

# We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

6,100

Open access books available

149,000

International authors and editors

185M

Downloads

Our authors are among the

154

Countries delivered to

TOP 1%

most cited scientists

12.2%

Contributors from top 500 universities



WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index  
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?  
Contact [book.department@intechopen.com](mailto:book.department@intechopen.com)

Numbers displayed above are based on latest data collected.  
For more information visit [www.intechopen.com](http://www.intechopen.com)



Chapter

# Predictive Data Analysis Using Linear Regression and Random Forest

*Julius Olufemi Ogunleye*

## Abstract

A statistical technique called predictive analysis (or analytics) makes use of machine learning and computers to find patterns in data and forecasts future actions. It is now preferred to go beyond descriptive analytics in order to learn whether training initiatives are effective and how they may be enhanced. Data from the past as well as the present can be used in predictive analysis to make predictions about what might occur in the future. Businesses can improve upcoming learning projects by taking actionable action after identifying the potential risks or possibilities. This chapter compares two predictive analysis models used in the predictive analysis of data: the Generalized Linear Model with Linear Regression (LR) and the Decision Trees with Random Forest (RF). With an RMSE (Root Mean Square Error) of 0.0264965 and an arithmetic mean for all errors of 0.016056967, Linear Regression did better in this analysis than Random Forest, which had an RMSE of 0.117875 and an arithmetic mean for all errors of 0.07062315. Through the hyper-parameter tuning procedure, these percentage errors can still be decreased. The combined strategy of combining LR and RF predictions, by averaging, nevertheless produced even more accurate predictions and will overcome the danger of over-fitting and producing incorrect predictions by individual algorithms, depending on the quality of data used for the training.

**Keywords:** data analysis, predictive data analysis, linear regression, random Forest, generalized linear model, decision trees

## 1. Introduction

Data analysis is the process of analyzing data to increase productivity and business growth. It involves steps like data cleansing, transformation, inspection, and modeling to perform market analysis, gather hidden data insights, enhance business studies, and generate reports based on the available data using tools like Tableau, Power BI, R and Python, Apache Spark, etc.

### 1.1 Predictive analysis

Predictive analytics also referred to as predictive analysis, is a subset of data analysis that focuses on creating future predictions from data. Other types of data analysis

exist, such as descriptive and diagnostic analysis, but predictive analysis is very well-liked in business analysis because it is crucial for making wise decisions. In any case, predictive analysis typically uses a variety of statistical models, techniques, and tools that all aid in understanding the patterns in datasets and making predictions. Data description and sorting are only a small part of predictive analysis. It largely relies on sophisticated models created to conclude from the data it encounters. To predict future trends, these models evaluate previous and present data using algorithms and machine learning. Depending on the particular requirements of people using predictive analysis, each model varies. Predictive analysis is very useful for assessing business decisions. This is because decisions effectively involve understanding their effects and basing them on projections of how a project, group, environment, or other entity will perform. A few typical fundamental models that are often used include:

- *Decision trees*: Use branching to illustrate the potential outcomes of each option or course of action.
- *Regression techniques*: Help in deciphering the connections between variables.
- *Neural networks*: Make use of algorithms to discover potential connections between data sets.

Prediction is a key component of data mining. Predictive analysis is a method for forecasting future patterns from current or historical data. As a result, businesses will be able to forecast future data trends. It can take many different forms, but some of the most advanced models make use of machine learning and artificial intelligence [1].

## **1.2 Models for predictive analysis**

Predictive analysis encompasses several different types of data analysis models. Most of these are regression models, which aim to determine the connections between two or more variables. They can aid in predicting the value of an unknown variable as the value of a known variable changes by recognizing the links between these variables.

### *i. Generalized Linear Model - Linear Regression*

The linear regression model is the most basic predictive analysis approach. In this approach, it is presumed that an unknown variable's value will scale linearly with a known variable's value. To track straightforward relationships and anticipate their future, such as expanding a customer base, linear regression models might be useful.

### *ii. Decision Trees - Random Forests*

Random forests are machine learning models that, among other things, can be used to model regression. They are appropriate for huge data collections with several variables and are made up of some decision trees.

### *iii. Neural Networks*

A cutting-edge tool for predictive analysis is neural networks. They are a collection of digital or biological neurons that talk to one another. A neural network changes shape and comes to new conclusions based on the data.

### 1.3 Predictive analysis tools

Aside from models, there are many specific tools available for conducting predictive analysis. These technologies aid in the discovery of connections that can be utilized to establish future predictions on data. They take on the bulk of the user's work by incorporating many statistical models used in the predictive analysis [2].

#### i. RapidMiner Studio

IBM provides a variety of predictive analytics technologies, including its premier SPSS Statistics software offering, as SaaS solutions. The system, which offers a variety of predictive analysis models, is primarily aimed at enterprise users.

#### ii. KNIME

Many of the functionalities of RapidMiner Studio are also available in the open-source data analysis tool KNIME. It appears to be made for more experienced users, though.

#### iii. IBM Predictive Analytics

A well-liked commercial tool for all types of predictive analysis is RapidMiner Studio. It aids in data collection, processing, and application of various statistical models to produce insightful results.

#### iv. SAP Predictive Analytics

SAP has a well-known SaaS product in the predictive analytics market. The developer of enterprise management software provides an analytics cloud for business users that is implemented similarly to IBM's.

## 2. Related works

### 2.1 Predictive analysis using linear regression with SAS (Bafna J., 2017)

According to Bafna J., a scalar dependent variable and one or more independent variables that are explanatory are connected using linear regression. The best-fitted straight line across the points in linear regression, one of the most widely used prediction methods, is referred to as a regression line. To demonstrate his thesis, the author gave the example of estimating people's weights based on their heights. The dependent variable in this situation is the weight, which needs to be predicted, and the independent variable is the height. The following outcomes were obtained using SAS' PROC REG to utilize linear regression to determine the relationship between two variables:

- The model has an R-squared score of 0.9541 (95.41%) > 0.7 (70%), suggesting that it suited the data well.
- With a P value of 0.00080.05, height was a significant variable in the model.

- To check for any outliers in the observations, the value of  $r$  was determined. If the value of  $r$  was greater than 2 or less than  $-2$ , the observations were considered outliers. (Note:  $-2 < r < 2$ .)

No observations deviated from the outliers range, leading the author to conclude that a major variable accounted for 95% of the person’s weight (height) [3].

## 2.2 Random forest model to identify factors associated with anabolic-androgenic steroid use (Manoochehri Z., Barati M., Faradmal J. and Manoochehri S., 2021)

Androgenic-anabolic steroids are one form of doping bodybuilders frequently take (AAS). In addition to breaking athletic ethics, using AAS would harm one’s physical and mental health. This study used a prototype willingness model to identify the key characteristics influencing AAS use among bodybuilders (PWM). A total of 280 male bodybuilders were chosen in 2016 utilizing multistage sampling from the bodybuilding clubs in Hamadan city for the analytical cross-sectional study. The data was then gathered through a self-administered questionnaire that included demographic data and PWM components, and a random forest model was also employed to evaluate the data. The most crucial elements in defining behavioral intention were behavioral willingness, attitude, and prior AAS usage. Additionally, BMI, attitude, subjective standards, and prototypes had the biggest impacts on predicting behavioral willingness to take AAS. Additionally, it was found that behavioral intention was more significant than behavioral willingness in predicting AAS usage. The findings indicate that, in comparison to the social reaction path, the reasoned action path has a stronger impact on predicting the use of AAS among bodybuilders. [4].

## 2.3 Linear regression analysis study (Kumari K. and Yadav S., 2021)

According to the authors, linear regression is a statistical method for determining the value of a dependent variable based on an independent variable and determining the relationship between two variables. It is a modeling method in which one or more independent variables are used to forecast a dependent variable, and, according to the authors, it is the most widely applied statistical method. The chapter provided an overview of the underlying ideas and examples of performing linear regression calculations using SPSS and Excel (**Table 1**).

Regression statistics	Values	Explanation
Multiple $R$	0.96332715	Correlation coefficient: 1 means perfect correlation and 0 means none
$R^2$	0.927999198	Coefficient of determination: How many points fall on the regression line. Here, 92% points fall within the line
Adjusted $R^2$	0.891998797	Adjusted $R^2$ : Adjusts for multiple variables, use with multiple variables
SE	516.3490153	
Observations	7	

**Table 1.**  
Summary output.

According to the table above, multiple R is the correlation coefficient, where 1 (one) denoted a perfect correlation, and 0 (zero) denoted a lack of correlation. The factors might account for 92% of the variation according to the R Square coefficient of determination. Adjusted R-squared was utilized because it was corrected for many factors. The best methods for figuring out the link between two variables, according to the authors, were correlation and linear regression. Correlation measures the strength of a linear relationship between two variables, whereas regression describes the relationship as an equation. In the essay, straightforward examples using SPSS and Excel were provided to illustrate linear regression analysis and urge readers to adopt these techniques to analyze their data [5].

### 3. Methods

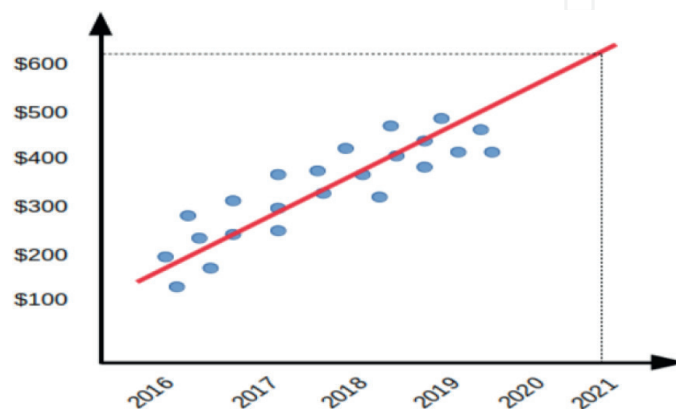
#### 3.1 Linear regression

A machine learning technique called linear regression enables the conversion of numerical inputs into numerical outputs and the fitting of a line through the data points. In other words, a method of modeling the relationship between one or more variables is called linear regression (**Figure 1**) [6]. From a machine learning perspective, this is done to accomplish generalization, which enables the model to forecast results for inputs it has never seen before. It is one of the most well-known concepts in statistics and machine learning, and since it is so crucial, it consumes a sizable chunk of almost every Machine Learning course [7].

$$y = mx + c.$$

where  $x$  is the score of the independent variable,  $m$  is the regression coefficient,  $c$  is the constant, and  $x$  is the independent variable, is the formula for every straight line on a plot.

The formula for this in machine learning is  $h(x) = w_0 + w_1 \cdot x$ , where  $x$  is the input feature,  $w_0$  and  $w_1$  are weights, and  $h(x)$  is the label (i.e.,  $y$ -value). The goal of linear regression is to identify the weights ( $w_0$  and  $w_1$ ) that produce the line that fits the input data the best (i.e.  $x$  features) [8].



**Figure 1.**  
Graphical illustration of a line (in red) generated by linear regression [3].

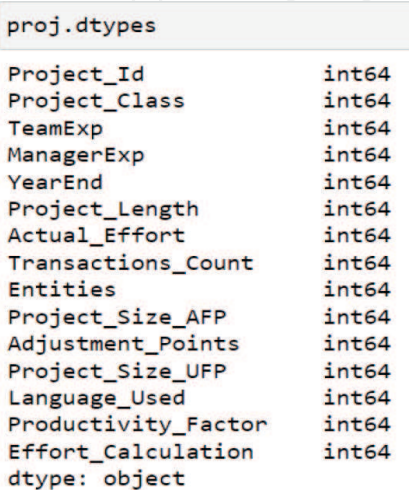
### 3.2 Random Forest

Machine learning methods for solving classification and regression issues include Random Forests. It uses ensemble learning, a method for solving complicated issues by combining a number of classifiers. The decision trees used in the random forest algorithm are numerous. The random forest algorithm creates a “forest” trained via bagging or bootstrap aggregation [9]. The accuracy of machine learning algorithms is increased by bagging, an ensemble meta-algorithm. Based on the predictions of the decision trees, the (random forest) algorithm determines the result. It makes predictions by averaging or averaging out the results from different trees [10]. The accuracy of the result grows as the number of trees increases. The decision tree algorithm’s drawbacks are eliminated by a random forest, which also decreases dataset overfitting and boosts precision. Without requiring numerous configurations in packages, it generates forecasts (like Scikit-learn) [11].

### 3.3 The random Forest Algorithm’s features

It overcomes the problem of overfitting in decision trees and is more accurate than the decision tree technique. In every random forest tree, a subset of characteristics is randomly chosen at the node’s splitting point, providing an efficient approach of addressing missing data [12].

The fundamental distinction between the random forest method and the decision tree algorithm is that the latter randomly selects the root nodes and groups the nodes (Figure 2). To produce the necessary forecast, the random forest uses the bagging approach [13]. Bagging entails using multiple samples of data (training data) as opposed to a single sample. Predictions are made using features and observations from a training dataset. Depending on the training information employed by the random forest algorithm, the decision trees generate various results. The highest ranking of these outputs will be chosen as the final output [7].



```
proj.dtypes
Project_Id          int64
Project_Class      int64
TeamExp            int64
ManagerExp         int64
YearEnd            int64
Project_Length     int64
Actual_Effort      int64
Transactions_Count int64
Entities           int64
Project_Size_AFP   int64
Adjustment_Points  int64
Project_Size_UFP   int64
Language_Used      int64
Productivity_Factor int64
Effort_Calculation int64
dtype: object
```

Figure 2.  
Data types.

### 3.3.1 Advantages of random forest

- It is capable of both classification and regression tasks.
- A random forest generates accurate predictions that are simple to comprehend.
- It has efficient handling of big datasets.
- Compared to the decision tree method, the random forest algorithm is more accurate at predicting outcomes [14].

### 3.3.2 Disadvantages of random forest

- More resources are needed for calculation when utilizing a random forest.
- It takes longer than a decision tree approach [15].

## 4. Discussions

In this chapter, the predictive analysis methods Linear Regression and Random Forest are compared. Data on software cost estimation was obtained from Kaggle, and the database contained details on the function point-measured size of the implemented program. To determine which model had the lowest error and anticipated the software cost, H2O AutoML was used. The expected performance of machine learning systems may be greatly impacted by erroneous and noisy input. Poor data quality, notably the significant occurrence of missing values and outliers, may result in inconsistent and incorrect conclusions. Therefore, a key stage in developing ML models is pre-processing data through selection, cleaning, reduction, transformation, and feature selection (**Table 2**).

The project dataset was divided into the train (80%) and test (20%) halves for modeling purposes using H2O. The first one was used to create models, while the second one was used to verify their capacity to estimate effort. H2O AutoML was used to apply two data mining prediction methods (Generalized Linear Models - Linear Regression (LR) and Decision Trees - Random Forest (RF)) for both dependent variables. In order to assess their potential utility for implementation inside companies, error and accuracy measures were contrasted. The error measurements used to evaluate the accuracy of software estimate models were *Mean Absolute Error (MAE)*, *Mean Squared Error (MSE)*, *Root Mean Squared Error (RMSE)*, and *Root Mean Squared Log Error (RMSLE)* (**Table 3**).

Linear Regression outperformed Random Forest, as shown in the graph above. Additionally, there is very little variation between the models' RMSLE and MAE. Therefore, it can be concluded that there were no significant errors. The difference in prediction accuracy between the algorithms was essentially small, and each one could be employed independently for the examination of the predictions. In conclusion, both models are quite good at making predictions. However, in this instance, linear regression outperformed the other model. If deployed for a specific company and trained using a homogeneous dataset, models may be more accurate (**Figure 3**).



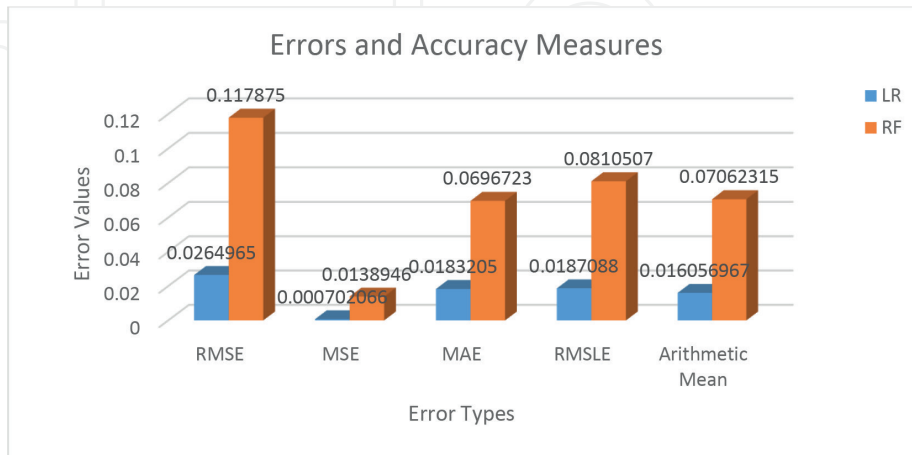
Project_ Id	Project_ Class	TeamExp	ManagerExp	YearEnd	Project_ Length	Actual_ Effort	Transactions_ count	Entities	Project_ Size_AFP	Adjustment_ Function_ Points	Project_ Size_UFP	Language	Productivity_ Factor	Effort_ Calculatlon
1	1	1	4	85	12	5152	253	52	305	34	302	1	20	6211
2	2	0	0	86	4	5635	197	124	321	33	315	1	29	9182
3	3	4	4	85	1	805	40	60	100	18	83	1	20	2013
4	4	0	0	86	5	3829	200	119	319	30	303	1	19	6107
5	5	0	0	86	4	2149	140	94	234	24	208	1	15	3592
6	6	0	0	86	4	2821	97	89	186	38	192	1	29	5409
7	7	2	1	85	9	2569	119	42	161	25	145	2	22	3479
8	8	1	2	83	13	3913	186	52	238	25	214	1	21	5007
9	9	3	1	85	12	7854	172	88	260	30	247	1	46	11,872
10	10	3	4	83	4	2422	78	38	116	24	103	1	31	3602
11	11	4	1	84	21	4067	167	99	266	24	237	1	24	6478
12	12	2	1	84	17	9051	146	112	258	40	271	1	62	15,994
13	13	1	1	84	3	2282	33	72	105	19	88	1	69	7261
14	14	3	4	85	8	4172	162	61	223	32	216	1	26	5743
15	15	4	4	85	9	4977	223	121	344	28	320	1	22	7678
16	16	3	2	85	8	1617	119	48	167	26	152	2	14	2269
17	17	4	3	85	8	3192	57	43	100	43	108	1	56	5600
18	18	4	4	86	14	3437	68	316	384	20	326	2	51	19,409
19	19	3	4	87	14	4494	100	386	395	21	340	2	45	17,751
20	20	4	2	86	5	840	58	34	92	29	86	1	14	1332
21	21	4	4	86	12	14,973	318	269	587	34	581	2	47	27,639
22	22	2	4	85	18	5180	88	170	258	34	255	1	59	15,187

Project_ Id	Project_ Class	TeamExp	ManagerExp	YearEnd	Project_ Length	Actual_ Effort	Transactions_ count	Entities	Project_ Size_AFP	Adjustment_ Function_ Points	Project_ Size_UFP	Language	Productivity_ Factor	Effort_ Calculation
23	23	2	4	86	5	5775	306	132	438	37	447	1	19	8266
24	24	4	1	87	20	10,577	304	78	382	39	397	1	35	13,291
25	25	1	4	86	8	3983	89	200	289	33	283	1	45	12,934
26	26	4	1	85	14	3164	86	230	316	33	310	1	37	11,626
27	27	2	0	86	6	3542	71	235	306	37	312	1	50	15,266
28	28	3	1	85	14	4277	148	324	472	39	491	1	29	13,640
29	29	4	4	85	16	7252	116	170	286	27	263	1	63	17,880
30	30	4	1	85	14	3948	175	277	452	37	461	1	23	10,197
31	31	4	3	86	6	3927	79	128	207	37	190	1	50	10,790

**Table 2.**  
 Sample of sourced data.

Algorithm	RMSE	MSE	MAE	RMSLE	Arithmetic Mean
LR	0.026497	0.000702066	0.018321	0.018709	0.016056967
RF	0.117875	0.0138946	0.069672	0.081051	0.07062315

**Table 3.**  
Errors and accuracy measures.



**Figure 3.**  
Graphical representation of the errors and accuracy measures.

## 5. Conclusion

Almost every field uses predictive analysis, even though it has drawn some criticisms. With more information, future outcomes can be predicted with relative accuracy. This makes it possible for organizations and businesses to make educated decisions to increase production. Learning the methods of predictive analysis has become essential for jobs in data science and business analysis since it has numerous applications in every conceivable industry. In this investigation, Random Forest had an RMSE of 0.117875 and arithmetic mean for all errors of 0.07062315, while Linear Regression had an RMSE of 0.0264965 and arithmetic mean of 0.016056967. Through the hyper-parameter tuning procedure, these percentage mistakes can still be decreased.

This study compares the Generalized Linear Model with Linear Regression and the Decision Trees with Random Forest models for predictive analysis. Additionally, a merged strategy was investigated, which used the arithmetic mean to combine the predictions of the two models. The outcomes demonstrated that distinct data mining techniques might be applied to make predictions. The combined strategy of combining LR and RF predictions by averaging nevertheless produced even more accurate predictions and will overcome the danger of over-fitting and producing incorrect predictions by individual algorithms, depending on the quality of data used for the training. To maintain accuracy in a project’s changing environment, it is important to remember that project management offices should ensure good input data quality and model updates.

## Acknowledgements

I, Julius Olufemi Ogunleye (the author), would like to express my gratitude to Ass. Prof. Zdenka Prokopova and Ass. Prof. Petr Silhavy for their support and

guidance in making this research work possible. This work was supported by the Faculty of Applied Informatics, Tomas Bata University in Zlín, under Projects IGA/CebiaTech/2022/001.

IntechOpen

IntechOpen


### **Author details**

Julius Olufemi Ogunleye  
Tomas Bata University in Zlin, Czech Republic

\*Address all correspondence to: [juliusolufemi@yahoo.com](mailto:juliusolufemi@yahoo.com)

### **IntechOpen**

---

© 2022 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/3.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. 

## References

- [1] Data Mining Techniques: Algorithm, Methods & top Data Mining Tools. Software Testing Help; March 2020. Available from: <https://www.softwaretestinghelp.com/data-mining-techniques/>
- [2] Steeneken F, Ackley D. A Complete Model of the Supermarket Business. BPTrends ■ January 2012
- [3] Bafna J. Predictive Analysis Using Linear Regression With SAS. Big Data Zone – DZone; 2017
- [4] Manoochehri Z, Barati M, Faradmal J, Manoochehri S. Random forest model to identify factors associated with anabolic-androgenic steroid use. BMC Sports Sci Med Rehabil. 2021;13(1):30
- [5] Kumari K, Yadav S. Linear regression analysis study. Curriculum in Cardiology—Statistics. 2021;4:33-36
- [6] Sumiran K. An overview of data mining techniques and their application in industrial engineering. Asian Journal of Applied Science and Technology. 2018;2:947-953
- [7] Mehmed K. Data Mining – Concepts, Models, Methods, and Algorithms. Edition – 2, Illustrated Edition. Wiley; 2011. ISBN 1118029127, 9781118029121
- [8] Varshini AGP, Kumari KA. Predictive analytics approaches for software effort estimation: A review. Indian Journal of Science and Technology. 2020;13:2094-2103
- [9] Nassif AB et al. Software development effort estimation using regression fuzzy models. Computational Intelligence and Neuroscience. 2019;2019:8367214
- [10] Azzeh MA, Nassif B, Banitaan S. Comparative analysis of soft computing techniques for predicting software effort based use case points. IET Software. 2018;12(1):19-29
- [11] Dejaeger K et al. Data mining techniques for software effort estimation: A comparative study. IEEE Transactions on Software Engineering. 2012;38(2):375-397. DOI: 10.1109/TSE.2011.55
- [12] Weiss GM, Davison BD. Data Mining. In: Bidgoli H, editor. Handbook of Technology Management. John Wiley and Sons; 2010
- [13] Berson A et al. An Overview of Data Mining Techniques. (Excerpts from the book 'Building Data Mining Applications for CRM' by Alex Berson, Stephen Smith, and Kurt Thearling). McGraw-Hill; 2005
- [14] Data Mining Techniques: Algorithm, Methods & top Data Mining Tools. Software Testing Help; April 2020. Available from: <https://www.softwaretestinghelp.com/data-mining-techniques/>
- [15] Kushwaha DS, Misra AK. Software Test Effort Estimation. (ACM SIGSOFT Software Engineering Notes – Page 3). May 2008;33(3)