PREDICTION OF COMPETENT STUDENTS HIGHER STUDIES AND BRIDGING THEM WITH FOUNDATIONS USING BIG DATA

Mrs. N.Kalyani, Mrs. V.Revathy

[1]Assistant Professor, [2]Assistant Professor

[1]Department of Computer Science and Engineering, [2]Department of Computer Science and Engineering,

[1]Star Lion College of Engineering and Technology,[2]Arasu Engineering College,

Thanjavur, Tamilnadu.

***Abstract* - Education is an important factor for developing the life of human and economy. In recent years many students drop out from their higher studies in India due to their financial problems. As per 2019 record in Tamil Nadu around 6 lakh people appeared for exam and the pass percentage was 91%. Engineering Colleges have filled with only 53% seats. This is merely due to financial problems and their personal issues. To assist them for their financial problems a lot of educational trust and organization are willing to help them. But there is no proper way to communicate them though the individual charitable trust and foundation have their own system to track but still they are not reachable to the rural area. The main objective of our project is to improve the education status of our nation by predicting the competent students and support them financially for their higher studies. In this paper data mining and machine learning are used with Big Data to predict the capability of students, who are competent to do higher studies and the predicted student list will be bridged with Charity foundation and individual sponsors. Machine Learning algorithm is used to predict the capability of the student. Big Data is used to improve the accuracy of the Prediction. Data mining is used to discover the relationship in the student data set and to analyze the required information for prediction. Finally the accuracy is perceived by using the logistic regression methodology to predict the values of students.**

Keywords: Prediction, Big data, Machine learning Algorithm, Data Mining.

I. INTRODUCTION

Generally, mining means the extraction of some valuable things from the earth. In computer science, data mining refers to taking out important information from a large amount of data. Data Mining is otherwise known as Knowledge Discovery or Knowledge Extraction [1]. Abhijit et.al (2007) emphasise the mining technology and methodology for data. Nowadays, data mining is used in almost every place where a huge amount of data is stored. Mostly information collected from data mining is used to predict future trends and it allows taking business decision. Financial Analysis, Biological Analysis, Scientific Analysis, Intrusion Detection, Fraud Detection, and Research Analysis these are some application where data mining is widely used

Scientific analysis

Our society is to collect huge amounts of scientific data that need to be analysed. Unfortunately, we can capture and store more new data faster than we can analyse the old data already accumulated.

Financial analysis

In many sector like banking there are huge amount of money could be transacted these could be done by data mining [2] Agathe et. al addresses these bank issues of data mining in their work.

Fraud detection

Data mining is used to identify the different types of fraud .here fraud could be classified into four types fraud, customer fraud, network fraud, computer based fraud [3] Parkavi et.al (2019) includes the fraud detection technology to provide security for the organization.

Machine learning is defined as a field of study that gives computers the capability to learn without being explicitly programmed. Machine-learning can learn itself without using any explicit program. Machine Learning working based on its previous experience and the process begins with feeding good status data and then training our machines by using machine learning models and different algorithms.

Financial service

This technology is used to identify the investment on trade. This system prevents the institution and organization from financial risk. Financial data prevented from financial fraud using machine learning techniques. Modelling and help their customer to make smart decisions

Virtual personal assistant

In personal assistant machine learning plays major role. It working based on previous involvement virtual assistants integrated into various platforms like Smart Speakers, Smartphone and Mobile Apps

Online fraud detection

Machine learning provides more secured online business for example PayPal using ml for protect their money from laundering. That company use set of tools for protect their money

Logistic regression

Here we use a logistic regression algorithm for prediction. Logistic regression is a supervised model this model maps input output like example pairs.

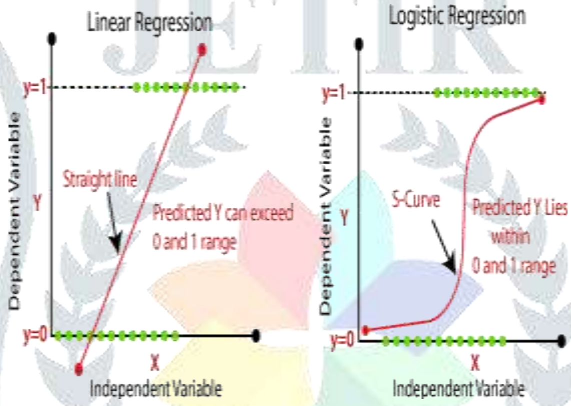


Fig 1 comparison of linear regression with logistic regression

It shows results in zero and one. One representing success and zero representing the failure. The dependent variable should be binary, only the meaningful variable should be used it is a supervised learning it means an input where the desired output already know (input output pair) Logistic Regression is a Machine Learning algorithm [9] Lixictao et.al (2019) includes the detailed study of logistic regression in their work, which is used for the classification problems, it is a predictive analysis algorithm and based on the concept of probability. The hypothesis of logistic regression tends it to limit the cost between 0 and 1

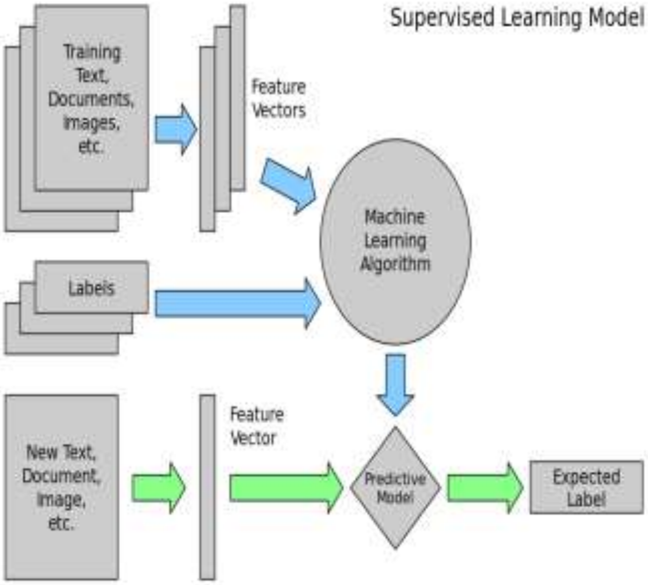

Supervised Learning Model
Training Text, Documents, Images, etc.
Feature Vectors
Machine Learning Algorithm
Labels
New Text, Document, Image, etc.
Feature Vector
Predictive Model
Expected Label

Fig 2 working methodology of machine learning

Imbalanced data set are used for classification. Imbalanced data set means the attribute which is used in our daily life like salary of each employee in an organization

$$\log\left[\frac{y}{1-y}\right] = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$$

The remaining part of this paper explains about related work, Proposed System, Experimental review, Conclusion and Future enhancement.

II. RELATED WORKS

The successful approaches for predicting the student competency and sponsor availability is based on the Logistic regression and Scikit techniques. To collect the details of student and sponsor, use pandas framework which is software library for python language, it manipulates and analyse the data. These data are transferred to pre-processing, where the data could be cleaned and logistic Regression takes place. The algorithm implements that results are entering into next phase called training. In this training phase scikit machine learning library train the data that trained data could be send to tensor flow is free open source library for machine learning specially used for train the data then finally the ever trained data are bridged that report will be integrated into cloud

Classification [3] is one of the most considerable supervised learning data mining technique used to classify predefined data sets the classification is mainly used in healthcare sectors for making decisions, diagnosis system and giving better treatment to the patients. Thyroid disease have also been detected worldwide and thus become a serious endocrine health problem and an issue of concern. k nearest neighbours, Support vector machine, Decision tree and Naïve Bayes are mainly used to classify the data. The Decision Tree can obtained highest accuracy of 98.89% over other classification techniques.

Classification is a classical problem in machine learning. In machine learning, the algorithm that can complete the classification task is usually called a classifier. Imbalanced datasets [9] typically refer to problems with classification problems where the distributions of classes are not equal. Imbalanced datasets are common in our daily life, such as spam detection, credit card fraud, and natural disaster detection.

Generally, there are two methods to deal with imbalanced dataset problems. One concentrates on the classifier, and another focuses on data. The operation of the logistic regression algorithm in order to generate the logistic regression classifier more powerful in classifying the minority class the analysis was implemented with Python program. If when the logistic regression model predicts an E-mail, the probability of spam is 0.7. The default threshold is 0.5. The threshold value 0.7 is greater than 0.5, it means that the model considers this email as spam

The growing trust on social network sites [7] causes people to generate massive data, also called big data. It is typically characterized by three computational issues namely volume, velocity and variety. Data mining provides a wide range of techniques for detecting useful knowledge from massive datasets like trends, patterns and rules. They discuss the different issues with social networks, different approaches, issues, current challenges and trends. They attempts to provide an overview of all the basics of social media analytics, possible approaches, scope of mining, issues and different challenges. Social media has various file formats like text, video, photographs, audio, PDF and PowerPoint.

So it's difficult to analysis the fraud news. Face book, are increasingly utilized by many people. These networks allow users to publish details about themselves and to connect to their friends.

There are different approaches for mining social media Graphical Approach Community Detection, Recommender System, Opinion Mining and Sentiment Mining.

A panda [12] is a software library for python program. It is machine learning framework which is used for manipulate the data like join, merge, Concatenation and data cleaning also. It import various file format like csv, excel etc. it is tool for reading or writing the data structure and its memory

Scikit-learn

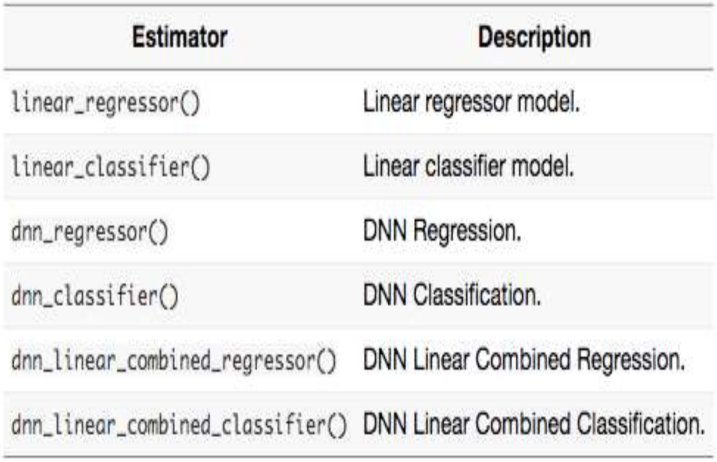
It is a machine learning library for python. [12] Geron et.al (2017) address the scikit learning model contains many useful tool machine learning including classification clustering and regression in their work.

Tensor flow framework

Tensor flow [12] is a data flow library it is a machine learning library also a maths library is detailed explained in their work. This flow is available on every OS and mobile platform like android and ios. This library calls the C++ construct and execute flow graph. This framework is build to deploy at scale and it debugs the program and run on both CPU and GPU

These are the algorithm very useful in Tensor flow

Table 1 Tensor flow algorithms



Estimator	Description
linear_regressor()	Linear regressor model.
linear_classifier()	Linear classifier model.
dnn_regressor()	DNN Regression.
dnn_classifier()	DNN Classification.
dnn_linear_combined_regressor()	DNN Linear Combined Regression.
dnn_linear_combined_classifier()	DNN Linear Combined Classification.

III. PROPOSED SYSTEM

This project includes modules namely Data classification module, Database module, Sponsor module, Student module, Logistic regression module, Bridging module, integration module. System architecture contains data set which is a combination of student database, sponsor database. Student data base can hold the collection of student's basic information about their education and their basic details related to integrate for higher education. Sponsor database is a collection of sponsor's details which can be preserved so confidentially for providing security for their datas which can hold the sponsor information. Pre-processing, Training, create report and integrate that report into cloud are the methodologies used in this proposed work.

Pre-processing is classified into three main stages, which is Data frame work, Data cleaning, deriving features for logistic regression. Data framework has been performed in the collection of student's data and need to check the student competency according to their data. If the student is competent to study medical that information is updated to the schema otherwise it checks other condition if they have enough competent to join engineering that information is send to schema otherwise it checks another condition whether that student able to join for any other course. Updated schema has a detailed list of competent students with their eligible for education. The schema is integrated into cloud and it is maintain by separate manager.

Data classification module

Classification of data in mining plays a vital role. In this module, segregation of the data which are needed for this project is collected from the district educational office. The data is classified by using the prescribed algorithm using the data format.

Database module

This database module working methodology includes two schemas. One is student database and another one is sponsor database. At student database we collect data from DEO office; those data are segregate from classification module. On sponsor module we collect the sponsor data that are willing to help.

Classified data are collected and send that data to preprocessing stage. At preprocessing stage data are cleaned and the features are derived. Logistic regression algorithm is working under preprocessing stage.tho find the competency of student. That competent student data set can be send to training phase. After completing the training phase testing phase held.

Sponsor module

In this module, we classified the different type of sponsors like who can help a student's complete higher studies or sponsors can able to help them to a limited amount. The identification of the sponsors can be made by the information collected. The details of the sponsors can be maintained confidentially for their security purposes.

Student module

In the student module, we classify the student details; analyze them who are capable to do higher studies and selecting the courses based on their major subject on higher secondary. Student list can be check whether the candidate is able to join doctor or engineering otherwise that student able to join any arts college.

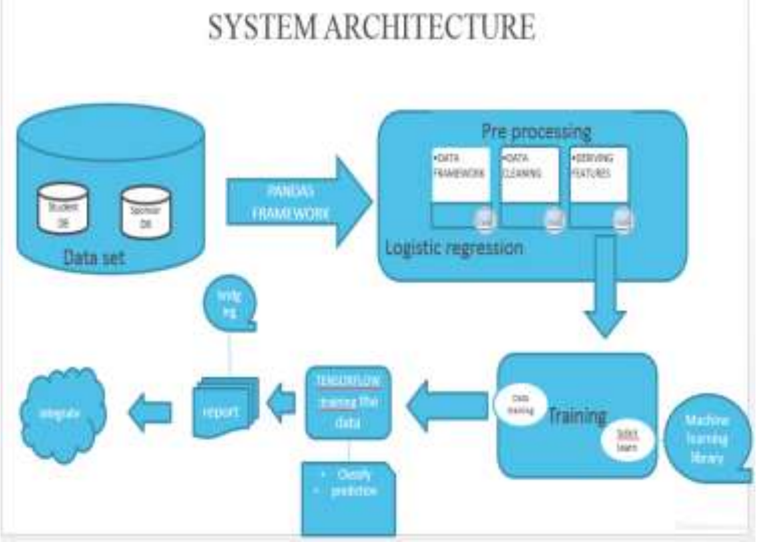


Fig 3 system architecture

These information is identified from there higher secondary course and that student mark history which is maintained in the student database module. Segregate that information as separate list to map them to the appropriate sponsors.

Logistic regression module

A machine learning algorithm called logistic regression which is used for a large amount of data. That cleaned data list can be process going under logistic regression. Logistic regression is a in its basic form uses to model although many more complex exist. In logistic regression are the parameters of a logistic model.

In this work, we pass the weighted sum of inputs through an activation function that can map values in between 0 and 1. Such activation function is known as sigmoid function and the curve obtained is called as sigmoid curve or S-curve.

Gini index is a measure of impurity or purity used while creating a decision tree in the CART (Classification and Regression Tree) algorithm. An attribute with the low Gini index should be preferred as compared to the high Gini index. It only creates binary splits, and the CART algorithm uses the Gini index to create binary splits. Gini index can be calculated using the below formula:

$$\text{Gini Index} = 1- \sum_j P_j^2$$

Bridging module

Bridging the details of students and sponsors based on the output of the Logistic regression module. The final report is called as bridging module. The final output shows number of student need money for their studies

Integration module

Integration module proceeds that the student and the sponsor details to be integrated into cloud. We use Amazon sage maker it is specialized for machine learning cloud developed by Amazon. The integrated information gives the required data whenever the user requests.

IV. EXPRIMENTAL REVIEW

Data mining and machine learning are used with Big Data to predict the capability of students, who are competent to do higher studies. The predicted student list will be bridged with Charity foundation and individual sponsors. We are analyzing the data from educational office (virtual data) to predict the competent student who are financially down.

This prediction model is working under the logistic regression algorithm. The large amount of data collected from DEO office and that data can be processing. We mine the necessary data from complete student details. We perform the data framework creation and cleaning of data which is removing the unnecessary features from the data and deriving the required data for the project. And it checks the condition if student able to join medical or engineering or any other course. Identifying the competent students and bridging them with charitable trusts and foundation. This system helps to find out the student who not able to study their higher studies and helping them financially. By using this project we improve the educational status. It predict more accurate because of using machine learning algorithm

A. Competitive student and Sponsor for Training data

Training data is taken to train the dataset with the variables for predicting the accurate values. The values should be considered as train data for training the remaining dataset. The 1/3 of the data is taken as training data and the remaining is taken to be the test data.

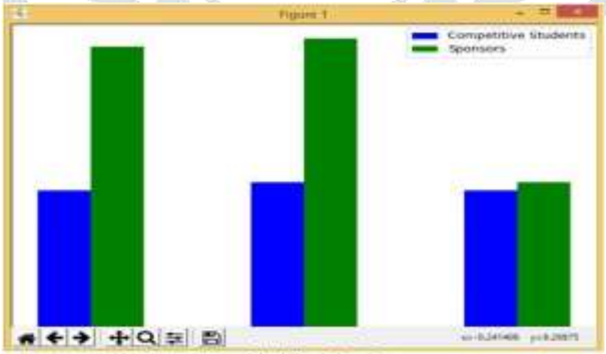


Fig 4 Training data

B. Competitive Student and Sponsor for Test Data

Test data has been given as input for the tool and the parameters need to check has been included in the logistic regression algorithm. The test data has been drawn as a graph to visualize the competitive students and sponsors that we have taken to test. Test data is splitted as arts, science and medicine according to the student competency.

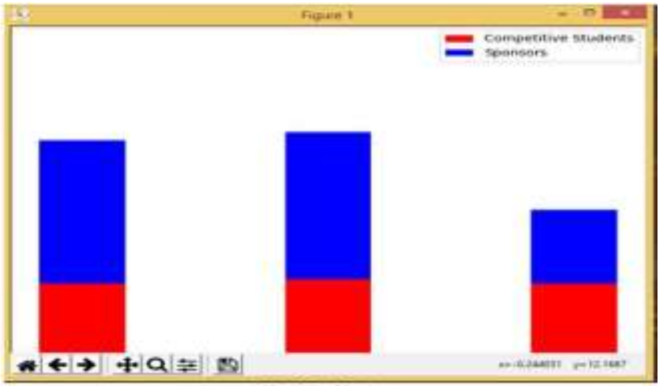


Fig 5 Test Data

C. Bridging

It is enlarging the boundaries towards the bridging of students with the sponsor details. The sponsor and their details have been maintained in the sponsor database and the student's information is maintained in the student's database and the two will be bridged according to that neediness of the student.

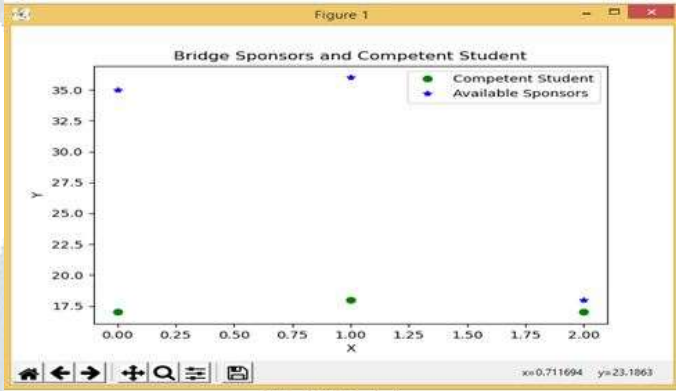


Fig 6 Bridging

D. Sponsors for Training Data

The sponsor details will be present in the sponsor database and it is utilized by the student database. The 1/3 of the information is taken as training data for training the database to produce the accurate results. And the testing sponsor data is bridged with testing student data with the splitted up dataset.

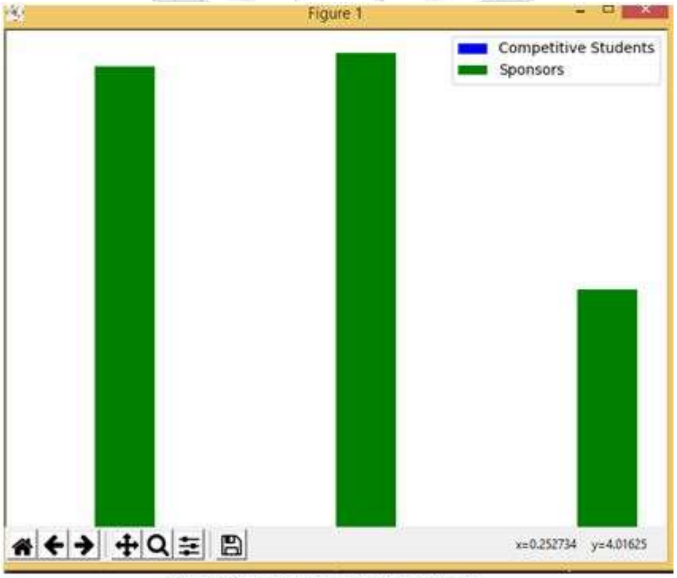


Fig 7 Sponsors Training Data

4.1 PERFORMANCE ANALYSIS

To evaluate the performance of the experiment we took the attributes as Accuracy, Recall, Precision, and F-Score to analyze the efficiency and prediction performance.

Table 2 Performance analysis of logistic and linear regression

Attribute	Logistic	Linear
Accuracy	0.88	0.7
Recall	0.7	0.602
Precision	0.826	0.712
F-score	0.739	0.612

Here, we compare the data set with previous algorithm linear regression with logistic regression algorithm. While comparing these two algorithm logistic algorithm [6] provide more accuracy than linear regression not only accuracy we get more F-score also. So we conclude that logistic is more suitable for this prediction model.

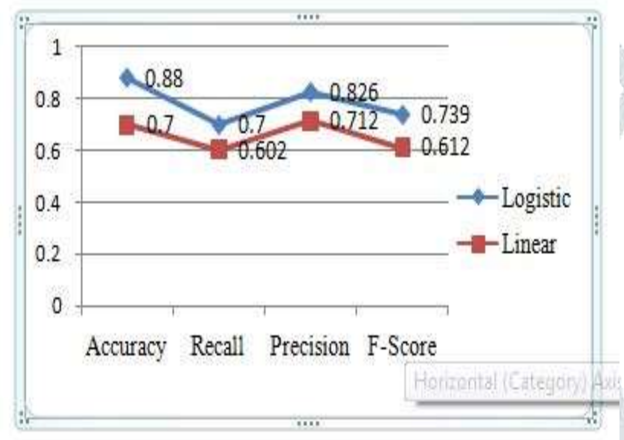


Fig 8 performance analysis

The logistic regression has the increasing accuracy and precision rate compared to the linear regression techniques like the values ranging from 0.88 in accuracy and 0.7 as recall value and 0.826 as Precision and 0.739 as F-Score. It has been taken for a large dataset and the values have been experimented.

V. CONCLUSION

This paper has presented the students who are all need sponsorship for their higher studies is predicted previously and sent them details to appropriate sponsor. The main objective of our project is everyone has the education, we successfully attain our goal. When this system is implemented as government project, using this project government identify previously how many doctors, engineers, and other professional peoples are passed out a certain year. So unemployment may reduce the large level. Due to this project, many jobs may develop like a database manager and cloud manager. And we conclude that logistic regression provides more accuracy.

VI. ACKNOWLEDGEMENT

I am thankful to the Almighty for providing me good health and environment to do this work. I thank Star Lion College of Engineering and Technology and Arasu Engineering College, Management, Authorities and all for their support. I thank Mr.M.Kannan M.E., (Ph.D), HOD/CSE , AEC for their support and providing the needful. I am also thankful to my colleagues for their constructive comments and suggestions. I thank my family members and Friends for their encouragement to do this wonderful work.

REFERENCES

[1] Abhijit A.Sawant and P.M.Chawan, ”Study of data mining technique for Financial Data Analysis”, International Journal of Engineering Science and Innovative Technology (IJESIT)

[2] Agathe MerceronBeuth “Learning Analytics: From Big Data to Meaningful Data Journal of Learning Analytics”, 2015.

[3] BibiAminaBegum, Parkavi.A, “Prediction of thyroid Disease Using Data Mining techniques” 5th International Conference on Advanced Computing & Communication Systems (ICACCS) 2019.

[4] Lakshmi B.N., Raghunandhan G.H., “A Conceptual Overview of Data Mining” Proceedings of the National Conference on Innovations in Emerging Technology, 2011.

[5] Dhwaani Parikh and Vineet Menon, “ Machine Learning Applied to Cervical Cancer Data”, I.J. Mathematical Sciences and Computing, 2019, (PP 53-64)

[6] George Siemens, Ryan S, Baker J.d, “Learning Analytics and Educational Data Mining” Communication and Collaboration

[7] Jyoti “More Issues in Mining Techniques in Social Media”,2017, IJSRCSEIT

[8] Keisuke Abe, “Data Mining and Machine Learning Applications for Educational Big Data in the university”,2019

[9] Lixictao, Pabhairbattachraya and yingquai “Improving Prediction Accuracy for Logistic Regression On Imbalanced Datasets”,2019

[10] Shlok Gilda, Pune Institute of Computer Technology, Pune, India, Evaluating “Machine Learning Algorithms for Fake News Detection”, 2017.

[11] Agarwal R, Srikant R,”Fast algorithms for mining association rules”, Proceedings of the 20th Int.Conf. Very Large Data Bases, VLDB, (PP 487-499), Morgan Kufmann, 1994.

[12] Geron A, Hands-on Machine Learning with Scikit-Learn and Tensor Flow, O’Reilly Media, Inc., 2017

[13] Abe k, ”Education support methods analyzing the school affairs data of the students in the university” in Proceedings of the 45th conference on Intelligent Systems Symposium of The Society of Instrument and Control Engineers,(PP. 1-6) 2018 (in Japanese)