPR (in %)	**64.34**	66.57	68.29	64.61	63.21
FPR (in %)	47.42	46.75	**41.71**	48.73	49.37
Precision (in %)	58.57	58.61	**61.43**	55.65	54.32
Miss Rate (in %)	**33.91**	37.61	40.85	41.72	41.54
Specificity (in %)	52.38	53.21	**58.24**	51.28	50.87

Table 6e gives the performance analysis of the Random Forest (Feature of Importance) Processed Dataset_1 using Random Forest (RF), K Nearest Neighbor (KNN), Gradient Boosting Tree (GBT), Neural Network and Naïve Bayes classifiers.

TABLE 6E: PERFORMANCE ANALYSIS OF THE RF (FEATURE OF IMPORTANCE) PROCESSED DATASET_1USING RF, KNN, GBT, NN AND NB CLASSIFIERS

Performance Metrics	Classification Techniques				
	RF	**KNN**	**GBT**	**NN**	**NB**
Accuracy (in %)	71.76	72.30	**72.87**	69.81	68.27
TPR (in %)	75.37	**76.37**	70.54	69.32	68.54
FPR (in %)	32.18	32.8	**24.22**	35.47	36.02
Precision (in %)	71.97	71.45	**78.97**	70.54	69.80
Miss Rate (in %)	24.63	**23.63**	29.46	31.74	32.36
Specificity (in %)	67.82	67.2	**75.78**	65.32	64.88

ISSN: 2278-4632
Vol-10 Issue-1 No. 1 January 2020

Result obtained for Dataset_2

Table 7a depicts the performance analysis of the original dataset_2 using Random Forest (RF), K Nearest Neighbor (K-NN), Gradient Boosting Tree (GBT), Neural Network (NN) and Naïve Bayes (NB) classifiers.

TABLE 7A: PERFORMANCE ANALYSIS OF THE ORIGINAL DATASET_2 USING RF, KNN, GBT, NN AND NB CLASSIFIERS

Performance Metrics	Classification Techniques				
	RF	KNN	GBT	NN	NB
Accuracy (in %)	43.97	44.98	**48.32**	42.86	41.75
TPR (in %)	51.26	47.68	**52.76**	46.35	45.87
FPR (in %)	63.8	57.67	**56.58**	64.32	65.52
Precision (in %)	46.11	**52.34**	50.84	45.96	44.83
Miss Rate (in %)	48.74	52.32	**47.24**	54.54	55.72
Specificity (in %)	36.2	42.33	**43.42**	35.19	34.69

Table 7b depicts the performance analysis of the feature selection methods for Chi-Square processed dataset_2 using Random Forest (RF), K Nearest Neighbor (K-NN), Gradient Boosting Tree (GBT), Neural Network (NN) and Naïve Bayes (NB) classifiers.

TABLE 7B: PERFORMANCE ANALYSIS OF THE CHI-SQUARE PROCESSED DATASET_2 USING RF, KNN, GBT, NN AND NB CLASSIFIERS

Performance Metrics	Classification Techniques				
	RF	KNN	GBT	NN	NB
Accuracy (in %)	69.34	70.94	**70.84**	67.43	66.83
TPR (in %)	73.05	75.50	**74.45**	71.16	70.61
FPR (in %)	35.31	33.75	**32.87**	36.22	37.64
Precision (in %)	69.21	69.77	**70.04**	67.32	65.98
Miss Rate (in %)	**29.65**	24.5	25.55	31.25	32.56
Specificity (in %)	64.91	66.25	**67.13**	62.34	61.48

Table 7c depicts the performance analysis of the Information Gain processed dataset_2 using Random Forest (RF), K Nearest Neighbor (K-NN), Gradient Boosting Tree (GBT), Neural Network (NN) and Naïve Bayes (NB) classifiers.

TABLE 7C: PERFORMANCE ANALYSIS OF THE INFORMATION GAIN PROCESSED DATASET_2 USING RF, KNN, GBT, NN AND NB CLASSIFIERS

Performance Metrics	Classification Techniques				
	RF	KNN	GBT	NN	NB
Accuracy (in %)	58.43	59.85	**59.73**	56.32	55.72
TPR (in %)	62.16	64.41	**63.34**	60.24	59.72
FPR (in %)	44.42	42.84	**43.78**	47.35	46.53
Precision (in %)	58.32	58.68	**61.13**	56.31	54.87
Miss Rate (in %)	**38.54**	35.56	36.67	40.34	41.64
Specificity (in %)	55.82	55.34	**56.24**	51.43	50.59

Table 7d depicts the performance analysis of the Gain Ratio processed dataset_2 using Random Forest (RF), K Nearest Neighbor (K-NN), Gradient Boosting Tree (GBT), Neural Network (NN) and Naïve Bayes (NB) classifiers.

TABLE 7D: PERFORMANCE ANALYSIS OF THE INFORMATION GAIN PROCESSED DATASET_2 USING RF, KNN, GBT, NN AND NB CLASSIFIERS

Performance Metrics	Classification Techniques				
	RF	KNN	GBT	NN	NB
Accuracy (in %)	57.34	58.74	**58.64**	55.43	54.81
TPR (in %)	61.27	63.32	**62.25**	59.13	58.61
FPR (in %)	45.53	43.75	**44.69**	48.43	47.44

Precision (in %)	57.43	57.79	**60.24**	55.42	53.78
Miss Rate (in %)	**39.45**	36.67	**37.78**	41.45	42.73
Specificity (in %)	54.71	54.45	**55.35**	50.54	49.68

Table 7e depicts the performance analysis of the Random Forest (Feature of Importance) processed dataset_2 using Random Forest (RF), K Nearest Neighbor (K-NN), Gradient Boosting Tree (GBT), Neural Network (NN) and Naïve Bayes (NB) classifiers.

TABLE 7E: PERFORMANCE ANALYSIS OF THE RANDOM FOREST (FEATURE OF IMPORTANCE) PROCESSED DATASET_2 USING RF, KNN, GBT, NN AND NB CLASSIFIERS

Performance Metrics	Classification Techniques				
	RF	**KNN**	**GBT**	**NN**	**NB**
Accuracy (in %)	71.67	71.47	**72.59**	69.78	68.92
TPR (in %)	**82.3**	74.90	71.19	69.24	67.72
FPR (in %)	31.91	32.31	**25.60**	33.42	34.54
Precision (in %)	72.76	71.92	**78.18**	69.53	68.25
Miss Rate (in %)	**17.7**	25.1	28.81	30.92	31.64
Specificity (in %)	68.09	67.69	**74.4**	65.75	64.18

Result obtained for Dataset_3

Table 8a depicts the performance analysis of the original dataset_3 using Random Forest (RF), K Nearest Neighbor (K-NN), Gradient Boosting Tree (GBT), Neural Network (NN) and Naïve Bayes (NB) classifiers.

TABLE 8A: PERFORMANCE ANALYSIS OF THE ORIGINAL DATASET_3 USING RF, KNN, GBT, NN AND NB CLASSIFIERS

Performance Metrics	Classification Techniques				
	RF	**KNN**	**GBT**	**NN**	**NB**
Accuracy (in %)	44.93	45.81	**50.16**	43.82	42.72
TPR (in %)	**55.11**	49.44	54.26	47.53	46.22
FPR (in %)	64.74	59.04	**54.40**	65.85	66.15
Precision (in %)	44.75	**52.80**	52.67	43.86	42.57
Miss Rate (in %)	**44.89**	50.56	45.74	52.65	53.83
Specificity (in %)	35.26	40.96	**45.6**	34.37	33.85

Table 8b depicts the performance analysis of the Chi-Square processed dataset_3 using Random Forest (RF), K Nearest Neighbor (K-NN), Gradient Boosting Tree (GBT), Neural Network (NN) and Naïve Bayes (NB) classifiers.

TABLE 8B: PERFORMANCE ANALYSIS OF THE CHI-SQUARE PROCESSED DATASET_3 USING RF, KNN, GBT, NN AND NB CLASSIFIERS

Performance Metrics	Classification Techniques				
	RF	**KNN**	**GBT**	**NN**	**NB**
Accuracy (in %)	68.81	67.11	66.19	64.28	63.22
TPR (in %)	67.35	69.42	70.57	65.46	64.53
FPR (in %)	28.79	35.32	38.08	39.19	40.43
Precision (in %)	74.80	67.58	64.97	63.86	62.67
Miss Rate (in %)	32.65	30.58	29.43	33.69	34.56
Specificity (in %)	71.21	64.68	61.92	60.81	59.57

Table 8c depicts the performance analysis of the Information Gain processed dataset_3 using Random Forest (RF), K Nearest Neighbor (K-NN), Gradient Boosting Tree (GBT), Neural Network (NN) and Naïve Bayes (NB) classifiers.

TABLE 8C: PERFORMANCE ANALYSIS OF THE INFORMATION GAIN PROCESSED DATASET_3 USING RF, KNN, GBT, NN AND NB CLASSIFIERS

Performance Metrics	Classification Techniques				
	RF	**KNN**	**GBT**	**NN**	**NB**

Accuracy (in %)	**57.92**	58.22	55.28	53.37	52.34
TPR (in %)	56.43	58.31	**60.46**	54.55	53.64
FPR (in %)	**37.88**	36.43	39.19	40.28	41.34
Precision (in %)	**63.91**	57.47	53.86	52.77	51.56
Miss Rate (in %)	41.57	41.69	**39.42**	44.78	45.63
Specificity (in %)	**60.12**	55.79	50.81	49.92	48.66

Table 8d depicts the performance analysis of the Gain Ratio processed dataset_3 using Random Forest (RF), K Nearest Neighbor (K-NN), Gradient Boosting Tree (GBT), Neural Network (NN) and Naïve Bayes (NB) classifiers.

TABLE 8D: PERFORMANCE ANALYSIS OF THE GAIN RATIO PROCESSED DATASET_3 USING RF, KNN, GBT, NN AND NB CLASSIFIERS

Performance Metrics	Classification Techniques				
	RF	**KNN**	**GBT**	**NN**	**NB**
Accuracy (in %)	**56.81**	57.32	54.19	52.46	51.45
TPR (in %)	55.65	57.53	**59.68**	53.73	52.86
FPR (in %)	**38.06**	37.65	40.32	41.40	42.56
Precision (in %)	**61.13**	56.69	52.08	51.95	50.78
Miss Rate (in %)	42.79	42.81	**40.64**	45.96	44.85
Specificity (in %)	**59.35**	56.91	49.05	48.74	47.88

Table 8e depicts the performance analysis of the Random Forest (Feature of Importance) processed dataset_3 using Random Forest (RF), K Nearest Neighbor (K-NN), Gradient Boosting Tree (GBT), Neural Network (NN) and Naïve Bayes (NB) classifiers.

TABLE 8E: PERFORMANCE ANALYSIS OF THE RANDOM FOREST (FEATURE OF IMPORTANCE) PROCESSED DATASET_3 USING RF, KNN, GBT, NN AND NB CLASSIFIERS

PERFORMANCE METRICS	CLASSIFICATION TECHNIQUES				
	RF	**KNN**	**GBT**	**NN**	**NB**
Accuracy (in %)	74.04	72.20	74.04	69.15	68.42
TPR (in %)	84.55	80.57	81.74	78.66	77.65
FPR (in %)	35.76	36.21	34.56	37.32	38.65
Precision (in %)	68.81	69.09	72.55	67.62	66.18
Miss Rate (in %)	15.45	19.43	18.26	20.63	21.57
Specificity (in %)	64.24	63.79	65.44	61.35	60.68

VI. CONCLUSIONS

Employee turnover has been identified as a pivotal factor to curb the growth of organizations. In this research work, the feature selection techniques like Chi-Square, Information Gain, Gain Ratio and Random Forest (Feature of Importance) are used to find the most relevant feature for improving the classification accuracy. The performance of these feature selection methods are analyzed with classifiers like Random Forest, K-Nearest Neighbor, Gradient Boosting Tree, Neural Network, and Naïve Bayes. Through the analyzing of the result obtained in this paper, it is clear that the Random Forest (Feature of Importance) processed datasets gives better performance with three classifiers like RF, KNN, and GBT. From the results obtained it is shown that the Chi-Square feature selection method also performs better after RF (Feature of Importance) with the same three classifiers, than the NB and NN. Using these feature selection methods, the classification accuracy of the HR analytics datasets can be improved.

REFERENCES

[1] O'Connell, Matthew, and Mei-Chuan Kung. "The Cost of Employee Turnover." *Industrial Management* 49.1 (2007).
[2] Subramony, Mahesh, and Brooks C. Holtom. "The long-term influence of service employee attrition on customer outcomes and profits." *Journal of Service Research* 15.4 (2012): 460-473.
[3] Farkiya, Rashmi. "A Study on Overview of Employee Attrition Rate in India." *Pioneer Journal* 7 (2014).
[4] Mishra, Sujeet N., Dev Raghvendra Lama, and Yogesh Pal. "Human Resource Predictive Analytics (HRPA) for HR management in organizations." *International Journal of Scientific & Technology Research* 5.5 (2016): 33-35.0.

ISSN: 2278-4632
Vol-10 Issue-1 No. 1 January 2020

[5] King, Kylie Goodell. "Data analytics in human resources: A case study and critical review." *Human Resource Development Review* 15.4 (2016): 487-495.
[6] Durairaj, M., and T. S. Poornappriya. "Why Feature Selection in Data Mining Is Prominent? A Survey." *International Conference on Artificial Intelligence, Smart Grid and Smart City Applications*. Springer, Cham, 2019.
[7] Durairaj, M., and T. S. Poornappriya. "Choosing a spectacular Feature Selection technique for telecommunication industry using fuzzy TOPSIS MCDM." *International Journal of Engineering & Technology* 7.4 (2018): 5856-5861.
[8] Poornappriya, T. S., and M. Durairaj. "High relevancy low redundancy vague set based feature selection method for telecom dataset." *Journal of Intelligent & Fuzzy Systems* 37.5 (2019): 6743-6760.
[9] Khalid, Samina, Tehmina Khalil, and Shamila Nasreen. "A survey of feature selection and feature extraction techniques in machine learning." *2014 Science and Information Conference*. IEEE, 2014.
[10] Chandrashekar, Girish, and Ferat Sahin. "A survey on feature selection methods." *Computers & Electrical Engineering* 40.1 (2014): 16-28.
[11] Xue, Bing, et al. "A survey on evolutionary computation approaches to feature selection." *IEEE Transactions on Evolutionary Computation* 20.4 (2015): 606-626.
[12] Win, Thee Zin, and Nang Saing Moon Kham. "Information Gain Measured Feature Selection to Reduce High Dimensional Data." Seventeenth International Conference on Computer Applications (ICCA 2019), 2019.
[13] Moran, Michal, and Goren Gordon. "Curious feature selection." *Information Sciences* 485 (2019): 42-54.
[14] Chiew, Kang Leng, et al. "A new hybrid ensemble feature selection framework for machine learning-based phishing detection system." *Information Sciences* 484 (2019): 153-166.
[15] Huang, Changqin, et al. "An efficient automatic multiple objectives optimization feature selection strategy for internet text classification." *International Journal of Machine Learning and Cybernetics* 10.5 (2019): 1151-1163.
[16] Singh, Ajeet, and Anurag Jain. "Adaptive credit card fraud detection techniques based on feature selection method." *Advances in Computer Communication and Computational Sciences*. Springer, Singapore, 2019. 167-178.
[17] Tsamardinos, Ioannis, et al. "A greedy feature selection algorithm for Big Data of high dimensionality." *Machine learning* 108.2 (2019): 149-202.
[18] Cutler, Adele, D. Richard Cutler, and John R. Stevens. "Random forests." *Ensemble machine learning*. Springer, Boston, MA, 2012. 157-175.
[19] Han, Eui-Hong Sam, George Karypis, and Vipin Kumar. "Text categorization using weight adjusted k-nearest neighbor classification." *Pacific-asia conference on knowledge discovery and data mining*. Springer, Berlin, Heidelberg, 2001.
[20] Li, Ping, Qiang Wu, and Christopher J. Burges. "Mcrank: Learning to rank using multiple classification and gradient boosting." *Advances in neural information processing systems*. 2008.
[21] Ghazikhani, Adel, Reza Monsefi, and Hadi Sadoghi Yazdi. "Online neural network model for non-stationary and imbalanced data stream classification." *International Journal of Machine Learning and Cybernetics* 5.1 (2014): 51-62.
[22] Ren, Jiangtao, et al. "Naive bayes classification of uncertain data." *2009 Ninth IEEE International Conference on Data Mining*. IEEE, 2009.
[23] https://www.kaggle.com/pavansubhasht/ibm-hr-analytics-attrition-dataset
[24] https://www.kaggle.com/rhuebner/human-resources-data-set
[25] https://www.kaggle.com/vjchoudhary7/hr-analytics-case-study#general_data.csv