Linear Attribute Projection and Performance Assessment for Signifying the Absenteeism at Work using Machine Learning

M. Shyamala Devi, Usha Vudatha, Sukriti Mukherjee, Bhavya Reddy Donthiri, S B Adhiyan, Nallareddy Jishnu

Abstract*: In recent times, with the technological advancement the industry and organization are transforming all their inflow and outflow operations into digital identity. At the outset, the name of the organization is also in the hands of the employee. One of the major needs of the employee in the working environment is to avail leave or vacation based on their family circumstances. Based on the health condition and need of the employee, the organization must extend their leave for the satisfaction of the employee. The performance of the employee is also predicted based on the working days in the organization. With this view, this paper attempts to analyze the performance of the employee and the number of working hours by using machine learning algorithms. The Absenteeism at work dataset from UCI machine learning Repository is used for prediction analysis. The prediction of absent hours is achieved in three ways. Firstly, the correlation between each of the dataset attributes are found and depicted as a histogram. Secondly, the top most high correlated features are identified which are directly fitted to the regression models like Linear regression, SRD regression, RANSAC regression, Ridge regression, Huber regression, ARD Regression, Passive Aggressive Regression and Theilson Regression. Thirdly, the Performance analysis is done by analyzing the performance metrics like Mean Squared Error, Mean Absolute Error, R2 Score, Explained Variance Score and Mean Squared Log Error. The implementation is done by python in Anaconda Spyder Navigator Integrated Development Environment. Experimental Result shows that the Passive Aggressive Regression have achieved the effective prediction of number of absent hours with minimum MSE of 0.04, MAE of 0.16, EVS of 0.03, MSLE of 0.32 and reasonable R2 Score of 0.89.

***Index Terms*: Machine Learning, MSE, MAE, R2 Score, Explained Variance Score and Mean Squared Log Error.**

Revised Manuscript Received on September 20, 2019.

M. Shyamala Devi, Associate Professor, Computer Science and Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Chennai, TamilNadu, India.

Usha Vudatha, III Year B.Tech Student, Computer Science and Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Chennai, TamilNadu, India.

Sukriti Mukherjee, III Year B.Tech Student, Computer Science and Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Chennai, TamilNadu, India.

Bhavya Reddy Donthiri, III Year B.Tech Student, Computer Science and Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Chennai, TamilNadu, India.

S B Adhiyan, III Year B.Tech Student, Computer Science and Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Chennai, TamilNadu, India.

Nallareddy Jishnu, III Year B.Tech Student, Computer Science and Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Chennai, TamilNadu, India.

I. INTRODUCTION

In machine learning, the prediction of the required attribute is done either by using regression or classification. However, it is mandatory for the organization to keep track of the employee records for updating the leave details for further process. The entire manpower of the organization is computerized for monitoring their activities. It is a known fact that the growth of the organization entirely depends on the welfare of the individual and their dedication. Any deviations in the health and stress of the employee directly affect the profit and the revenue of the organization

The paper is organized in which the literature survey is dealt with Section 2 followed by the preliminaries in the Section 3. Proposed work is discussed in Section 4 followed by the implementation and Performance Analysis in Section 5. The paper is concluded with Section 6.

II. RELATED WORK

A. Literature Survey

The university student admission is predicted based on the health and stress level of the student is done by using machine learning algorithm [1]. The dimensionality reduction is considered in predicting the target variable [2]. The various regression algorithms are used in predicting the performance of the student in the class [3]. The policy regulations and standards are also considered in predicting the target variable [4]. The prediction of the target variable is done through classification and then subjected to dimensionality reduction for the analysis of prediction of fall rate of the building roof [7]. The churn prediction of the customer is done through machine learning classification algorithms [6].

The machine learning feature selection and extraction methods can be used for the prediction of any factor in different application can be learnt through this article [7]–[14].

DOI: 10.35940/ijrte.C4405.098319

III. PERFORMANCE ASSESSMENT METRICS

The performance of any method can be accessed only by analyzing the effectiveness of the metrics. This metrics are used to analyze the real outcome of the method. Let us have a look on performance metrics that we have analyzed in this paper.

A. Mean Squared Error

Mean Squared Error is defined as the average squared difference between the predicted values and the original values. It is always non negative and values close to Zero or better. It is the Second moment of the error about the origin. The formula for finding the Mean Squared Error is given by,

$$\text{MSE}(y,\hat{y}) = \frac{1}{n_{\text{samples}}}\sum_{i=0}^{n_{\text{samples}}-1}(y_i-\hat{y}_i)^2.$$

The steps for finding the Mean Squared Error are shown below.

(1) Initially find the equation for the regression line.
(2) Predict the values of Y using the values of X obtained in step 1.
(3) Subtract the newly obtained Y values from the existing Y values. This gives the error.
(4) The error obtained is also known as the vertical distance of the given point from the regression line.
(5) Square the errors found in step 3.
(6) Sum the squares of the error values.
(7) Divide the sum obtained in step 5 by the total number of observations considered.

B. Mean Absolute Error

Mean Absolute Error is defined as the average distance between the given values and mean. Mean absolute error is always a non negative floating point. The best value is 0.0. The formula for finding the Mean Absolute Error is given by,

$$\text{MAE}(y,\hat{y}) = \frac{1}{n_{\text{samples}}}\sum_{i=0}^{n_{\text{samples}}-1}\left|y_i-\hat{y}_i\right|.$$

The steps for finding the Mean Absolute Error are shown below.

(1) Initially find the mean values of the testing data set.
(2) Find the difference between each value of X and mean value.This gives the error values.
(3) Add all the error values obtained in step 2.
(4) Divide the sum obtained in step 3 by the total number of observations considered.

C. R2 Score

R2 score is defined as the proportion of the variance in the dependent values that is predictable from the independent value(s) .R2 score lies between 0 and 100%. Two variables are said to be perfectly correlated if the R2 score is 100%. The formula for finding the R2 score is given by,

$$R^2(y,\hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}$$

The steps for finding the R2 score are shown below.

(1) Predict the y value of the i`th sample.
(2) Subtract each predicted value of y from the respective true value.
(3) Sum up all the values obtained in step 2.
(4) Find the mean value of y.
(5) Subtract mean value from each value of y.
(6) Sum up all the values obtained in step 5.
(7) Divide the value obtained in step 3 by the value obtained in step 6.
(8) Subtract the value obtained in step 7 from 1.

D. Explained Variance Score

Explained Variance Score measures the proportion to which the model accounts for the variation of the given data set. The possible value of Explained variance score is 1.0. The formula for finding the Explained Variance Score is given by,

$$explained\_variance(y,\hat{y}) = 1 - \frac{Var\{y-\hat{y}\}}{Var\{y\}}$$

The steps for finding the Explained Variance Score are shown below.

(1) Find the estimated target output of y and correct target output of y.
(2) Find the variance of the difference between estimated output of y and correct output and y.
(3) Find the variance of estimated target output of y.
(4) Divide the value obtained from step 2 by value obtained in step 3.
(5) Subtract the value obtained in step 4 from 1.

E. Mean Squared Log Error

Mean Squared Log Error is a regression metric corresponding to the expected value of the squared logarithmic error. where log is the natural logarithm of x. It is highly used when the target values are having exponential growth. The formula for finding the Mean Squared Log Error is given by,

$$\text{MSLE}(y,\hat{y}) = \frac{1}{n_{\text{samples}}}\sum_{i=0}^{n_{\text{samples}}-1}(\log_e(1+y_i)-\log_e(1+\hat{y}_i))^2.$$

The steps for finding the Explained Variance Score are shown below.

(1) Add 1 to each true value of y and find the logarithmic value.
(2) Add 1 to each predicted value of y and find the logarithmic value.
(3) Find the difference between the values obtained in step 1 and step 2.
(4) Sum up all the values obtained in step 3.
(5) Divide the value obtained in step 4 by total number of observations considered.

IV. PROPOSED WORK

In this proposed work, we have regression machine learning algorithms for predicting the absent probability of a person at workplace. Our contribution of predicting the probability of absence is achieved in three ways.

(i) Firstly, the correlation between each of the dataset attributes are found and depicted as a histogram correlation matrix.
(ii) Secondly, the top most high correlated features are identified which are directly fitted to the regression models like Linear regression, SRD regression, RANSAC regression, Ridge regression, Huber regression, ARD Regression, Passive Aggressive Regression and Theilson Regression.
(iii) Thirdly, the Performance analysis is done by analyzing the performance metrics like Mean Squared Error, Mean Absolute Error, R2 Score, Explained Variance Score and Mean Squared Log Error.

A. System Architecture

The architecture of proposed system is shown in Fig. 1

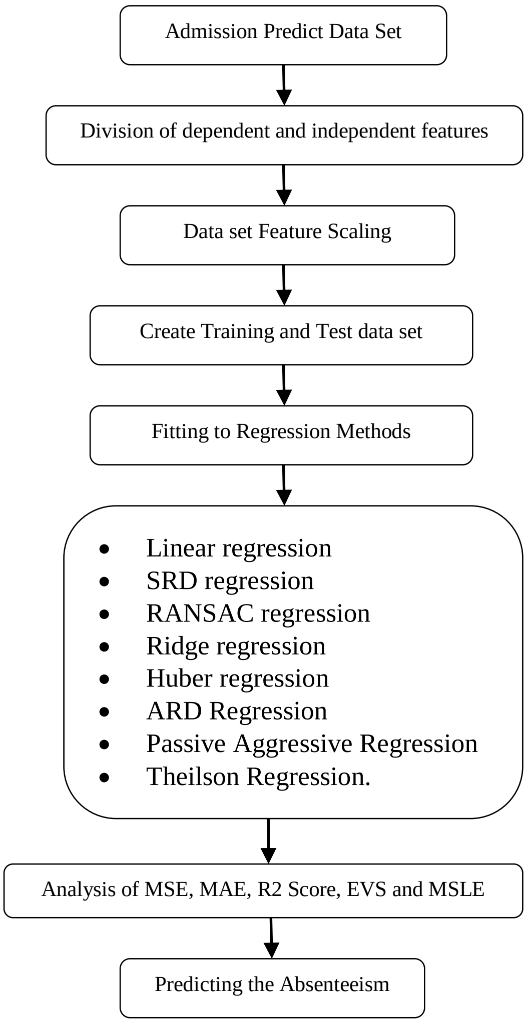
Admission Predict Data Set
Division of dependent and independent features
Data set Feature Scaling
Create Training and Test data set
Fitting to Regression Methods
• Linear regression
• SRD regression
• RANSAC regression
• Ridge regression
• Huber regression
• ARD Regression
• Passive Aggressive Regression
• Theilson Regression.
Analysis of MSE, MAE, R2 Score, EVS and MSLE
Predicting the Absenteeism

Fig. 1 System Architecture

V. IMPLEMENTATION AND PERFORMANCE ANALYSIS

A. Data Set Information

The Absenteeism at work Data Set from UCI Machine Learning Repository is used for implementation with 20 independent attribute and 1 Absenteeism time in hours dependent attribute and they are as follows,

(1) Individual identification (ID)
(2) Reason for absence (ICD). It has 28 categories (I to XXI) and is shown in Table. 2.
(3) Month of absence
(4) Day of the week (Monday (2), Tuesday (3), Wednesday (4), Thursday (5), Friday (6))
(5) Seasons (summer (1), autumn (2), winter (3), spring (4))
(6) Transportation expense
(7) Distance from Residence to Work (kilometers)
(8) Service time
(9) Age
(10) Work load Average/day
(11) Hit target
(12) Disciplinary failure (yes=1; no=0)
(13) Education (high school (1), graduate (2), postgraduate (3), master and doctor (4))
(14) Son (number of children)
(15) Social drinker (yes=1; no=0)
(16) Social smoker (yes=1; no=0)
(17) Pet (number of pet)
(18) Weight
(19) Height
(20) Body mass index
(21) Absenteeism time in hours (Target- Dependent Attribute)

B. Prediction of absenteeism in hours

Absenteeism at work Data Set is implemented to analyze the feature importance from the set of all attributes and is shown in Fig. 2.

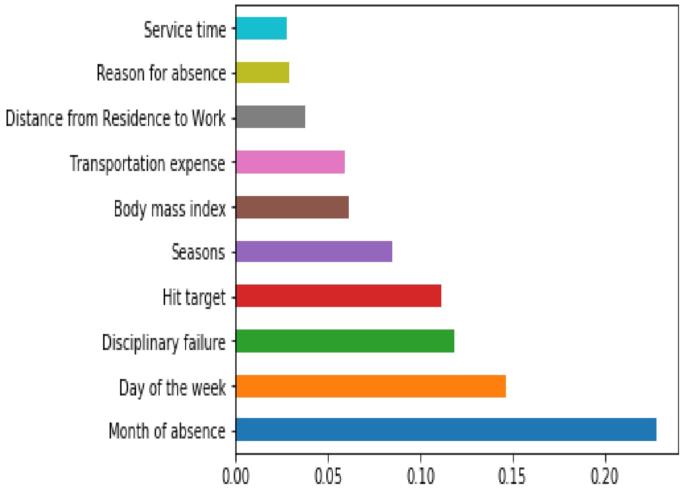

Fig. 2 Feature importance of attributes in the dataset

The Absenteeism at work Data Set is examined to find the correlation between each variable in the data set. The implementation is performed using python scripts and the obtained correlation matrix of the dataset is shown in Fig. 3.

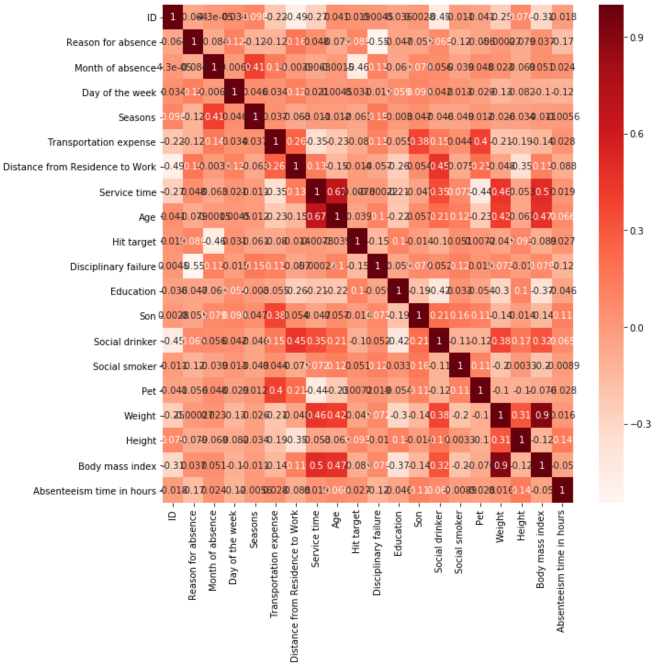

Fig. 3 Correlation matrix of Absenteeism at work Data Set

The data type of each attribute in the Absenteeism at work Data Set is given in Table. 1.

Table. 1. Data Type Information for Absenteeism at work Dataset

Attribute	Data Type Details
ID	REAL
Reason for absence	REAL
Month of absence	REAL
Day of the week	{5.0, 2.0, 3.0, 4.0, 6.0}
Seasons	{4.0, 1.0, 2.0, 3.0}
Transportation expense	REAL
Distance from Residence to Work	REAL
Service time	INTEGER
Age	INTEGER
Work load Average / Day	REAL
Hit target	REAL
Disciplinary Failure	{1.0, 0.0}
Education	REAL
Son	REAL
Social drinker	{1.0, 0.0}
Social Smoker	{1.0, 0.0}
Pet	REAL
Weight	REAL
Height	REAL
Body mass index	REAL
Absenteeism time in hours	REAL

Table. 2. Categories of Reason for Absence in the Dataset

Category	Mean Squared Error
I.	Certain infectious and parasitic diseases
II.	Neoplasms
III.	Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism
IV.	Endocrine, nutritional and metabolic diseases
V.	Mental and behavioural disorders
VI.	Diseases of the nervous system
VII.	Diseases of the eye and adnexa
VIII.	Diseases of the ear and mastoid process
IX.	Diseases of the circulatory system
X.	Diseases of the respiratory system
XI.	Diseases of the digestive system
XII.	Diseases of the skin and subcutaneous tissue
XIII.	Diseases of the musculoskeletal system and connective tissue
XIV.	Diseases of the genitourinary system
XV.	Pregnancy, childbirth and the puerperium
XVI.	Certain conditions originating in the perinatal period
XVII.	Congenital malformations, deformations and chromosomal abnormalities
XVIII.	Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified
XIX.	Injury, poisoning and certain other consequences of external causes
XX.	External causes of morbidity and mortality
XXI.	Factors influencing health status and contact with health services.
XXII.	Patient Follow-Up
XXIII.	Medical Consultation
XXIV.	Blood Donation
XXV.	Laboratory Examination
XXVI.	Unjustified Absence
XXVII.	Physiotherapy
XXVIII.	Dental Consultation

The top most high correlated features are identified which are directly fitted to the regression models like Linear regression, SRD regression, RANSAC regression, Ridge regression, Huber regression, ARD Regression, Passive Aggressive Regression and Theilson Regression. The mean squared error, mean absolute error, R2 score, Explained Variance Score and Mean Squared log error of each method is compared to analyze the performance and is shown in Table. 3 and Table. 4.

Table. 3. Comparison of Performance Metrics(MSE, MAE, R2 Score) for all Regression

Regression Methods	Mean Squared Error	Mean Absolute Error	R2 Score (%)
Linear Regression	0.10	0.23	66
SRD Regression	0.12	0.24	77
RANSAC Regression	0.14	0.22	79
Ridge Regression	0.11	0.26	54
Huber Regression	0.09	0.27	65
ARD Regression	0.07	0.21	72
Passive Aggressive Regression	0.04	0.16	89
Theilsen Regression	0.17	0.28	68

Table. 4. Comparison of Performance Metrics (EVS, MSLE) for all Regression

Regression Methods	Explained Variance Score	Mean Squared Log Error
Linear Regression	0.09	0.81
SRD Regression	0.08	0.65
RANSAC Regression	0.06	0.58
Ridge Regression	0.05	0.72
Huber Regression	0.07	0.63
ARD Regression	0.06	0.79
Passive Aggressive Regression	0.03	0.32
Theilsen Regression	0.1	0.84

The Performance Analysis of the metrics for all the regression is shown in Fig. 3. - Fig. 7.

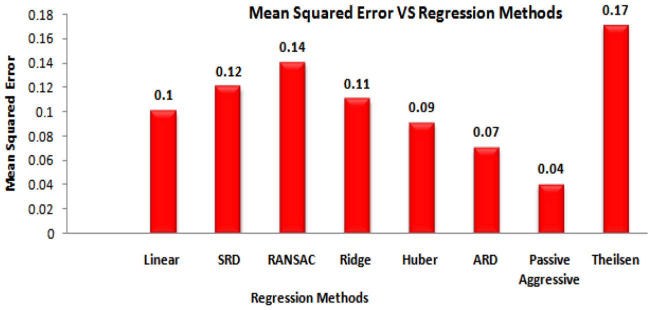

Fig. 4 MSE (VS) Regression Methods

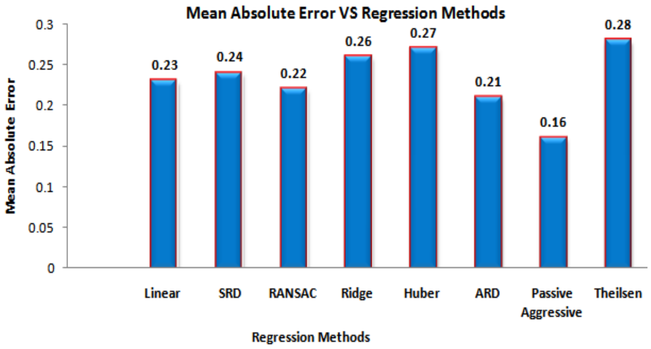

Fig. 5 MAE (VS) Regression Methods

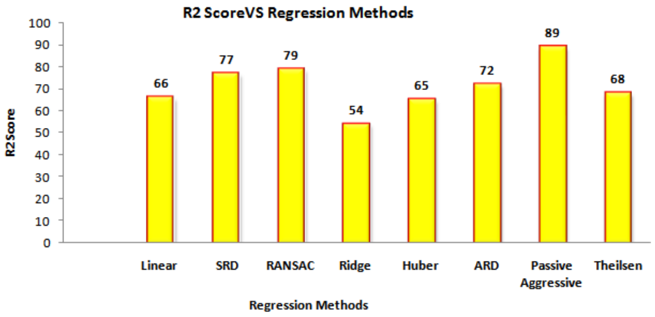

Fig. 6 R2 Score (VS) Regression Methods

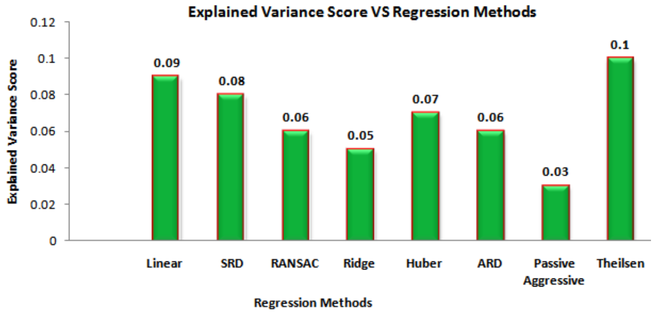

Fig. 7. EVS (VS) Regression Methods

.6

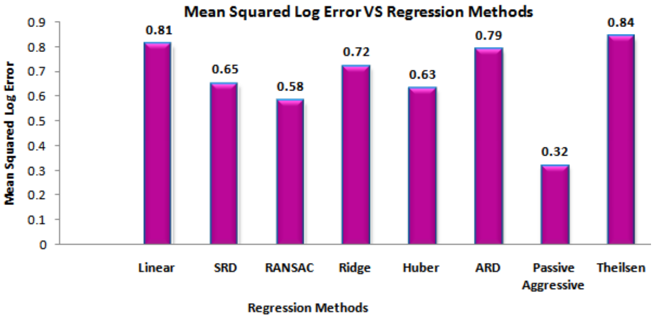

Fig. 8. MSLE (VS) Regression Methods

VI. CONCLUSION

This paper attempts to predict the number of absent hours of the employee based on various health conditions from the absenteeism at work dataset extracted from the UCI machine learning repository. The prediction of absent hours is done by using machine learning regression algorithms. The correlation between each of the dataset attributes are found and depicted as a histogram. Then the top most high correlated features are identified which are directly fitted to the regression models like Linear regression, SRD regression, RANSAC regression, Ridge regression, Huber regression, ARD Regression, Passive Aggressive Regression and Theilson Regression. Finally, the Performance analysis is done by analyzing the performance metrics like Mean Squared Error, Mean Absolute Error, R2 Score, Explained Variance Score and Mean Squared Log Error. Experimental Result shows that the Passive Aggressive Regression have achieved the effective prediction of number of absent hours with minimum MSE of 0.04, MAE of 0.16, EVS of 0.03, MSLE of 0.32 and reasonable R2 Score of 0.89.

Retrieval Number C4405098319/19©BEIESP
DOI: 10.35940/ijrte.C4405.098319

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

REFERENCES

1. Lau Lilleholt, Anders Aaby, and Guido Makransky, “ Students admitted to university based on a coginitive test and MMI are less stressed than students admitted based on CGPA”, Studies in Education Evaluation, Elsevier., vol. 61, Jun. 2019, pp. 170-175.
2. Joseph Janison, “Applying Machine Learning to Predict Davidson College's Admissions Yield”, proceedings of the ACM SIGCSE Technical Symposium., 2017.
3. S. Maldonado, G. Armelini, and A. Guevara, “Assessing university enrollment and admission efforts via hierarchical classification and feature selection”, Intelligent Data Analysis., vol. 21, no. 4, 2017.
4. William Eberle, Douglas Talbert, Erik Simpson, Larry Roberts, and Alexis pope, “ Using Machine Learning and Predictive Modeling to Assess Admission Policies and Standards”, Proceedings of the 9th Annual National Symposium., 2013.
5. M. Shyamala Devi, Rincy Merlin Mathew, and R. Suguna,”Attribute Heaving Extraction and Performance Analysis for the Prophesy of Roof Fall Rate using Principal Component Analysis”, International Journal of Innovative Technology and Exploring Engineering, vol. 8, no.8, June 2019, pp. 2319-2323.
6. R. Suguna, M. Shyamala Devi, and Rincy Merlin Mathew, ” Customer Churn Predictive Analysis by Component Minimization using Machine Learning”, International Journal of Innovative Technology and Exploring Engineering, vol. 8, no.8, June 2019, pp. 2329-2333.
7. Shyamala Devi Munisamy, and Suguna Ramadass Aparna Joshi, “Cultivar Prediction of Target Consumer Class using Feature Selection with Machine Learning Classification”, Springer’s book series “Learning and Analytics in Intelligent Systems, Springer, LAIS vol. 3, pp. 604-612, 2019.
8. Suguna Ramadass, and Shyamala Devi Munisamy, Praveen Kumar P, Naresh P, “Prediction of Customer Attrition using Feature Extraction Techniques and its Performance Assessment through dissimilar Classifiers”, Springer’s book series “Learning and Analytics in Intelligent Systems, Springer, , LAIS vol. 3, pp. 613-620, 2019.
9. R.Suguna, M. Shyamala Devi, Rupali Amit Bagate, and Aparna Shashikant Joshi, “Assessment of Feature Selection for Student Academic Performance through Machine Learning Classification”, Journal of Statistics and Management Systems, Taylor Francis, vol. 22, no. 4, 25 June 2019, pp. 729-739.
10. M. Shyamala Devi, Rincy Merlin Mathew, R. Suguna, “Feature Snatching and Performance Analysis for Connoting the Admittance Likelihood of student using Principal Component Analysis”, International Journal of Recent Technology and Engineering, Volume-8 Issue-2, 30 July 2019.pp. 4800-4807.
11. R. Suguna, M. Shyamala Devi, Rincy Merlin Mathew, “Customer Segment Prognostic System by Machine Learning using Principal Component and Linear Discriminant Analysis”, International Journal of Recent Technology and Engineering, Volume-8 Issue-2, 30 July 2019. pp. 6198-6203.
12. M. Shyamala Devi, Rincy Merlin Mathew, R. Suguna, ” Regressor Fitting of Feature Importance for Customer Segment Prediction with Ensembling Schemes using Machine Learning”, International Journal of Engineering and Advanced Technology, Volume-8 Issue-6, 30 August 2019
13. Rincy Merlin Mathew, R. Suguna, M. Shyamala Devi, “Composite Model Fabrication of Classification with Transformed Target Regressor for Customer Segmentation using Machine Learning”, International Journal of Engineering and Advanced Technology, Volume-8 Issue-6, 30 August 2019
14. R. Suguna, M. Shyamala Devi, Rincy Merlin Mathew, “Integrating Ensembling Schemes with Classification for Customer Group Prediction using Machine Learning”, International Journal of Engineering and Advanced Technology, Volume-8 Issue-6, 30 August 2019