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

Multiple Machine Learning Classifiers for Student's Admission to University Prediction

Anil B, Akram Pasha, Aman, Aman Kumar Singh, Aditya Kumar Singh

Abstract: ***Data is the most important asset for any organization which is further processed to produce useful information. Machine Learning and Big Data techniques are widely used for industrial sectors to generate useful patterns helpful for earning more profits and expand businesses. From the past few years, a lot of research works have been done by using Big Data techniques on educational data for improvement in Education System. Machine Learning and Big Data can be useful for predicting the students' admission, performance of teaching, performance of a student, identifying the group of students of similar behavior. However, the manual process of record checking is time consuming, tedious, and error prone; due to the inherent volume and complexity of data. In this study, the combination of linear and non-linear machine learning algorithms; Logistic Regression, Decision Tree, k-NN, and Naïve Bayes have been chosen to perform prediction of the target class for an unseen observation by polling. As the models built in this work are predicting the likelihood of a student taking up the admission into any university based on the student data collected by any marketing agency, the combined models are collectively called as the Admission Predictor. The administrative officials of any academic institution can use this kind of an application to explore and analyze the patterns that are affecting the student admission and come up with enhanced strategies to improve admission. Such an application not only plays a vital role in administration, but also help the management in reformulating the marketing criteria for overall development of academic institution.***

Keywords: Classification, Data Mining, Data Analytics, K-Fold Cross Validation, LDA, Machine Learning, PCA.

I. INTRODUCTION

Academic institutions across the globe produce and store the massive data related to student's admission, student's performance and so on. It is a routine task in such academic institutions to perform several types of analysis by orienting analysis with a particular objective of an institution. Screening of such massive student's data is too cumbersome willingness of a student to take up an admission to an institution is as good as predicting whether a customer is going to buy a product based on his purchase history. The admission predictor developed in this study uses the student's application data that includes many features including a class variable that has binary value. This class variable is true if the student had taken admission or false if he did not. Therefore, an attempt is made in this study to predict the likelihood of new students based on their features. We can find several such efforts made in the literature. Like the one reported in the work published in [3]. Using nominal and categorical attributes and past collected data this work is done at ease. Implementation of two different techniques on our data set; with that classification builds a predictive model and association rules which were used to find interesting hidden information in the student's records. This study will help the college/university to determine their direction and improve when necessary to cope up with their student admission to their college. It provides a beneficial tool to predict and evaluate those students who need attention and care and finds out any deviation before it happens and become a decrease in performance and reduce the failure rate.

The main aim of this project is to contribute towards smart prediction based on student data which can help in reducing congestion and improve quality of education. The college administration is the basic use of this application, this model predicts the binary outcome 'yes' / 'no' based on the history of admission data of students. The dataset used here is as of now synthetic and generated through the faker (python package). Some of the attributes that are used are 10^{th} class performance, 12^{th} class performance, competitive exam performance, college email clicked, college website visited, and so on.

The work proposed in this paper is developed using scikit-learn machine learning libraries [4] using Pyhton 3.6. The combination of both linear and non-linear machine learning algorithms are selected to be trained by the dataset. The machine learning models are tuned by performing iterative runs while training with several randomized splits to avoid over fitting and also to enhance its overall performance. As a result, the selection of the best 3 models is performed based on the evaluation metrics of the basic classifier.

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.
- It performs feature engineering by using feature selection and feature extraction techniques
- Trains 4 machine learning models with the datasets.
- Performs k-fold cross validation on 4 classification models built.
- Performs the critical performance analysis of the 4 classification models using several evaluation metrics.

Manuscript published on 30 May 2019.
***** Correspondence Author (s)
Anil B, School of Computing and Information Technology, Reva University, Bangalore, India
Akram Pasha, Associate Professor, School of Computing and Information Technology, Reva University, Bangalore, India.
Aman, School of Computing and Information Technology, Reva University, Bangalore, India.
Aman Kumar Singh, School of Computing and Information Technology, Reva University, Bangalore, India.
Aditya Kumar Singh, School of Computing and Information Technology, Reva University, Bangalore, India.

The rest 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. Section-III introduces the framework of the work proposed and what are all the different steps involved in getting the result. Ex-Data pre-processing, Training the model, etc. Section-IV discusses the experimental setup and results. Based on different approaches and methods to find out various results and performance of model, so that it would be easy to judge which approach is giving the best result. Example-k fold validation with different number of features and subset of dataset. Section-V defines the main aim and outcome of the model i.e. the prediction between acceptance or rejection of student admission into the organization based on previous dataset. Section-VI concludes the paper by presenting the future directions of this work.

II. RELATED WORK

From the previous few years of research, it is found that many machine learning classification models are used in predicting the target classes efficiently. In the work [5], the student admission prediction system having a large data set using the concept of distributed data mining. In the work [6], they studied on how a decision tree model can be used in studying the Student data and there benefits. Demo on the ID3 Decision Tree Algorithm usefulness on the student data for enrolment management was displayed in the work [7] [14] [15]. Also the work by [8] explains about how entrance exam marks plays a role in the admission of a student and how it can be used for admission purposes using decision support system. As demonstrated in the paper [9], they used few of the machine leaning algorithms and it's showed the impact on predicting the students admission based on their application. Similarly, the work by the team [10] gave an explanation about student enrolment prediction based on the student and college characteristics [11] [12] [13].

Various other researchers have used other machine learning algorithms like SVM, Naive Bayes to make a better learning model [16] [17]. There are few areas by these researchers which were never uncovered. Our goal is to cover those areas such as K-Fold Cross Validation, number of features selected and also the number of algorithms and how they react to these set of folds and number of features.

The Educational Institutes collect data of the students, this is very large in number. The number of students who actually take admission or meet the institute's requirements or number of seats available for admission is almost less than a 1% of the data collected for that year. To help reduce this humongous data to a small scale that can be targeted with minimum human resource, developing a predictor system that can predict the student possibility of taking the admission with all the requirements met based on the data collected by the institution for the past few years.
The major objectives of the work proposed is to:

- Develop a tool to ease the assessment of marketing data based on admissions
- Develop an alternative tool / replacement for CRM
- Help college/university to plan their curriculum and infrastructure according to the number of students seeking admission.

III. METHODOLOGY

The figure 1 shows the abstract view of the overall work proposed in this paper. The methodology that was incorporated in this paper has 5 important stages of development.

1. Data Pre-processing Stage
2. Feature Engineering Stage
3. k-Fold Cross Validation Data Splitting (Training and Test Set) Stage
   - Fitting the Models with Training Data
   - Testing the Models with Test Data
   - Evaluation of all the Models based on Classifier Performance metrics.
4. Polling for the final prediction using the top 3 models

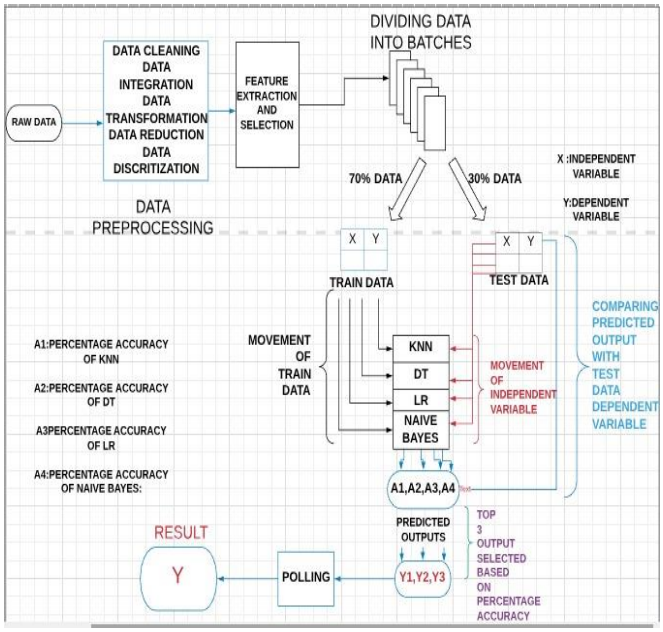

Fig-1: Proposed Framework

1. Data Pre-Processing:

It is a technique that transforms raw data into a machine-understandable format. Raw data may be incomplete. Hence, cannot be sent through a model. That would cause certain errors while prediction. That is the reason to pre-process data before passing it through the model.
The steps involved in data Pre-processing are:

- Read data
- Validating for missing values
- Validating for categorical data
- Standardize the data

2. Feature Engineering:

It is defined as creating new features or selecting features from the existing ones to improve model performance.
Like the name, address, etc. of the student data does not contribute to deciding whether the student will be taking the admission or not. Hence, these features are ignored while training the models.

3. k-Fold Cross Validation (Splitting Training and Test Set):

Training set is the subset of the data that is randomly chosen to train a model.

Test set is data set that is used to test a machine learning program after it has been trained on an initial training data set.

Steps:

- Shuffle the data set randomly
- Split the data set into K groups
- For each unique group:
  - Take the group as a hold out or test data set
  - The remaining groups as a training data set
  - Fit the models on the training set and evaluate it on the test set.
- Retain the evaluation scores and discard the models
- Use this scores to select the best model (Explained in Experimentation section).

Fitting the Models with Training Data:

The following is the overview of the basic machine learning algorithms that are used in this work:

- **K-Nearest-Neighbours (KNN):**

  The KNN algorithm is a robust and versatile classifier that falls into the family of supervised learning algorithms which is often used as a benchmark for more complex classifications.

  It is defined according to a distance metric between two data points. A popular choice is the Euclidean distance given by equation (1):

$$d(x, x') = \sqrt{(x_1 - x_1')^2 + (x_2 - x_2')^2 + \cdots + (x_n - x_n')^2}$$ ------- (1)

  More formally, given a positive integer K, an unseen observation x and a similarity metric d, KNN classifier performs the following two steps:

  - It runs through the whole dataset computing d between x and each training observation. let's call the K points in the training data that are closest to x the set A.
  - It then estimates the conditional probability for each class, that is, the fraction of points in A with that given class label. (Note I(x) is the indicator function which evaluates to 1 when the argument x is true and 0 otherwise)

$$P(y = j|X = x) = \frac{1}{K} \sum_{i \in A} I(y^i = j)$$ -------- (2)

  The equation (2) defines, that our input x gets assigned to the class with the largest probability.

- **Decision Tree (DT):**

  Decision Trees are a type of Supervised Machine Learning (that is you explain what the input is and what the corresponding output is in the training data) where the data is continuously **split** according to a certain parameter. The tree can be explained by two entities, namely decision nodes and leaves.

  Entropy controls how a Decision tree decides to split the data and impacts immensely on how a Decision Tree draws its boundaries, mathematically represented in equation (3):

$$Entropy = -\sum p(X) \log p(x)$$ -------- (3)

  The equation (4) defines the computation Gini score that gives an idea of how good a split is by how mixed the response classes are in the groups created by the split

.

$$G = \text{sum}\left(pk * (1 - pk)\right)$$ -------- (4)

  pk: proportion of same class inputs present in a particular group

- **Logistic Regression (LR):**

  Linear regression algorithms are used to predict/forecast values but logistic regression is used for classification tasks.

  Logistic regression algorithm also uses a linear equation with independent predictors to predict a value.

  Example of logistic regression equation is depicted in equation (5):

$$Y = \frac{e^{b0+b1*x}}{1+e^{b0+b1*x}}$$ -------- (5)

  The equation (5) defines, input values (x) are combined linearly using weights or coefficient values (referred to as the Greek capital letter Beta(b)) to predict an output value (y).

- **Naive Bayes:**

  Naive Bayes is a classification algorithm for binary (two-class) and multi-class classification problems. It's based on Bayes Theorem, whose formula is represented in equation (6):

$$P\left(\frac{A}{B}\right) = \frac{P\left(\frac{B}{A}\right) * P(A)}{P(B)}$$ -------- (6)

  The equation (6) defines, for two events, A and B, Bayes' theorem allows you to figure out P(A|B) (the probability that event A happened, given that test B was positive) from P(B|A) (the probability that test B happened, given that event A happened)

It is called naive Bayes because the calculation of the probabilities for each hypothesis are simplified to make their calculation tractable. Rather than attempting to calculate the values of each attribute value P(d1, d2, d3|h), they are assumed to be conditionally independent given the target value and calculated as P(d1|h) * P(d2|H) and so on.

Each of these models will be trained using the (70%) training set. Once the models are trained they are tested against the (30%) test set and accuracy is calculated.

Table-1: Evaluation of Classifiers				
	LR	**NB**	**DT**	**K-NN**
Accuracy	**91%**	84%	89%	87%

The table 1 shows the accuracies of the respective models LR (Logistic Regression), NB (Naïve Bayes), DT (Decision Tree) and KNN (K-Nearest Neighbour). It shows that the accuracy of Naïve Bayes is the least 84 and the accuracy of Logistic Regression is **91,** the highest.

5. Polling for the final prediction using the top 3 models:

Now based on the accuracy scores the top 3 models are selected and they are used to predict the output of an unknown data of student for which the model is not trained. Hence all 3 models will give their own predictions, among them the most common prediction is considered as the final predicted output.

IV. EXPERIMENTATION

DATA SET:

Universities collect data from various resources. Some of them are collected by making students fill forms and data from their junior college. Some of them are collected or bought from sources like coaching institutes and websites who collect thousands of students' data just for the purpose of selling it.

Considering all the above methods of collecting data, faker module of python is used to create a data set with enough attributes and large enough to fit for all the machine learning algorithms that will be used. This data set contains students' bio details, high school grades, entrance exam marks and also attributes that websites collect and consider it important. The data set has about 47,000 student details which are sufficient for this experimentation.

The table 2 depicts two sample of dataset consisting of one acceptance and one rejection set.

Data Pre-Processing:

- **Feature Selection:**
  It's really important to select the exact attributes that affect a student's admission. Each feature selected carries its own weight. The features selected are the students' marks, geotag entrance exam rank, websites related attributes.

Table-2: Sample Dataset		
	STUDENT 1	STUDENT 2
10 MARKS	80	91
12 MARKS	95	94
CET	0	7650
COMEDK	9323	0
WEBSITE VISITED	0	0
EMAIL CLICKED	1	0
ADMITTED	1	0

- **Feature Extraction:**
  It's also equally important to have a feature that is very dependent and extracted from a part of the data set (combination of few attributes). For example, by taking student's 10th marks and 12th marks for calculating its average and making it as one more dependent variable.

K- Fold Cross Validation:

Here, the process is jumbling the 70% of data (original sample), which is test data, partitioned into K folds. The folds are random and size of the folds are equal. This is done to improve the performance of the model. By considered these two, 5 and 10 as K for the number of folds.

Table-3: K=5 & N=All Features				
MODEL	**ACCURACY**	**PRECISION**	**RECALL**	**F-SCORE**
NB	57.33	0.666	0.57	0.58
LR	64.56	0.664	0.64	0.65
DT	**83.84**	0.839	0.83	0.83
K-NN	73.73	0.732	0.73	0.73

The table 3 shows the results (Accuracy, Precision, Recall, and F-Score) when considered **K = 5**, found that Decision Tree model had the highest accuracy and precision followed by the KNN model.

Table-4: K=10 & N=All Features				
MODEL	**ACCURACY**	**PRECISION**	**RECALL**	**F - SCORE**
NB	56.94	0.67	0.56	0.57
LR	67.81	0.66	0.67	0.67
DT	**82.46**	0.82	0.82	0.82
K-NN	73.55	0.72	0.73	0.73

The table 4 shows the results (Accuracy, Precision, Recall, and F-Score) when considered **K = 10**, found that accuracy and precision of the KNN increased compared to KNN model of **K = 5**, Decision tree model still had the highest accuracy and precision but decreased compared to when **K = 5**.

The total number of features which impact the models is 6 and the above results are when all the features were selected. To consider and validate which features have more impact on the model's results and also depending on the number of folds considered.

Table-5: K=5 & N=3 Features				
MODEL	ACCURACY	PRECISION	RECALL	F - SCORE
NB	73.44	0.75	0.734	0.74
LR	72.18	0.76	0.721	0.65
DT	**73.51**	0.73	0.735	0.73
K-NN	73.07	0.72	0.730	0.72

In the 6 features, when selecting the features as only three of them i.e. **N = 3** and setting the number of folds as 5 i.e. **K = 5.** The results for this is shown in the below table (5). This shows that for this combination of the features and the number of folds, the accuracy of Naïve Bayes and Logistic regression models have increased drastically and accuracy of Decision Tree model is very less compared to when all features were selected.

Table-6: K=5 & N=5 Features				
MODEL	ACCURACY	PRECISION	RECALL	F - SCORE
NB	56.81	0.65	0.56	0.57
LR	69.25	0.73	0.69	0.60
DT	**81.04**	0.81	0.81	0.81
K-NN	73.77	0.73	0.73	0.73

The table 6 shows the results (Accuracy, Precision, Recall, and F-Score) when selecting the features as five of them i.e. **N = 5** and setting the number of folds as 5 i.e. **K = 5.** The accuracy of the Naïve Bayes and Logistic regression models came back down, and accuracy of the Decision Tree model went back up to ~81%.

Table-7: K=10 & N=3 Features				
MODEL	ACCURACY	PRECISION	RECALL	F - SCORE
NB	**73.64**	0.75	0.736	0.74
LR	70.32	0.74	0.703	0.61
DT	73.31	0.73	0.733	0.73
K-NN	71.09	0.70	0.710	0.70

The table 7 shows the results (Accuracy, Precision, Recall, and F-Score) when selecting the features as three of them i.e. **N = 3** and setting the number of folds as 10 i.e. **K = 10.** The accuracy of all the models is similar to when number of folds were **K = 5**.

Table-8: K=10 & N=5 Features				
MODEL	ACCURACY	PRECISION	RECALL	F - SCORE
NB	57.03	0.67	0.57	0.58
LR	70.65	0.75	0.70	0.61
DT	**80.52**	0.80	0.80	0.80
K-NN	75.17	0.74	0.75	0.74

The table 8 shows the results (Accuracy, Precision, Recall, and F-Score) when selecting the features as three of them i.e. **N = 5** and setting the number of folds as 10 i.e. **K = 10.** The accuracy of all the models is similar to when number of folds was **K = 5**.

V. RESULT AND DISCUSSION

The main aim of this system is to classify new applications based on previous years' data of those students who got admitted or rejected in a particular university. Due to the number of increase in applications every year it is becoming human intensive task for university, which ends up in building this system. Given below are the steps for developing this system: After pre-processing the data, it is trained using different supervised classification models to classify applications into 'Accept' or 'Reject'. The different models used are Naïve Bayes, K-Nearest Neighbor, Decision Tree, and Logistic Regression. Each of these models was used to test the set of new applicants along with result to derive the accuracy.

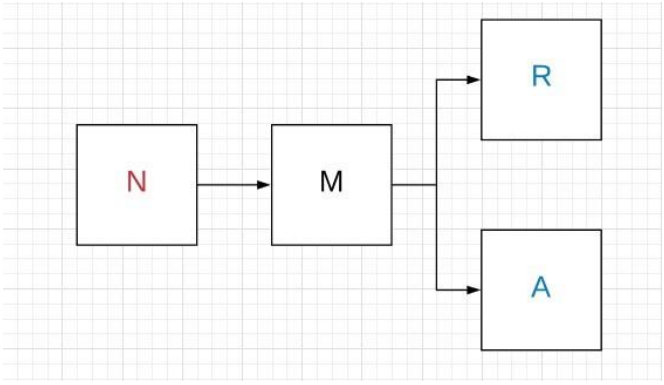

Fig-2: Student selection system

N: Represents the new applicants applying to the university
M: Different Models as mentioned
R: Class Reject
A: Class Accept

The highest error rate is in Naive Bayes of 35.06%. Followed by Logistic Regression of 12.56%. Naïve Bayes produced an error rate of 25.70 %. The best results is achieved by Decision Tree and K-Nearest Neighbor. Given below is the decision tree modelled as per our dataset.

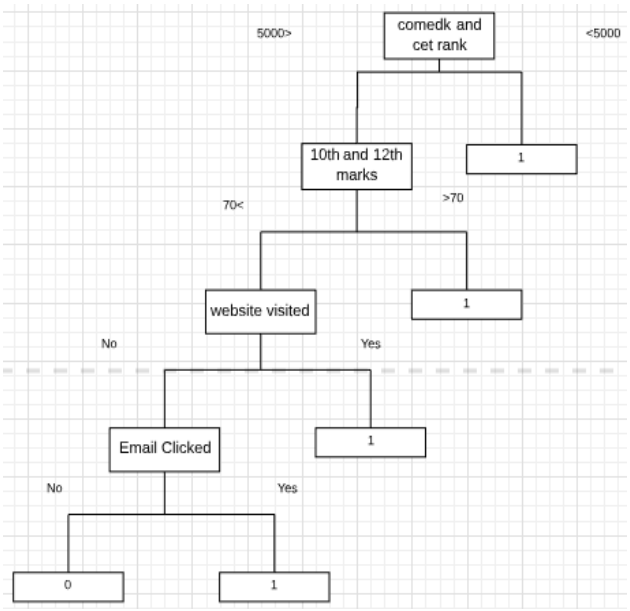

Fig-3: Decision tree

Here 1 represents 'Accept' and 0 represents 'Reject'.

VI. CONCLUSION AND FUTURE SCOPE

Data mining techniques and machine learning techniques have been contributed in various business related applications, boosting their business strategies with an improved decision making. Educational Institutions are also the owners of enormous of data, and getting the clear insight about the hidden patterns in the data would certainly improvise the overall educational system. Admission Prediction model for accomplishing the number of the student seeking admission to college/university has been presented. The prediction incorporated was purely based on student data, and such an approach towards data analysis would help in reducing congestion and improve quality of education across the nation. The predictive models built are not only automating the student admission prediction, but, it also categorizes a student with respect to their performance to choose the best suitable discipline of study. The work proposed in this paper can be helpful for any Educational System owning data. As the data stored grows exceptionally by any organization periodically, the educational system is not exceptional. Therefore, incorporation of distributed computing frameworks for such massive data, with enhanced security remains as the future work.

REFERENCES

1. IIE Report http://www.iie.org/Services/Project-Atlas/United-States/International-Students-In-US
2. Avrim L. Bluma, Pat Langley Selection of relevant features and examples in machine Learning.
3. Austin Waters and Risto Miikkulainen, "*GRADE, a statistical machine learning system developed to support the work of the graduate admissions committee at the University of Texas at Austin Department of Computer Science (UTCS).*"
4. Scikit-learn documentation https://scikit-learn.org/stable/documentation.html
5. Dineshkumar B Vaghela, Priyanka Sharma, "*Students' Admission Prediction using GRBST with Distributed Data Mining*" June 2015.
6. Elizabeth Murray, "*Using Decision Trees to Understand Student Data*".
7. Surjeet Kumar Yadav, Saurabh pal, "*Data Mining Application in Enrollment Management: A Case Study*", 5, March 2012.
8. Miren Tanna, "Decision Support System for Admission in Engineering Colleges based on Entrance Exam Marks", 11, August 2012 .
9. Jay Bibodi, Aasihwary Vadodaria, Anand Rawat, Jaidipkumar Patel, "*Admission Prediction System Using Machine Learning*".
10. Ahmad Slim, Don Hush, Tushar Ojah, Terry Babbitt,"*Predicting Student Enrollment Based On Student College* ".
11. Enrollment management in higher education - defining enrollment management, key offices and tasks in enrollment management, organizational models. Education Encyclopedia - StateUniversity.com, 2013.
12. Guyon and A. Elisseeff. An introduction to variable and feature selection. J. Mach. Learn. Res., 3:1157–1182, Mar. 2003.
13. D. Hossler and Bean, John P. The strategic management of college enrolments. San Francisco, Calif. : Jossey-Bass, 1st edition, 1990. Includes bibliographical references (p. 303-318) and index.
14. H.-A. Park. An introduction to logistic regression: From basic concepts to interpretation with particular attention to nursing domain. Korean Society of Nursing Science, 43(2), 2013.
15. Priyanga Chandrashekar, Kai Qian, Hossain Shahriar, Prabir Bhattacharya. "*Improving the accuracy of decision tree mining with data pre-processing*".
16. Python 2.7 Documentation-docs.python.org/2.7
17. Tinda Yang, Kai Qain, Dan-Chia-TienLo, Lixin Tao, "*Improve the prediction accuracy of naive bayes classifier using association rule mining*", 2016.

AUTHORS PROFILE

Anil B, Student, School of Computing and Information Technology, REVA University, Bengaluru, India, currently pursuing Bachelor of Technology from Reva University Bengaluru, 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.

Aman, Student, School of Computing and Information Technology, REVA University, Bengaluru, India, currently pursuing Bachelor of Technology from Reva University Bengaluru, India.

Aman Kumar Singh, Student, School of Computing and Information Technology, REVA University, Bengaluru, India, currently pursuing Bachelor of Technology from Reva University Bengaluru, India

Aditya Kumar Singh, Student, School of Computing and Information Technology, REVA University, Bengaluru, India, currently pursuing Bachelor of Technology from Reva University Bengaluru, India.