Prediction of Students' Performance based on Academic, Behaviour, Extra and Co-Curricular Activities

T. Jenitha
Department of Computer Science and Engineering, Mepco Schlenk Engineering College (Autonomous), Sivakasi, Tamil Nadu, India.
E-mail: jenitha@mepcoeng.ac.in

S. Santhi
Department of Computer Science and Engineering, Mepco Schlenk Engineering College (Autonomous), Sivakasi, Tamil Nadu, India.
E-mail: santhicse@mepcoeng.ac.in

J. Monisha Privthy Jeba
Department of Computer Science and Engineering, Mepco Schlenk Engineering College (Autonomous), Sivakasi, Tamil Nadu, India.
E-mail: privthy@mepcoeng.ac.in

Received October 30, 2020; Accepted November 30, 2020
ISSN: 1735-188X
DOI: 10.14704/WEB/V18SI01/WEB18058

Abstract

Since Academic institutions contain huge volume of data regarding students such as academic scores, scores in co and extracurricular activities, family annual income, family background and other supporting documents, predicting individual students performance in all aspects manually is a difficult task. The proposed work uses data mining techniques to identify students who are eligible for scholarships and other benefits. Students are classified into different categories by means of academic, behavior, extra and co-curricular activities. Machine Learning algorithms such as Naive Bayes, Decision Tree Classifier and Support Vector Machine are used for predicting the performance of the student. With the help of this proposed model parents and instructors can monitor student's performance and they can also provide essential technical and moral support. Also this helps in providing academic scholarship and training to the students to support them financially and to enrich their knowledge. It suggests the Academic Institutions to organize induction or training programmes at the beginning of the semester. Technical training, motivational talks, Yoga, etc are organized by the institutions by keeping in mind of students physical and mental health. Considering the e-learning platforms huge volumes of data and plethora of information are generated. In this work, various learning models are constructed and their accuracies are compared to analyse which algorithm out-performs.

Keywords

Data Mining, Decision Tree, Naive Bayes, Support Vector Machine.

Introduction

Universities today have very complex system and are working in highly competitive environment. In order to develop strategic plan and action plan it is very important for modern universities to deeply analyze the performance of each student. Performance of Student's is one of the important keys of the marketing crusade and helps to reach the promising potential of the university.

By analyzing student's academic record, behavior, awards and achievements in extra and co-curricular activities universities can take necessary action to achieve their mission and vision also retain in the top position. Students can also be benefitted by getting scholarship and value added programs. Parents and teachers can give support and moral strength to the students. Every student can also do self analysis which helps them to go along with their dream.

Performance of Students is one of the most important criteria for evaluating higher educational institutions. The institutions are ranked based on the students' academic performances. Previous researches say that there are lots of reviews and discussions on Student's performance are still going on. Primary and higher academic institutions are operating on high pace to produce industry ready students and academicians to a secure a place for themselves in this world. The main aim of these academic institutes is on generating graduates with good intellectual skill as well as good co and extracurricular Skills. It is necessary to track how the student is performing in a particular field and in which field they require more training. By using academic database, the authorities can decide what should be the action taken for forthcoming semester to improve the performance of the students. They may implement the action plan in the beginning of every semester as a special program or which can be a continuous process throughout the course. However analyzing such a voluminous data is a difficult tasks that too increasing in day by day. Machine learning techniques are used to simplify the prediction. Many algorithms are used to predict student's performance, each one has its own advantages and disadvantages, and also they need to be compared to give accurate result. This work uses Naive Bayes' algorithm, Decision Tree classifier and Support Vector Machine techniques. These algorithms are mainly used to evaluate and analyze the performance of students.

Data mining is the knowledge based discovery process used to extract or mine voluminous data. It tries to bring out the inherent structures, patterns, anomalies, correlation and changes from voluminous information stored in internal databases and in data warehouses or other information repositories available. The raw and unstructured data need to be transformed into useful information and knowledge for further analysis. Decision Tree Support, Machine Learning, Deep Learning, Neural Networks, Artificial Intelligence and Business Management are some of the fields where data mining is used. Data mining algorithms are used on very large amount of data for discovering hidden patterns and relationship between them which is helpful in the process of decision making. Even though knowledge discovery and data mining are typically same, data mining is actually an important part of the information discovery process. This requires the step by step procedure for extracting useful information from raw data.

Classification algorithm is one of the data mining techniques which groups data into predefined labels. It comes under supervised learning in which rules are generated and training and testing data are categorized under predefined group. There are two phases in classification technique. The first phase is the learning phase, where the classification rules are generated and training data is analyzed. The second phase is the classification phase, where test data are classified into predefined groups based on the generated rules. The classification algorithms need target classes, the output labels are created for the attribute “performance”, which has two values either “Good” or “Bad”.

Clustering works on the principle of grouping data attributes in such a way that the data within the cluster are more similar to each other than to those in other clusters. Several fields such as pattern recognition, image processing, information retrieval and machine learning use clustering as a famous technique for analyzing statistical data. There are several methods in clustering which are differ in cluster attributes and the method used to find the elements of the clusters in an efficient manner.

As we have two predefined classes “Good” and “Bad” for performance we have used classification technique instead of clustering technique, an unsupervised learning technique in which data are discovered from examples of unknown a priori.

Regression method is used to find the correlation between one or more dependent and independent variables. Most of the problems require complicated predictive models. This model can be used while predicting results of regression and classification. Neural networks also can produce results for both classification and regression.

Evaluating which data mining technique is suitable for the particular application is totally depends on the end users. Generally performance of each method is verified by evaluating the accuracy of the output data. Accuracy is measured by determining how much percentage of date is correctly classified and how many are misclassified.

Related Work

Education is one of the powerful fields where various techniques of data mining are used. Due to the massive amount of students' data the research in the academic field is rapidly increasing which can be used to find an informative pattern showing students' learning behaviour. Astha Soni et.al [1] proposed a model for predicting the performance of first year Computer Science students. The data collected contain the students' population, history of academic records and personal information. Three models namely the Decision Tree, Naive Bayes, and Rule Based classification are built from the data in order to generate the best academic performance prediction model for students. This method is not fit for small size data due to missing data and data incompleteness. In addition to the academic score students attendance is also considered as the most performance affecting factor.

Predicting the consolidated performance of each student is a significant criterion of higher educational institutions. It will help the institute to predict the performance of the students [2] more accurately. The study revealed that the Multilayer Perception is more accurate than the other algorithms. It is potentially time consuming and technically complex process and also unable to predict whether a student gets placed or not based only on their academic performance.

Edin et.al [3] developed a model which can obtain the output based on students’ academic success. Predicting students performance is done to take precautionary steps against student failure in exams or dropouts in courses is a very significant form of analysis. There may be many factors influencing student’s discontinuation of course. Leraning models are constructed to identify and predict these dropouts. Classification method has been used for prediction but the experiment cannot be extended with more distinctive attributes.

Kamini et.al [4] discussed the mathematical modelling and algorithms which are based on Support Vector Machine are used for learning. SVMs were developed for binary classification. SVM is a supervised learning method that analyses data and identifies patterns used for classification and regression analysis. The SVM classifier is fed with input data and predicts, for each given input, there are two possible classes form the output, making the SVM a binary linear classifier. Classification maps an input data item into one of several predefined output categories. These algorithms output classifiers, for example, in the form of binary outputs or rules which can be used to make the classifier to learn further.

A support vector machine outputs a hyperplane or set of hyperplanes in higher dimensions. It separates the space into two planes. The largest distance to the nearest training data points of any class forms good separation, since in general if the margin is larger it lowers the errors of the classifier.

An attribute is called as predictor variable, and an attribute which is selected to define the hyperplane is called as feature. The process of choosing an appropriate way of selecting attributes is known as feature selection. A set of features that describes one test case is called as vector. So the goal of SVM modelling is to find the optimal hyperplane that separates cluster of vectors in such a way that cases with one category of the target variable are at one side of the plane and other categories are at another side of the plane.

A classification task consists of the training and testing data in some proportions. Every instance in the training data has a target value and several attributes. SVM model predicts the target value of data instances in the testing data set. If data is linear, a hyperplane may be used to separate the data. But often data are non-linear and inseparable. In such situations, kernels are used to map non-linear data to high dimensional space.

Transforming the data into high dimensional space a similarity matrix can be built based on the dot product. If the feature set is chosen correctly, pattern recognition can be done. Though new kernels are being proposed, the following are basic kernels are properly used. They are,

1. Linear
2. Polynomial
3. Radial Basis Function
4. Sigmoid Function

Radial Basis Function kernel has more advantages relatively. Unlike the linear kernel, this kernel can map the samples in a non-linear fashion to high order dimensions nonlinearly so it can adapt to scenarios when the relation between class labels and attributes is nonlinear. Support Vector Machine acts as one of the best approaches to data modelling and classification. But it should deal with

1. Having more than two predictor variables.
2. Separating the points with non – linear curves.
3. Handling the clusters which are not completely separated.
4. Handling classifications other than binary categories.

Predicting student’s performance in order to prevent or take precautionary steps against student failure in exams or dropouts in courses is very important these days. There may be many factors influencing student’s discontinuation of course. In this paper, a classification method is used. Decision tree classifiers are used for solving the class imbalance problem is also discussed.

Baradwaj. et al [5] proposed a method in which decision tree classifiers are used to classify the students based on their academic, personal, and family data. The results obtained are not always balanced. Resolving is used as an effective technique to solve the class imbalance problem. Interpretable results with good accuracy have been obtained.

The performance of classifiers can be evaluated using a confusion matrix consisting of True positive, true negative, false positive and false negative values. This matrix contains the information about the actual and predicted values. A large number of evaluation measures can be obtained by using confusion matrix.

Educational Data Mining (EDM) is a new research area in which data mining and machine learning methods in interdisciplinary education field. It uses data mining and machine learning to explore data from academic databases to find out structure and pattern that predicts the behaviour and performance of students. Its goal is to collect more data on educational settings to improve educational outcomes and produce skilled personals to the society.

For academic institutions it is a great concern to predict the behaviour and performance of the students using EDM. As a result, it would give warning to the students whoever at risk by examining their score, and help them to avoid such failures by providing required counselling and training.

However, several factors make the process of measuring student’s academic performance as a challenging one. The factors influencing students’ performance are personal, demographics, educational and family background, psychological, academic progress and other environmental factors. These factors are interrelated and are directly or indirectly connected with students' behaviour, profession and performance. The interrelationship between these factors is complex and is connected in a non linear way.

There are many machine learning related techniques available to predict student's academic performance. Proposed a model which used multiple linear regression model and support vector machine (SVM) to predict individual and overall student academic performance. Amjad et.al [6] predicted student's performance by combining SVM and a fully connected neural network to improve the classification accuracy. Ali et.al [7] constructed traditional artificial neural networks models to predict students academic performance. Gray et.al [8] used regression method to predict the students' marks in higher education systems. Proposed a model using decision trees and SVM to predict students academic performance with huge amount of data.

Several data mining techniques are used for analysing students performance. Academic institutions contain huge amount of students' academic database containing score and other personal details. These student databases along with other relevant attributes like family background, family income, etc are taken into consideration. This helps the academicians to identify such students and feed them economically and technically to lift them from risk. The pattern is developed by analysing students' current score with previous score using data mining classification algorithms such as Naive Bayes, Decision Tree classifier and Support Vector Machine.

The prediction model helps the parents and teachers to monitor students' regular performance and provide necessary moral and technical support. This work has more significance and is more beneficial to students and parents. Parents can know about their wards and provide them with the necessary counselling and guidance at the early phase itself. This helps the economically poor students to continue their studies by providing scholarship and funds through charity. Also it helps technically poor students to improve their skill by organising skill development programs. So this predictive model as a whole reduces the risk of students' failure in academic courses and dropouts due to external factors. With the help of this model organization can get details of students and their strength and weakness in academic and other activities, so that the institute can provide raining and cal arrange some brain storming and motivational session in order to kindle their effort towards success.

One of the techniques in data mining is classification algorithm which helps to classify the data into predefined classes. Classification is a supervised learning algorithm in which rules are generated based on training and testing data set. The data is classified under any one of the predefined groups. This process includes two phases. The first one is the learning phase, in which the rules are generated and training data is analysed. The second one is the classification phase. The test set is classified into target groups according to the generated rules. Since classification algorithms requires target classes based on values of input attributes, the component "performance" is included for all students, for which the two possible values are either "Good" or "Bad". The model is trained using training dataset and the same is tested using test data and the accuracy is calculated.

Problem Formulation

The predictive function needs to be learned from the three classification models proposed. The popular models are selected from the existing literature. The details of various methods are as follows:

Decision tree is a very common technique which is used for experimental analysis. Each node represents a test on an attribute and the branch node represents a chosen solution among the various alternatives. The leaf node signifies the chosen outcome/class-label. It is widely used as a supervised learning decision support tool. It predicts the value of the target variable form by learning rules the data attributes. The cost of using the decision trees is logarithmic. It deals with both numerical and categorical data.

Naive Bayes classifiers are based on Bayes theorem. It is based on the assumption that each attribute is independent of each other. This factor might not be true for all real world situations. It is a probabilistic learning model and can be used for multi-class classification. The algorithm calculates the class and conditional probabilities. There are many variants of Naïve Bayes classifiers. The Naive Bayes model can be easily learnt from the training data. It can handle binary, nominal and categorical data.

Support Vector Machine is another simple machine learning algorithm and can be used for both regression and classification tasks. It produces significant accuracy and is highly preferable. SVM's objective is to find a k-dimensional hyperplane that separates the given data points.

Individual Feature Analysis

The impact of each feature available in the data is calculated for predicting the student's behaviour. Out of the 48 features available 20 features are selected by the dimensionality reduction process. In the proposed work three classifiers (NB, SVM and Decision Tree) which are discussed above are used. The features selected by the models are analysed to understand which features represent the student performance best. It was inferred that the other Extra-Curricular features also played an important role. The feature “RespectElder” was found out to have a lower performance by all three models. The features the Academic Performance and Placement Performance have added significant value to the prediction. They seem to increase the overall accuracy of the classification. The next best feature in the feature set was found to be “AttendanceAbove75”with an F1-score of 0.77.The system’s performance is greatly increased by selecting only the best performing features from the feature set.

Comparisons

The performances of the three classifiers are compared to understand which one models the data better. The results of the analysis concludes that the Support Vector Machine model out performs the other two. SVM produces an accuracy of 83.33% which is comparatively better than the other two (Decision tree and Naive Bayes) prediction models. The initial analysis showed that the student’s academic performance is not only dependent on the marks but the extracurricular features also had significant impact on the academic performance of students. The dataset is further divided into 5 partitions to predict the outcomes for scholarship, students’ academic performance, students’ behaviour, technical skills and overall performance. Similar data are grouped into clusters or subclasses using the method called clustering. The clusters are formed in predefined numbers and the mean of it is obtained using Euclidean distance method.

As the result of analysis, students academic performance is contributed by students’ personal information, family income, family status, culture, attitude, extra and co-curricular activities and mental stability. In addition to academic score students should have aptitude skills required for placement includes quantitative and qualitative aptitude marks, group discussion marks, technical skill, and mock interviews response as well as coding. Indulging in bad habits like using un parliamentary words, smoking or gambling also affects the students’ academics as well as behaviour. Teachers can monitor student’s regular performance by analysing their score in various fields and send information to the parents. Also motivate the students to go in a right direction to achieve their goal. They can also identify in what specific area the student needs attention and provide necessary training and motivation.

Proposed Work

The architectural design for the Prediction of Students' Performance using Support Vector Machine Algorithm is shown in Figure1. In the above module design data is collected and stored in a dataset which contains information about students like student marks, student behavior, student's details of participation in extra and co-curricular activities. Using support vector Machine Algorithm (SVM) we predict whether the student performs good or not in the academics based on the past records of student's performance. It also predicts whether the student behaves good or not and whether a student does active participation in extra and co-curricular activities. This algorithm also predicts the accuracy score whether it correctly predicts the result or not. Then, we predict the overall performance of the student by combining the above three prediction (Academics performance prediction, Behavior prediction, Extra and co-curricular activities prediction). It determines the performance of the student by categorizing as "Poor", "Average" and "Excellent".

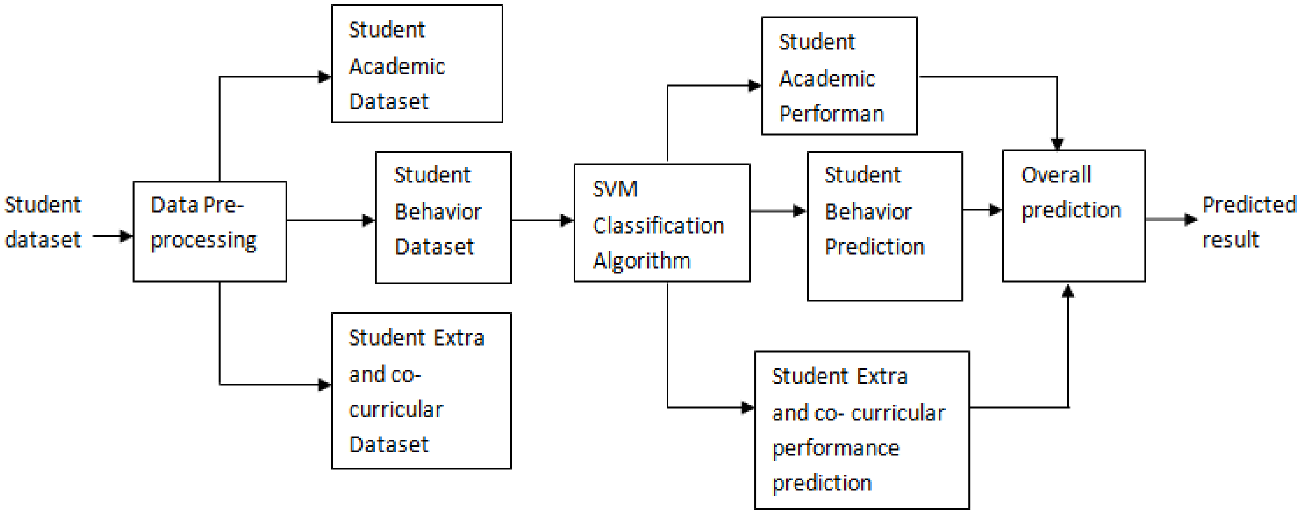

Student dataset
Data Pre-processing
Student Academic Dataset
Student Behavior Dataset
Student Extra and co-curricular Dataset
SVM Classification Algorithm
Student Academic Performan
Student Behavior Prediction
Student Extra and co- curricular performance prediction
Overall prediction
Predicted result

Figure 1 Module diagram for Student Performance Prediction

Modules Description

The various modules involved in the work are:

i. Academic performance Prediction
ii. Behaviour Prediction
iii. Extra and co-curricular activities prediction

i) Academic Performance Prediction

Student's performance is considered as the most important component in higher educational institutions. This is because of the fact that most of the academic institutions are based on the finest record of students' academic performances. Primary and professional institutions are working on high pace to generate skilled people in this competitive world. These Educational institutes focus on producing skilled graduates with high academic score. This contains the basic details of every student along with academic records based on the grade point average (gpa) marks obtained by the student and the Disciplinary actions.

The students are encouraged by providing scholarship for their academic activities. The attribute includes student Name, gpa, RollNumber, Scholarship amount, attendance, behavioural performance. This can be done by using this project. Along with their academic activities, their disciplinary actions are also counted. If they have good discipline they are eligible for scholarship. Also they should have good marks.. The analysis clearly shows that the students' academic performance not totally depend upon academic score but also depend upon disciplinary actions. The proposed model helps the parents and teachers to monitor student's performance in various fields and extend their support towards their development.

Students with marks good academic records like the gpa, interest in studies, interest in attending and learning extra courses, good concentration in co-curriculars are predicted to have good academic records.

Pseudocode

```
  //Input: Training and Testing Students Datasets having the binary values for categorizing
marks, attendance percentage, interest in studies and interest in learning extra courses
  //Output: Predicting academic performance
  //Process: With Students' Roll Number, every other details are fetched from the database.
  functionacademicPerformancePrediction()
  {
  if(marks > 75) then //Disciplinary Action Validation
  {
  if( attendance > 75% and interest_instudy == “true” and extra_courses == “true”)
  Print Academic performance as “Good”
  }
  else
  {
  Print Academic performance as “Bad”
  }
  }
```

ii) Behaviour Prediction

In this section, based on the behavioral criteria like handling emotions, handling anger, patience testing, care for parents, spending money in useful manner, etc. the student's behavioral performance is evaluated. A data set consisting of fields like handling emotions, handling anger, patience testing, care for parents, spending money in useful manner. From this set the marks are aggregated and judged. This contains the basic details of every student along with handling emotions, handling anger, patience testing, care for parents, spending money in useful manner. It contains comments about the student's performance in coding and aptitude in order to improve their placement preparations. The attribute includes Roll number, name, and behavioral prediction results.

Students' performance is predicted based on their positive attitude and behavior. The performance analysis model was created, to measure the performance of students in the areas such as "Academic", "Behavior" and "Extra and co-curricular activities". The analysis report suggests that what are all the skills the students should improve. It helps the students to concentrate on technical skills and soft skills. It focuses on generating graduates with good behavior. Generating disciplined, well equipped, ethical and highly qualified students is the aim of educational institutions. In this regard, they need to keep track of students performance in terms of behavior, ethics, shareability confidence and trustability. By using Educational data mining the performance of students behavior is analysed based that, an institute come forward to provide necessary support to the students to mould and strong their internal characteristics.

Pseudocode

```
//Input: Training and Testing Students Datasets having the details about students' behavior.
//Output: Students' overall performance level in the behavior perspective.
//Process: With the Students' roll number, every other details are fetched from the database.
function behaviorPrediction()
{
if(anger==”false” and patience==”true” and fight_with_friends==”false” an
bad_habits==”false” and emotional==”false” and spend_money_useful==”true”)
{
    Print Behavioral performance as “Good”
}
else
{
    Print Behavioral performance as “Bad”
}
}
```

iii) Extra and Co-Curricular Activity Prediction

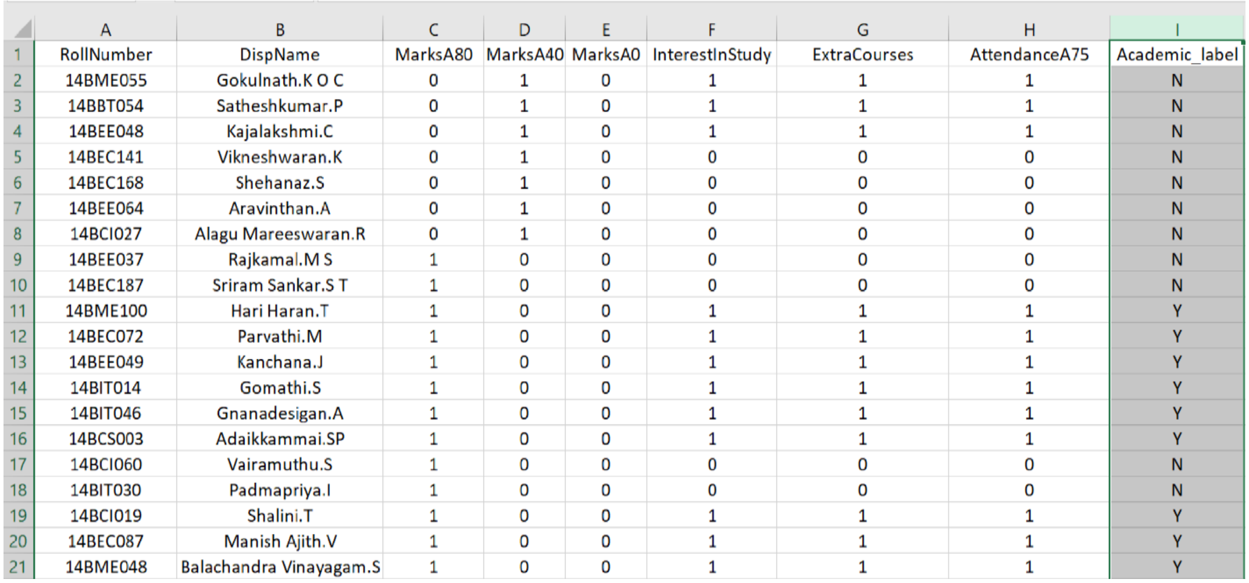
In this section, based on the active involvement of the students in extra and co-curricular activities, the student's extra and co-curricular activities performance is evaluated. A data set consisting of fields having different extra and co-curricular activity club. From this set the marks are aggregated and judged. This contains the basic details of every student along with having different extra and co-curricular activity club. It contains comments about the student's performance in extra and co-curricular activities. The attribute includes Roll number, name, extra and co-curricular activities prediction results.

Students' performance is analyzed based on their extra and co-curricular activities. The performance model was created, to measure the performance of the students in Extra and co-curricular activities. This model helps the students to check themselves where they are lagging in the fields of sports and co-curricular forums. Academic institutions not only concentrate on academic score but also concentrate on students involvement in sports, seminars, symposiums, online contests, professional club activities, etc. By keeping this in mind, institutions organize annual athletic meet, zone level competitions, conferences, symposiums, cultural festivals and other technical events. This helps the students to equip their talents in various fields. Also, institutes provide scholarship based on sports achievements and awards to recognize their talents. This internally helps the students to improve their skills and getting placements.

Pseudocode

```
//Input: Training and Testing Students Datasets having the details about students' participation in extra and co-curricular activities.
//Output: Students' overall performance level in the Extra and co-curricular participation.
//Process: With the Students' roll number, every other details are fetched from the database.
function extraCocurricularActivityPrediction()
{
if(count(co-curricular)>=3 and count(extra-curricular)>=2)
{
        Print Extra_Cocurricular performance as "Good"
}
else
{
        Print Extra_Cocurricular performance as "Bad"
}
}
```

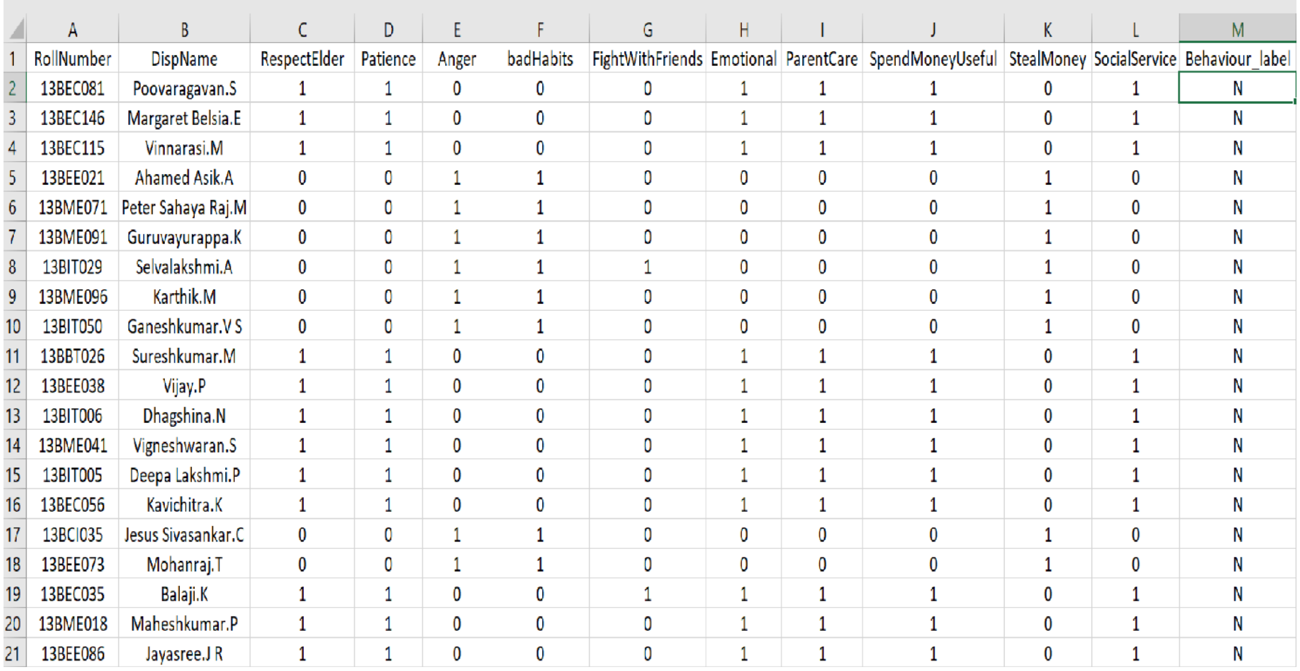
Results and Discussion



	A	B	C	D	E	F	G	H	I
1	RollNumber	DispName	MarksA80	MarksA40	MarksA0	InterestInStudy	ExtraCourses	AttendanceA75	Academic_label
2	14BME055	Gokulnath.K O C	0	1	0	1	1	1	N
3	14BBT054	Satheshkumar.P	0	1	0	1	1	1	N
4	14BEE048	Kajalakshmi.C	0	1	0	1	1	1	N
5	14BEC141	Vikneshwaran.K	0	1	0	0	0	0	N
6	14BEC168	Shehanaz.S	0	1	0	0	0	0	N
7	14BEE064	Aravinthan.A	0	1	0	0	0	0	N
8	14BCI027	Alagu Mareeswaran.R	0	1	0	0	0	0	N
9	14BEE037	Rajkamal.M S	1	0	0	0	0	0	N
10	14BEC187	Sriram Sankar.S T	1	0	0	0	0	0	N
11	14BME100	Hari Haran.T	1	0	0	1	1	1	Y
12	14BEC072	Parvathi.M	1	0	0	1	1	1	Y
13	14BEE049	Kanchana.J	1	0	0	1	1	1	Y
14	14BIT014	Gomathi.S	1	0	0	1	1	1	Y
15	14BIT046	Gnanadesigan.A	1	0	0	1	1	1	Y
16	14BCS003	Adaikkammai.SP	1	0	0	1	1	1	Y
17	14BCI060	Vairamuthu.S	1	0	0	0	0	0	N
18	14BIT030	Padmapriya.I	1	0	0	0	0	0	N
19	14BCI019	Shalini.T	1	0	0	1	1	1	Y
20	14BEC087	Manish Ajith.V	1	0	0	1	1	1	Y
21	14BME048	Balachandra Vinayagam.S	1	0	0	1	1	1	Y

Figure 2 Academic Performance dataset

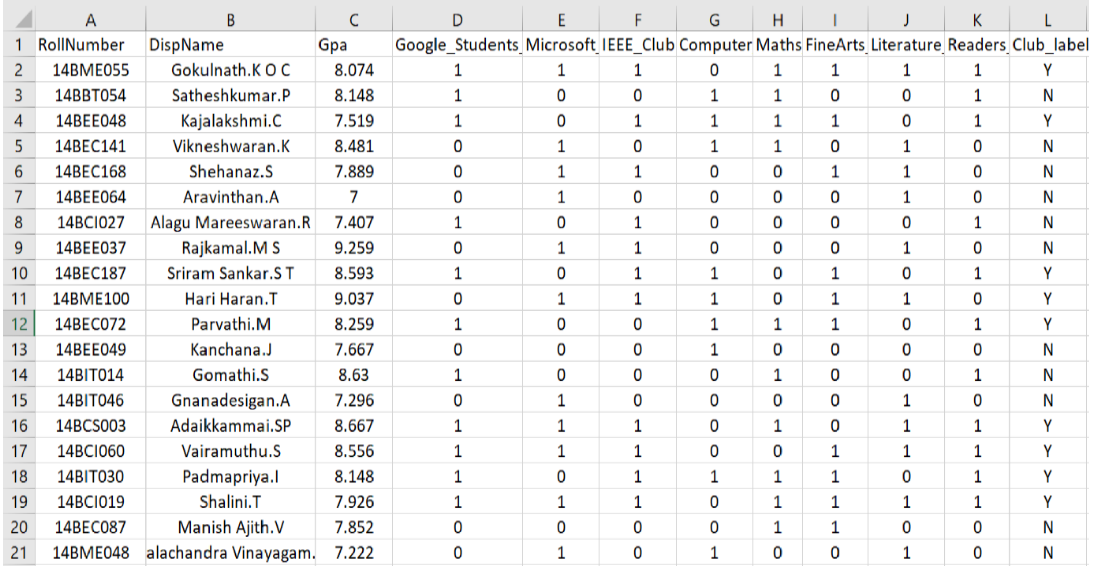
Figure2 shows the academic dataset that is being obtained after processing the original dataset which has 8 attributes that are “MarksA80”, “MarksA40”, “MarksA0”, “InterestInStudy”, “ExtraCourses”, “AttendanceA70”, “RollNumber” and “Name”.



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	RollNumber	DispName	RespectElder	Patience	Anger	badHabits	FightWithFriends	Emotional	ParentCare	SpendMoneyUseful	StealMoney	SocialService	Behaviour_label
2	13BEC081	Poovaragavan.S	1	1	0	0	0	1	1	1	0	1	N
3	13BEC146	Margaret Belsia.E	1	1	0	0	0	1	1	1	0	1	N
4	13BEC115	Vinnarasi.M	1	1	0	0	0	1	1	1	0	1	N
5	13BEE021	Ahamed Asik.A	0	0	1	1	0	0	0	0	1	0	N
6	13BME071	Peter Sahaya Raj.M	0	0	1	1	0	0	0	0	1	0	N
7	13BME091	Guruvayurappa.K	0	0	1	1	0	0	0	0	1	0	N
8	13BIT029	Selvalakshmi.A	0	0	1	1	1	0	0	0	1	0	N
9	13BME096	Karthik.M	0	0	1	1	0	0	0	0	1	0	N
10	13BIT050	Ganeshkumar.V S	0	0	1	1	0	0	0	0	1	0	N
11	13BBT026	Sureshkumar.M	1	1	0	0	0	1	1	1	0	1	N
12	13BEE038	Vijay.P	1	1	0	0	0	1	1	1	0	1	N
13	13BIT006	Dhagshina.N	1	1	0	0	0	1	1	1	0	1	N
14	13BME041	Vigneshwaran.S	1	1	0	0	0	1	1	1	0	1	N
15	13BIT005	Deepa Lakshmi.P	1	1	0	0	0	1	1	1	0	1	N
16	13BEC056	Kavichitra.K	1	1	0	0	0	1	1	1	0	1	N
17	13BCI035	Jesus Sivasankar.C	0	0	1	1	0	0	0	0	1	0	N
18	13BEE073	Mohanraj.T	0	0	1	1	0	0	0	0	1	0	N
19	13BEC035	Balaji.K	1	1	0	0	1	1	1	1	0	1	N
20	13BME018	Maheshkumar.P	1	1	0	0	0	1	1	1	0	1	N
21	13BEE086	Jayasree.J R	1	1	0	0	0	1	1	1	0	1	N

Figure 3 Student Behavior dataset

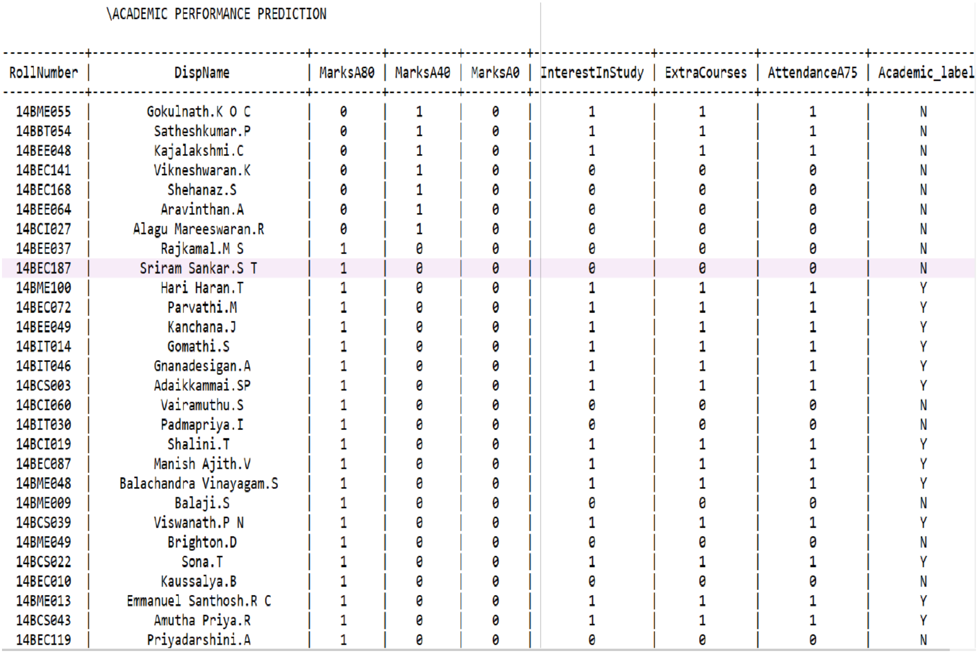
Figure 3 shows the behavior dataset that is being obtained after processing the original dataset which has 12 attributes that contains “RollNumber”, “Name” along with 8 binary attributes testing the behavior of a student.



	A	B	C	D	E	F	G	H	I	J	K	L
1	RollNumber	DispName	Gpa	Google_Students_	Microsoft_	IEEE_Club	Computer	Maths	FineArts_	Literature_	Readers_	Club_label
2	14BME055	Gokulnath.K O C	8.074	1	1	1	0	1	1	1	1	Y
3	14BBT054	Satheshkumar.P	8.148	1	0	0	1	1	0	0	1	N
4	14BEE048	Kajalakshmi.C	7.519	1	0	1	1	1	1	0	1	Y
5	14BEC141	Vikneshwaran.K	8.481	0	1	0	1	1	0	1	0	N
6	14BEC168	Shehanaz.S	7.889	0	1	1	0	0	1	1	0	N
7	14BEE064	Aravinthan.A	7	0	1	0	0	0	0	1	0	N
8	14BCI027	Alagu Mareeswaran.R	7.407	1	0	1	0	0	0	0	1	N
9	14BEE037	Rajkamal.M S	9.259	0	1	1	0	0	0	1	0	N
10	14BEC187	Sriram Sankar.S T	8.593	1	0	1	1	0	1	0	1	Y
11	14BME100	Hari Haran.T	9.037	0	1	1	1	0	1	1	0	Y
12	14BEC072	Parvathi.M	8.259	1	0	0	1	1	1	0	1	Y
13	14BEE049	Kanchana.J	7.667	0	0	0	1	0	0	0	0	N
14	14BIT014	Gomathi.S	8.63	1	0	0	0	1	0	0	1	N
15	14BIT046	Gnanadesigan.A	7.296	0	1	0	0	0	0	1	0	N
16	14BCS003	Adaikkammai.SP	8.667	1	1	1	0	1	0	1	1	Y
17	14BCI060	Vairamuthu.S	8.556	1	1	1	0	0	1	1	1	Y
18	14BIT030	Padmapriya.I	8.148	1	0	1	1	1	1	0	1	Y
19	14BCI019	Shalini.T	7.926	1	1	1	0	1	1	1	1	Y
20	14BEC087	Manish Ajith.V	7.852	0	0	0	0	1	1	0	0	N
21	14BME048	alachandra Vinayagam.	7.222	0	1	0	1	0	0	1	0	N

Figure 4 Student Extra and Co-curricular dataset

Figure 4 shows the extra and co-curricular dataset that is being obtained after processing the original dataset which has 10 attributes that includes “RollNumber”, “Name” and 5 fields of co-curricular clubs and 3 fields of extra-curricular clubs which are all binary attributes.

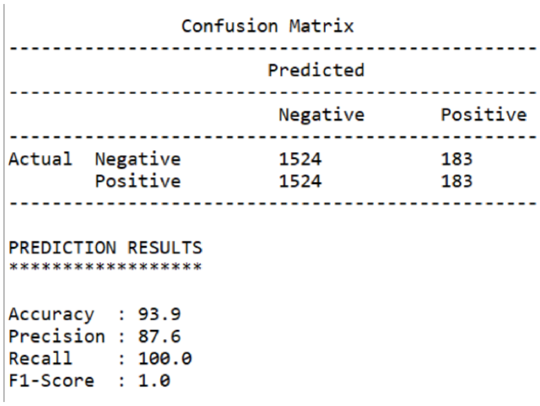


```
              \ACADEMIC PERFORMANCE PREDICTION

-----------+--------------------------------+----------+----------+----------+-----------------+--------------+---------------+---------------
RollNumber |            DispName            | MarksA80 | MarksA40 | MarksA0  | InterestInStudy | ExtraCourses | AttendanceA75 | Academic_label
-----------+--------------------------------+----------+----------+----------+-----------------+--------------+---------------+---------------
 14BME055  |         Gokulnath.K O C        |    0     |    1     |    0     |        1        |      1       |       1       |       N
 14BBT054  |          Satheshkumar.P        |    0     |    1     |    0     |        1        |      1       |       1       |       N
 14BEE048  |          Kajalakshmi.C         |    0     |    1     |    0     |        1        |      1       |       1       |       N
 14BEC141  |          Vikneshwaran.K        |    0     |    1     |    0     |        0        |      0       |       0       |       N
 14BEC168  |            Shehanaz.S          |    0     |    1     |    0     |        0        |      0       |       0       |       N
 14BEE064  |           Aravinthan.A         |    0     |    1     |    0     |        0        |      0       |       0       |       N
 14BCI027  |       Alagu Mareeswaran.R      |    0     |    1     |    0     |        0        |      0       |       0       |       N
 14BEE037  |           Rajkamal.M S         |    1     |    0     |    0     |        0        |      0       |       0       |       N
 14BEC187  |        Sriram Sankar.S T       |    1     |    0     |    0     |        0        |      0       |       0       |       N
 14BME100  |           Hari Haran.T         |    1     |    0     |    0     |        1        |      1       |       1       |       Y
 14BEC072  |            Parvathi.M          |    1     |    0     |    0     |        1        |      1       |       1       |       Y
 14BEE049  |            Kanchana.J          |    1     |    0     |    0     |        1        |      1       |       1       |       Y
 14BIT014  |            Gomathi.S           |    1     |    0     |    0     |        1        |      1       |       1       |       Y
 14BIT046  |          Gnanadesigan.A        |    1     |    0     |    0     |        1        |      1       |       1       |       Y
 14BCS003  |          Adaikkammai.SP        |    1     |    0     |    0     |        1        |      1       |       1       |       Y
 14BCI060  |           Vairamuthu.S         |    1     |    0     |    0     |        0        |      0       |       0       |       N
 14BIT030  |           Padmapriya.I         |    1     |    0     |    0     |        0        |      0       |       0       |       N
 14BCI019  |            Shalini.T           |    1     |    0     |    0     |        1        |      1       |       1       |       Y
 14BEC087  |          Manish Ajith.V        |    1     |    0     |    0     |        1        |      1       |       1       |       Y
 14BME048  |     Balachandra Vinayagam.S    |    1     |    0     |    0     |        1        |      1       |       1       |       Y
 14BME009  |             Balaji.S           |    1     |    0     |    0     |        0        |      0       |       0       |       N
 14BCS039  |          Viswanath.P N         |    1     |    0     |    0     |        1        |      1       |       1       |       Y
 14BME049  |            Brighton.D          |    1     |    0     |    0     |        0        |      0       |       0       |       N
 14BCS022  |              Sona.T            |    1     |    0     |    0     |        1        |      1       |       1       |       Y
 14BEC010  |           Kaussalya.B          |    1     |    0     |    0     |        0        |      0       |       0       |       N
 14BME013  |       Emmanuel Santhosh.R C    |    1     |    0     |    0     |        1        |      1       |       1       |       Y
 14BCS043  |          Amutha Priya.R        |    1     |    0     |    0     |        1        |      1       |       1       |       Y
 14BEC119  |          Priyadarshini.A       |    1     |    0     |    0     |        0        |      0       |       0       |       N
```

Figure 5 Academic Performance Prediction

Figure 5 shows the output of "academic performance prediction" for the student. Here, the class label is generated by using Support Vector Machine (SVM) Algorithm on the basis of the marks obtained by the student as "GPA" and other attributes like "InterestInStudies", "ExtraCourses" and "Attendance". The output label "Academic_label" is generated based on these attributes. A student is predicted to be "Good" at academics if "gpa" is above 7.5, "Attendance" percentage is above 75% and has interest in studies.



```
                Confusion Matrix
-------------------------------------------------
                        Predicted
-------------------------------------------------
                         Negative       Positive
-------------------------------------------------
Actual  Negative         1524           183
        Positive         1524           183
-------------------------------------------------

PREDICTION RESULTS
******************

Accuracy  : 93.9
Precision : 87.6
Recall    : 100.0
F1-Score  : 1.0
```

Figure 6 Accuracy Results

Figure 6 shows the accuracy results of "academic performance" for the student. The confusion matrix of the academic performance results generated is also displayed. The prediction parameters are accuracy, recall, precision and F1-score.

Performance Analysis

The Performance of the entire project is analyzed and the accuracy is found as 90%. The analysis is done for 10,000 students and their performance is predicted and classified. Out of 10,000 students, the classifications for 19,000 students have been correctly classified. Figure 7 shows the accuracy graph.

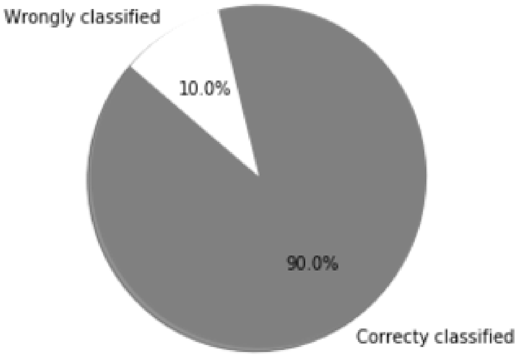


Figure 7 Accuracy Graph

Conclusion and Future Enhancement

In our work, an effort is made to find the student's performance from their performance in various fields using Classification algorithms such as Decision trees, Naive Baye's and SVM. It helps us to identify the students performance under different category based on that suggests for scholarships, training and other related activities. A feature map is constructed by taking into consideration of student's academic history, family income, status, family expenditure and personal information. Selection of dominant features is unavoidable as it gives a subset of features. By using various classification algorithms it is observed that our analysis result is very effective for the above said features. It is observed that students' personal information and family details have very strong impact on overall performance due to inherent reasons. Meta data analysis on student's databse encouraged us to carry out further examination to improve the results. Since we have direct and indirect features influencing students performance level the proposed model is very much useful for educational institutions to review the performance of the students and suggests them to take necessary steps to improve.

The experiment is carried out by using only three classification algorithms such as decision trees, Naive Bayes and SVM with more relevant features. Further the accuracy can be improved if more features are added and more classification algorithms are used and the comparative study is done.

References

Astha, S., Vivek, K., Rajwant, K., & Hemavathi, D. (2018). Predicting Student Performance using Data Mining Techniques. *International Journal of Pure and Applied Mathematics, 19*(12), 221-227.

Romero, C., & Ventura, S. (2007). Educational data mining: A survey from 1995 to 2005. *Expert systems with applications*, *33*(1), 135-146.

Sembiring, S., Zarlis, M., Hartama, D., Ramliana, S., & Wani, E. (2011). Prediction of student academic performance by an application of data mining techniques. *In International Conference on Management and Artificial Intelligence IPEDR*, *6*(1), 110-114.

Kabakchieva, D., Stefanova, K., & Kisimov, V. (2015). Analyzing University Data for Determining Student Profiles and Predicting Performance. *Conference Proceedings of the 4th International Conference on Educational Data Mining (EDM 2011), 34*, 347-348.

Çalışkan, Ö. (2019). Readiness for organizational change scale: Validity and reliability study. Educational Administration: Theory and Practice, 25(4), 663-692.

Saa, A.A. (2016). Educational data mining & students' performance prediction. *International Journal of Advanced Computer Science and Applications*, *7*(5), 212-220.

Özgenel, M., & Bozkurt, B. N. (2019). Two factors predicting the academic success of high school students: Justice in classroom management and school engagement. Educational Administration: Theory and Practice, 25(3), 621-658.

Kesik, F., & Aslan, H. (2020). Metaphoric expressions of the students about the concept of happiness. Educational Administration: Theory and Practice, 26(2), 303-354.