JOURNAL OF CRITICAL REVIEWS Student Placement analysis and prediction for improving the education standards by using Supervised Machine Learning Algorithms

Article *in* Journal of Critical Reviews · May 2020

CITATIONS
0

READS
262

4 authors, including:

Sikhinam Nagamani
Rajiv Gandhi University of Knowledge Technologies
5 PUBLICATIONS **47** CITATIONS

SEE PROFILE

JOURNAL OF CRITICAL REVIEWS

ISSN- 2394-5125VOL 7, ISSUE 14, 2020

Student Placement analysis and prediction for improving the education standards by using Supervised Machine Learning Algorithms

S. Nagamani[1], K. Mohan Reddy[2], UmaBhargavi[3], S. RaviKumar [4]

[1,2,3,4] Lakireddy Bali Reddy College of Engineering, Mylavaram, Andhra Pradesh, India.

ABSTRACT: The main goal of all educational institutions is to provide students with employment opportunities in accordance with their core subjects. Reputation and annual admissions of an organization always hang on the placements it delivers to a student. This is one of the major factors that all the institutions heavily strive to strengthen their placement cell which have a prominent role in development of the institution. It is highly advantageous if there is any assistance for this section to place its students. The principle aim is to use the previous and present academic data records of students which could lead to the prediction of the individual's placement selection. Data required is collected from the institution on which algorithms are applied. Initial stage is to pre-process the data that has been gathered, which is followed by application of classification algorithms such as Support Vector Machine and Random Forest. Results obtained can vary with each algorithm and this comparison is done among accuracy, precision and recall values which will help to recognize the best between two algorithms.

I. INTRODUCTION

Educational institutions are now rising in large numbers. The aim of each higher education institution is to get their students a well-paid job for their students through their placement cell. One of the biggest challenges facing higher learning institutions today is to boost the positioning performance of scholars.Placements are considered to be very necessary for each university [1]. The college's basic progress is assessed by the student's presence on campus. Every student admits to the college by seeing percentage of placements at the college. In this context, thus the approach is to anticipate and evaluate the need for placement in colleges, which helps to construct colleges and students to enhance their placements. In this selection method, the likelihood that undergraduates will be put in a company by applying classification algorithms such as Random Forest and Support Vector Machine predicts. The main purpose of this model is to predict whether the student he/she is put in campus recruitment or not. In this reason the data considered is student academic background as the total number, backlogs and credits. The algorithms are based on student data from the preceding year.

There are certain factors that any organization follows during the student placement selection process such as student's profile that includes his/her academic percentages of SSC, Intermediate, and undergraduate. Some other features like technical skills, programming skills, aptitude, and reasoning are included in the selection process. Here the factors namely student's personal profile is not considered. This proposal focuses on managed learning is more explicitly predictive examination, in which future prediction can be made.Many classification algorithms are applied on the student's data, among them Random Forest algorithm gives the best accuracy when compared with other ML algorithms like support vector machine, linear logistic regression, decision tree. For example, accuracy, analysis, and F-measurement, there are commonly used reminders to evaluate the robustness of machine learning calculations. In order to improve the analysis of the prediction, the data indicators have been adjusted by choice of component and custom variable creation.

ISSN- 2394-5125VOL 7, ISSUE 14, 2020

1.1 Related Works:

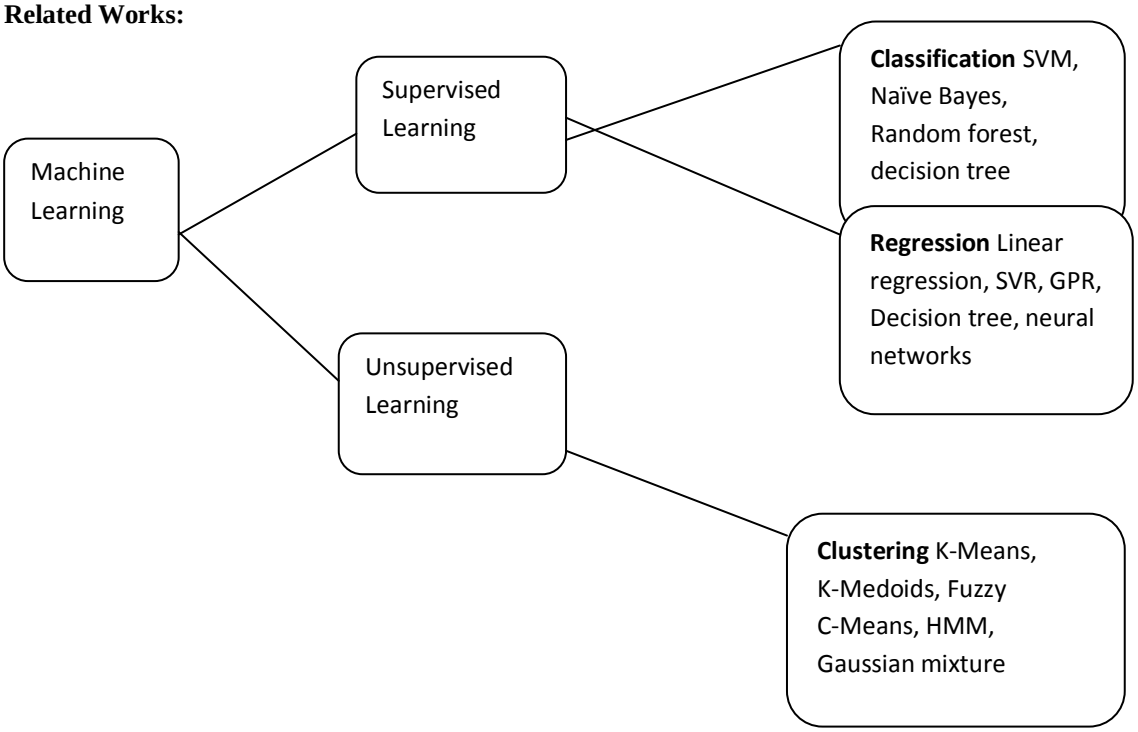

Figure1. Machine Learning Algorithms

2. LITERATURE SURVEY:

This paper utilises Bayesian classification algorithm in predicting the student's placement and thus by calculating the probability of student getting placed in a company. Here the accuracy obtained is about 76.28%. [2]

During the period 2002, Luan used Logistic regression algorithm to determine which student can easily place in company during campus recruitment. By using regression model Luan obtained accuracy 80.33%. [3]

Understudies trying for specialized instruction by and large select instructive foundations with great reputation in grounds positions. Numerous multiple times the notoriety of such establishment is dictated by the compensation bundles offered by scouts to its understudies. Right now is appropriate to explore and recognize those variables that may impact the understudy grounds position risks in specialized instruction. The State of Andhra Pradesh which has a high convergence of specialized instruction organizations was picked as the investigation region. A cautious audit of writing lead to the ID of six theoretical determinants of understudy grounds situation in specialized training. An arbitrary example 250 MBA understudy's situation information were assembled from various organizations and six indicators paired calculated relapse model was fitted to the information to evaluate the chances for the understudy grounds situation. Assessed Results of the investigation demonstrate that the odds of grounds situation are affected by four indicators: CGPA, Specialization in PG, Specialization in UG and Gender. [4] [5]

Ajay Shiv Sharma, Swaraj Prince and Keshav Kumar presented the development of placement predictor (PPS) with help of logistic regression model in which binary values are obtained in the form of results that leads to the accuracy of 75.78% [6]

ISSN- 2394-5125VOL 7, ISSUE 14, 2020

Student placement prediction model using gradient boosted tree algorithm was proposed by OktarianiNurulPratiwi. (August 2013) which gives the best accuracy but the processing time for this model was greater. [7]

With the development of voluminous measure of information in instructive foundations', the need is to mine the huge dataset to deliver some valuable data out of it. Right now centred on to frame a choice emotionally supportive network for the instructive organizations' which can assist them with knowing about the situation plausibility of understudies. Our exploration isn't constrained to discover arrangement probability however we did staggered examination on understudy execution dataset which will foresee that what level of talk with process an understudy is probably going to pass. For this we have applied Naïve Bayes and Improved Naïve Bayes which is incorporated with help highlight choice strategy to get the forecast. Information examination was finished utilizing NetBeans and WEKA. For this our proposed method gave preferred precision over existing guileless Bayes which was 84.7% and innocent Bayes gave 80.96% precision.

Anticipating the presentation of an understudy is a decent worry to the upper instruction organizations. The reason for situation the executive's framework is to change the present manual framework by the help of electronic programming framework satisfying their needs, so their important information/data is put away for a more drawn out time with basic getting to and control of information. Understudy's scholarly accomplishments and their position in grounds determination is a troublesome issue in current manual framework. Observing the understudy's advancement for their grounds arrangement helps in checking the understudy's movement inside the scholastic environment. The point of associations is to supply prevalent chances to their understudies. This proposed understudy expectation framework is most significant methodology which can be used to separate the understudy information/data on the premise of the understudy execution. Overseeing position and instructing records in any bigger association is very extreme in light of the enormous number of understudies. This framework can order the understudy information effortlessly and can be valuable to a few instructive associations. There are a few order calculations and science based systems which can be taken almost as great resources for arranging the understudies' data set in the training field. In Our framework, Gullible Bayes, SVM, KNN calculation is applied to foresee understudy execution which can encourage distinguishing execution of understudies and furthermore gives proposal to improve execution for understudies, for example, we are going to group the understudy's information set for arrangement and non-placement classes dependent on that outcome, training associations can give better preparing than their understudies. In view of information got by framework, understudy's presentation is examined in various perspectives to check the accomplishments of the understudies through their exercises and proposes improvement for better arrangement.[15] "Data Mining Approach for Predicting Student and Institution's Placement Percentage", Professor. Ashok M Assistant Professor Apoorva A,2016 International Conference on Computational Systems and Information Systems for solutions.

METHODOLOGY:

The whole approach is represented by the following flowchart.

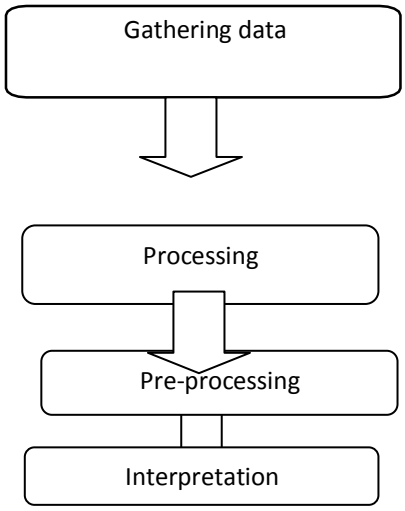

Figure2. Random Forestalgorithm flowchart technique

ISSN- 2394-5125VOL 7, ISSUE 14, 2020

1.1 Data gathering

The example information has been gathered from our school arrangement division which comprises of the considerable number of records of earlier years understudies. The dataset gathered comprise of more than 700 cases of understudies.

1.2 Preprocessing

Information preprocessing is a strategy that is utilized to change over rough information into a clean dataset. The information is accumulated from various sources is in unpolished organization which isn't likely for the examination.

Pre-handling for this methodology makes 4 basic yet powerful steps.

1.2.1 Attribute selection

Some of the attributes which were not relevant to the experimental goal were ignored in the initial dataset. Attributes namely name, roll no, gender were not used for the prediction purpose. The principle qualities utilized for this investigation are credit, technical skills, programming skills, and percentage acquired in X, intermediate, and undergraduate.

1.2.2 Cleaning missing values

Now and again the dataset contain missing qualities. We should be prepared to deal with the issue when we run over them. Clearly you can delete the total source code of data, but what if you accidentally delete main information? After all, we won't have to attempt to do that. One in everything about principal steady intend to deal with the issue is to require a mean of the considerable number of estimations of a similar segment and have it to replace the missing information.

The library utilized for the undertaking is called Scikit Learn preprocessing. It contains a class called Imputer which will assist us with dealing with the missing information.

1.2.3 Training and Test data

Parting the Dataset into Training set and Test Set

Presently the following stage is to part our dataset into two. We will prepare our AI models on our preparation set, i.e. our AI models will attempt to see any connections in our preparation set and afterward we will test the models on our test set to inspect how precisely it will foresee. A general principle of the thumb is to allocate 80% of the dataset to preparing set and in this manner the staying 20% to test set.

1.2.4Feature Scaling

The last advance of information pre preparing is including scaling. It is a technique used to institutionalize the scope of independent variables.

Be that as it may, for what reason is it essential? A ton of AI models depend on Euclidean separation. In the event that, for instance, the qualities in a single section (x) is a lot higher than the incentive in another segment (y), (x2-x1) squared will give a far more prominent incentive than (y2-y1) squared. So distinctly, one square qualification rules over the other square differentiation. In the AI conditions, the square distinction with the lower and encouragement in contrast with the far more projectingvalue will nearly be treated as though it doesn't exist. That is the reason it's important to change every one of our factors into a similar scale. There are a few different ways of scaling the information. One way is called Standardization which might be utilized. For each perception of the chose section, our program will apply the recipe of institutionalization and fit it to a scale.

1.3 Processing:

ISSN- 2394-5125VOL 7, ISSUE 14, 2020

Processing in this paper's sense is applying machine learning algorithms like random forest, support vector machine to the data to find the best results.

1.3.1 Support vector Machine:

The support vector machine algorithm is used for both regression and classification models. This is based on the principle of' decision planes' in which hyperplanes are used to define a group of objects. The data sets can be conveniently differentiated with the help of a line called a decision boundary.

SVM libraries are packed with some popular kernels namely polynomial, radial basis function and sigmoid. It works in the following steps:

1. Mapping data to a high dimensional feature space such that data points can be classified even though the data is not linearly separated.
2. The separator could be drawn as a hyperplane.
3. Characteristics of new data can be used to predict the group to which a new record belongs to.

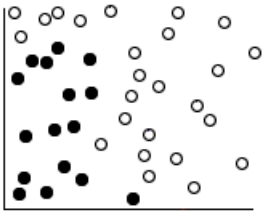

Figure3. both groups are separated with a curve. (Original dataset)

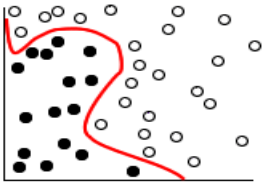

Figure4. Boundary between two categories defined by a hyperplane after transformation.

1.3.2 Random forest

The randomforests calculation can similarly be thought of as an outfit technique in AI. The contribution to arandomforest calculation is a data set comprising of records, with qualities. Irregular subsets of the info are made. On every one of the arbitrary subset made, a choice tree will be built. The last class of a test record will be chosen by the calculation which utilizes the dominant part vote method.

Each hierarchy is developed utilizing the accompanying calculation:

1. First, begin with random sample selection from a given data set.
2. After this, the algorithm will create a decision tree for every sample data. So, that any decision tree will get the prediction result.
3. Here, selection takes places for every prediction result.
4. Finally, most desired predicted value will be treated as the final prediction result.

ISSN- 2394-5125VOL 7, ISSUE 14, 2020

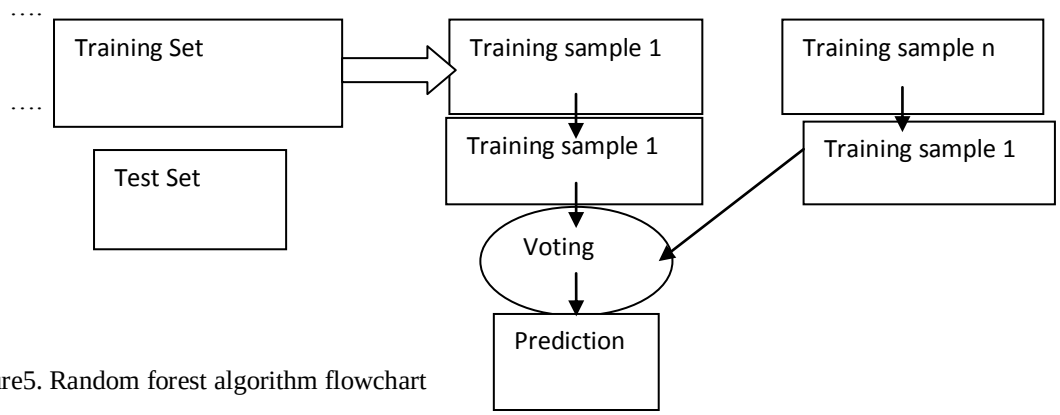

Figure5. Random forest algorithm flowchart

Results and Discussion:

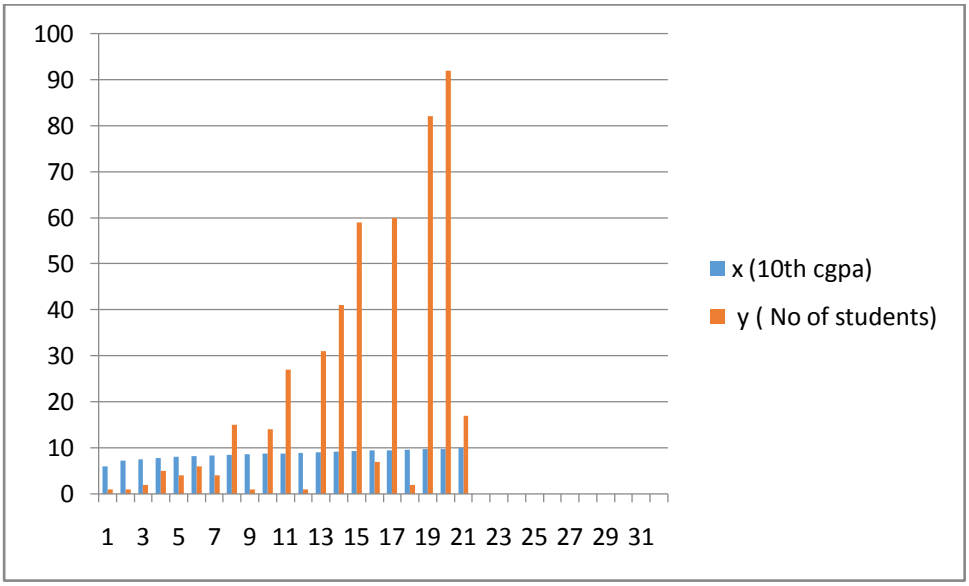

Figure6. 10th class grade points

In the above figure X- axis represents the CGPA of students obtained in SSC. And Y-axis indicates the number of students.

ISSN- 2394-5125VOL 7, ISSUE 14, 2020

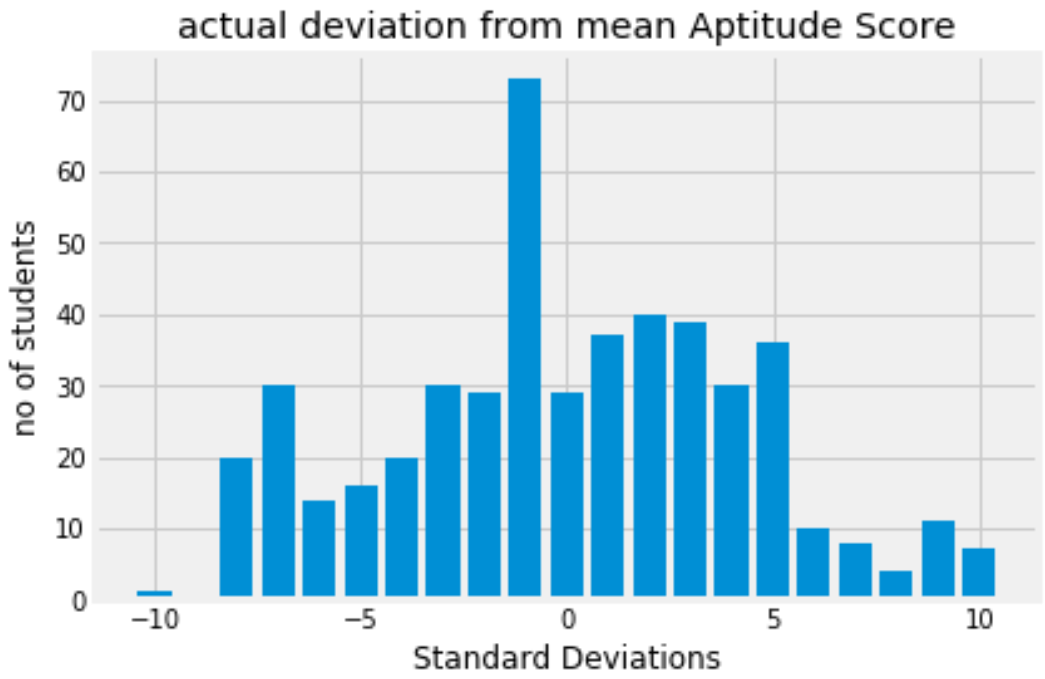

Figure7.

The actual deviation from mean aptitude score can be seen in the above figure in which mean is calculated for the aptitude such that the difference between the actual value and difference value is shown in terms of graph.

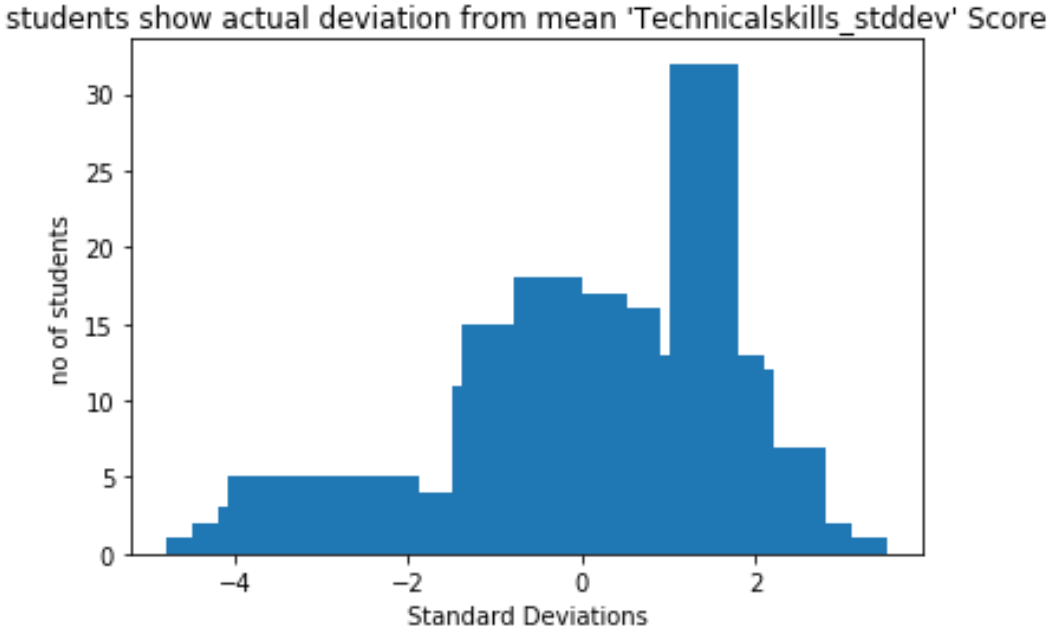

Figure8. The above diagram represents students score deviation from mean of technical skills.

$$\boldsymbol{d}\mathrm{f}[\text{'Technicalskills}'_{\text{stddev}} = [\mathrm{s} - \mathrm{df}[\text{'Technicalskills'}].\mathrm{mean}() \text{ for s in df}[\text{'Technicalskills'}]]$$

$$df['Technicalskills\_stddev'].mean()$$

ISSN- 2394-5125VOL 7, ISSUE 14, 2020

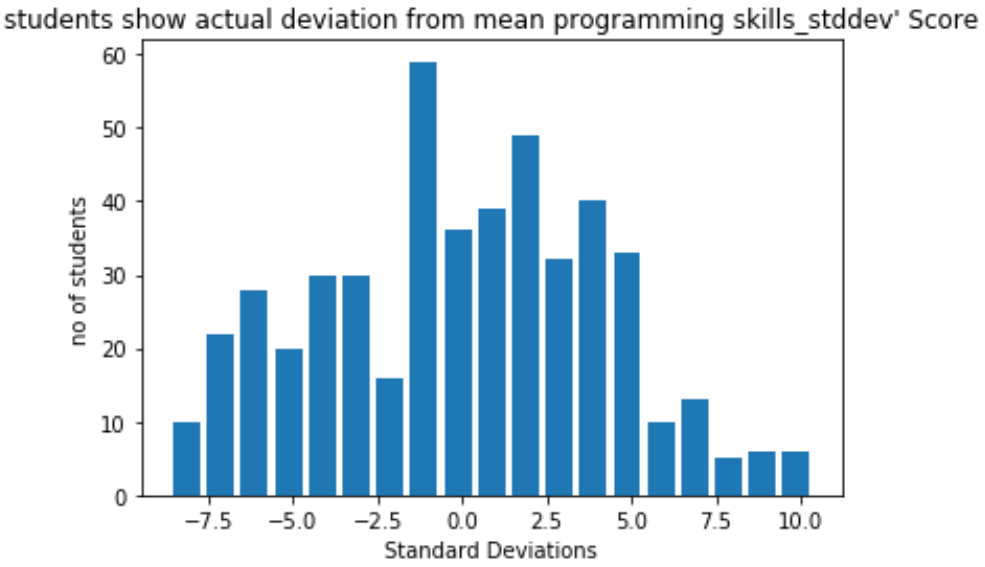

Figure9. Above mentioned figure depicts the mean variance of students from standard deviation of programming skills.

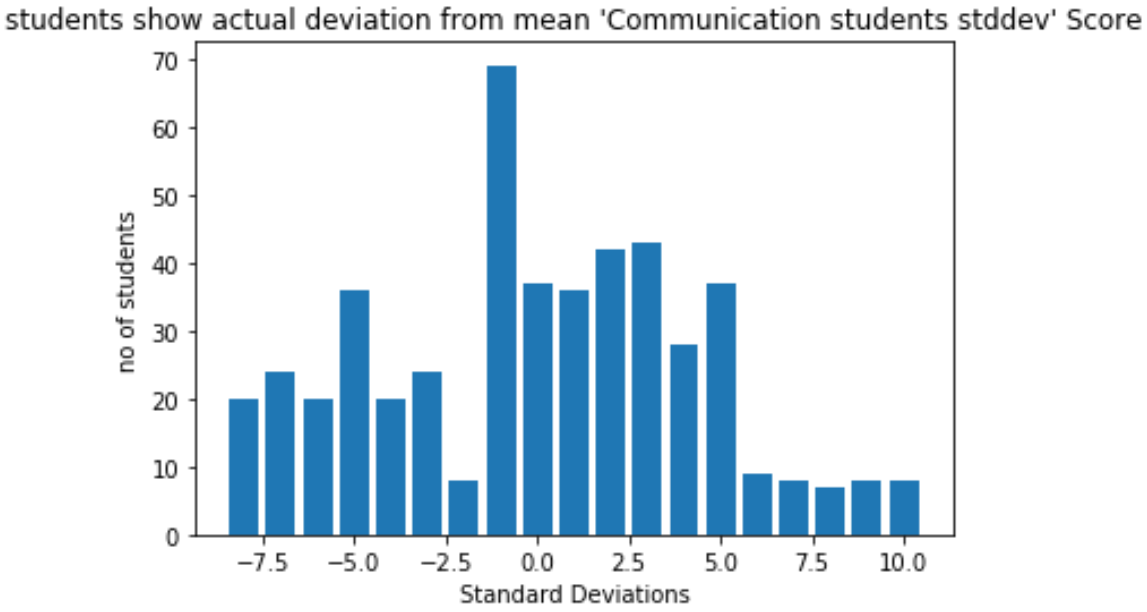

Figure10. It represents the standard deviation of communication skills of the students from actual mean of overall students.

$$df['Communication\_stddev'] = [s - df['Communication\ '].mean()\ for\ s\ in\ df['Communication\ ']]$$

df['Communication_stddev'].mean()

ISSN- 2394-5125VOL 7, ISSUE 14, 2020

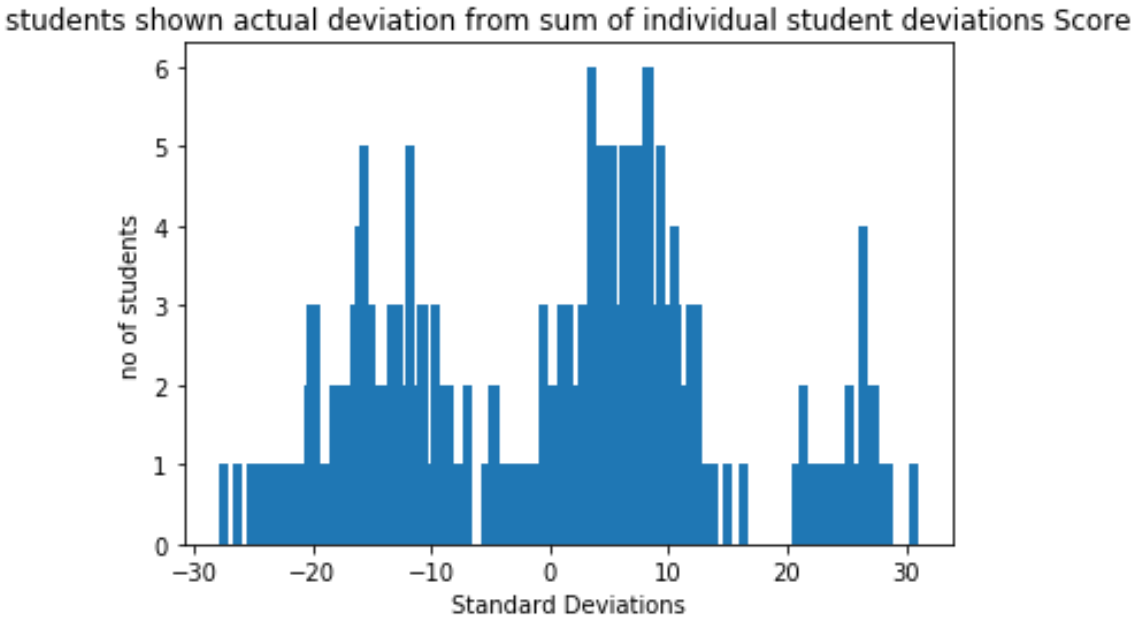

Figure11. The above mentioned figure indicates the sum of individual student deviation scores from the actual mean deviation.

Result:

The informational index utilized for is further spitted into two sets comprising of two third as preparing set and 33% as testing set. Among the two calculations applied arbitrary timberland demonstrated the best outcomes. The proficiency of the two methodologies is thought about as far as the precision.

The exactness of the forecast model/classifier is characterized as the complete number of accurately anticipated/grouped occurrences. Precision is given by utilizing following equation:

True positive is an outcome where the model accurately identifies the positive class. Likewise, true negative is an outcome where the model correctly predicts the negative class.

False positive is an outcome where the model incorrectly predicts the positive class. On the other hand false negative is an outcome where the model incorrectly predicts the negative class.

$$\text{Accuracy}=\left(TP+\frac{TN}{TP}+FN+FP+TN\right)*100$$

Algorithms	Accuracy	Precision	Recall
Support Vector Machine	82.85%	0.877	0.81
Random Forest	85.14%	0.92	0.87
Linear Regression	80.05%	0.63	0.78
K-Means	75%	0.59	0.75

ISSN- 2394-5125VOL 7, ISSUE 14, 2020

Naïve Bayes	76.23%	0.88	0.81

Table1. Comparison of the performances of machine learning algorithms such as support vector machine, random forest, linear regression, k-means, and Naïve Bayes.

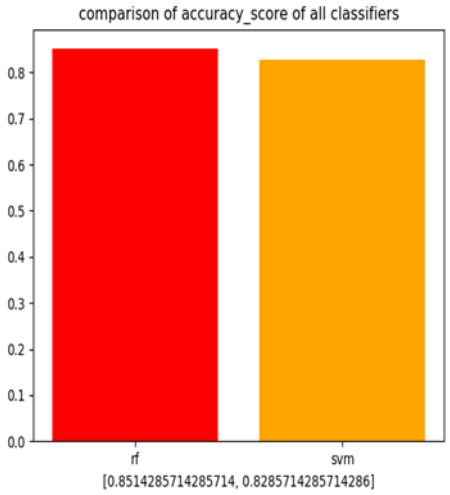

Figure12. Random forest algorithm gives the best accuracy i.e. 85.14%. Whereas, support vector machine algorithm gives the accuracy 82.85%.

$$acc\_rf = metrics.accuracy\_score(rf\_pred, y\_test)$$

$$acc\_svm = metrics.accuracy\_score(sv\_pred, y\_test$$

Future Enhancement:

The future enhancements of the project are to focus on the parameters according to the company that is going to recruit students instead of testing all the parameters. We can also eliminate some of unwanted attributes so that the processing time can be reduced and also within short span we can get the result.

Conclusion:

The campus placement activity is incredibly a lot of vital as institution point of view as well as student point of view. In this regard to improve the student's performance, a work has been analysed and predicted using the classification algorithms Support vector machine and the Random forest algorithm to validate the approaches. The algorithms are applied on the data set and attributes used to build the model. The accuracy obtained after analysis for Decision tree is 85.14% and for the Support vector machine is 82.85%. Hence, from the above said analysis and prediction its better if the Random Forest algorithm is used to predict the placement results.

References:

[1] "Student Placement Analyzer: A Recommendation System Using Machine Learning" 2017 International Conference on advanced computing and communication systems (ICACCS-2017), Jan 06-07,2017, Coimbatore, INDIA

[2] Patricio Garcia,AnaliaAmandi,SilviaSchiaffino, MarceloCampo, "Evaluating Bayesian networks' precision for detecting students' learning styles", Computers & Education ,49, pp.794-808,2007.

[3] Luan, Jing. "Data Mining and Knowledge Management in Higher Education-Potential Applications" (2002).

[4]. Ajay Kumar Pal and Saurabh Pal, "Classification Model of Prediction for Placement of Students", I. J. Modern Education and Computer Science, 2013, 11, 49-56

[5] G.Vadivu, K.Sornalakshmi, "Applying Machine Learning Algorithms for Student Employability Prediction Using R". Int. J. Pharm. Sci. Rev. Res., Vol.43(1), No. 11, pp. 38-41, 2017.

[6]Ajay Shiv Sharma, Swaraj Prince, Shubham Kapoor, Keshav Kumar "PPS-Placement prediction system using logistic regression" IEEE international conference on MOOC,innovation and Technology in Education(MITE), December 2014.

[3]. MangasuliSheetal B, Prof. Savita Bakare "Prediction of Campus Placement Using Data Mining AlgorithmFuzzy logic and K nearest neighbour" International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 6, June 2016 .

[5]. Jai Ruby, Dr. K. David "Predicting the Performance of Students in Higher Education Using Data Mining Classification Algorithms - A Case Study" International Journal for Research in Applied Science & Engineering Technology (IJRASET) Vol. 2,Issue 11,November 2014.

[6]. Ankita A Nichat, Dr.Anjali B Raut "Predicting and Analysis of Student Performance Using Decision Tree Technique" International Journal of Innovative Research in Computer and Communication Engineering V0l. 5, Issue 4, April 2017.

[7]. OktarianiNurulPratiwi "Predicting Student Placement Class using Data Mining" IEEE International Conference 2013.

[8] Purushottama Rao K., Koneru A., Naga Raju D. (2019) OEFC Algorithm—Sentiment Analysis on Goods and Service Tax System in India. In:Mallick P., Balas V., Bhoi A., Zobaa A. (eds) Cognitive Informatics and Soft Computing. Advances in Intelligent Systems and Computing, vol 768.Springer,Singapore

[9] K. Lavanya, L. S. S. Reddy and B. Eswara Reddy, "Modelling of Missing Data Imputation using Additive LASSO Regression Model in Microsoft Azure", Journal of Engineering and Applied Sciences,2018,Vol 13,Special Issue 8,pp:6324-6334.(SCOPUS)

[10] Rama DeviBurri, Ram Burri, Ramesh ReddyBojja, SrinivasaraoBuraga"Insurance claim Analysis using Machine learning Algorithms, "International journal of innovative technology and Exploring Engineering (IJITEE), Volume-8,Issue-6S4, April-2019, ISSN: 22278-3075