Efficient Data Mining Model for Employees Churn Prediction and Safety Measure

Anusha viswanadapalli[1], J.Vijayaraj[2], S.Shivaprasad[3,*], Vanga Mohan Aditya Reddy[4], K.Parthiban[5], P.Balamurugan[6]

[1,5]Faculty of Computer Science and Engineering, VFSTR (Deemed to be University), Guntur, Andhra Pradesh, India

[3]Faculty of Computer Science and Engineering, S.R.University, Warangal, Telangana, India

[2]Research Scholar, Department of Computer Science and Engineering, PEC, Puducherry, India

[4]Student of Computer Science and Engineering, VFSTR (Deemed to be University), Guntur, Andhra Pradesh, India

[6]Faculty of Mechanical Engineering, Kalasalingam Academy of Research and Education, Anand Nagar, Krishnnakoil, India

ABSTRACT

Employee Churn which is otherwise called representative turnover is an exorbitant issue for organizations. The genuine expense for supplanting a worker can frequently be very enormous. In this work, we aimed to understand why and when employees are most likely to leave a company i.e the probability of an active employee leaving the organization and the key factors of an employee leaving the organization. For this purpose, we created such standard dataset where we include those attributes that are helpful for our analysis to predict the factors that are responsible for an employee to leave a company. The attributes we used in the dataset are satisfaction level, last evaluation, a number of projects, monthly average hours, amount of time spend in the company, employees left the company, promotions in last 5years, departments, salary. Further, under these attributes, we include 603 data samples. It is also useful to the company to retain the employees' safety and secure without losing them in the organization for a long time. We applied various Machine Learning models such as, Logistic Regression Classifier, Random Forest Classifier, SVM to check that our dataset is resulting with accurate values or not and which model is predicting the best. Thus, after applying all the models to the dataset, the Random Forest Classifier is giving more accuracy that is about 97.2% when compared to all the other classification models. This Random Forest Classifier correctly depicts the factors responsible for an employee leaving the company.

Key Words: Factors, Employee left, Classifiers, Logistic Regression, Random Forest, SVM

I.Introduction

Employee Churn is also known as Employee turnover, which means that when employees leave the organization, it can also be called as terminates. The churn or turnover is considered as both the calculation of rates of people leaving the organization and as well as the individual terminates themselves. Employee turnover or churn also refers to the percentage of employees who leave a company or organization and then replaced by new employees. It is a very cost problem for any company. Suppose if we consider, accompany is investing between 4 weeks and 3 months training for the newly joined employees. This investment would be a loss for the company, if the newly joined employees decided to leave the company in the first year itself. Like this, so many problems will arise due to newly joined employees. Employee churn can be affected by salary, job satisfaction, work conditions, promotions, etc.., There are many advantages of retaining an employee rather than hiring a new one. Like hiring new employees has its own cost of hiring and training cost. New employee also takes time to acquire similar skills as the experience employee. So, because of these reasons, it is important to retain valuable employees. Hence we are going to predict the factors for an employee leaving the

company. We use machine learning models to predict the factors responsible for an employee leaving the company and also to find the accuracy of the project. So, finally this project can give an idea or an input to the company as to what are the steps to be taken to retain the employees and who are at high risk of leaving.

2.Literature Survey

Employee churn can be determined as leak or departure of creative central from the company [2]. The analysis is said as voluntary turnover. The analysis on voluntary turnover studies [3].

It was found that the strong predictors of voluntary turnover are satisfaction level, salary, promotion last 5 years, time spend in company, last evaluation, average monthly hours. And there are some personal factors such as age, marital status, education, gender also play a role in voluntary employee churn [4], [5], [6], [7]. The main factors are satisfaction level, salary, average monthly hours, time spend company, promotion last 5 years plays an important role in employee churn prediction [9-12].

High employee turnover or employee churn has very severe effect on the organization. It is difficult to replace employees with same skill and same talent; it also affects the ongoing work in the company and the productivity of the existing employees in the company. Acquiring new employees in the company has its own cost such as hiring cost and training cost of the company[8]. Organization will take this problem into consideration and will solve this problem by applying different machine learning algorithms to predict the employee churn and also will make them to take the necessary actions required to retain employee churn[13-26].

3.Proposed Methodology

First here we have a problem statement; we will be creating a dataset to our project by collecting the data. The data is stored as a csv (comma separated values) file. After creating the data we will be sending the data to the data preprocessing.

Here we will be carrying out our project in two phases. In the first phase we will be applying the basic methods such as Data Visualization, Cluster Analysis, and Correlation Analysis. By applying these methods we can draw the conclusions like what are the factors that are responsible an employee leaving the company.

In the second phase, after predicting the factors that are responsible for an employee to leave a company we are going to check how accurate are those, First to check that we should split some of the data to the training phase and testing phase. Here we are sending 70% of the data to the training phase and remaining 30% to the testing phase. Here splitting the data into training phase and testing phase we are going to find the accuracy. We will find accuracy using three methods namely Logistic Regression Classifier, Random Forest Classifier, SVM (Support Vector Machine) Classifier. Now we will compare the accuracy obtained from the three methods and by comparing we get the best accuracy for the Random Forest Classifier and we will be declared it as the best model.

Finally, we will construct the confusion matrix by using three methods namely Logistic Regression Classifier, Random Forest Classifier and SVM (Support Vector Machine) Classifier and we will also be calculating the precision and recall values.

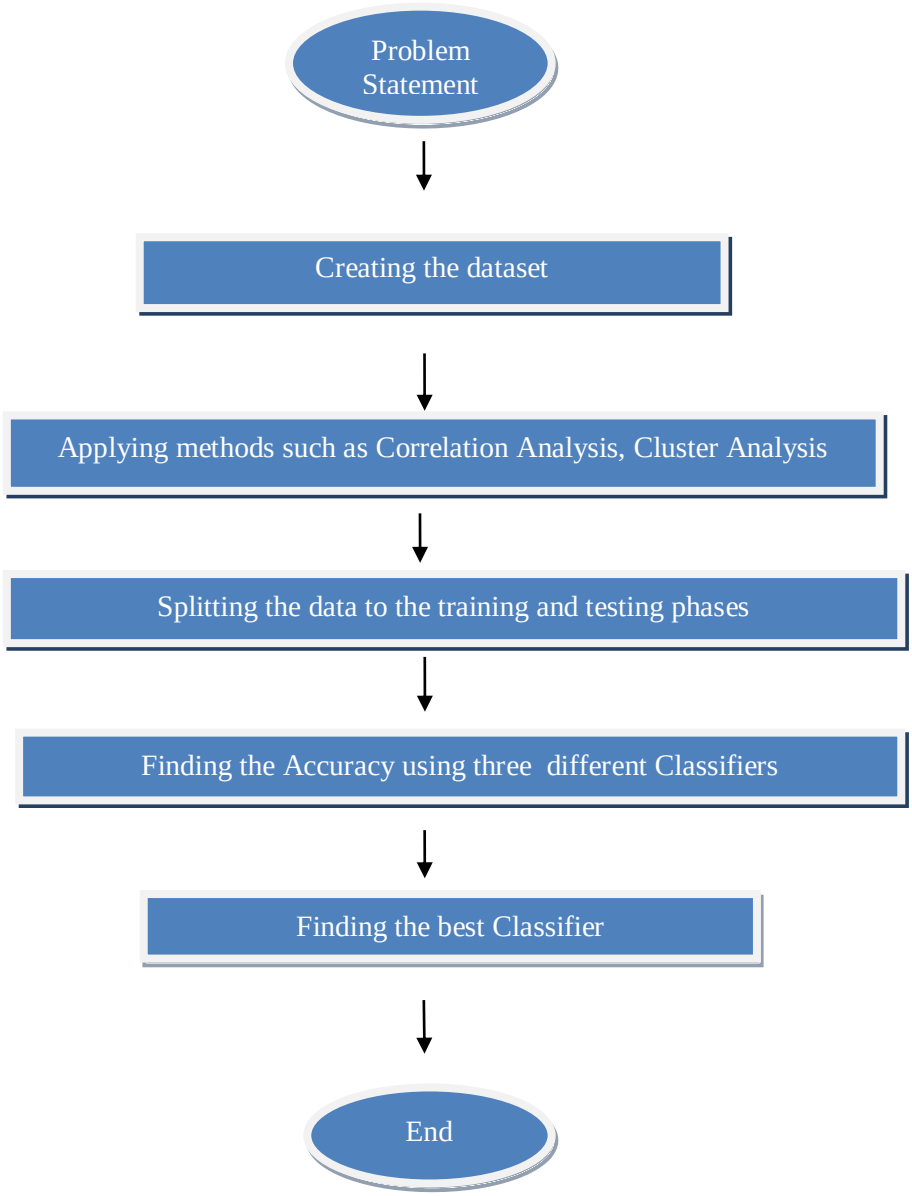


Fig.1 The proposed Methodology of our project is represented in Flow Chaw Chart

3.1

Database Creation

Data can originate from different sources, and needs to be checked before it can be put to use. This can be done by directly importing files that may already be available in .csv or .xlsx formats.

```
satisfaction level          float64
last evaluation             float64
number of project             int64
average monthly hours         int64
time spend company            int64
work accident                 int64
left                          int64
promotion last 5years         int64
Departments                  object
salary                       object
dtype: object
```

Fig. 2. Database creation

	A	B	C	D	E	F	G	H	I	J
1	satisfaction level	last evaluation	number of project	average monthly hours	time spend company	work accident	left	promotion last 5years	Departments	salary
2	0.36	0.41	3	167	10	0	0	0	support	low
3	0.71	0.78	4	227	2	0	0	0	support	low
4	0.94	0.9	4	144	4	0	0	0	support	low
5	0.51	0.76	4	140	3	0	0	0	support	low
6	0.83	0.48	4	220	3	1	0	0	support	low
7	0.22	0.62	3	180	3	0	0	0	support	low
8	0.66	0.89	4	173	4	0	0	0	support	low
9	0.14	0.58	3	179	5	0	0	0	support	low
10	0.16	0.96	5	137	5	1	0	0	technical	low
11	0.81	0.78	3	165	3	0	0	0	technical	high
12	0.73	0.94	3	177	3	0	0	0	technical	low
13	0.7	0.58	5	168	3	0	0	0	management	low
14	0.62	0.73	3	245	4	0	0	0	IT	low
15	0.5	0.83	5	258	2	0	0	0	IT	low
16	0.7	0.88	3	159	2	0	0	0	IT	high
17	0.53	0.73	3	163	3	1	0	0	IT	low
18	0.87	0.9	3	174	2	0	0	0	IT	low
19	0.59	0.6	3	214	2	1	0	0	product_mng	low
20	0.94	0.67	4	191	3	1	0	0	product_mng	high
21	0.2	0.53	5	272	5	0	0	0	product_mng	low
22	0.42	0.44	3	183	2	0	0	0	product_mng	medium
23	0.43	0.66	4	135	2	0	0	0	IT	high
24	0.43	0.76	6	154	2	0	0	0	management	high
25	0.77	0.86	5	238	3	0	0	0	management	high

Fig. 3. Database table

Here we included the attributes such as, satisfaction level, last evaluation, a number of projects, monthly average hours, amount of time spend in company, employees left the company, promotions in last 5years, departments, salary. We classify the employees like 1employee left the company and for employee does not leave the company and we collected totally "603" values from various sources..

3.2 Data Pre-processing

Datasets in any data mining applications can have missing data values. These missing values can get propagated due to lack of communication among the parameters in a data collection system. These missing values can affect the performance of a data mining system, and it should be noticed.

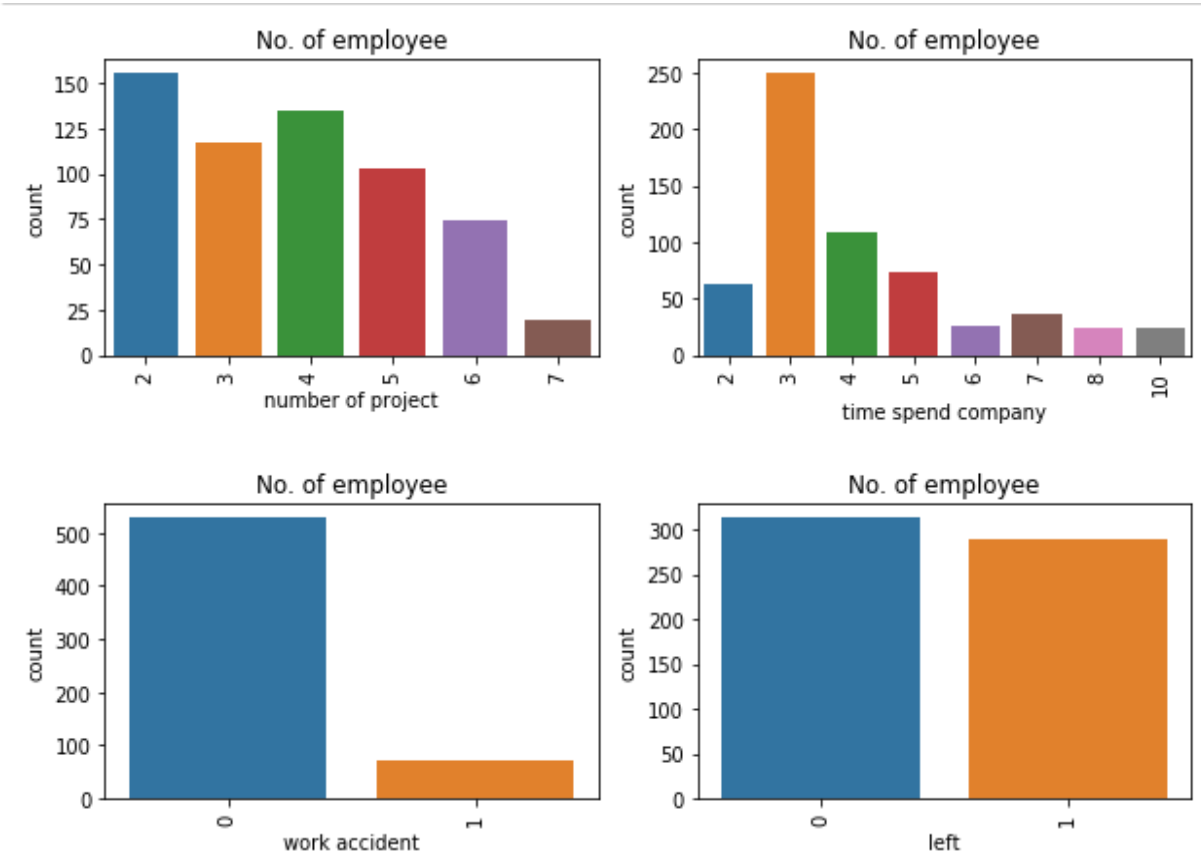
3.3 Proposed Methods

Phase 1:

The methods that we applied in phase 1 are Data Visualization, Cluster Analysis, and Correlation Analysis.

Data Visualization:

Data Visualization is the visual elements like charts, graphs and maps, data visualization tools provide an accessible way to see and understand trends, outliers and patterns in data. Data Visualization is another form of visual art that grabs our interest and keeps eyes on the message



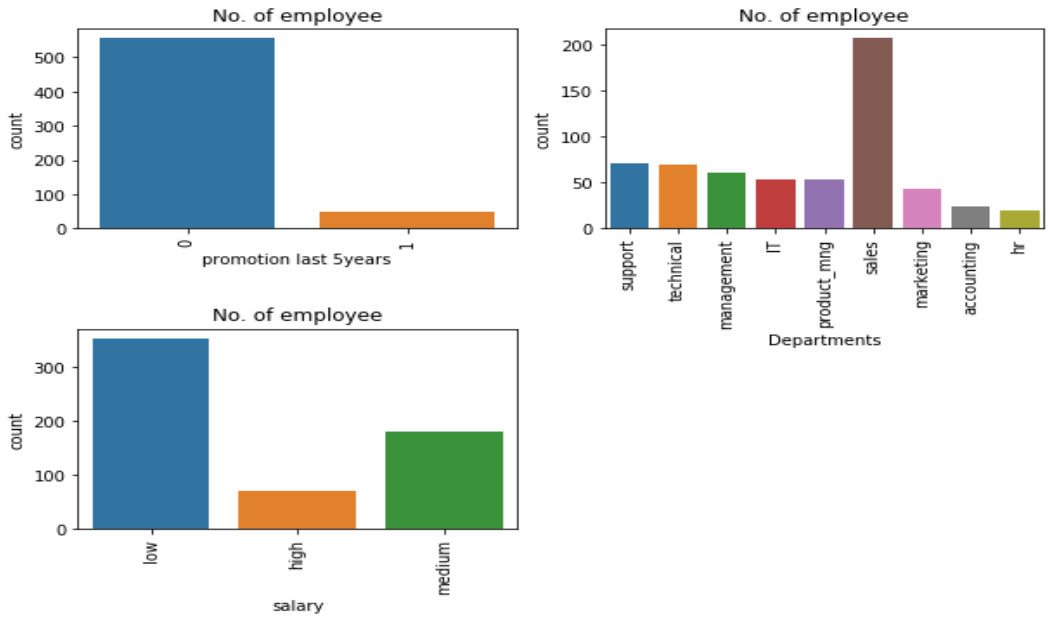


Fig. 4 Visualization of database created

1.Here in our project we apply data visualization techniques on attributes like satisfaction level, last evaluation, number of project, monthly average hours, amount of time spend in company, employees left the company, promotions in last 5years, departments, salary..

The some of the conclusions are:

As we can in the screenshots 1. The number of projects is generally 3 to 4

2. The number of promotions in last 5 years is very less.

3. Most of the employees are in the sales category of the department

4. Most of the salary is in the range between from low to medium.

2.Here we are performing Data Visualization on the people who left the company and who do not leave the company.

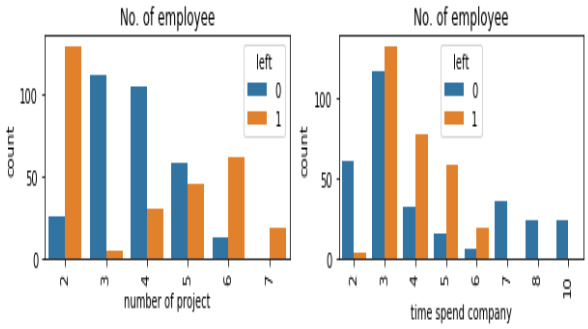
By drawing and analyzing the charts the some of the conclusion are:

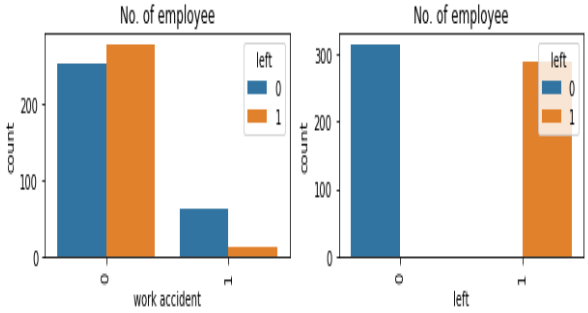
1. The employees who have number of projects from 6 to 7 are left more.

2. The person who spends 5 years in a company is having more chances of leaving.

3. People who did not get promotions left the company more.

4. The people who are having low salary are left more.





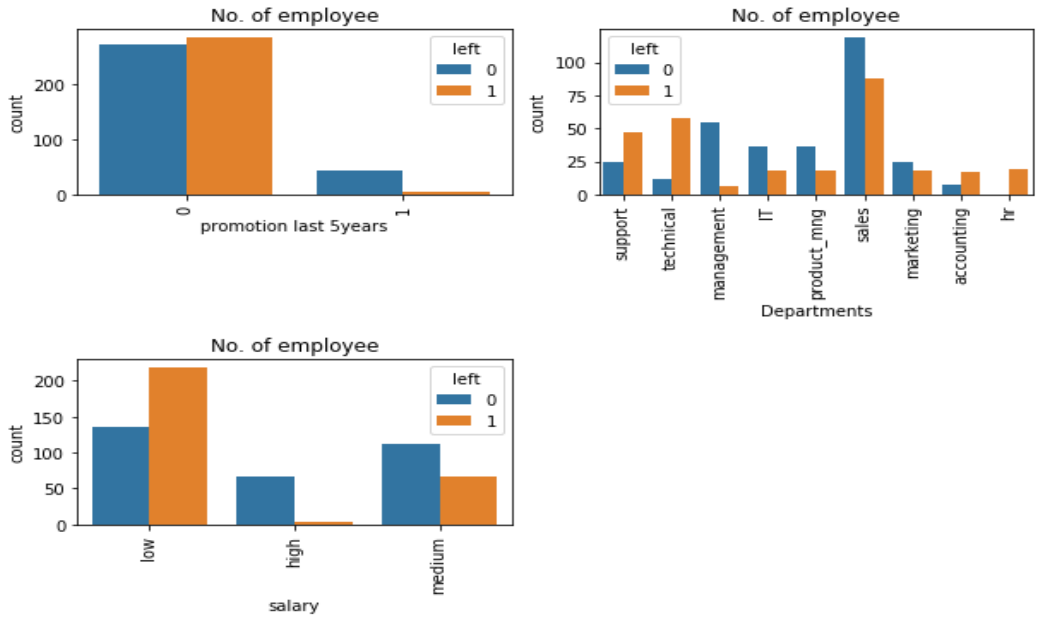


Fig.5 Statistical description of different attributes who left and stay in different Departments

3.4 Correlation Analysis:

A Correlation is a number between -1 and +1 that measures the degree of association between two attributes (call them as X and Y).A positive value for the correlation implies a positive association .In this case large values of X tend to be associated with large values of Y and small values of X tend to be associated with small values of Y.A negative value for the correlation implies that a negative or inverse association .In this case large values of X tend to be associated with small values of Y and small values of X tend to be associated with large values of Y.

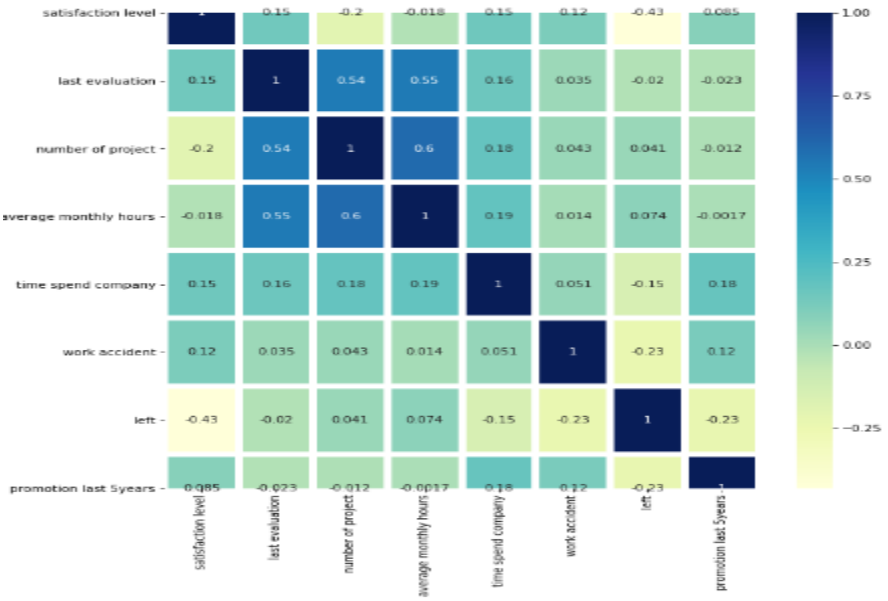


Fig 6.Correlation analysis

In our project the use of correlation analysis is:

Looking at the correlation matrix we can see that the people who left the company the highest negative correlation is with satisfaction level. Which implies as

satisfaction level increases the number of people who left the company decreases.

3.5 Cluster Analysis:

Clustering is an example for unsupervised learning.

.

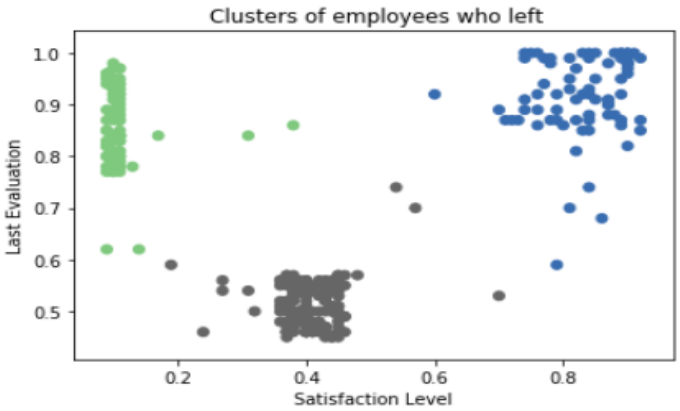


Fig. 7. Cluster analysis

Here by using cluster analysis method for our employee churn prediction project we draw some conclusions like:

There are three categories of employees:

Employees with high satisfaction and high performance.

Employees with low satisfaction and high performance

Employees with low satisfaction and low performance

Here in our project we use K-means Clustering and here k value = 3

Here we apply cluster analysis on the satisfaction level of employees who left the company.

Hence, these are methods we applied in the phase 1.

By applying these methods such as Data Visualization, Correlation Analysis and Cluster Analysis, we are going to predict the factors that are responsible for an employee leaving the company.

Hence, in phase 1, the factors are predicted and then we move to phase2.

Phase 2:

Here in phase 1, we predicted the model and in phase 2 we are going to check how accurate the model is and which is the best model. In this, first we need to split the data into training test and testing set.

3.6 Training and Testing a model:

Training and testing means that we need to send some of our code to training phase and testing phase. Here we are going to test our model. Here we split some our data to training phase and testing phase. Here we give more amount of our data to training phase and less amount of data to testing phase.

After a model has been processed by using the training set, you test the model by making predictions against the test set. Because the data in the testing set already contains known values for the attribute that you want

to predict, it is easy to determine whether the model's guesses are correct.

```
In [18]: from sklearn import preprocessing
         le = preprocessing.LabelEncoder()
         emp_df['salary']=le.fit_transform(emp_df['salary'])
         emp_df['Departments']=le.fit_transform(emp_df['Departments'])

In [19]: y=emp_df.left
         X = emp_df.drop('left',axis=1)
         from sklearn.model_selection import train_test_split
         X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)
```

Fig. 8. Testing and training model

Here our test size=0.3 means that we are sending 30% of our data to testing and remaining 70% of our data to training.

In 'y' we storing the employees left

In 'X' we are storing all column values except left column values.

This is about training and testing in our project.

After sending, some data to training and testing phases we are going to calculate the accuracy of the model by using three Classifier namely Logistic Regression Classifier, Random Forest Classifier, SVM (Support Vector Machine) Classifier.

3.7 Logistic Regression Classifier:

Logistic regression is a classification algorithm used to assign observations to a discrete set of classes. Now, Here we have the input and output prepared. We are going to create and define our classification model and we are going to represent it with an instance of the class Logistic Regression. The logistic regression model takes real-valued inputs and makes a prediction as to the probability of the input belonging to the default class (class 0). If the probability is > 0.5 we can take the output as a prediction for the default class (class 0), otherwise, the prediction is for the other class (class 1)model.

Logistic Regression has several optional parameters which are used to define the behavior of the model and approach.

The parameters are: *penalty *tol *fit_intercept *dual *c *intercept_scaling *class_weight *random_state *solver *max_iter *multi_class *verbose *warm_start *n_jobs *l1_ratio

```
In [21]: #Logistic Regression Classifier
         from sklearn.linear_model import LogisticRegression
         from sklearn import metrics
         logreg = LogisticRegression()
         logreg.fit(X_train, y_train)

Out[21]: LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
                            intercept_scaling=1, l1_ratio=None, max_iter=100,
                            multi_class='warn', n_jobs=None, penalty='l2',
                            random_state=None, solver='warn', tol=0.0001, verbose=0,
                            warm_start=False)

In [22]: from sklearn.metrics import accuracy_score
         print('Logistic regression accuracy: {:.3f}'.format(accuracy_score(y_test, logreg.predict(X_test))))

         Logistic regression accuracy: 0.773
```

Here we can observe the parameters of the logistic Regression Classifier and by using these parameters we can know the nature and behavior of the model. Here, the accuracy that we got using logistic Regression Classifier is 0.773.

3.8 Random Forest Classifier:

A random forest is a Meta estimator that fits various choice tree classifiers on different sub-tests of the dataset and utilizations averaging to improve the prescient precision and power over-fitting.Random forest, like its name implies, consists of a comprises of an enormous number of individual choice trees that work as a gathering. Each individual tree in the random forest spits out a class prediction and the class with the most votes becomes our model's prediction.

The parameters that are observed in Random Forest Classifier are:

*n_estimators	*criterion	*max_depth
*min_samples_split	*min_samples_leaf	*min_weight_fraction_leaf
*max_features	*max_leaf_nodes	*min_impurity_decrease
*min_impurity_split	*bootstrap	*oob_score
*n_jobs	*random_state	*verbose
*warm_start	*class_weight	

```
In [23]: #Random Forest Classifier
         from sklearn.ensemble import RandomForestClassifier
         rf = RandomForestClassifier()
         rf.fit(X_train, y_train)

Out[23]: RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini',
                                max_depth=None, max_features='auto', max_leaf_nodes=None,
                                min_impurity_decrease=0.0, min_impurity_split=None,
                                min_samples_leaf=1, min_samples_split=2,
                                min_weight_fraction_leaf=0.0, n_estimators=10,
                                n_jobs=None, oob_score=False, random_state=None,
                                verbose=0, warm_start=False)

In [24]: print('Random Forest Accuracy: {:.3f}'.format(accuracy_score(y_test, rf.predict(X_test))))

         Random Forest Accuracy: 0.972
```

Here we can observe the parameters of the Random Forest Classifier and by using these parameters we can know the nature and behavior of the model. Here the accuracy we got using Random Forest Classifier is 0.972

3.9 SVM (Support Vector Machine) Classifier:

Support Vector Machine (SVM) is a supervised Algorithm which can be used for both classification and regression challenges. However, it is generally utilized in grouping issues. In the SVM calculation, we plot every information thing as a point in n-dimensional space (where n is number of highlights

you have) with the estimation of each component being

the estimation of specific arrange. At that point, we perform characterization by finding the contrast

*c *cache_size
*coef *decision_sunction_shape
*gamma *kernel
*probability *random_state
*tol *verbose

between two classes quite well

The parameters that are observed in SVM Classifier are:

*class_weight
*degree
*max_iter
*shrinking

```
In [25]: #SVM Classifier
         from sklearn.svm import SVC
         svc = SVC()
         svc.fit(X_train, y_train)

Out[25]: SVC(C=1.0, cache_size=200, class_weight=None, coef0=0.0,
             decision_function_shape='ovr', degree=3, gamma='auto_deprecated',
             kernel='rbf', max_iter=-1, probability=False, random_state=None,
             shrinking=True, tol=0.001, verbose=False)

In [26]: print('Support vector machine accuracy: {:.3f}'.format(accuracy_score(y_test, svc.predict(X_test))))

         Support vector machine accuracy: 0.912
```

Here these are the parameters that we got using SVM Classifier and these parameters are used to know the nature and behavior of the model. The accuracy that we got using SVM Classifier is 0.912.

Here we found accuracy by using the three methods respectively.

3.10 Comparison between the three Classifiers:

Here we are using three classifiers named: 1.Logistic Regression Classifier

2. Random Forest Classifier

3. SVM Classifier.

By comparing the accuracy of the three Classifiers, in Logistic Regression we got accuracy as “0.773” and in Random Forest Classifier we got accuracy as “0.972” and finally in SVM Classifier we got accuracy as “0.912”.

Here, by comparing all the three Classifier we can observe that Random Forest Classifier is having more accuracy. So, here we can conclude that for our project Employee Churn Prediction Random Forest Classifier is the best Classifier method.

Now, we are constructing the Confusion Matrix and calculating the Precision and Recall Scores:

Confusion Matrix:

Confusion matrix is used for performance Measurement of the Classification Problem and where the output can be in two or more classes. Confusion matrix is a table with 4 different combinations of predicted and actual values.

The 4 different Combinations are:

1. True Positive: Predicted Positive and it is true.
2. True Negative: Predicted Negative and it is true.
3. False Positive: Predicted Positive and it is false
4. False Negative: Predicted Negative and it is false.

Precision and Recall Scores:

Precision: The proportion of positive identifications that are actually correct is called Precision.

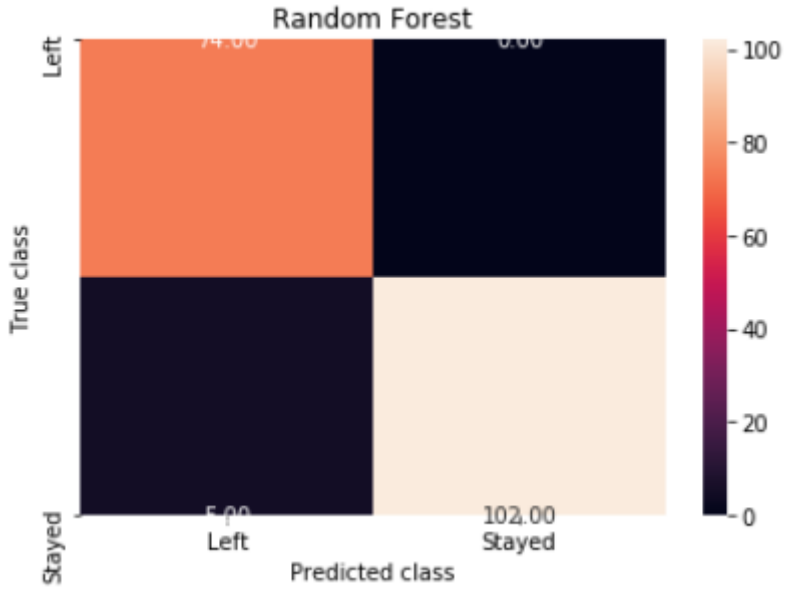
Precision= (True Positive)/ (True Positive + False Positive)

Recall: The proportion of Actual Positives identified correctly is called Recall.

Recall= (True Positive)/ (True Positive + False Negative)

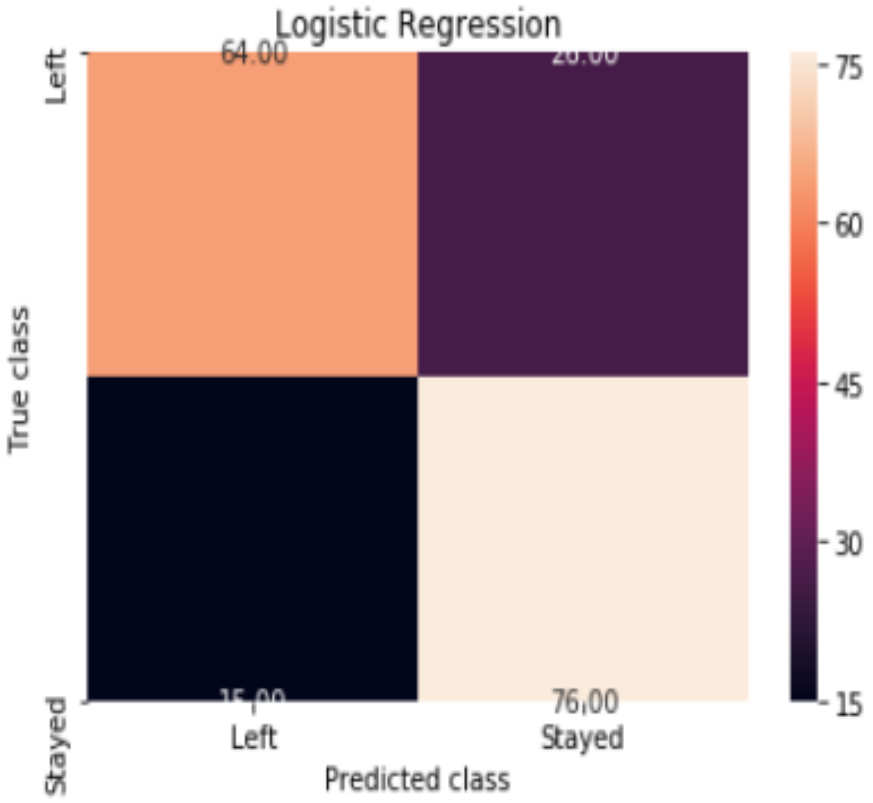
Random Forest Classifier:

	precision	recall	f1-score	support
0	0.95	1.00	0.98	102
1	1.00	0.94	0.97	79
accuracy			0.97	181
macro avg	0.98	0.97	0.97	181
weighted avg	0.97	0.97	0.97	181



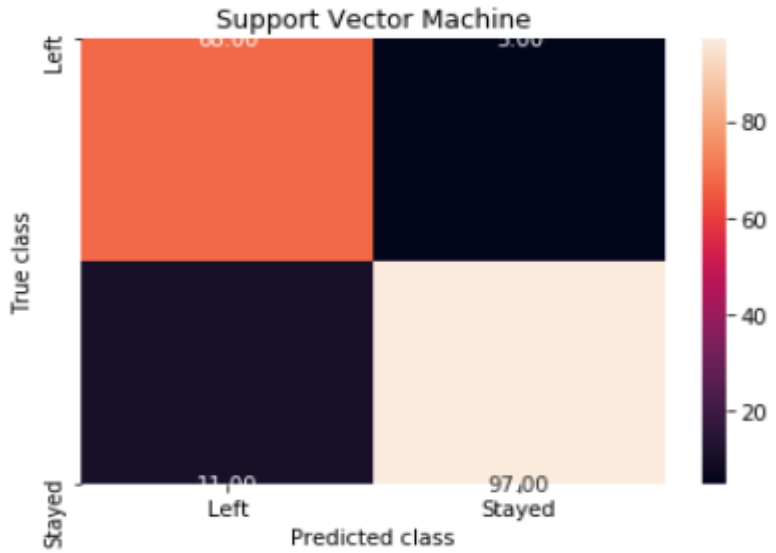
Logistic Regression Classifier:

	precision	recall	f1-score	support
0	0.84	0.75	0.79	102
1	0.71	0.81	0.76	79
accuracy			0.77	181
macro avg	0.77	0.78	0.77	181
weighted avg	0.78	0.77	0.77	181



SVM (Support Vector Machine) Classifier:

	precision	recall	f1-score	support
0	0.90	0.95	0.92	102
1	0.93	0.86	0.89	79
accuracy			0.91	181
macro avg	0.91	0.91	0.91	181
weighted avg	0.91	0.91	0.91	181



4. Results

Identifying no of Departments and salary:

```
In [10]: print(emp_df['Departments'].unique())
         print(emp_df.salary.unique())

['support' 'technical' 'management' 'IT' 'product_mng' 'sales' 'marketing'
 'accounting' 'hr']
['low' 'high' 'medium']
```

Fig. Number of Departments and salary

Identifying the number of people left:

```
In [9]: emp_df['left'].value_counts()

Out[9]: 0    314
        1    289
        Name: left, dtype: int64
```

Fig. number of people left

Identifying the percentage of people left:

```
In [11]: print(emp_df.left.value_counts()/len(emp_df)*100)

         0    52.072968
         1    47.927032
         Name: left, dtype: float64
```

Fig. percentage of people left

Identifying the some of the factors responsible for an employee leaving the company:

```
In [12]: emp_df.groupby('left').mean()
```

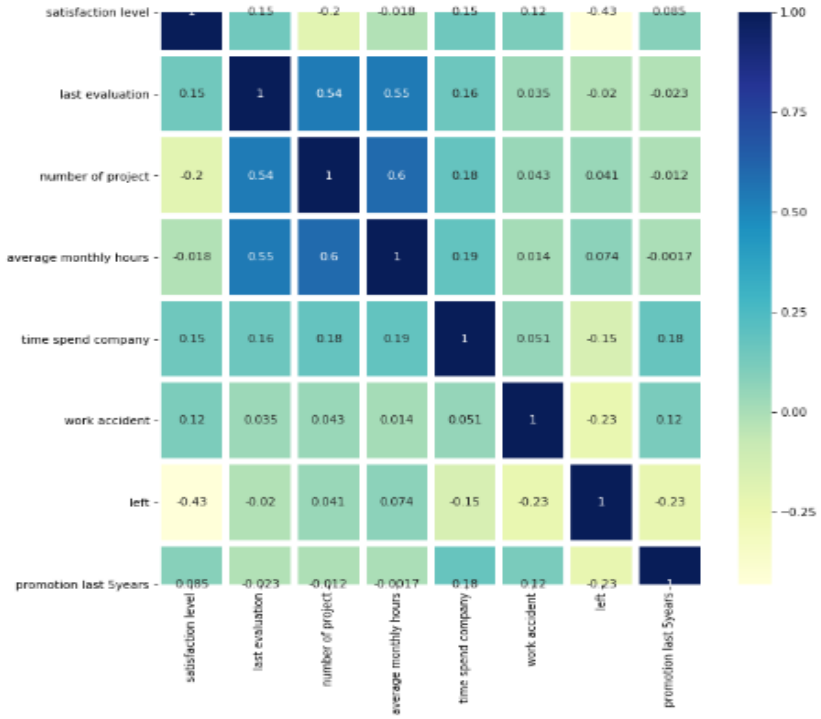
Out[12]:

left	satisfaction level	last evaluation	number of project	average monthly hours	time spend company	work accident	promotion last 5years
0	0.663248	0.717834	3.745223	199.27707	4.439490	0.194268	0.136943
1	0.433460	0.710657	3.865052	207.33910	3.854671	0.041522	0.013841

Fig. Factors responsible for an employee leaving the company

Here we can observe that the employees who left the company has low satisfaction level, worked more hours and low promotion rate.

```
In [15]: correlation_matrix = emp_df.corr()
plt.figure(figsize=(10,10))
sns.heatmap(correlation_matrix,annot=True,cmap="YlGnBu",linewidths=5.0)
plt.show()
```



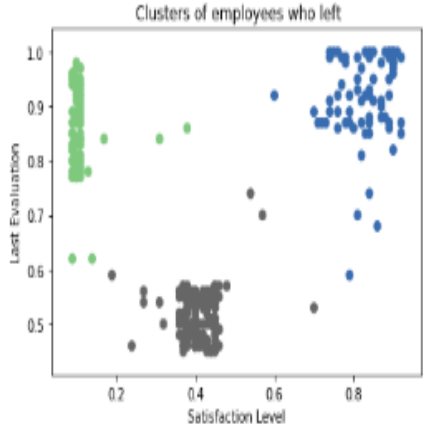
Identifying the main factor responsible for an employee to leave a company:

In, this we can observe that the satisfaction level is the main factor responsible for an employee leaving the company.

Clustering:

```
In [16]: from sklearn.cluster import KMeans
left_emp =  emp_df[['satisfaction level', 'last evaluation']][emp_df.left == 1]
kmeans = KMeans(n_clusters = 3, random_state = 0).fit(left_emp)
```

```
In [17]: left_emp['label'] = kmeans.labels_
plt.scatter(left_emp['satisfaction level'], left_emp['last evaluation'], c=left_emp['label'],cmap='Accent')
plt.xlabel('Satisfaction Level')
plt.ylabel('Last Evaluation')
plt.title(' Clusters of employees who left')
plt.show()
```



Identifying accuracy through Logistic Regression:

```
In [21]: #Logistic Regression Classifier
         from sklearn.linear_model import LogisticRegression
         from sklearn import metrics
         logreg = LogisticRegression()
         logreg.fit(X_train, y_train)

Out[21]: LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
                            intercept_scaling=1, l1_ratio=None, max_iter=100,
                            multi_class='warn', n_jobs=None, penalty='l2',
                            random_state=None, solver='warn', tol=0.0001, verbose=0,
                            warm_start=False)

In [22]: from sklearn.metrics import accuracy_score
         print('Logistic regression accuracy: {:.3f}'.format(accuracy_score(y_test, logreg.predict(X_test))))

         Logistic regression accuracy: 0.773
```

Identifying accuracy through Random Forest Classifier

```
In [23]: #Random Forest Classifier
         from sklearn.ensemble import RandomForestClassifier
         rf = RandomForestClassifier()
         rf.fit(X_train, y_train)

Out[23]: RandomForestClassifier(bootstrap=True, class_weight=None, criterion='gini',
                                max_depth=None, max_features='auto', max_leaf_nodes=None,
                                min_impurity_decrease=0.0, min_impurity_split=None,
                                min_samples_leaf=1, min_samples_split=2,
                                min_weight_fraction_leaf=0.0, n_estimators=10,
                                n_jobs=None, oob_score=False, random_state=None,
                                verbose=0, warm_start=False)

In [24]: print('Random Forest Accuracy: {:.3f}'.format(accuracy_score(y_test, rf.predict(X_test))))

         Random Forest Accuracy: 0.972
```

Identifying the Accuracy using SVM (Support Vector Machine Classifier):

```
In [25]: #SVM Classifier
         from sklearn.svm import SVC
         svc = SVC()
         svc.fit(X_train, y_train)

Out[25]: SVC(C=1.0, cache_size=200, class_weight=None, coef0=0.0,
             decision_function_shape='ovr', degree=3, gamma='auto_deprecated',
             kernel='rbf', max_iter=-1, probability=False, random_state=None,
             shrinking=True, tol=0.001, verbose=False)

In [26]: print('Support vector machine accuracy: {:.3f}'.format(accuracy_score(y_test, svc.predict(X_test))))

         Support vector machine accuracy: 0.912
```

Table.1 Accuracies of different models

S.No	Model	ACCURACY
1	Logistic Regression	**77.3**
2	Random Forest	**97.2**
3	SVM	**91.2**

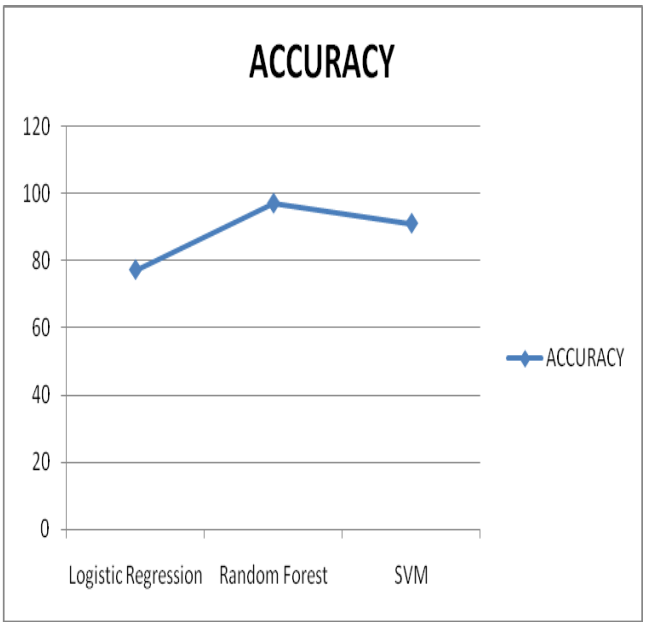


Fig. Accuarcies of model

In this we apply Logistic Regression Classifier, Random Forest Classifier and SVM Classifier.

By comparing the accuracy of the three Classifiers, in Logistic Regression we got accuracy as 77.3% and in Random Forest Classifier we got accuracy as 97.2% and finally in SVM Classifier we got accuracy as 91.2%. Here, by comparing all the three Classifier we can observe that Random Forest Classifier is having more accuracy.

Conclusion

In this work, we apply different classification methods to find out the factors responsible for an employee to leave company are such as lo salary, low satisfaction rate, low promotion rate, high number of projects. It is also useful to the company to retain the employees safety and secure without losing them in the organization for a long time. We applied Logistic Regression Classifier, Random Forest Classifier and SVM Classifier. By comparing the accuracy of the three Classifiers, in Logistic Regression we got accuracy as 77.3% and in Random Forest Classifier we got accuracy as 97.2% and finally in SVM Classifier we got accuracy as 91.2%. Random forest will be given good accuracy compare to all other methods. In future, Apply deep learning models to exactly to predict the churn of employees.

References

1.M. Stoval and N. Bontis, "Voluntary turnover: Knowledge management –Friend or foe?", Journal of Intellectual Capital, 3(3), 303-322, 2002.

2. Latchoumi, T. P., & Sunitha, R. (2010, September). Multi agent systems in distributed datawarehousing.

In 2010 International Conference on Computer and Communication Technology (ICCCT) (pp. 442-447). IEEE.

3.J. L. Cotton and J. M. Tuttle, "Employee turnover: A meta-analysis and review with implications for research", Academy of management Review, 11(1), 55-70, 1986.

4.L. M. Finkelstein, K. M. Ryanand E.B. King, "What do the young (old) people think of me? Content and accuracy of age-based metastereotypes", European Journal of Work and Organizational Psychology, 22(6), 633-657, 2013.

5.B. Holtom, T. Mitchell, T. Lee, and M. Eberly, "Turnover and retention research: A glance at the past, a closer review of the present, and a venture into the future", Academy of Management Annals, 2: 231-274, 2008

6. Ranjeeth, S., Latchoumi, T. P., & Victer Paul, P. (2019). Optimal stochastic gradient descent with multilayer perceptron based student's academic performance prediction model. Recent Advances in Computer Science and Communications. https://doi. org/10.2174/2666255813666191116150319.

7.C. von Hippel, E. K. Kalokerinos and J. D. Henry, "Stereotype threat among older employees: Relationship with job attitudes and turnover intentions", Psychology and aging, 28(1), 17, 2013

8. D. G. Allen and R. W. Griffeth, "Test of a mediated performance –Turnover relationship highlighting the moderating roles of visibility and reward contingency", Journal of Applied Psychology, 86(5), 1014-1021, 2001.

9. D. Liu, T. R. Mitchell, T. W.Lee, B. C. Holtom, and T. R. Hinkin, "When employees are out of step with coworkers: How job satisfaction trajectory and dispersion influence individual-and unit-level voluntary turnover", Academy of Management Journal, 55(6), 1360-1380, 2012.

10. Loganathan, J., Latchoumi, T. P., Janakiraman, S., & parthiban, L. (2016, August). A novel multi-criteria channel decision in co-operative cognitive radio network using E-TOPSIS. In Proceedings of the International Conference on Informatics and Analytics (pp. 1-6).

11. B. W. Swider, and R. D. Zimmerman, "Born to burnout: A meta-analytic path model of personality, job burnout, and work outcomes", Journal of Vocational Behavior, 76(3), 487-506, 2010.

12. T. M. Heckert and A. M. Farabee, "Turnover intentions of the faculty at a teaching-focused university", Psychological reports, 99(1), 39-45, 2006.

13. Ranjeeth, S., Latchoumi, T. P., & Paul, P. V. (2020). Role of gender on academic performance based on different parameters: Data from secondary school education. Data in brief, 29, 105257.

14. Ezhilarasi, T. P., Dilip, G., Latchoumi, T. P., & Balamurugan, K. (2020). UIP—A Smart Web Application to Manage Network Environments. In Proceedings of the Third International Conference on Computational Intelligence and Informatics (pp. 97-108). Springer, Singapore.

15. Loganathan, J., Janakiraman, S., & Latchoumi, T. P. A Novel Architecture for Next Generation Cellular Network Using Opportunistic Spectrum Access Scheme. Journal of Advanced Research in Dynamical and Control Systems,(12), 1388-1400.

16. Loganathan, J., Janakiraman, S., Latchoumi, T. P., & Shanthoshini, B. (2017). DYNAMIC VIRTUAL SERVER FOR OPTIMIZED WEB SERVICE INTERACTION. International Journal of Pure and Applied Mathematics, 117(19), 371-377.

17. Sekaran, K., Rajakumar, R., Dinesh, K., Rajkumar, Y., Latchoumi, T. P., Kadry, S., & Lim, S. (2020). An energy-efficient cluster head selection in wireless

sensor network using grey wolf optimization algorithm. TELKOMNIKA, 18(6), 2822-2833.

18. Balamurugan, K., Uthayakumar, M., Sankar, S., Hareesh, U. S., & Warrier, K. G. K. (2018). Effect of abrasive waterjet machining on LaPO 4/Y 2 O 3 ceramic matrix composite. Journal of the Australian Ceramic Society, 54(2), 205-214.

19. Bhasha, A. C., & Balamurugan, K. (2019). Fabrication and property evaluation of Al 6061+ x%(RHA+ TiC) hybrid metal matrix composite. SN Applied Sciences, 1(9), 977.

20. Balamurugan, K., Uthayakumar, M., Ramakrishna, M., & Pillai, U. T. S. (2020). Air Jet Erosion Studies on Mg/SiC Composite. Silicon, 12(2), 413-423.

21. Gowthaman, S., Balamurugan, K., Kumar, P. M., Ali, S. A., Kumar, K. M., & Gopal, N. V. R. (2018). Electrical discharge machining studies on monel-super alloy. Procedia Manufacturing, 20, 386-391.