IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 23, Issue 5, Ser. I (Sep. –Oct. 2021), PP 5-58
www.iosrjournals.org

Comparison of Machine Learning Algorithms in Predicting Students Performance

Mrs. Geetha N[1], Dr. Piyush Kumar Pareek[2]
[1]*Computer Science Engineering, East west Institute of Technology, Bengaluru, India,*
[2]*Computer Science Engineering, East west Institute of Technology, Bengaluru, India.*

Abstract
The contribution of Machine Learning Algorithms in the field of education is very huge, especially in the prediction of student performance [15]. Many colleges are failing to address the problems that the students face in their academic performance as well as in their career. This paper aims at defining the right metric for the engineering student's success and defining the better statistical mechanisms in analyzing the performance of students. This also helps mentors to choose the appropriate way in success of their students. In technical education like engineering, most of the colleges won't give much of priority to the non-academic aspects, which is the main aspect in getting the desired results and provide the student satisfaction. The machine learning algorithms helps at defining the better mechanisms to analyze student performance by using the sophisticated predictive techniques [13]. Mentoring, attendance, career interest, native place, their previous academic performance are the key attributes for the better performance of the students and even their placement in companies. This paper compares the different machine algorithms in predicting student's performance. This paper also helps the mentors and organization in getting their better results in academic as well as their student's placement.

Keywords: *Education Data Mining, Machine Learning Algorithms, Prediction.*

Date of Submission: 28-09-2021 Date of Acceptance: 12-10-2021

I. Introduction

Students' academic performance and success are the major concerns to the management, faculties, stakeholders like parents, government especially in the higher education because the student's success will have direct impact on the development of any nation in the current era. In this paper many variables like previous grades, student's native place, career interest, urban/rural, sex, age, electives, cbsc/non-cbsc, attendance, student motivation and mentors counselling attendance, CGPA, SGPA are considered.

In order to start the process, for prediction of academic performance of an engineering student, first the primary data is collected, upon different indicators, both academic and non-academic aspects are considered for evaluating the academic success of the engineering graduate students and is cleaned. The machine learning (ML) algorithms are suitable for predicting and analysing. The various methods are used in cleaning the primary data [discussed in the paper: "Quantitative analysis of educational data mining" [13]]. A framework was also developed for the prediction of student performance with the cleaned dataset obtained from engineering universities [14]. In this paper, the comparison of 5 important ML algorithms is illustrated, and it has been observed that the Naïve Bayes algorithm achieved a stronger result with the right indicators. SPSS an IBM tool also helps in choosing the right metrics for the analysis. Based on the proposed model, it was clear that the attendance, past academic performance, career interest, motivation to the students especially in the 1st, 2nd and 3rd of engineering were the most important and influential variables. [10].

II. Literature Survey

The stronger results will be achieved by considering the factors like family relationships, parents education, especially the mothers education in the study Yalcin Ozkan[10], in this the model is developed using neural networks and then compared with the 6 various machine learning algorithms.

Many studies have been conducted to highlight the usefulness of "Data Mining" approaches in higher education like technical/non-technical education, indicating that this is the best approach and concept for obtaining relevant and accurate information regarding student behavior and learning effectiveness [4].

Abeer and Elaraby [2] did research on developing categorization criteria and forecasting students' achievement in a particular course programme based on the activities and behavior of the students which is collected previously. They [1] used numerous features from the university database to process and analyse the data during a six-year period (2005–10). This study predict the students' final grades in the chosen course, as well as “assist students in improving their performance, identifying those students who needed special attention to reduce failing ratio and taking appropriate action at the right time” [1].

Several investigations have been conducted under the proposed study object. Bhardwaj, for instance, employed the Nave Bayes algorithm to prediction in education system using thirteen variables [5]. The survey is conducted on BCA students from university in Faizabad, India, those who has completed their final exam in 2010. The extraction of data is done by using different substantial approach.

Tek Bist Bithari and sharan Thapa[12] the prediction of engineering students based on the different factors, family background, past educational records, and other factors and applied different techniques of data mining like SVM, Decision Tree, regression and ensemble method to get the good results of their research.

The success of the students is directly proportionate to the academic motivation to the student, Ivana Durdevic Babic[9] has discussed about the academic motivation of the students and he used 3 ML classifiers like, neural networks, SVM and decision tree. The researcher has taken the data for the course LMS and found the efficient classification model among three is the Neural network model.

Aimad Qazdar[11] in this research, he proposed a framework based on the ML algorithms at the Morocco high school for the performance prediction of some group students in the science department . The student data set is collected from the SMS-Masaar, a learning management system. He has considered 2nd semester and national examination marks in the prediction of student’s in his proposed model..

Okereke GE[7] in his paper a small set of data is considered for his analysis and used Decision Tree algorithm in training and testing of the dataset and two dissimilar dataset. He observed that the more accuracy is based in the dataset used and trained and not the classification algorithm.

In the research conducted by Chhaya saraf[6] in the medical college to improve the student’s performance. The 53 1st semester students were selected (selection for poor performance) and the mentoring program has conducted and the feedback has taken from the students. This mentoring program helped the students in achieving higher mean score of the results in the next internal assessment among 98.1% slow achievers.

Jyoti Kumari[8] in her research used traditional ML algorithms like linear regression, k-nearest neighbour algorithm, decision tree algorithm , Bayes algorithm, and also analyzed for the student placements in companies based on their SGPA and CGPA and observed that the Naïve Bayesian algorithm gives the more accuracy in predicting GPA of student.

III. Comparison of Ml Algorithms With Results

More than 2000 student data is collected from the 2 different universities in the Bangalore. The main variables chosen are marks, attendance, mentoring information, urban or rural background, CBSC or Non-CBSC, bridge course pre and post-test marks, pre-university marks, sex, age, grades obtained in the exams in the first year engineering and possible values are separated in to FCD, Distinction, 1st class, 2nd class and fail. Their attendance is one of the important variable in the future prediction. The data is collected from different sources, different departments and learning management systems, and cleaned using various cleaning techniques/methods and the machine learning algorithms are applied for the analysis.

3.1 DECISION TREE LEARNING

Decision Tree Learning (DTL) is a well-known supervised learning technique that use decision trees. It is the Learning model, which is a tree-like hierarchical decision-making paradigm. They aid in the identification of a strategy that leads to a final choice. Both regression and classification problems can be solved using Decision Tree Learning. This algorithm's primary requirement is that the data be discrete. The entire dataset is initially treated as the root node, from which the tree is constructed. Nodes are produced by recursively distributing the attributes in the dataset to build the tree. In the identification of the particular attribute in each levels for the root. Two techniques namely Information gain and Gini Index is used, and ID3 algorithm was used.

3.1.1 Attribute Selection Measures:

Information Gain:

The entropy means the amount of impurity in the system, considering different aspects. In this information theory, the impurity in a group of systems. The average entropy computes after split of the dataset. The decision tree algorithm ID3 (Iterative Dichotomiser) makes use of gained information.

$$\boxed{\text{Info(D)} = - \sum_{i=1}^{m} pi \log_2 pi}$$

Where, Pi is the probability that an arbitrary tuple in D ∈ class Ci.

Where, $$\boxed{\text{Info}_A\text{(D)} = \sum_{j=1}^{v} \frac{|Dj|}{|D|} \text{ X Info}(D_j)}$$

$$\boxed{\text{Gain(A)} = \text{Info(D)} - \text{Info}_A\text{(D)}}$$

Info(D) is an average amount of required information to identify the class label of a given tuple D.

|Dj|/|D| is the weight of a given jth partition.

InfoA(D) is an expected information required for the analysis and it is based on the partitioning of A to classify tuple from D.

A, Attribute which has highest information gain.

Gain(A), is taken by splitting attribute at node N().

Gini index

To create the split points Gini method is used in Decision tree algorithm.

$$\boxed{\text{Gini(D)} = 1 - \sum_{i=1}^{m} \text{Pi}^2}$$

Pi=probability and arbitrary tuple in D ∈ class C,

$$\boxed{\text{Gini}_A\text{(D)} = \frac{|D1|}{|D|} \text{Gini}(D_1) + \frac{|D2|}{|D|} \text{Gini}(D_2)}$$

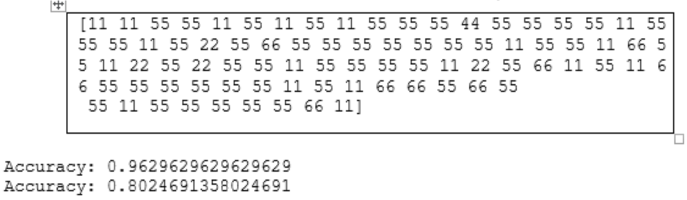
The Gini index (GiniA)of D is:

A small amount of Gini index taken as the dividing point for each pair of neighboring values.

$$\boxed{\Delta Gini(A) = Gini(D) - Gini_A(D).}$$

Sample Results

The decision tree algorithm ID3 (Iterative Dichotomiser) makes use of gained information and here are results:



```
[11 11 55 55 11 55 11 55 11 55 55 55 44 55 55 55 55 11 55
55 55 11 55 22 55 66 55 55 55 55 55 55 55 11 55 55 11 66 5
5 11 22 55 22 55 55 11 55 55 55 55 11 22 55 66 11 55 11 6
6 55 55 55 55 55 55 11 55 11 66 66 55 66 55
 55 11 55 55 55 55 55 66 11]

Accuracy: 0.9629629629629629
Accuracy: 0.8024691358024691
```

Fig 1: ID3 results

3.2 CLASSIFICATION

This is a supervised learning algorithm that classifies the given data into categories. Such learning models work by surmise a function from a labelled training dataset. The training datasets are composed of training examples, input vector pair and an expected output value pair. The algorithm when given a new observation, decides as to which category it belongs to by comparing it with the training examples. For instance, the Naïve Bayes classifier was used in this project.

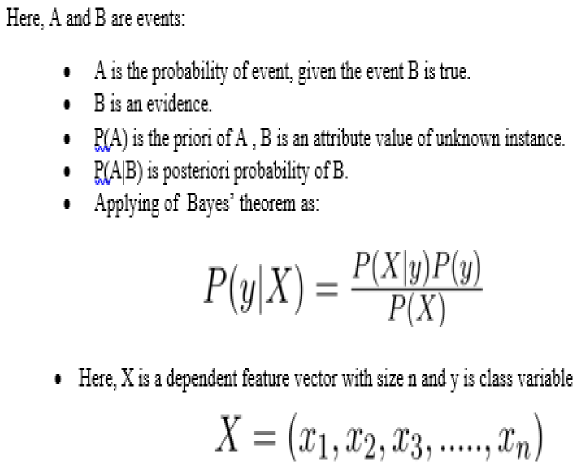
3.2.1 Naïve Bayes Classification

The Naive Bayes classifier is a collection of classification algorithms and it is based on Bayes' Theorem. In Naïve Bayes a family of algorithms, every pair of features are classified and are independent, but they share a common principle.

Bayes' Theorem

Bayes' theorem is mathematical equation:

$$\boxed{P(A|B) = \frac{P(B|A)P(A)}{P(B)}}$$



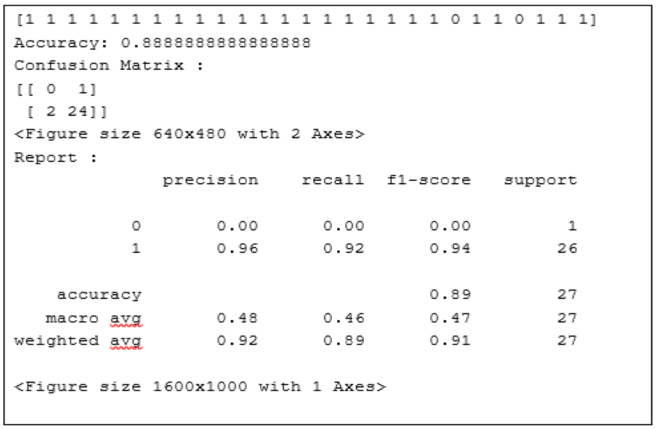
Here, A and B are events:

- A is the probability of event, given the event B is true.
- B is an evidence.
- P(A) is the priori of A , B is an attribute value of unknown instance.
- P(A|B) is posteriori probability of B.
- Applying of Bayes' theorem as:

$$P(y|X) = \frac{P(X|y)P(y)}{P(X)}$$

- Here, X is a dependent feature vector with size n and y is class variable

$$X = (x_1, x_2, x_3, ..., x_n)$$

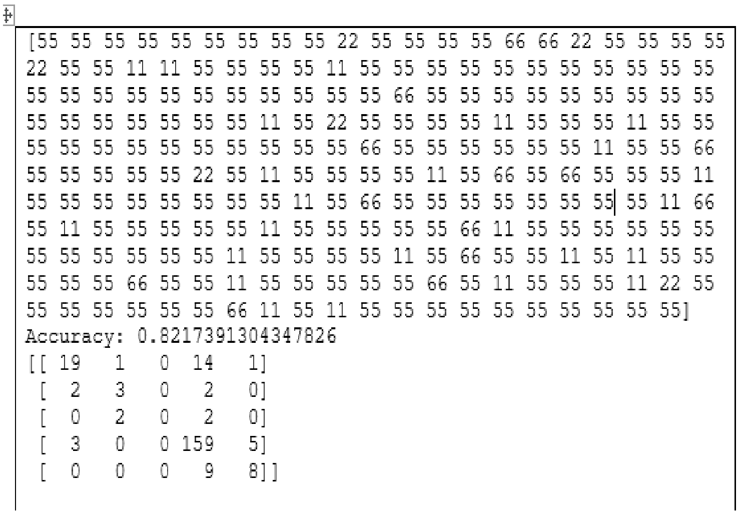


```
[1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 0 1 1 1]
Accuracy: 0.8888888888888888
Confusion Matrix :
[[ 0  1]
 [ 2 24]]
<Figure size 640x480 with 2 Axes>
Report :
              precision    recall  f1-score   support

           0       0.00      0.00      0.00         1
           1       0.96      0.92      0.94        26

    accuracy                           0.89        27
   macro avg       0.48      0.46      0.47        27
weighted avg       0.92      0.89      0.91        27

<Figure size 1600x1000 with 1 Axes>
```

Fig 2: Results Naïve Bayes

3.3 LOGISTIC REGRESSION

This is a classification ML algorithm which is helpful in predicting the probability of a categorical binary variable or dependent variable. In the logistic regression, yes or no are the values of the dependent variable i,e a binary variable and these values are coded as 1 (true, yes, success, etc.) or 0 (false, no, failure, etc.).



```
[55 55 55 55 55 55 55 55 55 22 55 55 55 55 66 66 22 55 55 55 55
22 55 55 11 11 55 55 55 55 11 55 55 55 55 55 55 55 55 55 55 55
55 55 55 55 55 55 55 55 55 55 55 66 55 55 55 55 55 55 55 55 55
55 55 55 55 55 55 55 11 55 22 55 55 55 55 11 55 55 55 11 55 55
55 55 55 55 55 55 55 55 55 55 66 55 55 55 55 55 55 11 55 55 66
55 55 55 55 55 22 55 11 55 55 55 55 11 55 66 55 66 55 55 55 11
55 55 55 55 55 55 55 55 11 55 66 55 55 55 55 55 55 55 55 11 66
55 11 55 55 55 55 55 11 55 55 55 55 55 66 11 55 55 55 55 55 55
55 55 55 55 55 55 11 55 55 55 55 11 55 66 55 55 11 55 11 55 55
55 55 55 66 55 55 11 55 55 55 55 55 66 55 11 55 55 55 11 22 55
55 55 55 55 55 55 66 11 55 11 55 55 55 55 55 55 55 55 55 55]
Accuracy: 0.8217391304347826
[[ 19   1   0  14   1]
 [  2   3   0   2   0]
 [  0   2   0   2   0]
 [  3   0   0 159   5]
 [  0   0   0   9   8]]
```

Fig 3: Results 1- Logistic Regression

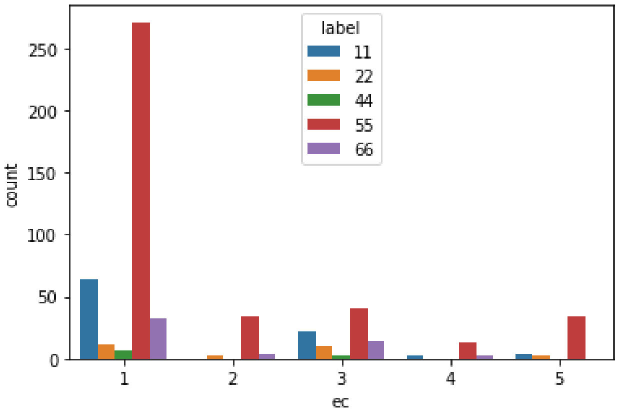


Fig 4: Results 2- Logistic Regression

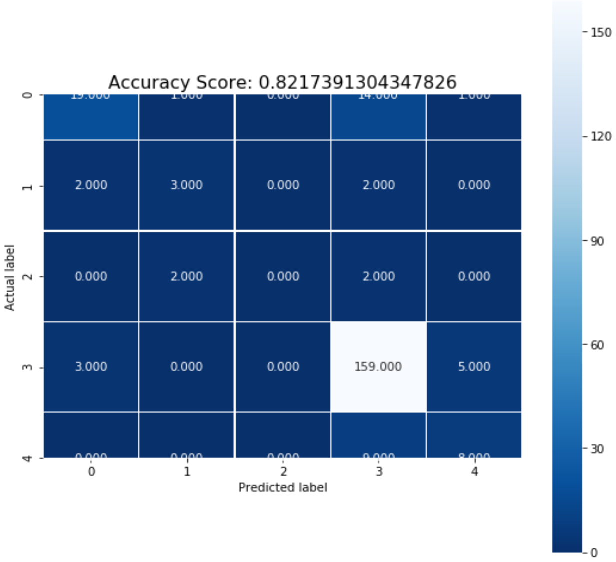
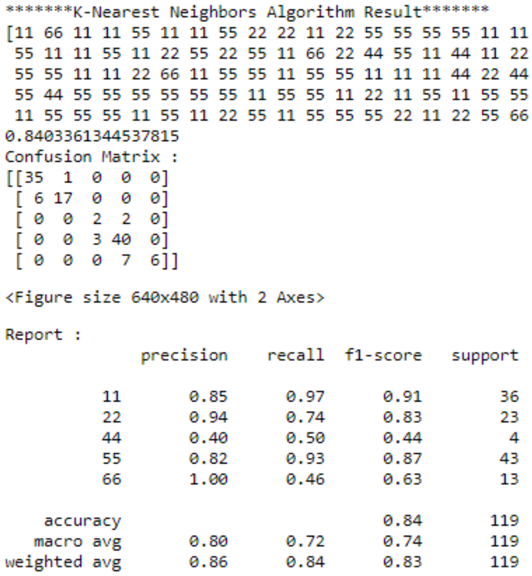


Fig 5: Results 1- Logistic Regression

3.4 K-NN ALGORITHM:

The K-nearest neighbors (KNN) is comparatively easy for the implementation in its basic form as it is a supervised ML algorithm. This also performs the complex tasks like classification in few situations. Sometimes this algorithms named as lazy learning algorithm, which has no fixed training phase. The training of the data can be done for entire data. The most useful feature is non-parametric, KNN doesn't assume anything about the underlying data.



```
*******K-Nearest Neighbors Algorithm Result*******
[11 66 11 11 55 11 11 55 22 22 11 22 55 55 55 55 11 11 11 22 11 55 55 11
 55 11 11 55 11 22 55 22 55 11 66 22 44 55 11 44 11 22 22 55 66 55 22 55
 55 55 11 11 22 66 11 55 55 11 55 55 11 11 11 44 22 44 11 55 22 55 11 55
 55 44 55 55 55 55 55 55 11 55 55 11 22 11 55 11 55 55 11 11 11 11 55 55
 11 55 55 55 11 55 11 22 55 11 55 55 55 22 11 22 55 66 55 22 11 11 66]
0.8403361344537815
Confusion Matrix :
[[35  1  0  0  0]
 [ 6 17  0  0  0]
 [ 0  0  2  2  0]
 [ 0  0  3 40  0]
 [ 0  0  0  7  6]]

<Figure size 640x480 with 2 Axes>

Report :
              precision    recall  f1-score   support

          11       0.85      0.97      0.91        36
          22       0.94      0.74      0.83        23
          44       0.40      0.50      0.44         4
          55       0.82      0.93      0.87        43
          66       1.00      0.46      0.63        13

    accuracy                           0.84       119
   macro avg       0.80      0.72      0.74       119
weighted avg       0.86      0.84      0.83       119
```

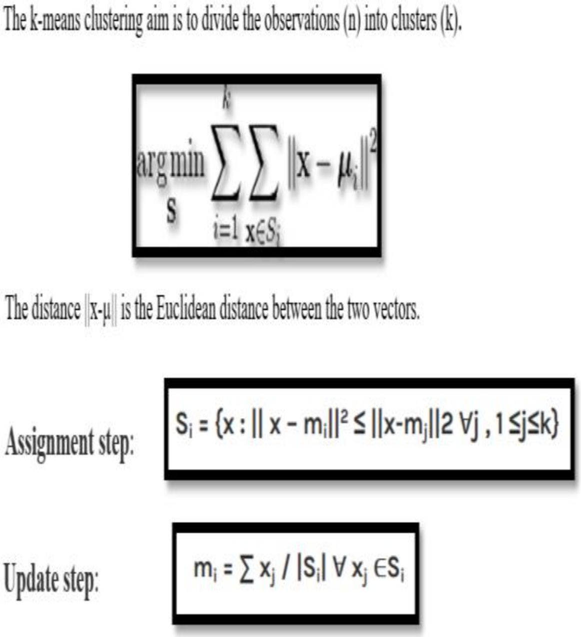
Fig 6: K-NN Results

3.5 CLUSTERING

The Clustering is an unsupervised algorithm which works on grouping. The grouping is done among the data points in a way that it points in the same group which have more similarity than that of another group. This can be achieved by comparing these features of the data points and mapping. K-Means clustering algorithm was used and the results are obtained.

3.5.1 *K-Means Clustering*

K-means clustering is a vector quantization approach that originated in signal processing (SP) and for clustering analysis in data mining.



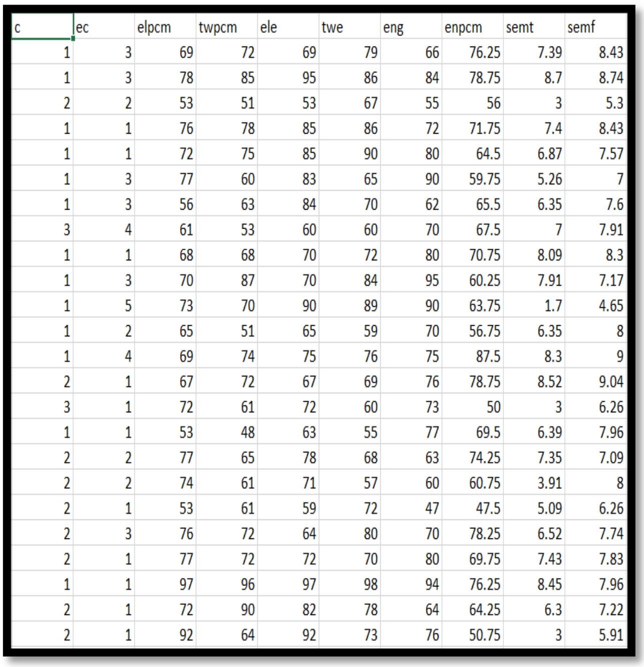
The k-means clustering aim is to divide the observations (n) into clusters (k).

$$\boxed{\arg\min_{S} \sum_{i=1}^{k} \sum_{x \in S_i} \|x - \mu_i\|^2}$$

The distance $\|x-\mu\|$ is the Euclidean distance between the two vectors.

Assignment step: $\boxed{S_i = \{x : \| x - m_i\|^2 \leq \|x-m_j\|2 \; \forall j \, , 1 \leq j \leq k\}}$

Update step: $\boxed{m_i = \sum x_j / |S_i| \; \forall \; x_j \in S_i}$



c	ec	elpcm	twpcm	ele	twe	eng	enpcm	semt	semf
1	3	69	72	69	79	66	76.25	7.39	8.43
1	3	78	85	95	86	84	78.75	8.7	8.74
2	2	53	51	53	67	55	56	3	5.3
1	1	76	78	85	86	72	71.75	7.4	8.43
1	1	72	75	85	90	80	64.5	6.87	7.57
1	3	77	60	83	65	90	59.75	5.26	7
1	3	56	63	84	70	62	65.5	6.35	7.6
3	4	61	53	60	60	70	67.5	7	7.91
1	1	68	68	70	72	80	70.75	8.09	8.3
1	3	70	87	70	84	95	60.25	7.91	7.17
1	5	73	70	90	89	90	63.75	1.7	4.65
1	2	65	51	65	59	70	56.75	6.35	8
1	4	69	74	75	76	75	87.5	8.3	9
2	1	67	72	67	69	76	78.75	8.52	9.04
3	1	72	61	72	60	73	50	3	6.26
1	1	53	48	63	55	77	69.5	6.39	7.96
2	2	77	65	78	68	63	74.25	7.35	7.09
2	2	74	61	71	57	60	60.75	3.91	8
2	1	53	61	59	72	47	47.5	5.09	6.26
2	3	76	72	64	80	70	78.25	6.52	7.74
2	1	77	72	72	70	80	69.75	7.43	7.83
1	1	97	96	97	98	94	76.25	8.45	7.96
2	1	72	90	82	78	64	64.25	6.3	7.22
2	1	92	64	92	73	76	50.75	3	5.91

Fig 7: Dataset used for K-Means

3.6 PREDICTION

Prediction is a technique used to forecast the possible future values of a system by analysing the given training dataset. The training dataset comprises of training examples. These training examples can consist of the past values of the system. The basic technique used for prediction is regression. Regression is a technique in inferential statistics that gives inferences by statistically analysing one or more dependent and independent variables. For complex prediction one can used more sophisticated techniques such as logistic regression, Classification and Regression Tree (CART) etc. For instance, Linear Regression was used.

3.6.1 Linear Regression

The Linear regression predict a dependent variable (y) value, for an independent variable (x).

The equation-for line :

Y= **mx + b**

The three main evaluation metrics of Linear Regression are:

Mean Absolute Error (MAE):

$$\mathrm{MAE} = \frac{1}{n}\sum_{j=1}^{n} |y_j - y_j|$$

Mean Squared Error (MSE) :

$$MSE = \frac{1}{N}\sum_{i}^{n}(Y_i - y_i)^2$$

Root Mean Squared Error (RMSE):

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2}$$

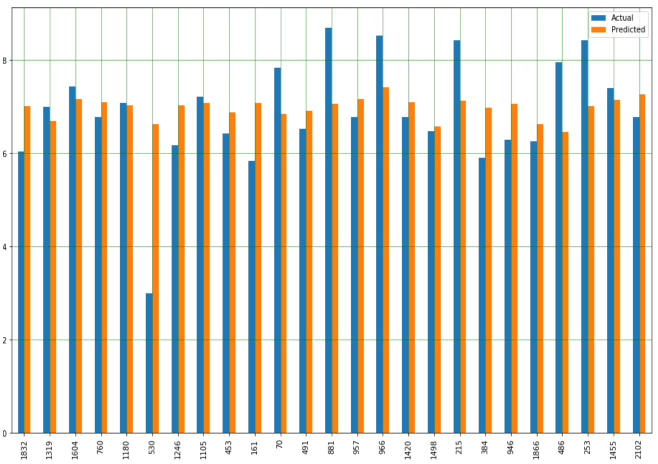


Fig 8: Figure: Sample Results with Mean Absolute Error: 0.9

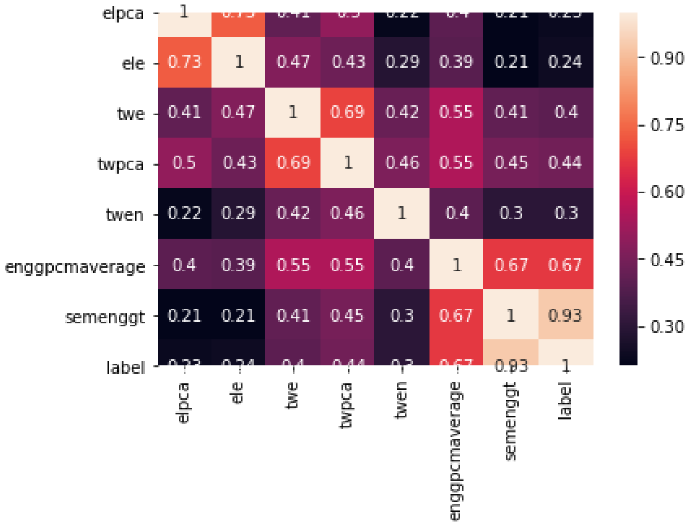


Fig 9: Results –Linear Regression

IV. Conclusion

In this paper the different machine learning (ML) algorithms are compared in predicting the student's performance. It has been observed that the attendance, motivation and counselling to the student's in the right time, career interest produces the fair results in their academic performance and even their placements in the engineering colleges. This paper also compares the results between different ML algorithms. It is observed that the Naïve Bayes algorithms will give good results compared to others. This paper helps the mentors to council and motive the students in right approach for their success. On the tip of iceberg, this paper serves the best purpose for the further research in the field of EDM.

References

[1]. Ahmed, A.B.E.D. and Elaraby, "Data Mining: A prediction for Student's Performance Using Classification Method. World Journal of Computer Application and Technology", 2(2), pp.43-47, I.S., 2014.

[2]. S.N. Madhu, "Discovery of students' academic patterns using data mining techniques" International Journal of Computer Science and Engineering", Vol 4, no. 06, 2012.

[3]. Bhardwaj, B.K. and Pal, S., 2012. "A prediction for performance improvement using classification. (IJCSIS) International Journal of Computer Science and Information Security", Vol. 9, No. 4, April 2011.

[4]. Ramaswami and R. Bhaskaran, "A Study on Feature Selection Techniques in Educational Data Mining," J. Comput., vol. 1, no. 1, pp. 7–11, 2009.

[5]. V. Kumar, "An Empirical Study of the Applications of Data Mining Techniques in Higher Education," International. J. Adv. Computer Science Appl., vol. 2, no. 3, pp. 80–84, 2011.

[6]. Chhaya Saraf, Afreen Begum H Itagi, "Role of mentoring in improving academic performance among low achievers in a medical college of Chhattisgarh, India", Journal of Education Technology in Health Sciences, 4(3):107-111, September-December, 2017.

[7]. Okereke GE, Mamah CH, Ukekwe EC and Nwagwu HC, "A Machine Learning Based Framework for Predicting Student's Academic Performance", Physical Science & Biophysics Journal, MEDWIN PUBLISHERS ISSN: 2641-9165, Volume 4 Issue 2, July 13, 2020.

[8]. Jyoti Kumari, K. Ramalakshmi, "A Comparison Of Machine Learning Techniques For The Prediction Of The Student's Academic Performance", Emerging Trends in Computing and Expert Technology, January 2020.

[9]. Ivana Durdevic Babic," Machine learning methods in predicting the student academic motivation", Croatian Operational Research Review 443 CRORR 8(2017), 443–461, November 30, 2017

[10]. Yalçın Özkan1," Prediction of Student Performance By Deep Learning Algorithm", 7th International Conference on "Innovations in Learning for the Future": Digital Transformation in Education, Future-Learning 2018, September 11-14, İstanbul

[11]. Aimad Qazdar1,"A machine learning algorithm framework for predicting students performance" ,case study- baccalaureate students in Morocco", Education and Information Technologies · November 2019, Springer Science, Business Media, LLC, part of Springer Nature 2019.

[12]. Tek Bist Bithari, Sharan Thapa, and Hari K.C. ,"PREDICTING ACADEMIC PERFORMANCE OF ENGINEERING STUDENTS USING ENSEMBLE METHOD", TECHNICAL JOURNAL Vol 2, No.1, October 2020 Nepal Engineers' Association, Gandaki Province ISSN : 2676-1416.

[13]. N., Geetha, Piyush Kumar Pareek, "Quantitative Analysis of Student Data Mining", May 17, 2019, Available at SRN: https://ssrn.com/abstract=3510975.

[14]. Smitha Rao M.S., Pallavi M., Geetha N. (2019) Conceptual Machine Learning Framework for Initial Data Analysis. In: Peng SL., Dey N., Bundele M. (eds) Computing and Network Sustainability. Lecture Notes in Networks and Systems, vol 75. Springer, Singapore. https://doi.org/10.1007/978-981-13-7150-9_6

[15]. Siwar Chibani, Francois-Xavier, "Machine learning approaches for the prediction of materials properties", APL Materials,vol 8,4th Aug 2020.

Mrs. Geetha N, et. al. "Comparison of Machine Learning Algorithms in Predicting Students Performance." *IOSR Journal of Computer Engineering (IOSR-JCE)*, 23(5), 2021, pp. 50-58.