International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-8 Issue-5S, May 2019

Comparative Study of Multiple Machine Learning Algorithms for Students' Performance Data for Job Placement in University

Athreya Shetty B, Akram Pasha, Amith Singh S, Shreyas N I, Adithya R Hande

Abstract: ***In the era of data evolution, many organizations have taken the lead in storing the data in huge data repositories. Analysis of data comes with several challenges since the time the data is captures till the insights are inferred from the data. Accentuating the accuracy of data analysis is of paramount importance as many critical decisions are totally dependent on the outcomes of the analysis. Machine learning has been found as the most effective and most preferred tool in the literature for in-memory data analytics. Universities mostly collect the statistical data related to the students that is only either used quantitatively or sparsely analyzed to gain the insights that could be useful for the authorities to enhance the percentage of placements in campus drives held through early analysis of such data accurately. The work proposed in this paper formulates the problem of predicting the likelihood of a student getting placed in a company as a binary classification problem. Then it makes an effort to train and perform the empirical study of following multiple machine learning algorithms with the placement data; Logistic Regression, Naïve Bayes, Support Vector Machine, K-Nearest Neighbor and Decision Tree. The machine learning classification models are built to predict the probabilities of a student getting placed in a company based on the student's academic scores, achievements, work experience (internship), and many other relevant features. Such an analysis helps the university authorities to dynamically create plans to enhance the unlikely students to be placed in a company participating in the campus recruitment held in the university. To improve these models and to avoid the models from overfitting to the training data, strategies like K-Fold cross-validation is applied for various values of k. The machine learning models selected are also compared for its efficiency by employing the supervised and unsupervised feature extraction techniques such as PCA and LDA. The Decision Tree model with K as 10 for cross-validation and PCA has outperformed all the other models producing the accuracy of 72.83% with satisfactory support and recall during experimentation. The application focuses on the targeted group of students, to eventually improve the probability of students getting placed during campus recruitment drives held in the university.***

Index Terms**: *Data Mining, Data Analytics, Classification, Machine Learning, K-Fold Cross Validation, PCA, LDA***

I. INTRODUCTION

There is a need to improve the skill sets of the students attending the campus placements [1] [2] [3]. Study on their previous data with machine learning algorithms tells us the necessary measures to be taken to improve the quality of the students and increase the probability of getting placed[4][5][6]. The methods that are implemented and used in the universities are placement cell where they have all the placement data regarding the students and interact with them to improve the placements [7][8][9]. This placement data can be used more efficiently with the implementation of machine learning models to improve the placements as well as to predict whether a student might get placed or not and the relevant reasons [10][11][12][13][14]. After many experimentation and research we have found out the most efficient machine learning model suited for the task with comparison to many other machine learning prediction algorithms (Decision Tree) [15][16][17][18].

During the course of this experimentation and research, we have selected 5 basic linear and non-linear machine learning models, which are Logistic Regression, Naïve Bayes, Support Vector Machine, K-Nearest Neighbour and Decision Tree. Along with these models, in order to improve the result, we have used K-Fold Cross Validation, PCA and LDA. We have calculated the model accuracy, precision, recall and f1-score to support our result and the efficiency of the model. The major contributions of the work proposed are:

- It incorporates the entire life cycle of the machine learning model development from data collection to the visualization of results.
- Performs cleaning of data by taking care of null values and outliers.
- Performs the transformation of data to make it suitable for fitting into a machine learning model using data scaling techniques
- Trains 5 machine learning models.
- Performs k-fold cross validation on 5 classification models built.
- Performs the critical performance analysis of the 5 classification models using several evaluation metrics.

The major sections of the paper is organized in the following structure. Section-II discusses the related work in the field of data analytics using machine learning classifiers.

Manuscript published on 30 May 2019.
***** Correspondence Author (s)
Athreya Shetty B, School of Computing and Information Technology, REVA University, Bangalore, India.
Akram Pasha, School of Computing and Information Technology, REVA University, Bangalore, India.
Amith Singh S, School of Computing and Information Technology, REVA University, Bangalore, India.
Shreyas N I, School of Computing and Information Technology, REVA University, Bangalore, India.
Adithya R Hande, School of Computing and Information Technology, REVA University, Bangalore, India.

Section-III introduces the framework of the work proposed. Section-IV discusses the experimental setup and results. Section-V concludes the paper by presenting the future directions of this work.

II. RELATED WORK

In the year 2016, the work of [19], conducted the research on student placement prediction. The models used were Logistic Regression, K-NN and SVM. The key attributes considered were academic details, communication skills, team work and programming skills. In the work of [20], placement prediction based on the psychological state of the student was studied and presented. They have used the real time data of the final year student with Neural Network and Decision Tree Classifier for prediction. A student placement analyzer was built in the work of [21], where in, machine learning to identify the students whose technical as well as interpersonal skill were targeted for improvement. In the work of [22] a classification and prediction model was built for improving the accuracy of student placement prediction. The clustering techniques used were canopy, EM, Farthest First, Hierarchical Cluster and SimpleKMeans. Student performance data was used in the work of [23] to predict the kind of company the student will fit into. They have used Decision Tree algorithm and the clustering techniques on the data using data mining tools like WEKA. In the year 2017, the work of [24] had presented a paper on Educational Dropout Prediction using machine learning techniques such as SVM. They were able to find out the attributes that were required to predict student dropouts.

The prediction of student's likelihood of getting placed in the campus drive conducted in the university can be formulated as a binary classification problem, where in, the predictor is expected to classify the unseen observations to a target class of whether a student is placed or not.

III. METHODOLOGY

The data collection mechanism was targeted to collect sufficient amount of data required to build the robust classifier. We can see utilization of such a mechanism in the work of [9]. The data collection is the first step in any data analytics process. We intended to collect the data that is sufficiently huge and contains all the required information about the student that would contribute to a classification model. We sceptically designed the google form and collected the data from all the present students of REVA University pursuing B. Tech, 2015-2019 engineering batch in Computer Science and Engineering and the Placement Department of REVA University. The attributes of our interest include the student's grades, achievements, awards and many more. There are several machine learning algorithms that are available for prediction. Out of many such algorithms, we chose the combination of linear and non-linear classification algorithms for our comparative research and study of the placement prediction; Logistic Regression, Naïve Bayes, SVM, K-Nearest Neighbours and Decision Tree algorithms, to benchmark the machine learning model having highest accuracy of prediction.

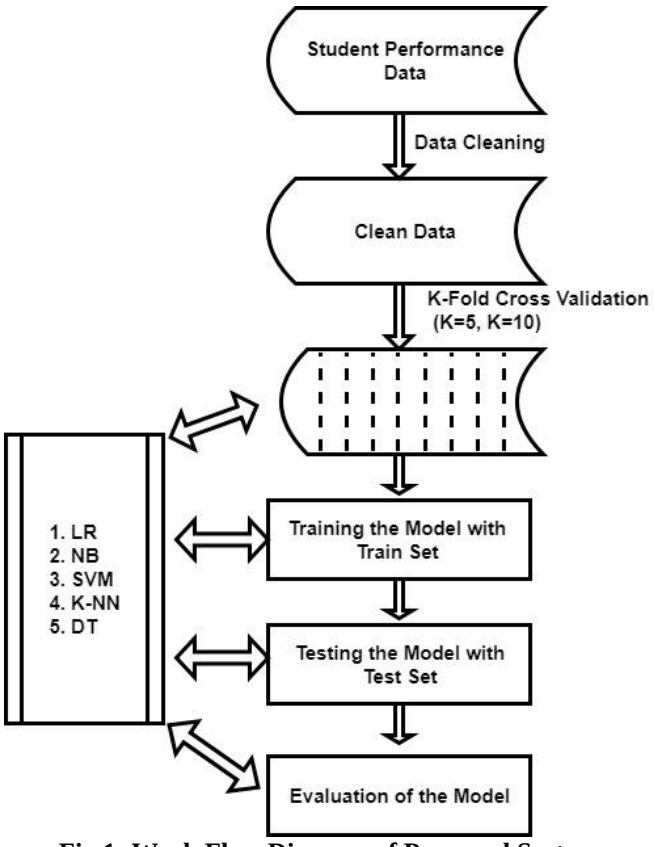

Fig 1: Work Flow Diagram of Proposed System

Principal Component Analysis (PCA) is used as one of the greatly used feature extraction techniques that reduces the dimension of the data without losing much of the information. It is generally used along with the classification models to work on the dimensionally reduced datasets. It is more often used for object recognition, computer vision, data compression etc. Logistic Regression is one of the fundamental techniques that is used in classifying the datasets in to groups based on the non-linearity features. Logistic regression is used for classification and not for regression. Support Vector Machine (SVM) classifier is a linear supervised classifier such as any other classifier like logistic regression that uses kernel functions to construct the hyperplane that separated the two classes of data. The difference is that it has a margin-based loss function. Decision Tree is the most commonly use ensemble machine learning algorithm that builds the classifier using the attributes contribution in creating the branches of tree. It is used in statistics and data analysis for predictive models. Decision Tree breaks down the dataset into smaller subsets. It can handle both categorical and numerical data.

The initial step in developing the classification models, is to pre-process [9] the data that involves cleaning of data, handling the null values, conversion of string values to numerical values, finding the important features and many more pre-processing techniques to convert the data into an understandable format.

Problem Statement:

The machine learning model used in this work is a binary classification problem that predicts the class of a new observation based on the observations that were trained to a model.

With the class having binary values {0, 1}, mathematically, the machine learning models can be defined as follows.

The Logistic Regression Classifier creates the hypothesis as shown in equation (1) with hΘ(x) = sigmoid (Z)

$$Z = WX + B ----(1)$$

The Naïve Bayes Classifier creates the hypothesis as shown in equation (2)

$$P\left(\frac{A}{B}\right) = \frac{P\left(\frac{B}{A}\right)P(A)}{P(B)} ----(2)$$

The Support Vector Classifier creates the hypothesis as shown in equation (3)

$$x.y = x_1y_1 + x_2y_2 = \sum_{i=1}^{2}(xi.yi) ----(3)$$

The K-Nearest Neighbours Classifier creates the hypothesis as shown in equation (4)

$$y = \frac{i}{k}\sum_{i=1}^{2} yi ----(4)$$

IV. EXPERIMENTATION AND DISCUSSION OF RESULTS

The classification models chosen are developed using scikit-learn v0.20.3 [38] documentation and developed using python 3.6.

The classification models tend to give biased results on the training datasets. Therefore, to avoid over fitting of data, we incorporated k-fold cross validation, a statistical method that divides the training and test data sets to estimate the overall performance of machine learning algorithms.

The predictive machine learning models categorize the students whether a student gets placed or not in the campus drive based on the history of students performance so far. The data collected contains various features related to a student during his or her course of education.

Applying the predictive machine learning algorithm on the data collected and processed. The below data shows the accuracy of different models.

Table 1: Evaluation of Classifiers

	LR	NB	SVM	KNN	DT
Accuracy	60.5	45.41	48.66	55.42	**72.83**

The table 1 shows the accuracies of the respective models LR (Logistic Regression), NB (Naïve Bayes), SVM (Support Vector Machine), KNN (K-Nearest Neighbour) and DT (Decision Tree). It shows that the accuracy of Naïve Bayes is the least 45.41 and the accuracy of Decision Tree is **72.83,** the highest.

The table-1 also shows the accuracy of models after repetitive modification of the model and the data. Naïve Bayes is a probabilistic supervised machine learning model that is used for any classification task that performs the classification based on the probability estimations using Bayes Theorem. The assumption made in this model is that the features are independent. Since the features in the placement datasets are not independent, the accuracy of the Naïve Bayes model results to be the least compared to the rest. Naïve Bayes model works extremely well on a dataset where all the features are independent. They are mostly used in spam filtering, sentiment analysis, recommendation system etc. It is fast and easy to implement but the biggest disadvantage is that it requires the predictor to be independent.

Decision tree is a supervised machine learning algorithm that is most used in both regression and classification tasks to get the stable models. The performance of the Decision Tree is found most efficient on both categorical and continuous input and output variables. The algorithm is completely invariant and remains stable to scaling of the data. Decision tree works extremely well when the attributes in the data sets are of different scales or mix of binary and continuous. The dataset used in the current study matches exactly with the requirement of the Decision Tree algorithm resulting with highest accuracy as against the other models used.

K-Fold Cross Validation:

Cross validation is a statistical method primarily used to combat the bias-variance trade off found in any machine learning model. It is a strategy that estimates the overall performance of a machine learning model. In this current study, it is applied on various machine learning models to compare the performances of all the models and to choose the best model that suits for the data set used.

The procedure has a single parameter called k that refers to the number of sets that a given data sets are to be split while training any machine learning model. The specific value of k is chosen so as to divide the data sets into acceptable percentage of splits in terms of training and testing data.

Table 2: K-Cross Validation for K=5

K=5	LR	NB	SVM	KNN	DT
Accuracy	62.74	44.18	53.52	59.34	**63.85**
Precision	80.76	52.5	56.75	63.32	70.27
Recall	86.0	25.99	84.00	78.0	74.0
F1-Score	83.29	34.76	67.73	68.89	72.08

The table 2 shows the results (Accuracy, Precision, Recall, F1-Score) when the machine learning models are applied over the data set with the K-Fold cross validation for K=5. The highest accuracy obtained is for Decision Tree which is 63.85 and the least accuracy result is obtained for Naïve Bayes which is 44.18.

The table 3 shows the results (Accuracy, Precision, Recall, F1-Score) when the machine learning models are applied over the data set with the K-Fold cross validation for K=10. The highest accuracy obtained is for Decision Tree which is 67.63 and the least accuracy result is obtained for Naïve Bayes which is 41.38.

Table 3: K-Cross Validation for K=10

K=5	LR	NB	SVM	KNN	DT
Accuracy	63.61	41.38	54,72	54.86	**67.63**
Precision	69.97	31.66	57.88	63.54	72.55
Recall	83.99	20.0	84.0	70.0	76.0
F1-Score	76.34	24.51	68.53	66.61	74.23

Principal Component Analysis (PCA):

Principal Component Analysis (PCA) is used as one of the greatly used feature extraction techniques that reduces the dimension of the data without losing much of the information.
Simple matrix operations from linear algebra and statistics are used in this method to calculate a projection of the original data into the same number of fewer dimensions.

Table 4: PCA, K=5

K=5	LR	NB	SVM	KNN	DT
Accuracy	58.35	47.82	47.89	55.42	**69.78**
Precision	63.38	52.50	56.75	63.32	69.74
Recall	86.00	25.99	84.00	78.00	78.00
F1-Score	58.92	40.67	45.42	54.18	64.28

The table 4 shows the results (Accuracy, Precision, Recall, F1-Score) when the machine learning models are applied over the data set with the K-Fold cross validation for K=5 and Principal component analysis. The highest accuracy obtained is for Decision Tree which is 69.78 and the least accuracy result is obtained for Naïve Bayes which is 47.82.

Table 5: PCA, K=5

K=5	LR	NB	SVM	KNN	DT
Accuracy	60.50	45.41	48.66	52.08	**72.83**
Precision	63.39	31.66	57.88	63.54	77.78
Recall	84.99	20.00	83.99	70.00	78.00
F1-Score	59.91	35.21	46.36	49.63	69.03

The table 5 shows the results (Accuracy, Precision, Recall, F1-Score) when the machine learning models are applied over the data set with the K-Fold cross validation for K=5 and Principal component analysis. The highest accuracy obtained is for Decision Tree which is 72.83 and the least accuracy result is obtained for Naïve Bayes which is 45.41.

Linear Discriminant Analysis:

Logistic regression is a classification algorithm mostly suitable for binary classification problems. Whereas Linear Discriminant Analysis is a preferred linear classification technique when the number of decision values in a decision variables are more than two.

Table 6: LDA, K=5

K=5	LR	NB	SVM	KNN	DT
Accuracy	58.35	47.82	47.53	55.42	**63.42**
Precision	63.38	52.50	56.75	63.32	71.86
Recall	86.00	25.94	84.00	78.00	78.00
F1-Score	58.92	40.67	45.42	54.18	67.60

The table 6 shows the results (Accuracy, Precision, Recall, F1-Score) when the machine learning models are applied over the data set with the K-Fold cross validation for K=5 and Linear Discriminant analysis. The highest accuracy obtained is for Decision Tree which is 63.42 and the least accuracy result is obtained for SVM that is 47.53.

Table 7: LDA, K=10

K=5	LR	NB	SVM	KNN	DT
Accuracy	59.50	45.41	48.66	52.08	**68.91**
Precision	63.39	31.66	57.88	63.54	77.58
Recall	83.99	20.00	83.99	70.00	76.00
F1-Score	59.91	85.21	46.46	49.63	68.07

The table 7 shows the results (Accuracy, Precision, Recall, F1-Score) when the machine learning models are applied over the data set with the K-Fold cross validation for K=10 and Linear Discriminant Analysis. The highest accuracy obtained is for Decision Tree which is 68.91 and the least accuracy result is obtained for Naïve Bayes which is 45.41.

The accuracy of the model depends highly upon the vast data collected. Also for the better dataset the with Random Forest model, we can have better accuracy. When new features are added to the dataset, the accuracy will automatically increase. After modifying the model again and again and also by adding new features, there can be huge improvement in the student performance, placement results as well as the predictions made by the model. The general tendency of students getting placed were the students with good SGPA score, 10^{th} and 12^{th} score including the people with additional achievements and internship experience. Extra training can provided to the pre final year students. With the available data and resources, students can be provided with very particular training instead of a general common training to everyone. Extra classes can conducted to improve the semester scores of the students to meet the company criteria's. Also the model is used to predict the reason for not getting placed.

The small dataset and the Entropy in Decision Tree had greater impact in getting high results for Decision Tree in our current work. The Entropy in Decision Tree controls how a Decision Tree decides to split the data and impacts immensely on how a Decision Tree draws its boundaries. The entropy in a Decision Tree is calculated using the equation (5), where p(X) indicates the probability of the observations in a dataset.

$$Entropy = -\sum p(X)\ log\ p(X) - - - -(5)$$

V. CONCLUSION AND FUTURE SCOPE

The analytics of placement data is one of the tasks that can be exercised in the field of educational data mining. This paper collects the students' performance data and trains 5 machine learning classification models. The classification outcome of the work proposed gives us the detailed picture of the students' performance that would make them placed. The accuracy of the Decision Tree algorithm was found to be 72.83 which is the highest among all and in most cases. The algorithm with the least accuracy was Naïve Bayes which is 45.41 respectively. The vast amount of student at university, state, or at country level can form the big data. Then, the development of distributed computing classification models remains a future direction of this work.

ACKNOWLEDGMENT

The authors acknowledge the management and School of C&IT personnel of REVA University, Bengaluru for providing data support and all other facilities to carry out this project.

REFERENCES

1. Klopfenstein K, Thomas MK. The link between advanced placement experience and early college success. Southern Economic Journal. 2009 Jan 1:873-91.
2. Berger DM. Mandatory assessment and placement: The view from an English department. New Directions for Community Colleges. 1997 Dec;1997(100):33-41.
3. Morgan DL, Michaelides MP. Setting Cut Scores for College Placement. Research Report No. 2005-9. College Board. 2005.
4. Clement, A. and Murugavel, T., 2015. English for Employability: A Case Study of the English Language Training Need Analysis for Engineering Students in India. *English language teaching*, *8*(2), pp.116-125.
5. Albraikan A, Hafidh B, El Saddik A. iAware: A Real-Time Emotional Biofeedback System Based on Physiological Signals. IEEE Access. 2018;6:78780-9.
6. Miller, Ryan, Hang Ho, Vivienne Ng, Melissa Tran, Douglas Rappaport, William JA Rappaport, Stewart J. Dandorf, James Dunleavy, Rebecca Viscusi, and Richard Amini. "Introducing a fresh cadaver model for ultrasound-guided central venous access training in undergraduate medical education." *Western Journal of Emergency Medicine* 17, no. 3 (2016): 362.
7. Reeves LM, Schmorrow DD, Stanney KM. Augmented cognition and cognitive state assessment technology–near-term, mid-term, and long-term research objectives. InInternational Conference on Foundations of Augmented Cognition 2007 Jul 22 (pp. 220-228). Springer, Berlin, Heidelberg.
8. Sharma AS, Prince S, Kapoor S, Kumar K. PPS—Placement prediction system using logistic regression. In2014 IEEE International Conference on MOOC, Innovation and Technology in Education (MITE) 2014 Dec 19 (pp. 337-341). IEEE.
9. Jeevalatha T, Ananthi N, Kumar DS. Performance analysis of undergraduate students placement selection using decision tree algorithms. International Journal of Computer Applications. 2014 Jan 1;108(15).
10. Kabra RR, Bichkar RS. Performance prediction of engineering students using decision trees. International Journal of computer applications. 2011 Dec;36(11):8-12.
11. Kuzilek J, Hlosta M, Herrmannova D, Zdrahal Z, Wolff A. OU Analyse: analysing at-risk students at The Open University. Learning Analytics Review. 2015 Mar 18:1-6.
12. Mishra T, Kumar D, Gupta S. Mining students' data for prediction performance. In2014 Fourth International Conference on Advanced Computing & Communication Technologies 2014 Feb 8 (pp. 255-262). IEEE.
13. Pandey M, Sharma VK. A decision tree algorithm pertaining to the student performance analysis and prediction. International Journal of Computer Applications. 2013 Jan 1;61(13).
14. Taruna S, Pandey M. An empirical analysis of classification techniques for predicting academic performance. In2014 IEEE International Advance Computing Conference (IACC) 2014 Feb 21 (pp. 523-528). IEEE.
15. Choudhary R, Gianey HK. Comprehensive Review On Supervised Machine Learning Algorithms. In2017 International Conference on Machine Learning and Data Science (MLDS) 2017 Dec 14 (pp. 37-43). IEEE.
16. Kumar RS, KP JK. ANALYSIS OF STUDENT PERFORMANCE BASED ON CLASSIFICATION AND MAPREDUCE APPROACH IN BIGDATA. International Journal of Pure and Applied Mathematics. 2018;118(14):141-8.
17. Lakshmanan GT, Li Y, Strom R. Placement strategies for internet-scale data stream systems. IEEE Internet Computing. 2008 Nov;12(6):50-60.
18. Thangavel SK, Bkaratki PD, Sankar A. Student placement analyzer: A recommendation system using machine learning. In2017 4th International Conference on Advanced Computing and Communication Systems (ICACCS) 2017 Jan 6 (pp. 1-5). IEEE.
19. Giri A, Bhagavath MV, Pruthvi B, Dubey N. A Placement Prediction System using k-nearest neighbors classifier. In2016 Second International Conference on Cognitive Computing and Information Processing (CCIP) 2016 Aug 12 (pp. 1-4). IEEE.
20. Halde RR, Deshpande A, Mahajan A. Psychology assisted prediction of academic performance using machine learning. In2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT) 2016 May 20 (pp. 431-435). IEEE.
21. Thangavel SK, Bkaratki PD, Sankar A. Student placement analyzer: A recommendation system using machine learning. In2017 4th International Conference on Advanced Computing and Communication Systems (ICACCS) 2017 Jan 6 (pp. 1-5). IEEE.
22. Shukla M, Malviya AK. Modified Classification and Prediction Model for Improving Accuracy of Student Placement Prediction. Available at SSRN 3351006. 2019 Mar 12.
23. Pruthi K, Bhatia P. Application of Data Mining in predicting placement of students. In2015 International Conference on Green Computing and Internet of Things (ICGCIoT) 2015 Oct 8 (pp. 528-533). IEEE.
24. Kumar M, Singh AJ, Handa D. Literature survey on educational dropout prediction. International Journal of Education and Management Engineering. 2017 Mar 1;7(2):8.
25. Python 2.7 Documentation-docs.python.org/2.7 Scikit Learn Machine Learning in Python- www.scikit-learn.org

AUTHORS PROFILE

Mr. Athreya Shetty B is pursuing Bachelor of Technology in REVA University, Bangalore, India.

Prof. Akram Pasha, Associate Professor, School of Computing and Information Technology, REVA University, Bengaluru, India owns Bachelor of Engineering from Mysore University, Master of Engineering from Bangalore University and currently pursuing Ph. D from Visveswaraya Technological University, Belagaum, India. His major areas of research interests are applications of Computational Intelligence techniques on Data analytics.

Mr. Amith Singh S is pursuing Bachelor of Technology in REVA University, Bangalore, India.

Mr. Shreyas N I is pursuing Bachelor of Technology in REVA University, Bangalore, India.

Mr. Adithya R Hande is pursuing Bachelor of Technology in REVA University, Bangalore, India.