IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

FACTORS INFLUENCING STUDENTS DROPOUT MACHINE LEARNING APPROACH

V. Lavanya , Dr.K. Santhi Sree

M.Tech Student, Professor of Computer Science and engineering
Computer Science and Engineering
School of Information Technology
Jawaharlal Nehru Technological University Hyderabad
Hyderabad, India

Abstract: Over the years the number of students dropping out of school is growing rapidly. High enrolment rates have become a major threat to many educational institutions or universities. The student enters the institution with many dreams and expectations. When expectations do not meet or certain factors such as demographics will begin to degrade them from their registered system. It is a major threat to all educational institutions. Various process of size reduction, including feature selection and feature removal. Feature selection is a step-by-step process used to select an appropriate attribute from a given attribute set. In the process of feature extraction, it involves the conversion of high-density data with low corresponding size. Feature selection includes things like academics, personal characteristics, psychological factors, health issues, teacher perspective, student behaviour.

The project we introduce a student dropout prediction method by using K Nearest Neighbour, Random Forest, Naive-Bayes, Decision Tree in Python language. Student-related information is collected from a variety of sources as large data is required to predict things. The data collected contains detailed information including basic information, parental educational background, family-student relationships, community performance and student learning. Machine learning strategies are applied to selected features that target students and our main task is to apply different techniques to selected features after which to apply different metrics to each algorithm. Comparisons are made for machine learning algorithms for each metric and the model that provides the best results is considered the best. There are many factors that affect a student's ability to do school, as mentioned above. Early prediction of resignation helps an organization to keep students from the right curriculum.

Keywords: Dropout, Identification, Factors, Machine Learning Algorithms, Prediction.

I. INTRODUCTION

The term dropout indicates the termination of a student from school, college or any other educational institute without fulfilling the registered course. Dropout means that "Any student who leaves from school or any other educational institution for any reason before completion of a registered program of studies without transferring to another elementary or institution". Each year, more than a million students will leave the school or any other educational institution without full filling the course in various universities or schools that are around 8,000 students every academic year. Dropout of the student is mainly influenced by personal factors as well as by the institution also. My project is useful for University-based or School-based to examine the student behaviour, that leads to a dropout or not in the early stage and can prevent the dropout from taking necessary action towards the dropout reason. The existing method is very time consuming and not very accurate and focuses on only specific factors. The proposed method is a combined approach that takes into consideration factors such as demographics, academic performance, health issues, place of residence etc. which increases the accuracy and implements methods that reduce the time taken for prediction.

Machine learning is the technique of building models from the data. Machine learning is a field of computer science that is pertained to building algorithms that will be useful for collecting some chunk of data based on a certain phenomenon. The data collected may come from nature, generated by humans or it may be generated by a few of the algorithms. Machine learning scrutinizes the study and disposition of algorithms that may learn against the data and frame assumptions on the available data. Learning is an illustrative phenomenon for all living beings. Humans have made an idea that why can't non-living things should involve in learning, this proposal gave a path to the introduction of machines. The invention of this ideology enabling the machine that can learn like humans have come true. Naturally, machines are not intelligent but humans are making it possible.

Machine learning offers an efficient alternative to the conventional engineering flow when development cost and time are the fundamental concerns when the problem appears to be too complex to study in its full generality. On the other side, this approach has the considerable disadvantage of providing generally sub-optional performance interpretability to the solution. Computers follow certain programming instructions which help to figure out positive complications and to achieve the required task. So different algorithms are used by different one to perform the same task. The implemented algorithms use a sequence of instructions that are going to execute from one state to another. The efficiency of the program can be measured with a quickness. Machine learning techniques will automatically detect the change by the users and starts to flag them without manually telling them to do so. We use machine learning to solve problems that are very complex.

In order to identify the machine learning tasks we will be using certain criteria:

- We should take a function that maps input to outputs.
- A large amount of data should be created.
- The task should be able to clearly define goals and metrics.
- The decision made should be able to clearly explain how we have made the decision.
- The function learned must not change constantly.
- No specialized dexterity, physical skills, or mobility is required.

1.1SCOPE OF THE PROJECT:

- To study the nature of students who join the school and their educational background which is helpful for the institution to improve their knowledge.
- To Build a model that can be used to predict whether the student is likely to drop out or not.
- To apply different machine learning algorithms to the dataset.
- To propose the best classifier that predicts the student dropout.

II. IMPLEMENTATION PROCEDURE

In the proposed methodology we are using extensive and descriptive viewpoints and include a wide scope of approaches which are officially recognized as possessing certain qualifications or meeting certain standards. It also involves practical assessment to construct validity including structural aspects.

2.1 Data Acquisition

Data used for predicting the students dropout rate is being collected from various online sources. The implementation of the system is done by gathering the information about students from two different schools of Uttar Pradesh. We have gathered the data regarding students' information, their academic performance and their relationship with their family members which mainly influence the students in order to continue their higher education. It consists of the details of the 10th studying students of two different schools.

2.2 Data Cleanup

The data collected may contain some missing values, duplicate rows and some unrelated data. The errors can be handled by:

- Deleting the duplicate rows from the dataset.
- If a row exists with fewer features having null or wrong values then the mean of the values is taken and it is filled in those columns.

2.3 Data Pre-processing

The process is to identify the required independent variables for predicting the dropout of students and to predict the binary dependent variable 'PASSED' using the independent variables in data pre-processing. The dataset here we are using for predicting students dropout contains information about their academic performance and their relationship with their family members which mainly influence the students in order to continue their higher education.

To predict the dropout of students, the dataset is split into training and testing. At any point of time splitting has 80% training rate and 20% testing rate. The predicted value will be 1 if the student is likely to drop out of school for higher education and 0 if continued.

2.4 Feature Selection

It is the way of determining the inappropriate features in our dataset with which the accuracy of the model can be achieved. If we are unable to identify the inappropriate feature then it may affect the execution of the model. In order to make machine learning algorithms work properly, we need to select the appropriate features. The intention of feature selection includes: Models can be simplified so that it becomes uncomplicated for the researchers to illustrate. Time taken to train the models should be very less. Bypassing the curse of measure the magnitude of the data is decreased.

Methods used for feature selection are:

Filter method: In this method statistical measure is used for selective features. It takes very less imputation time. By selecting a specific metric with which we can identify the irrelevant attributes and refine the inappropriate attributes from the redundant columns referring to the model. Based upon the feature score each and every column are rated.

Wrapper method: To evaluate a combination belonging to features and elect model performance scores a predictive model is used. In the wrapper approach, desperate solutions are made in a form to select the features on the search problem. The working of the model lays on the result of the classifier.

Embedded method: Here learning algorithms are integrated as a segment of the feature collection. A decision tree is the utmost typical embedded technique. In the Decision tree algorithm, each and every step is a firm of features that are divided into smaller subsets.

In our dataset, there will be many features for a particular student. All of these may not be needed to predict the correct result, so we are going for the feature selection. Here we are going to analyze each feature and only the features which are mostly affecting the result are going to be considered.

III. ALGORITHMS USED

Model deployment infers to the provision and interaction of software components in the system to achieve a predefined intension. Algorithms show us a way to interact, manipulate and transfer data to the system. Algorithms are programs that computers can understand. Machine learning algorithms can be written in Java, Python, R. Each of these languages provide a wide variety of libraries with which the algorithms can be operated. With these languages, we can make the system interactive and can find solutions to many problems. Machine learning algorithms are different as we need to give input and output then it will automatically build a model. The generated models are self-explanatory as the data is increasing the algorithm becomes more complicated which increases the accuracy of the model with a large volume of data. The algorithms we are using for this task are K Nearest neighbour, Decision tree, Random forest, Naive Baye's.

3.1 K-Nearest Neighbour

K-Nearest Neighbour is one of the supervised machine learning models which is easy to implement. To quantify the conditional distribution with the greatest estimated probability of Y given X we can use KNN. 'K' is the positive integer given on the test x_0, KNN will identify the 'k' points which are near to the training data x0, represented by N0. Which estimates the conditional probability for class jn as a fraction of points in N_0 whose response value is equal to j.

$$P_r(\boldsymbol{Y = j/X = x_0}) = \boldsymbol{1K \sum i \in N_0 I(Y_i = j)}$$

We will be implementing the k-nearest neighbour algorithm by using knn(). KNN algorithm follows two main steps for predictions they are first thing is to fit the model and the next is to make predictions on the model.

There are four inputs for the functions

1) For all the test data we need to make predictions and are represented ad "test".
2) For all training data, its predictors are represented in the matrix and are sent as "train".
3) For all the training observations the vector for class labels is represented as train.
4) For a value of 'k's the number of nearest neighbours generated are used by the classifier.

3.2 Decision Tree

A decision tree is one of the flexible methods for approximating distinct functions which are represented in tree format. Among the inductive inference algorithm, decision trees are the most successfully used one. Some of the characteristics of the decision tree are:

As the attributes in the dataset are fixed the best attribute to select as a root is the one that contains a small number of possible values. As target function mostly has Boolean values as output decision tree methods can easily learn functions with two possible output values. Naturally, decision trees generally represent disjunctive interpretation, If there are any errors in the data decision trees are strong enough to handle those. If the training data is containing any missing data then it can estimate the missing value based on the other sample nodes and the best decision is going to be taken by the algorithm.

ID3 algorithm: ID3 is the basic decision tree algorithm used for constructing a tree in a top-down approach. Among the given fixed attributes ID3 algorithm selects the perfect attribute as the root of the tree and a descendent of the node is generated for each possible value of the attribute all the values in the column are sorted in descending order. In order to select the best attribute as the root node some measure should be applied here we are using information gain works well on the given attributes for the target classification.

3.3 Random Forest

In order to make predictions in the random forest model, we can train a group of decision tree classifiers on each and every subset of the training data. We need to consider all the models and from that models, we should predict the class that gets more points. Such samples of different trees combined to generate a model are called random forests which is a more powerful learning algorithm. If the predictions are independent of one another then this ensemble method works well. To improve the accuracy of the ensemble model we should train the data by using various algorithms.

Advantages:

- Overfitting of the data is reduced with which the accuracy of the model is increasing.
- Classification and regression problems can be solved.
- Produces accurate results with categorical and continuous values.
- If any missing values are present in the data it will automatically fill with the mean of the data.
- Use a rule-based approach for normalizing the data.

Disadvantages:

- As numerous models are built high computational power is required.
- For training the trees a lot of time is required.
- We can't determine the significance of individual variables.

3.4 Bernoulli Naive Bayes

For calculating the clear probability for the hypothesis we are going for Bayesian learning. These algorithms performance is high when we compare with other algorithms and they provide us with a clear way that we can understand the solution to the problem and we can accurately find the solutions to any kind of problem by changing the probabilities. For each and every training example in the dataset, we can increase or decrease the probability that is estimated by a hypothesis that will be valid with which flexibility in learning the algorithm is achieved with each single training data. Previous knowledge about the data will helps us to resolve the probability. With Bayesian methods, probabilistic predictions can be made. If any new training data has arrived it can be classified by linking multiple predictions. Difficulties that encounter with bayesian learning are we should have a keen knowledge about the different probabilities ie we should have some background knowledge about the data. The cost of the optimal hypothesis is needed in advance. With this, we can calculate the probabilities from the given set of data points.

$$P^{(H/D)} = \frac{p(H)\,p(D/H)}{p(D)}$$

where P(h/D) is the posterior probability. 'h' is the initial probability on the training data. P(D/h) is the probability by observing the training data. The probability P(h/D) decreases as P(D) increases, as 'D' is independent of h. P(h/D) increases with P(h) as it is dependent on P(h). We are using this Bayes theorem to make a connection of the machine learning problems by taking sample training data to predict the target function by referring to the hypothetical space.

IV. DATABASE DESIGN

4.1 My SQL

My SQL is a structured query language that is mainly used for accessing and managing the records in the database. The database is used to store the collection of records. SQL is used to create, modify and extract the data from the database. We can also restrict the user in order to access the database. The database is used to organize the data into tables and to establish a relationship with all the data points. To implement a relation with the computer storage system and to manage users we need to implement SQL with the operating system. It also provides the facility for testing the data backup and also for network access facilities. For many database-driven web applications, MySQL is used. Dual-licensing distribution is being offered to the clients by using my SQL server. Cross-platform support is available with my SQL. A huge number of extensions are available with it. Stored procedures, Triggers, cursors, schema and data definition and manipulation languages and many more features are provided. All the transactions are stored with the save points. It is also provided with a built-in data library. For each and every table multiple storage engines are available which are flexible with the application.

Several reasons for the lack of interoperability between database systems include:

- The complexity and size of the SQL standard mean that most developers do not support the whole standard.
- The standard does not define database performance in many key areas.
- The SQL standard specifies the syntax to be used for the database.

However, the general definition of grammatical semantics is poorly defined, leading to ambiguity.

Most database vendors have large customer bases available; when a new version of the SQL standard contradicts the previous behaviour of merchant data, the merchant may not want to violate the backward compatibility. A small trade incentive is available to vendors to make the conversion of database providers easier. Users who test data software often set other factors such as higher performance priorities than standard compliance.

V. RESULTS

The project involves a lot of solid research from a data-based perspective. In addition, the corrections and preliminary consideration involved in making this trial are also mentioned. This section describes the test materials, flow and results obtained during the test of predicting students dropout.

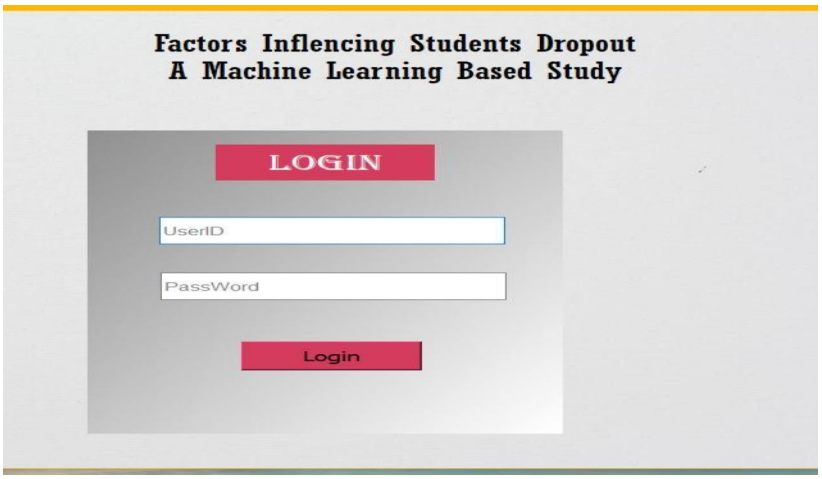

FIG 7.1: LOGIN SCREEN

When we run the application the login screen will appear. So we need to login to the application to predict the dropout status of the student. Login by using the username and password if the username and password match then only we can access the application if they are not matching with the database then it displays an error message as invalid login credentials.

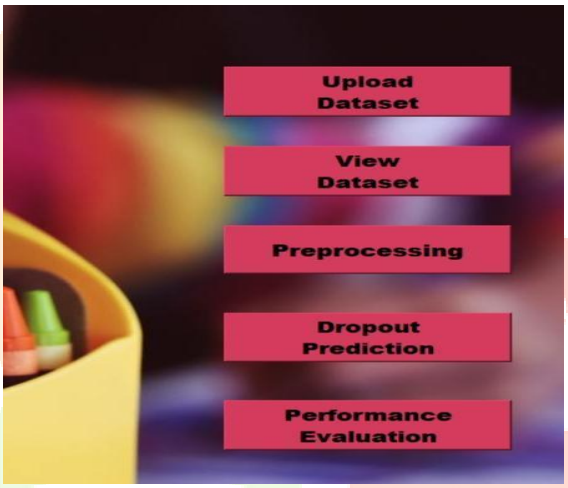

FIG 7.2: ADMIN DASHBOARD

If we are provided with the valid credentials then the welcome user window is being displayed with the following buttons as upload dataset, view dataset, preprocessing, dropout prediction, performance evaluation. We can choose the choice ass per our convenience.

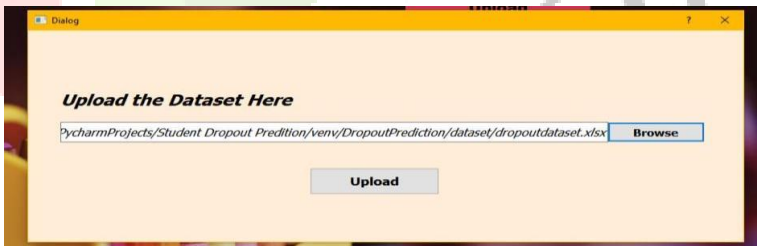

FIG 7.3: UPLOAD THE DATASET

To upload the dataset we need click the upload dataset as shown in fig 7.2 then the screen will appear as shown in fig 7.3. We can browse the system and select the path or we can manually type the path of the data present in the system. After successfully selecting the path click on the upload then a pop up will display showing data uploaded successfully

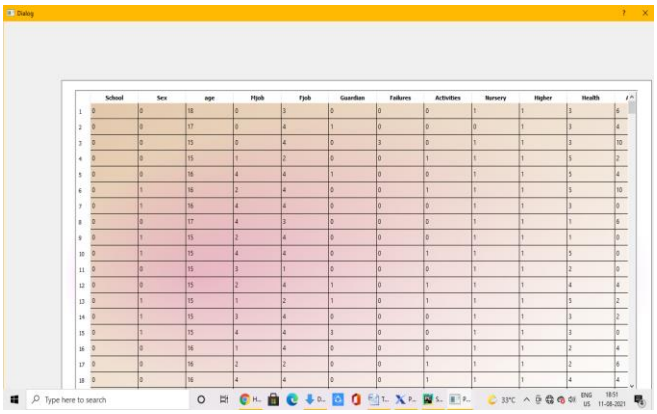

FIG 7.4: VIEW OF THE DATASET

The view of the view dataset is displayed in fig 7.4 which displays the dataset uploaded. The data contains only numerical values.

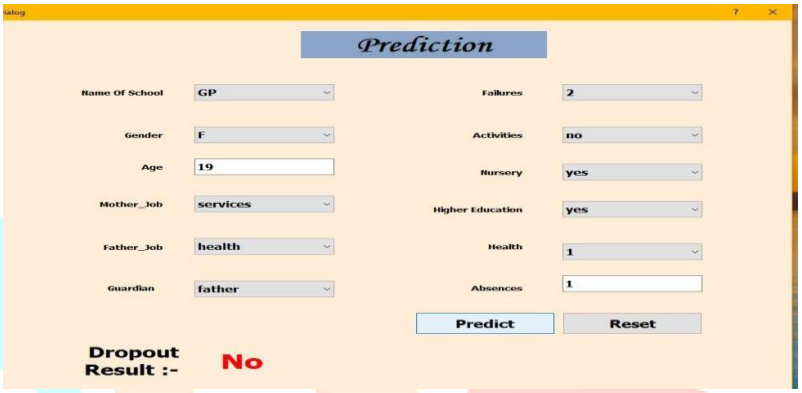

FIG 7.6: PREDICTION RESULT

The prediction result is displayed with a binary value 'Yes' or 'No'. If the result is 'No' the student is going to continue the studies. If the result is 'Yes' the student will be dropping out of studies.

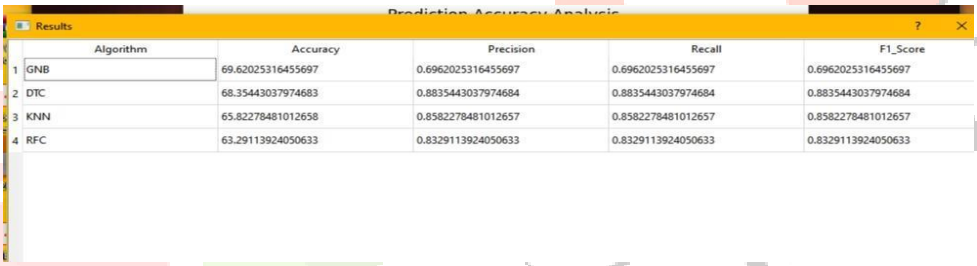

FIG 7.7: PERFORMANCE OF ALGORITHMS

Evaluation metrics are used to measure the performance of the algorithms applied. Here we are using metrics like accuracy, precision, recall, f1-score. By observing the above result we can say that for all the applied algorithms precision, recall, f1-score results are same but the accuracy results are different. So, we are going to take into consideration the accuracy measure as best measure for evaluating the results.

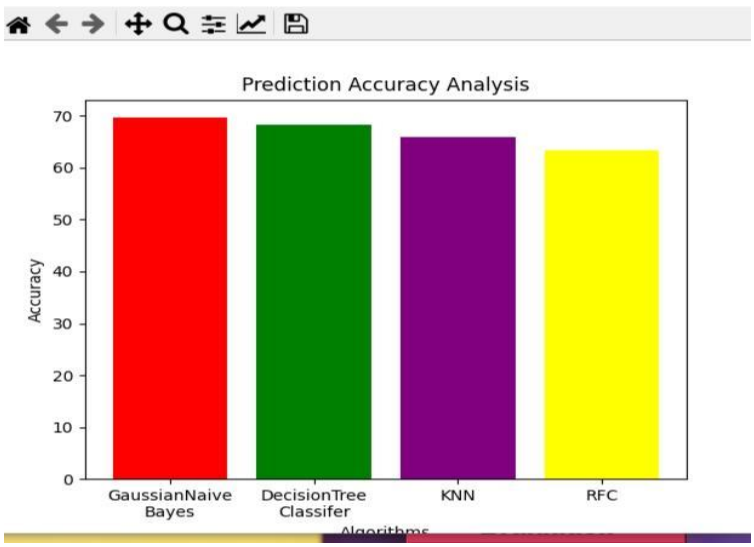

FIG 7.8: PERFORMANCE EVALUATION

Plotting the accuracy measure of the algorithms we can say that Gaussian naive bayes algorithm is performing well than others and the results that are produced by using Gaussian naive bayes are accurate.

VI. CONCLUSION

Considering the recent years, the number of student dropouts from the educational institute is increasing day by day. It not only affects the educational institution but also affect the future of the student. It is a major threat to all educational institutions. The dropout reason may vary; it depends on the factors that lead to choosing the student to commit dropout. In this project, we focus on the reason behind the student dropout. So that, data collection plays an important role in this project. The collected data are evaluated by the various techniques under data pre-processing. The data collected by various resources shows many factors like Academics, Demographical factors, Psychological factors, Health issues etc. plays important role in student dropout. As mentioned that student dropout is a major threat to all educational institutions. This model will help to identify the student who is going to drop out. While identifying them at an early stage will prevent the student drop out and can monitor and give valuable counselling to change the mind of the student from the dropout. It will also show the right path for the student to achieve their dreams.

REFERENCES

[1] "Release 1.2.0". 26 December 2020. Retrieved 15 January2021.

[2] "License – Package overview – pandas 1.0.0 documentation". pandas. 28 January 2020. Retrieved 30 January 2020.

[3] Scikit-learn.org. (2019). sklearn.preprocessing.OneHotEncoder — scikit-learn 0.22 documentation. [online] Available at: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html.

[4] Krish Naik (2019). Feature Engineering-How to Perform One Hot Encoding for Multi Categorical Variables. YouTube. Available at: https://www.youtube.com/watch?v=6WDFfaYtN6s&list=PLZoTAELRMXVPwYGE2PXD3x0bfKnR0cJjN&ab_channel=Krish Naik [Accessed 10 Sep. 2020].

[5] Carlos Márquez-Vera, Cristóbal Romero Morales, and Sebastián Ventura Soto "Predicting School Failure and Dropout by Using Data Mining Techniques",IEEE Journal of Latin-American Learning Technologies, Vol.8, No. 1, February 2013.

[6] F. Araque, C. Roldán, and A. Salguero, "Factors influencing universitydropout rates," Computer education.

[7] "Release 3.3.3". 12 November 2020. Retrieved 14 November2020.

[8] "Matplotlib: Python plotting — Matplotlib 3.2.0 documentation". matplotlib.org. Retrieved 2020-03-14.

[9] P.Cortez and A. Silva. Using Data Mining to Predict Secondary School Student Performance. In A. Brito and J. Teixeira Eds., Proceedings of 5th [23]Future Business TEChnology Conference (FUBUTEC 2008) pp. 5-12, Porto, Portugal, April, 2008, EUROSIS, ISBN 978-9077381-39-7.

[10]Ayesha, S. , Mustafa, T. , Sattar, A. and Khan, I. (2010) 'Data Mining Model for Higher Education System', European Journal of Scientific Research, vol. 43, no. 1.

[11] Bharadwaj B.K. and Pal S. "Mining Educational Data to Analyze Students" Performance", International Journal of Advance Computer Science and Applications (IJACSA), Vol. 2, No. 6, pp. 63-69.

[12]Techopedia (2019). Production Environment. Techopedia. https://www.techopedia.com/definition/8989/production-environment

[13] Opeyemi, Bamigbade (2019). Deployment of Machine learning Models Demystified (Part 1). Towards Data Science.

[14] Acharya, A., Sinha, D.: Early prediction of student performance using machine learning techniques. Int. J. Comput. Appl. 107(1), 37–43 (2014)