**St. Clair Ace Acumen college**

Logo

Description automatically generated

**Title: To Understand why employees leave a company and to predict the next person who is going to leave.**

Gurpreet Singh (0775814)

Manav Singh (0775064)

DAB402: CAPSTONE PROJECT

Prof: Savita Seharawat

Table of Contents

[Abstract 4](#_Toc100404073)

[Introduction 5](#_Toc100404074)

[DATA TYPE DESCRIPTION: 5](#_Toc100404075)

[Literature review 7](#_Toc100404076)

[Managing Human talent 7](#_Toc100404077)

[Classification for Prediction 9](#_Toc100404078)

[Decision Tree 11](#_Toc100404079)

[Methodology 12](#_Toc100404080)

[Descriptive Analysis 14](#_Toc100404081)

[DATA TYPES 14](#_Toc100404082)

[Processing the Dataset 18](#_Toc100404083)

[CorrelationMatrix 19](#_Toc100404084)

[Visualization 20](#_Toc100404087)

[Employee Salary Turnover Distribution 20](#_Toc100404088)

[Employee Department Turnover Distribution 21](#_Toc100404089)

[Data Preprocessing 25](#_Toc100404090)

[Training and testing 26](#_Toc100404091)

[Summarizing the class distribution. 26](#_Toc100404092)

[Optimal Model 26](#_Toc100404093)

[K-means Clustering 26](#_Toc100404094)

[ROC graph 28](#_Toc100404095)

[Classifications 29](#_Toc100404096)

[Conclusion 30](#_Toc100404097)

[Reference 31](#_Toc100404098)

# Abstract

Nowadays, in the era of globalisation the percentage of being get employed increased and to make there future better or the way they think is seems to be bit easy. So, regards to this when any employee see that he/she is not growing weather in the case of income or the post they are working on they simply start doing their research to find a new possible window with which they can escape and go for good opportunity which ultimately leads the employee to leave the company for which they are working. A Company only holds its name as company if it has their employee with them, in other words without employee there is nothing to be name as company. The theme of my project will be Classification. The name of my data source is “Human Resources Analytic”, it contains 14,999 entries along with 10 attributes and with no null values. The datatype of the variables are divided in 2 float, 6 integer and 2 object.

Keywords: Outlier, Data mining and Decision Tree.

Research Questions: -

1. what factors contribute most to employee turnover?
2. Creation of model that can predict if a certain employee will leave the company or not?
3. How to stop employee turnover and to increase employee productivity and growth?

Tools: - Python

**https://github.com/JOHN-MANAV/CAPSTONE-PROJECT**

## Introduction

The Human Resource Analytics dataset is a simulated dataset taken from Kaggle and the main motive of this dataset is to understand the main problem because of which the most experienced and best employees leaving the company. So, we use data mining method to discover the pattern from the dataset and to make the decision. Data mining analyze the data and helps to bring up the hidden factors so that useful patterns and information can be generated. There are several classification techniques in data mining such as the decision tree, neural network, rough set theory, biasing theory, and fuzzy, Artificial Immune system (AIS), support Vector Machine (SVM), Genetic algorithm and nearest neighbour. In data mining task classification and prediction is among the popular task for knowledge discovery. There are few more dataset like our dataset which approximately do the same thing but with different process and techniques. It is becoming more often so HR department must do take some steps in advance based on existing or previous dataset. There are many ways, or we can say many techniques by which we can solve the problem of employee’s turnover ratio.

# DATA TYPE DESCRIPTION:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| FEATURES | TYPE | | DESCRIPTION | | |
| Satisfaction level | Floating | | An employee’s level of satisfaction in percentage. | | |
| Last evaluation | Floating | | An employee’s evaluation score given in the percentage. | | |
| Number project | Numeric | | The amount of the projects that employees have done. | | |
| Average monthly hours | Numeric | | The number of hours which employees worked in the month. | | |
| Time spend company | Numeric | | The numbers of the years of a worked at company. | | |
| Work accident | Numeric | | Whether an employees had an accident or not.  0 = No  1 = Yes | | |
| Promotion last 5years  Department | Numeric  Objects | | Whether an employee had promotion within the company in last five years.  0 = No  1 = Yes  The type of department in which employee worked, sales, technical, support, IT, Product, marketing, accounting, HR, management. | | |
| Salary | Objects |  |  | The type of the salary which employees got and which range, Low, Medium, High. |
| Left | Numeric |  |  | The employees whether left the company or stayed.  1 = leave  0 = Not leave |

Table 1: Data types of description

# Literature review

In my learn about part, there are my methodology to resolve it. First phase consist of the introduction involving my dataset’s subject and in literature evaluate I have described some associated work on Human Resource (HR)/Human Resource dataset which encompass classification for prediction, the choice tree method whereas different part furnish some end result and dialogue after that this will quit will my conclusion phase which will direct towards the future of the dataset like how HR branch can retailer there organization by using maintaining there personnel with there organisation. The methodology which I have used for my facts evaluation is as given below:  
Original Dataset, Data preprocessing, Feature selection, Training dataset ,Models: RF, logistic and choice tree, Classification. There are some related works which resemble with my work is as Listed below.

## Managing Human talent

If we speak about any organization, for there HR branch it’s a fundamental phase proper now for them to reserve their proficient personnel inside the organization. HR branch plans to control and organize the proficient worker inside the organization. It is additionally a way to exhibit leadership inside the company whose continuity motivate the branch in advancing. Talent administration is very essential and want some interest from HR professional. The authors are indebted to Madam Hanita Yusuf, Assistant Registrar at Universiti Teknologi MARA Terengganu for her permission to use the present employees’ contrast marks as coaching information set in this study.

Data mining is section of the procedure in Knowledge Discovery in Database (KDD). Data mining is a step in KDD and presently receives wonderful attention and is recognized as a newly rising analysis tool. Recently, records mining has given a super deal of issue and interest in the facts industry and in society as a whole. This is due to the extensive accessibility of good sized amount of information and the vital wishes for turning such statistics into beneficial records and knowledge. Data mining has a number of tasks such as classification and prediction; notion description; association; cluster analysis; outlier analysis; style and comparison analysis; statistical analysis and others. Computer utility interfaces with statistics mining tool can assist executives to make extra informative and goal decisions. Besides that, it can assist managers to retrieve, summarize and analyze associated facts to make wiser and greater informative decisions. There are very few studies associated to prediction utility in HR the use of this approach. However, this strategy is pretty popular in HR personnel choice problems. From the literature study, prediction applications in HRM are rare and there are some examples such as to predict the length of service, sales premiums, to reap persistence indices of insurance plan dealers and to analyze mis-operation behaviors of the operators. Over the years, information mining has worried in various techniques together with statistics, SVM, AIS, neural network, decision tree, genetic algorithm, and visualization techniques. Data mining has been utilized in many fields such as finance, marketing, manufacturing, health care, customer relationship and etc. However, its software in HRM is rare.  
Recently, with the new demand and expanded visibility, HRM seeks a extra strategic position by using turning to data mining methods. This can be achieved via figuring out generated patterns from the existing facts in HR databases as useful knowledge. Thus, this learn about concentrates on identifying the patterns that relate to the talent. The patterns can be generated by the use of some of the most important data mining techniques, such as the clustering technique which is used to list the employees with comparable characteristics, to group the performances and etc. From the association technique, patterns that are discovered can be used to companion the employee’s profile for the most excellent program/job, associated with employee’s mindset to performance and etc. In classification and prediction, the sample can be used to predict the percentage accuracy in employee’s performance, behavior, and attitudes, predict the overall performance growth for the duration of the overall performance period, and also become aware of the high-quality profile for distinctive worker and etc.

The matching of records mining troubles and talent management desires is very crucial. Therefore, it is very important to decide the suitable facts mining techniques. In HRM, there are some hobbies on solving HRM problems the usage of statistics mining approach. There are very few discussions on the uses of information mining related to employee’s overall performance prediction, assignment assignment, employee’s recruitment and many others. Due to these reasons, this find out about tries to use the statistics mining method for employee’s overall performance prediction as one of the strategies to predict the human brain in an organization. The cause of this find out about is to determine the employees’ performance through predicting their overall performance primarily based on the past experience know-how via previous overall performance assessment data. In this study, the classification technique will be used for human talent prediction.

## Classification for Prediction

Classification and prediction are amongst the methods that can produce sensible decision. Currently, many classification and prediction methods have been proposed by way of researchers in computing device learning, pattern recognition, and statistics. In this study, we are focusing on classification techniques in facts mining as part of computing device studying process.  
The classification technique has two phases; the first phase is studying manner the place the training data are analyzed by using the classification algorithm. Learned mannequin or classifier is represented in the structure of classification rules. The 2d section is classification process, where the take a look at statistics are used to estimate the accuracy of classification model or classifier. If the accuracy is viewed acceptable, the mannequin can be utilized to the new records to comprehend the prediction result. There are many methods that can be used for classification such as choice tree, Bayesian methods, Bayesian network, rulebased algorithms, neural network, guide vector machine, association rule mining, k-nearest-neighbor, case-based reasoning, genetic algorithms, tough units and fuzzy logic. In this study, our dialogue focuses on the three fundamental classification methods i.e. choice tree, neural network and k-nearest-neighbor. Decision tree and neural network are observed beneficial in creating predictive models in many fields. D. Decision Tree Techniques Decision tree can produce a model with rules that are human-readable and interpretable. The classification challenge using decision tree technique can be performed without complex computations and the approach can be used for both non-stop and express variables. This method is appropriate for predicting express consequences and less suitable for utility with time series data. Decision tree classifiers are quite famous strategies due to the fact the building of tree does now not require any domain professional knowledge or parameter placing and is terrific for exploratory understanding discovery. Currently, there are lots lookup that employed selection tree methods such as in electrical energy strength consumption, prediction of breast cancer, accident frequency, personnel selection, job attitudes and others. It is noted that, the choice tree is amongst the effective classification algorithms. Some of choice tree classifiers are C4.5/C5.0/J4.8, NBTree, SimpleCart, REPTree, BFTree and others. Association rule mining, k-nearest-neighbor, case-based reasoning, genetic algorithms, rough sets and fuzzy logic. In this study, our discussion focuses on the three principal classification strategies i.e. decision tree, neural community and k-nearest-neighbor. Decision tree and neural community are observed useful in creating predictive fashions in many fields. D. Decision Tree Techniques Decision tree can produce a model with rules that are human-readable and interpretable. The classification challenge using choice tree method can be carried out besides problematic computations and the approach can be used for both continuous and specific variables. This approach is suitable for predicting specific outcomes and less terrific for utility with time sequence data. Decision tree classifiers are pretty popular methods because the development of tree does now not require any domain specialist understanding or parameter placing and is fantastic for exploratory information discovery. Currently, there are tons lookup that employed selection tree methods such as in electrical energy strength consumption, prediction of breast cancer, accident frequency, personnel selection, job attitudes and others. It is cited that, the decision tree is among the effective classification algorithms. Some of selection tree classifiers are C4.5/C5.0/J4.8, NBTree, SimpleCart, REPTree, BFTree and others.

### Decision Tree

Classifier The C4.5 approach is one of the selection tree households that can produce both choice tree and rulesets; and construct a tree for the reason of enhancing prediction accuracy. Besides that, C4.5 models are effortless to apprehend as the regulations that are derived from the method have a very simple interpretation. The C4.5 / C5.0 / J48 classifier is among the most popular and powerful selection tree classifiers . C5.0 and J48 are the elevated variations of C4.5 algorithms. WEKA toolkit package deal has its own version known as J48. J48 is an optimized implementation of C4.5 rev. eight C4.5 creates an initial tree the use of the divide-and-conquer algorithm as described below:  
1) If all the instances in S belong to the equal class or S is small, the tree is a leaf labeled with the most conventional category in S.  
2) However, select a test based totally on single attribute with two or greater outcomes. Make this take a look at the root of the tree with one branch for every outcome of the test, partition S into corresponding subsets S1, S2, according to the outcome for each case, and practice to the identical method recursively to each subset.  
Usually there are many assessments that can be chosen in the remaining step. C4.5 makes use of two heuristic criteria to rank viable tests: information acquire that uses attribute decision measure, which minimizes the total entropy of the subset {Si}, and the default acquire ratio that divides information achieve by way of the records supplied by the check consequences

# MethodologyFlow chart of the training and testing process

Figure 1: Methodology

## Descriptive Analysis

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Column1** | **count** | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** |
| satisfaction\_level | 14999 | 1 | 0 | 0 | 0 | 1 | 1 | 1 |
| last\_evaluation | 14999 | 1 | 0 | 0 | 1 | 1 | 1 | 1 |
| number\_project | 14999 | 4 | 1 | 2 | 3 | 4 | 5 | 7 |
| average\_montly\_hours | 14999 | 201 | 50 | 96 | 156 | 200 | 245 | 310 |
| time\_spend\_company | 14999 | 3 | 1 | 2 | 3 | 3 | 4 | 10 |
| Work\_accident | 14999 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| left | 14999 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| promotion\_last\_5years | 14999 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |

Table 2: Descriptive Analysis

Total number of entries are 15,000 as python counting starts from 0 so that’s why in data descriptive analysis it shows 14,999.

### DATA TYPES

|  |  |  |
| --- | --- | --- |
| FEATURES | TYPE | DESCRIPTION |
| Satisfaction level | Floating | An employee’s level of satisfaction in percentage. |
| Last evaluation | Floating | An employee’s evaluation score given in the percentage. |

Figure 2: DATA TYPE

**Satisfaction level**: - It shows the percentage of satisfaction level among the employees’. So, satisfaction means that how much a employee is satisfied by his/her job in the time period they spent in the organisation.

**Last evaluation**: - It represent the percentage level of evaluation of every employee. In every organisation every individual employee is evaluated based on there work done in the time period they served in organisation.

|  |  |  |
| --- | --- | --- |
| FEATURES | TYPE | DESCRIPTION |
| Department | OBJECT | The type of department in which employee worked, sales, technical, support, IT, Product, marketing, accounting, HR, management. |
| Salary | OBJECT | The type of the salary which employees got and which range, Low, Medium, High. |

Table 3: DATA TYPES

We can see in the above Table 3, in our dataset there are two objects.

**Department**: - It represent the different field in which employees are working like Sales department, IT department and Marketing field etc.

**Salary**: - It represent that in which category the employees are falling. Like in our dataset there are only 3 type of salary category given as “Low, Medium, High”

|  |  |  |
| --- | --- | --- |
| FEATURES | TYPE | DESCRIPTION |
| Number project | Numeric | The amount of the projects that employees have done. |
| Average monthly hours | Numeric | The number of hours which employees worked in the month. |
| Time spend company | Numeric | The numbers of the years of a worked at company. |
| Work accident | Numeric | Whether an employees  had an accident or not.  0 = No  1 = Yes |
| Promotion last 5years | Numeric | Whether an employee had promotion within the company in last five years.  0 = No  1 = Yes |
| Left | Numeric | The employees whether left the company or stayed.  1 = leave  0 = Not leave |

Table 4: DATA TYPES

**Number Project**: - This attribute shows the total number of projects done by every individual employee.

**Average monthly Hours**: - This the total amount of Average Monthly Hours employee spent on work by every employee.

**Time spent in company**: - This represents that how many years a employees worked in the organisation.

**Work accident**: - This about the work hazards. Like some time, some accidents happen on work it can be because of employee or it can be natural.

**Promotion in last 5 years**: - It shows whether an employee get promoted in between the time spent by employee in the organisation.

**Left**: - It shoes weather the employee given in dataset stayed or left in the organisation after the 5 years.

# Processing the Dataset

To start the step of preprocessing the dataset is necessary to import some useful Python libraries.

* NumPy: Is a fundamental package to use linear algebra and random number capabilities.
* Pandas: Is a package to work with relational data as tables.
* Matplot and Seaborn is a package used to create plots and data visualization.

After that we have Load the Data set to the online platform (Google Collab) and we have seen that there are 15,000 entries and 10 attributes. Which further shows that our dataset contains 6 integers, 2 float and 2 object attributes.

Furthermore, for our convenience we have changed the Attributes names as mentioned.

'Satisfaction\_level’:  'satisfaction'

                        'last\_evaluation' :  'evaluation'

                        'number\_project' :  'projectCount'

                        'average\_montly\_hours' :  'averageMonthlyHours'

                        'time\_spend\_company' :  'yearsAtCompany'

                        'Work\_accident' :  'workAccident'

                        'promotion\_last\_5years' :  'promotion'

                        'sales'   : 'department'

                        'left'  : 'turnover'

CorrelationMatrixChart

Description automatically generated

The correlation is a very useful statistical analysis that describes the degree of relationship between two variables. Let´s see the table above for what relationship are in the data.

* Negative correlation of (-0.39) between satisfaction level and the employees that left the company.
* The highest positive correlation is between number of projects and average monthly hours (0.42).
* Last evaluation is high correlated to number project (0.35) and average monthly hours (0.34).
* Work accident have a low negative correlation (-0.15) and salary (-0.16) with employees that left.

From the heatmap, there is a positive (+) correlation between projectCount, averageMonthlyHours, and evaluation. Which could mean that the employees who spent more hours and did more projects were evaluated highly.

For the negative (-) relationships, turnover and satisfaction are highly correlated. we assume that people tend to leave a company more when they are less satisfied

# Visualization

We can see that in the visualization part on turnover with different attributes. By the visualizing the graphs we can be known about the outputs which are shown in the form of axis. Thus, we can assume about the people who have more turnover at specific period. So, this part is vital informative for people to understand and analyze the output. The use of visualize mot only to visualize but also to pass the knowledge to others person. We can state that this is knowledge transfer and information process. In our dataset, we have graphs which provides us the distribution information about the turnover. Turnover means employee left the company and left is shown by “1” and those who stayed is represented by “0”.

## Employee Salary Turnover Distribution

Chart, bar chart

Description automatically generated

Figure 3: Employee Salary Turnover Distribution

In the above graph, we can see that we have employee salary turnover distribution. So, we can get the information about the organization which sector has more turnover. By looking the above graph, we can say that the most people left the company because of the low salary or medium salary. We can see barely with high salary people had not left the organization.

### Employee Department Turnover Distribution

Chart, bar chart

Description automatically generated

Figure 4: Employee Department Turnover Distribution

The above graph of Employee Department Turnover Distribution, the sales, technical, and support department were the top 3 departments to have employee turnover. Whereas the management department had the smallest amount of turnover

#### Employee Department Turnover Distribution

Chart, bar chart

Description automatically generated

Figure 5: Employee Department Turnover Distribution

* More than half of the employees with 2,6, and 7 projects left the company
* Majority of the employees who did not leave the company had 3,4, and 5 projects
* All the employees with 7 projects left the company
* There is an increase in employee turnover rate as project count increases.

##### Stayed or left the organization

Chart, bar chart

Description automatically generated

Figure 6: Stayed or left the organization

This bar graph provides the information that the number of people who have an organization are 3571. And the account of people who stayed in the company are less as compared to the left.

###### Salaries by department

Chart, bar chart

Description automatically generated

Figure 7: Salaries by department

In the above graph, we can analyze that we have 10 department which are represented with different colors so that we can identify easily when we compare them each other.

* Most of the employees of the sales department have low or medium salaries, this may be due that in some companies the sales commission is paid separately.
* Technical department is in the second place where most of the employees receives low and medium salaries.

Satisfaction according to the project done in the number of years by the employee

Graphical user interface

Description automatically generated

Figure 8: satisfaction according to the project done in the number of years by the employee

From the employees that left with high performance had 4 or more years in the company and working on 5 or more projects had the Low satisfaction level, worked more hours and they Haven´t been promoted in the last five years.

After the graph part I have converted our Object attribute into Categorical attribute for further use in the analysis. Earlier we have “Sales” and “Department” attribute as Object and after converting we have converted our Object attribute into “Categorical” attribute.

# Data Preprocessing

In data preprocessing we have done scaling for our integer attributes and created the dummy values. Data Preprocessing is a approach that is used to convert the raw statistics into a smooth statistics set. In other words, on every occasion the information is gathered from extraordinary sources it is amassed in uncooked layout which is no longer possible for the analysis.

Need of data preprocessing:

For attaining higher outcomes from the utilized model in Machine Learning tasks the structure of the facts has to be in a ideal manner. Some distinct Machine Learning model wants records in a exact format, for example, Random Forest algorithm does not guide null values, consequently to execute random woodland algorithm null values have to be managed from the unique raw statistics set.  
Another issue is that the statistics set must be formatted in such a way that extra than one Machine Learning and Deep Learning algorithm are performed in one data set, and first-rate out of them is chosen.

# Training and testing

After processing of our dataset, we moved to the train-test split procedure which is used to estimate the performance of machine learning algorithms when they are used to make predictions on data. The observations in the training set structure the trip that the algorithm makes use of to learn. In supervised gaining knowledge of problems, each commentary consists of an determined output variable and one or more found enter variables.

The Test set is a set of observations used to evaluate the overall performance of the model using some performance metric. It is necessary that no observations from the training set are blanketed in the take a look at set. If the test set does include examples from the education set, it will be tough to check whether the algorithm has learned to generalize from the training set or has really memorized it.  
  
A program that generalizes well will be in a position to effectively function a mission with new data. In contrast, a program that memorizes the education statistics by using learning an overly complex mannequin should predict the values of the response variable for the training set accurately, but will fail to predict the price of the response variable for new examples. Memorizing the coaching set is known as over-fitting. A application that memorizes its observations may also not perform its undertaking well, as it should memorize relations and buildings that are noise or coincidence. Balancing memorization and generalization, or over-fitting and under-fitting, is a hassle frequent to many desktop studying algorithms. Regularization may be applied to many models to reduce over-fitting.

# Summarizing the class distribution.

We have done before sampling and After sampling of our dataset to stable our dataset because earlier our dataset was unstable.

# Optimal Model

By Optimization we mean tuning the model into the best performance model.

# K-means Clustering

# K-means clustering algorithm computes the centroids and iterates till we it finds most reliable centroid. It assumes that the quantity of clusters are already known. It is additionally called flat clustering algorithm. The quantity of clusters recognized from information by means of algorithm is represented by using ‘K’ in K-means. In this algorithm, the records points are assigned to a cluster in such a manner that the sum of the squared distance between the information points and centroid would be minimum. It is to be understood that less version within the clusters will lead to greater comparable information points inside same cluster.

Chart, scatter chart

Description automatically generated

Figure 9: Clustering

In the above figure, we have created three cluster with different colors. Each color represent its information.

Blue Color: These are the people who are hardworking but sad employee.

Red color: These are the people who are not hardworking and sad employee.

Green: These are the people who are hardworking and happy employee.

# ROC graph

# AUC-ROC curve is one of the most usually used metrics to consider the performance of computer getting to know algorithms in the instances the place we have imbalanced datasets. In this article we see ROC curves and its related ideas in detail.

The AUC value lies between 0.5 to 1.

Value between 0.5-0.6 of AUC denotes a bad classifier

Whereas value between 0.7 to 0.8 is considered to be acceptable

Exact 1 denotes an excellent classifier.

More the positive rate with less negative rate in ROC graph means higher performance model for future test regarding this dataset. I have best ROC curve of Random Forest.

A picture containing graphical user interface

Description automatically generated

Figure 10: ROC GRAPH

# Classifications

Models Train test split

0 Logistic Model 0.780965

1 Random Forest Model 0.975623

2 Decision Tree Model 0.946172

# Conclusion

# In the closing I have concluded that the employee who left the organization have labored for much less than 150hr/months or when they overwork. Like, for extra than 250hrs/month. Those employee who fall under the class of low or high evaluation have been viewed in extra turnover rate. If I talk about the quantity of project, these who have performed 2,6 or 7 mission are at most important risk for turnover. Employee satisfaction, yearsAtCompany, and contrast were the three biggest elements in figuring out turnover.

# Reference

1. H. Jantan, A. R. Hamdan, and Z. A. Othman, “Human talent prediction in hrm using c4. 5 classification algorithm,” International Journal on Computer Science and Engineering, vol. 2, no. 8, pp. 2526–2534, 2010.
2. <https://image.slidesharecdn.com/machine-learning-and-data-mining-14-evaluation-and-credibility-28463/95/machine-learning-and-data-mining-14-evaluation-and-credibility-16-728.jpg>
3. Y. Feng, et al., "Knowledge discovery in traditional Chinese medicine: State of the art and perspectives," Artificial Intelligence in Medicine, vol. 38, pp. 219-236, 2006. [2] E. Frank, Hall, M., et al., "Data mining in bioinformatics using Weka," Bioinformatics Application Note, vol. 20, pp. 2479-2481, 2004.
4. C. Combes, Meskens, N., Rivat,, C. & Vandamme J.P., "Using a KDD process to forecast the duration of surgery," International Journal of Production Economics, vol. 112, pp. 279-293, 2008.
5. C. L. Chang, & Chen, C.H., "Applying decision tree and neural network to increase quality of dermatologic diagnosis," Expert Systems with Applications, vol. 36, pp. 4035-4041, 2009.
6. A. S. Chang, & Leu, S.S., "Data mining model for identifying project profitablility variables," International Journal of Project Management, vol. 24, pp. 199-206, 2006.
7. B. Ekasingh, Ngamsomsuke, K., et al., "A data mining approach to simulating farmers' crop choices for integrated water resources management," Journal of Environmental Management, vol. 77, pp. 315-325, 2005.