# MACHINE LEARNING BASED STUDENT PERFORMANCE ANALYSIS SYSTEM

**R.Karthikeyan[1] S.Satheesbabu[2] P.Gokulakrishnan[3]**
[1]Professor,PSNA College of Engineering and Technology, Dindigul, India,
Email: karthikeyan@psnacet.edu.in
[2] Associate Professor, PSNA College of Engineering and Technology, Dindigul, India
[3] Professor, PSNA College of Engineering and Technology, Dindigul, India

**Abstract—**The academic output of the student is normally stored in various formats in the student administration system (files, documents, records, photographs and other formats). These data can be collected for valuable knowledge from the students. However, it is difficult to analyze the increasing amount of data of students through conventional statistical techniques and database management tools. For universities to gather valuable information, a tool is therefore required. This helpful knowledge can be used to predict the success of students. Leistungs analyze learning results is a framework that aims for success in the areas of student interest at various levels and dimensions. This paper proposes a complete structure as a rule-based recommendation method not only for analyzing and forecasting the success of students but also for presenting their reasons. The proposed system analyzes demographic details for pupils, studies and psychological features so that students, teachers and parents can collect all possible knowledge. To seek maximum accuracy in academic predictions across a range of powerful techniques of data mining. The system successfully recognizes the limitations of the student and makes adequate recommendations. The practical case study on 200 students indicates the excellent performance of the proposed system compared to the current framework.

**Keywords—**RecommenderSystem,PerformanceAnalysis,Statistical Techniques

## 1 INTRODUCTION

Students are essential advantages of producing high quality graduates who excel in academia, practical know-how, self-development and creative thought for any school, college and other educational institution. In order to do this it is necessary to evaluate the output of students in any college, school or any other institution of education. By performing diverse tests, evaluations and other indicators, academic success can be assessed. However, student to student academical performance may vary, as each student's performance is different.

The number of universities/institutions in higher education has multiplied over the past decade. Every year, they generate a large number of graduates. Universities and institutes can better pursue the pedagogy, but still face the problem of dropouts, low-ranking students and unemployed.

To comprehend and analyze poor performance variables is a dynamic and continuous mechanism that hides knowledge from academic performance and student behaviour, both past and present. Strong tools are necessary to scientifically evaluate and forecast student results. While collected by universities and organisations, the data remains unused and does not enhance the performance of the students by taking any decisions or policy.

If universities can determine low performance factors earlier and can predict the behavior of students, this information can help them take effective steps to increase student performance. This situation will win all university/institute stakeholders such as administrators, teachers, students, parents. They will also win. Students would be able to recognize and strengthen their deficiencies in advance. Teachers should schedule their courses according to the students' needs and can offer these students better guidance. Parents in such institutions will be confident of their ward results. Management will develop better policies and plans to improve these students' success by additional facilities. In the end, it will help to produce competent people and, consequently, sustainable development for the country.

A data mining research and prediction have shown remarkable results with respect to fraud identification, consumer behavior forecast, financial market, lending evaluation, prediction of failure, intrusion evaluation and real-estate assessment. In the education system too, it can be very successful. It is a very useful tool for exposing secret patterns and important information that can otherwise be sought and comprehended with statistical methods.

## 2 LITERATURE SURVEY

Many researchers used exclusive statistics and mining techniques in educational institutions for predicting student knowledge. Some techniques have been examined and algorithms have been implemented, particularly decision-making trees and regression techniques, but also a number of ideas and ideas have finally been provided.

In G.N. Pandey, Sonali Agarwal and M. D. Tiwari's study, a group of classifiers have been added to the data set taken and attempts have been made to determine the best classifier among the classifiers involved. The right or optimal solution is analyzed by means of a comparative analysis. They thought of a data cube. This data cube includes names, verbal skills and math scores. The student's marks are located in a different cell separately. The student's classification study takes these 3 characteristics and 2000 documents. WEKA software is used in order to preprocess and classify data. Missing values are managed very effectively by WEKA software. It also helps to handle data comfortably. Classification is regulated, which means class labels are present.

Different data mining methods including cluster analysis, outline analysis, and classification were implemented in the work of Mounika Goyal and RajanVokra in their research. Based on their findings, the students were divided into 4 groups using the clustering and classification techniques. Therefore, in higher education, they used clustering methods. They also examined the usefulness of OLAP and OLTP in society and how OLTP promotes learning from students. They also investigated the use of biometric systems, integrated payment gateways, message integration, contact through e-mail, access via mobile mail, etc. Property cost is low, safety is high and solution scalable are the key advantages they suggest. OLAP is only for retrieval purposes, the data model is multi-dimensional, the star schema is followed, management and managers are the standard user. For OLTP it is used to update, the data model for organization relationships is used, structured schemes are used, and common users include workers within a company. The complicated method is simplified by a data warehousing system. The construction of the data warehouse is divided into four sections: data sources, data storage, information supply. Different methods of data mining are then introduced.

K. VenkataLakkshmi explored the improved academic domain of data mining studies for student graduate data gathered from Villupuram University of Engineering and Technology in the research paper by J. K. JyothiKalpana. The data covers a total of five years [2008-2013]. The strategies they used are centre-dependent clustering, based on density and distribution. The used software applies the techniques in the data set of the engineering student. They studied the clustering algorithms based on the centre, the density and distribution. These algorithms were applied to student clusters, and they tried to increase their efficiency. The students' output was evaluated using a simple k-means algorithm.

A decision tree model was used by Al-Radaideh et al to predict the final degree in 2005 of students studied C + + at Yarmouk University, Jordan. The methods ID3, C4.5 or the NaïveBayes were used for three different classifications. The results of the results showed a better forecast than other simulations for the Decision Tree models.

They proposed a methodology that was based on the CRISP-DM which includes the cross-branch data mining technique, in the research paper of Dr. Muhammad Reza Beikzadeh and SomnukPhon. The researchers' principal objective is to analyze the success of students in one of the major courses. They took the decision tree on data mining. This decision tree is one of the techniques implemented in the IBM miner tools for categorizing data mining. The numerous advantages of higher education data mining procedures were also stated in this study. They also listed the processes necessary to identify the students according to their qualifications and take the necessary measures.

## 3 TECHNOLOGY TACK

### Python:

Python is a general programming language, interpreted, high-level. With its noticeable use of important whitespace, Python's design philosophy emphasizes code readability. Its structures of the language and object-oriented approach are intended to allow programmers to write simple, logical code for large and small projects. Python is collected with complex typing and garbage. It supports several paradigms of programming, including method, object, and functional programming. Because of its robust standard library, Python is often defined as "batteries including".

### Anaconda:

The programming languages of Python and R (data science, machine learning software, large scale data processing, predictive analytics, etc.) which Anaconda is a free and open-source distribution to simplify package management and deployment. The package versions are administered by the conda framework for package management. Datascience Package for several operating systems is included in the Anaconda distribution.

Anaconda Navigator is an Anaconda delivery Graphical User Interface (GUI) for the user to run applications and control conda packages, environments, and channels without command-line commands. You can browse Anaconda Cloud or a local Anaconda Repository for packages. Install the packages into the environment and upgrade the packages.

### Jupyter:

A Web-based collaborative computing environment for building Jupyter notebook documents – which are a kind of computational notebook. Jugyter Notebooks (formerly IPython

SSN (Print): 2204-0595
ISSN (Online): 2203-1731

Notebooks). The term "notebooks" may refer to a large number of different entities, specifically the web application Jupyter, the Jupyter Python server or, depending on the context, Jupyter document format. The Jupyter Notebook document is the versioned JSON document, with a sequence of input/output cells that may include code, word(s). The maths and the rich media typically end with the ".ipynb" extension. This document contains an ordered list of input/output cells.

## Libraries:

1. *Pandas:*

Pandas is a Python programming language software library for data processing and testing. It provides data structures and operations to manipulate numerical tables and time series in particular. It is primarily used as data frames for machine learning.

### NumPy:

NumPy is a library for the Python programming language that supports large, multi-dimensional arrays and matrices, as well as a large group of high-level mathematical functions. It aims at Python, a non-optimizing byte code interpreter C Python ,'s reference implementation. For this edition of Python the mathematical algorithms are always slower than those that have been compiled. The slowness issue is solved in part by supplying multidimensional arrays, functions and operators that work effectively with the arrays and that require the rewrite of certain code with NumPy, mostly internal loops.

### SkLearn:

Scikit-learn is a Python programming free software learning machine (formerly Scikits.learn and also known as sklearn). The libraries NumPy and SciPy in Python have been developed to deal with different classification algorithms, regression and clustering, including vector machines, random forests, gradient boosts, K-means and DBSCAN.

### Matplotlib:

Matplotlib is a Python programming library and its NumPy extension of numeric mathematics. It offers the object-oriented API for integrating plots into applications using GUI-toolkits for general purposes such as Tkinter, wxPython, Qt or GTK+. A procedural pylab interface based on a state machine is also available (such as OpenGL), which closely resembles MATLAB but discourages its use.

### Seaborn:

Seaborn is a matplot lib-based Python data display library. It offers an appealing, insightful statistical graphics interface.

### Flask:

Flask is a Python-written micro web platform. It is known as a microframe because no special tools or libraries are required. There are no database abstraction layer, validation of the type or any other components that provide common functions from pre-existing libraries. However, Flask supports extensions that can incorporate applications functionality in Flask itself. For object-relational mappers there are extensions, form validation, upload handling, different open authentication technologies and a number of popular tools. Extensions are often more often revised than the main Flask.

SSN (Print): 2204-0595
ISSN (Online): 2203-1731

## IMPLEMENTATION

### 4.1 ProblemIdentification:

Pointing the effectiveness of the student performance analysis method is essential to problem and data comprehension. The project goals and goals are defined prior to the system creation, problems and data understanding. Issue detection and review of the current processes are focused on their efficacy and performance. If the issues are found, solutions to each problem are identified and gathered by reading and reviewing the research papers concerned.

### 4.2 DataSets:

Data sets typically approach a single database table capability, in which each column or table's field is a given variable, and in which each row or table's record is a match for each student of the query data. The overview of the dataset is given below.

### 4.3 ExploratoryDataAnalysis:

We carry out a descriptive analysis and evaluate the target variable during this process. We also checked how many classes and a number of other (high cardinality) variables were available. I also visualized the target variable in a histogram as a good way to understand the data distribution for parameter adjustment.

### 4.4 DataPrepossessing:

This module performs preprocessing. Data cleaning, data incorporation, transformation, reduction of data are preprocessing techniques. We are washing, processing and reducing our project. In the purification process, we tape incomplete values, contradictory values and zero values.

### 4.5 DataCleaning

During this step, we drop these high variables as a precursor to the preprocessing step. The data in the dataset is cleaned first. Transformation is now implemented, i.e. Raw data is translated to the full use of data. Data Cleaning for missing rows, false columns, erroneous file format and a null value field should be performed Data Cleaning.

### 4.6 DataTransformation:

The goal variable is removed from the entire data set and transformed into a single-hot model matrix. Often, it is essential to process data in a sparse matrix format for some algorithms. This phase is automated when creating models by other statistical tools like R. I attributed the lack of data values to 0. In order to prevent variables at different scales from having a strong effect on the coefficients, I have scaled the continuous variables by min-max normalisation.

### 4.7 DividingdatasetasTestingandTrainingSets:

The dataset needs to be classified into two distinct data units: data sets and datasets. This is achieved during computer and data mining to acquire one's expertise. We typically use 30% of the data as a test set and the remainder 70% as a training set.

### 4.8 ModelCreation:

In our project, we use three algorithms to pick the appropriate model algorithm. They are the logistic return, the random forest and the descent of the stochastic gradient.

Random forests or random decision forests are an ensemble method of learning for classification, retrogradation and other tasks that operate by building a variety of decision-making bodies during training, and producing class mode (classification) or medium (regression) of individual trees. Random decision forests correctly override their preparation habits for decision-makers. It uses a method usually referred to as bagging, called the Bootstrap Aggregation. The basic concept is not to rely on individual decision treaties but to combine several decision treaties in deciding the end result.

Regression is a statistical model that uses a logistic role in its basic form, although several more complicated extensions exist, in the model of a binary-dependent variable. Regression analysis estimates the parameters of the logistic model (or logit regression) (a form of binary regression). Mathematically, a binary logistic model features a dependent variable with two possible values, namely pass/fail expressed by an indicator variable, with the labels '0' and '1' for both values. Log-odding is a linear combination of the "1" value, which is the logarithm of the odds; each of the independent variables may be a binary variable (two classes, coded by the indicator variable) or an independent variable; (any real value). The associated probability for the labeled "1" value may be between 0 (certainly the value of "0") and 1 (certainly the value of "1," which is why the marking is essential.

Only when a decision threshold is shown is logistic regression a classification technique becomes. The threshold setting is a key feature of the logistical regression and depends on the issue of classification itself. The determination of the threshold value is primarily determined by the precision and reminder values. We prefer both accuracy and remember 1, but this is rarely the case.

Stochastic gradient decrease (often SGD) is an iterative way to refine an objective function with the necessary properties of smoothness (e.g. differentiable or subdifferentiable). The gradient optimization can be seen as a stochastic approximation, as it replaces the actual gradient (computed by an estimate from the total of data) (calculated from a randomly selected subset of the data). This reduces the computational burden and achieves faster iterations in trade at a slightly lower convergence rate, particularly with Big Data applications.

It is a simple yet efficient approach to discriminatory learning of linear classifiers under convex loss functions, such as (linear). While SGD has long been involved in the culture of machine learning, only recently in the sense of large-scale learning it has received significant attention.

In the field of text classification and natural language processing, SGD has been successfully introduced for the sizeable problems associated to machine learning. Since the data is spacious, the classificators in this module scale up to more than 10^5 workout examples and over 10^5 features.

### 4.9 Evaluation

The model assessment is an important part of the process of model growth. It helps to identify the right model for our data and how well the selected model performs in the future. Here we evaluate the models based on the test data and choose the one with the best results.

SSN (Print): 2204-0595
ISSN (Online): 2203-1731

## 5 SYSTEM DESIGN

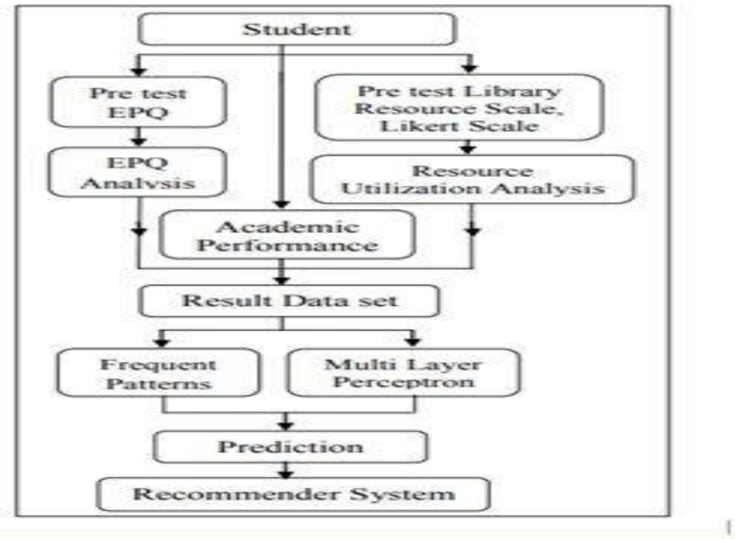
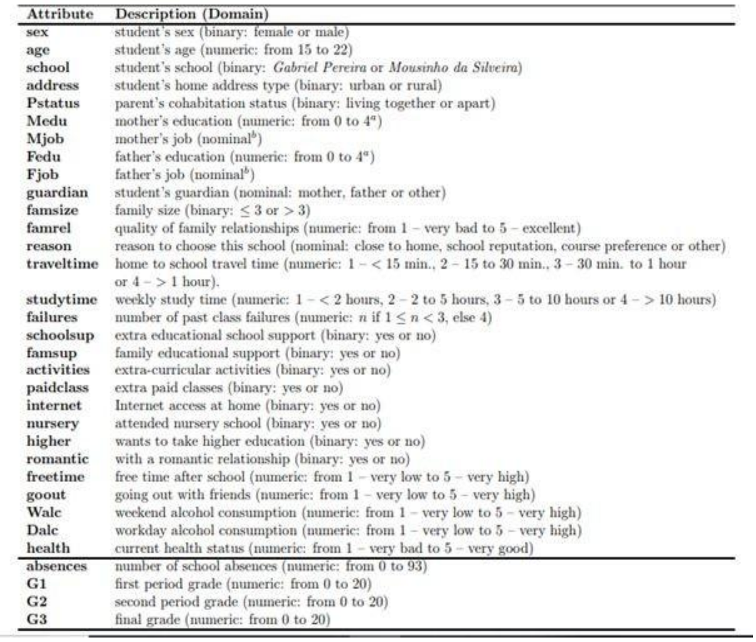


**Figure 1:** System architecture design

## 5.1 DATASET DESCRIPTION



| Attribute | Description (Domain) |
|---|---|
| **sex** | student's sex (binary: female or male) |
| **age** | student's age (numeric: from 15 to 22) |
| **school** | student's school (binary: *Gabriel Pereira* or *Mousinho da Silveira*) |
| **address** | student's home address type (binary: urban or rural) |
| **Pstatus** | parent's cohabitation status (binary: living together or apart) |
| **Medu** | mother's education (numeric: from 0 to $4^a$) |
| **Mjob** | mother's job (nominal$^b$) |
| **Fedu** | father's education (numeric: from 0 to $4^a$) |
| **Fjob** | father's job (nominal$^b$) |
| **guardian** | student's guardian (nominal: mother, father or other) |
| **famsize** | family size (binary: $\leq 3$ or $> 3$) |
| **famrel** | quality of family relationships (numeric: from 1 – very bad to 5 – excellent) |
| **reason** | reason to choose this school (nominal: close to home, school reputation, course preference or other) |
| **traveltime** | home to school travel time (numeric: 1 – < 15 min., 2 – 15 to 30 min., 3 – 30 min. to 1 hour or 4 – > 1 hour). |
| **studytime** | weekly study time (numeric: 1 – < 2 hours, 2 – 2 to 5 hours, 3 – 5 to 10 hours or 4 – > 10 hours) |
| **failures** | number of past class failures (numeric: $n$ if $1 \leq n < 3$, else 4) |
| **schoolsup** | extra educational school support (binary: yes or no) |
| **famsup** | family educational support (binary: yes or no) |
| **activities** | extra-curricular activities (binary: yes or no) |
| **paidclass** | extra paid classes (binary: yes or no) |
| **internet** | Internet access at home (binary: yes or no) |
| **nursery** | attended nursery school (binary: yes or no) |
| **higher** | wants to take higher education (binary: yes or no) |
| **romantic** | with a romantic relationship (binary: yes or no) |
| **freetime** | free time after school (numeric: from 1 – very low to 5 – very high) |
| **goout** | going out with friends (numeric: from 1 – very low to 5 – very high) |
| **Walc** | weekend alcohol consumption (numeric: from 1 – very low to 5 – very high) |
| **Dalc** | workday alcohol consumption (numeric: from 1 – very low to 5 – very high) |
| **health** | current health status (numeric: from 1 – very bad to 5 – very good) |
| **absences** | number of school absences (numeric: from 0 to 93) |
| **G1** | first period grade (numeric: from 0 to 20) |
| **G2** | second period grade (numeric: from 0 to 20) |
| **G3** | final grade (numeric: from 0 to 20) |

## 6 CONCLUSION

The system takes raw table data as input and process the data Using Random Forest Classifier, Logistic Regression Classifier and Stochastic Gradient Descent Classifier with the help of Python Language and Anaconda IDE. The input data is analyzed based on various factors such as Romantic Status, Alcohol Consumption, Parents Education Level, Frequency of Going Out, Desire of Higher Education, and Urbanvs. Rural Students. The process includes various modules such as Data Pre-Processing, Data Transformation,Data Cleaning, and Divide & Testing with Training Sets. Each Model produce distinct output and the best model is selected as final output.

Thus, this system helps parents and educational institution to evaluate the students' performance based on various aspects such as alcoholic, free time, study time, etc.

## REFERENCES

[1] A. Sonali Agarwal, G.N.Pandey and M.D.Tiwari ,"Data Mining in Education: Data Classification and Decision Tree Approach", International Journal of e-Education, e-Business,e-Managementande-LearningVol.2,No.2,April2012.

[2] MounikaGoyal and RajanVokra, "Applications of Data MininginHigherEducation",InternationalJournalonComputerScience Issues (IJCSI), ISSN (Online): 1694- 0814 , Vol 9, Issue2,No1,March2012

[3] Q. A. AI-Radaideh, E. W. AI-Shawakfa, and M. I. AI-Najjar,"Mining student data using decision trees", International Arab Conference on Information Technology(ACIT 2006), YarmoukUniversity,Jordan,2006.

[4] KarishmaKhan, JanhaviPatil" Analysis of Intelligent System for Student's Performance using E-learning Approach",2018.

[5] D. Liu, H. Sun and L. Huang, "Research Performance Evaluation for Measuring Efficiency with Data Envelopment Analysis Method," *2018 Seventh International Conference of Educational Innovation through Technology (EITT)*, Auckland, New Zealand, 2018, pp. 254-257, doi: 10.1109/EITT.2018.00058.

[6] I. A. J. Mhlanga, N. M. Ahmad, A. Azman and S. F. Abdul Razak, "Performance Analysis of the Effect of a Combiner on a MapReduce Job," *2018 IEEE Student Conference on Research and Development (SCOReD)*, Selangor, Malaysia, 2018, pp. 1-5, doi: 10.1109/SCORED.2018.8711046.

[7] P. Sokkhey and T. Okazaki, "Comparative Study of Prediction Models on High School Student Performance in Mathematics," *2019 34th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC)*, JeJu, Korea (South), 2019, pp. 1-4, doi: 10.1109/ITC-CSCC.2019.8793331.

[8] M. Wati, N. Novirasari, E. Budiman and Haeruddin, "Multi-Criteria Decision-Making for Evaluation of Student Academic Performance Based on Objective Weights," *2018 Third International Conference on Informatics and Computing (ICIC)*, Palembang, Indonesia, 2018, pp. 1-5, doi: 10.1109/IAC.2018.8780421