STUDENT MARKS PREDICTOR

A Project Report submitted in the partial fulfillment
of the requirements for the award of degree of

BACHELOR OF TECHNOLOGY

in

Computer Science & Engineering / Information Technology

Submitted by

AYUSH RANJAN SRIVASTAVA
(Roll No. 1752510012)

PRACHI SHARMA
(Roll No. 1752510030)

VANDANA TIWARI
(Roll No. 1752510051)

VARUN KUMAR SINGH
(Roll No. 1752510053)

Under the guidance of

Mr. Ashok Kumar Rai
(Assistant Professor)

Computer Science And Engineering
BUDDHA INSTITUTE OF TECHNOLOGY
(Affiliated to Dr. A.P.J. Abdul Kalam Technical University, Lucknow)
CL-1, Sector-7, GIDA, GORAKHPUR

January, 2021

STUDENT MARKS PREDICTOR

A Project Report submitted in the partial fulfillment
of the requirements for the award of degree of

BACHELOR OF TECHNOLOGY

in

Computer Science & Engineering / Information Technology

Submitted by

AYUSH RANJAN SRIVASTAVA
(Roll No. 1752510012)

PRACHI SHARMA
(Roll No. 1752510030)

VANDANA TIWARI
(Roll No. 1752510051)

VARUN KUMAR SINGH
(Roll No. 1752510053)

Under the guidance of

Mr. Ashok Kumar Rai
(Assistant Professor)

Computer Science And Engineering
BUDDHA INSTITUTE OF TECHNOLOGY
(Affiliated to Dr. A.P.J. Abdul Kalam Technical University, Lucknow)
CL-1, Sector-7, GIDA, GORAKHPUR

January, 2021

CERTIFICATE

Certified that Mr. Ayush Ranjan Srivastava, Ms. Prachi Sharma, Ms. Vandana Tiwari, Mr. Varun Kumar Singh are students of B.Tech. (CS) of BUDDHA INSTITUTE OF TECHNOLOGY, Gorakhpur have carried out the project work presented in this report entitled **"Students Marks Predictor"** for the award of Bachelor of Technology degree from Dr. A.P.J. Abdul Kalam Technical University, Lucknow under my supervision.

Mr. Manish Kr. Gupta
Head of Department
Department of Computer Science & Engineering

Place : Gorakhpur

Date: 29 July 2021

DECLARATION

We declare that the project report entitled "STUDENT'S MARKS PREDICTOR", submitted for the partial fulfillment of the requirement for the Bachelor's Degree in Computer Science & Engineering, to CSE department, Buddha Institute of Technology, GIDA, Gorakhpur, affiliated to Dr. A.P.J. Abdul Kalam Technical University, Lucknow comprises our original work.

AYUSH RANJAN SRIVASTAVA

PRACHI SHARMA

VANDANA TIWARI

VARUN KUMAR SINGH

CSE Department
Buddha Institute of Technology
GIDA, Gorakhpur

Date: 29 July 2021

ACKNOWLEDGEMENTS

I am very much indebted to **Mr. Ashok Kumar Rai**, for his consistent support, excellent guidance, encouragement which enabled me to obtain insight into the project work.

I am grateful to **Manish Kr. Gupta**, for his valuable guidance, suggestions and support which helps me to carry out the project work in time.

I extend my sincere gratitude to **Mr. Sacchidanand Chaturvedi,** for their valuable support to carry out the project work.

I am also thankful to faculty members of CSE Department for their consistent encouragement and support..

I would like to extend my gratitude to my my parents, brothers, relatives and friends for their patience, support and encouragement throughout the work. Lastly

I wish to thank almighty god for showering blessings.

Ayush Ranjan Srivastava
Prachi Shrama
Vandana Tiwari
Varun Kumar Singh

LIST OF FIGURES

FIGURE NO.	TITLE	PAGE NO.
1.1	Machine Learning	12
1.2	Linear Regression	16
1.3	Implementation Flow	28
1.4	Linear Regression Snapshot	29
1.5	SVM Code Snapshot	30
1.6	Workflow	33

LIST OF TABLES

TABLE NO.	TITLE	PAGE NO.
1.1	Error Rate Comparison	26
1.2	Experimental Results	29
1.3	Model Comparision	32

ABSTRACT

This paper focuses on predicting student performance using personalized analytics based on educational databases. Predicting student performance becomes more challenging due to the huge amount of databases. We present two different approaches to improving student achievements. Both approaches are validated on one course which was offered to students of the school of computing between the years of 2019 and 2020. The first approach is based on regression algorithms to predict student performance. Regression is a data mining function that predicts a number. The main goal is to find how well a student can perform in the programming language by predicting grades based on their school background and performance in semester exams. In the model build (training) process, a regression algorithm estimates the value of the dependent variable as a function of the predictors in the build data based on the independent variables. These relationships between predictors and target are summarized in a model, which can then be applied to a different dataset in which the target values are unknown. The Second approach is to find the error rate of regression algorithms by using root mean square error. The obtained results reveal that the school background also plays a major role in predicting grades. Finally, we can identify the students who are at risk and provide better additional training for the weak students.

***Key Words*:** Marks Prediction, Students, Regression, Model,Algorithm.

TABLE OF CONTENTS

CERTIFICATE BY GUIDE(S)	**3**
DECLARATION	**4**
ACKNOWLEDGEMENTS	**5**
LIST OF FIGURES	**6**
LIST OF TABLES	**7**
ABSTRACT	**8**
TABLE OF CONTENTS	**9**
CHAPTER 1: INTRODUCTION	**10-16**
CHAPTER 2: OBJECTIVE	**17-18**
CHAPTER 3: LITERATURE REVIEW	**19-24**
CHAPTER 4: METHODOLOGY	**25-28**
4.1: IMPLEMEMTATION	**25**
4.2: COMPARING APPROACHES	**25**
4.3: ROOT MEAN SQUARE ERROR	**26**
CHAPTER 5: RESULT	**29-31**
CHAPTER 6: CONCLUSION AND FUTURE SCOPE	**32-34**
6.1: CONCLUSION	**32**
6.2: FUTURE SCOPE	**34**
REFERENCES	**35**
PUBLICATION & CERTIFICATES	**36-50**

Chapter 1

INTRODUCTION

Over the past 35 years, a vast amount of knowledge has been accumulated on text mining for Information Retrieval (IR). Using automated text mining algorithms to discover knowledge from natural language texts provides numerous challenges but also offer unique possibilities. One of the most natural forms of storing information is in the form of natural language texts .This can be easily interpreted by a human but it is still a great challenge for computers to derive meaning from this data. However, computers do offer an important advantage over human capabilities: computing power. This means that computers can find patterns, which are non-trivial recurrences, within data faster and more accurate than their human counterpart, but this can only be done if the structure of the data is known. Natural language does contain implicit grammatical structure, but these structures are deeply complex and vary across different languages. The main aim of this project is to use data mining methodologies to study students' performance in the courses. Data mining provides many tasks that could be used to study the student performance. In this research, the classification task is used to evaluate students' performance and as there are many approaches that are used for data classification, the decision tree method is used here. Information like Attendance, Class test, Seminar and Assignment marks were collected from the student's management system, to predict the performance at the end of the semester. This paper investigates the accuracy of data mining classification methods for predicting student performance.

Education is very important issue regarding development of a country. The main objective of educational institutions is to provide high quality education to its students. One way to accomplish this is by predicting student's academic performance and thereby taking early steps to improve student's performance and teaching quality.

This system aims to predict student's marks using linear regression. The idea behind this analysis is to predict the marks of students by their studying hours.Through this project we can determine:

How many hours need to do the study to get 99% marks

If I will do study x() hours per day so how much marks I will get

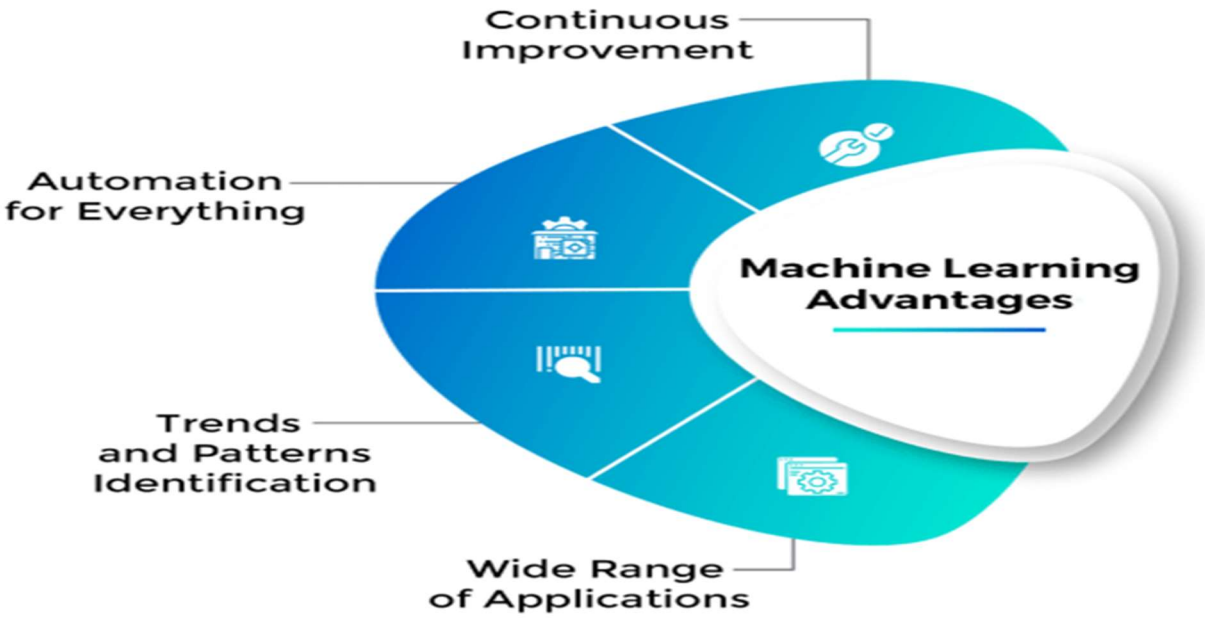
Through these points the school can determine the performance of the student.

To implement this model we are using Jupyter Notebook which is an open web source application. The model is deployed on the python framework called Flask. The data set is taken from Kaggle, which provides data for free.

The economic success of any country highly depends on making higher education more affordable and that considers one of the main concerns for any government. One of the factors that contributes to the educational expenses is the studying time spent by students in order to graduate. For example, the loan debt of the American students has been increased due to the failure of many students in getting graduated on time [1]. Higher education is provided for free to the students in Iraq by the government. Yet, failing of graduating on time costs the government extra expenses. To avoid these expenses, the government has to ensure that the student graduate on time. Machine learning techniques can be used to forecast the performance of the students and identifying the at risk students as early as possible so appropriate actions can be taken to enhance their performance. One of the most important steps when using these techniques is choosing the attributes or the descriptive features which used as input to the machine learning algorithm. The attributes can be categorized into GPA and grades, demographics, psychological profile, cultural, academic progress, and educational background [2]. This research introduces two new attributes that focus on to the effect of using the internet as a learning resource and the effect of the time spent by students on social networks on the students' performance. Four machine learning techniques, fully connected feed forward Artificial Neural Network, Naïve Bayes, Decision Tree, and Logistic Regression, have been used to build the machine learning model. ROC index has been used to compare the accuracy of the four models. The dataset used to build the models is collected from the students at the College Of Humanities during 2015 and 2016 academic years using a survey and the student's grade book. The dataset has the information of 161 students. The activities of this research include feature engineering to create the students dataset, data collecting, data preprocessing, creating and evaluating four machine learning models, and finding the best model and analyzing the results.

Internet has a wide door to an innovative style of gaining knowledge. The amount of information available therein exceeds that of any physical library. There are many uses of

internet but academic purpose conquers the highest desirable position as far as students are concerned. Generally, the college-going students have an immense variety of subjects in the core area which is sometimes difficult for the students to understand. So they have chosen internet where there is a lot of information provided about the area of their study. Nowadays internet plays a major role in and around the people and makes them depend on it for each and everything. This is mainly affected by student's community where they use internet for various purpose like getting notes for studies, doing the assignment and other related activity and also for communications. Almost every engineering student is required to take the corresponding engineering course which is highly impacted over the area which they have chosen. The courses contain all the essential and basic subjects from all the engineering subjects. So there is a chance of engineering students using internet for the various purpose including the entertainment along with the studies. Prediction of student's academic performance has long been observed and considered to be an important research topic in many disciplines because it aids the educator and learners. Educators can use the predicted results to identify the number of students who will do well, averagely or poorly in a particular class to take measures accordingly. Many areas can be chosen to arrive for the prediction of student's academic performance.



Measuring academic performance of students is more challenging since student's academic performance depends on diverse factors like personal, socio-economic, psychological and other environmental variables. The scope of this paper is to predict the student grades using the best algorithms with high accuracy. There is a critical need to develop innovative approaches that ensure students become graduates in a timely fashion and are well trained and workforce ready in their field of study. From the paper, The higher secondary education is important in a student‟s life because it is one of the main factors that are going to decide the future of the student based on their mark in the higher secondary examination, they are going to get the college education and the field of information technology based on their knowledge in programming. We present methods that draw on techniques from recommender systems to accurately predict students‟ programming course grades. Data mining provides many tasks that are used to predict the student‟s performance. In this paper, the regression task is used to evaluate the performance of a student and as there are many approaches that are used for data regression, the multi-linear regression and SVM, random forest and regression tree methods was used here. For this study, recent real world data has collected. Information[refer table1] like the medium of study, syllabus, intermediate background etc., were collected. mathematics marks are also collected to know the logical ability of the student and English marks to know the communication and understanding level of the student This study is more useful for identifying weak students in the programming at the beginning of the semester and the identified students can be assisted by the educators so that their performance is better in future. This study investigates the accuracy of some regression techniques for predicting performance of a student.

Today every educational institution handles and deals with large amount of student data which can be beneficial for a number of reasons. One of the important application of such data is predicting student performance. Such a prediction can be useful not only for the students but also for teachers/mentors. Mentors can provide special assistance to the students who are on the verge of failing. In order to determine which category a student lies, such data can be quite helpful. This application can be used by any prominent school or colleges. It can be used to predict the pointer ranges or percentage range for future semester exams. These ranges can be predicted using a number of data mining algorithms such as classification algorithms, rule-based algorithms, ensemble methods, and neural networks. The main aim of this project is the selection of features that show a strong relationship with a target attribute that is to be predicted from a high

dimensional dataset. We have evaluated and compared the number of algorithms such as decision tree, random forest, support vector machine, naive Bayes and neural networks by applying them on the dataset The rest of the paper provides an explanation on nature of neural networks along with the results of our evaluation.

Engineering dynamics is a fundamental sophomore-level course that nearly all engineering students majoring in aerospace, mechanics, and civil engineering are required to take (Ibrahim, 2004; Rubin & Altus, 2000; Zhu, Aung, & Zhou, 2010). The course cultivates students' ability to "visualize the interactions of forces and moments, etc., with the physical world" (Muthu & Glass, 1999). It is an essential basis for many advanced engineering courses such as advanced dynamics, machine design, and system dynamics and control (Biggers, Orr, & Benson, 2010; Huang & Fang, 2010). However, engineering dynamics is also regarded as one of the most challenging courses for undergraduates (Self, Wood, & Hansen, 2004).

The course requires students to have solid mathematical skills and a good understanding of fundamental concepts and principles of the field. Many students perform poorly in or even fail this course. The mean score of the final comprehensive exam in the dynamics class is below 70 out of 100 at Utah State University in 2009. On average, only 53% of the engineering dynamics questions were answered correctly in the Fundamentals of Engineering (FE) Examination in U.S. in 2009 (Barrett et al., 2010). Pedagogical and instructional interventions can improve student academic performance by building up a more solid foundation and enhancing students' learning of engineering concepts and principles (Etkina, Mestre, & O'Donnell, 2005). For example, interventional process of constructing knowledge can help students to relate (and, later, integrate) new information to prior knowledge and achieve complex learning goals 2 (Etkina et al., 2005; Royer, 1986).

Students may be able to construct a hierarchical structure of knowledge and gain better understanding of the principles after training (Dufresne, Gerace, Hardiman, & Mestre, 1992). To achieve better learning outcomes, the choice of instructional interventions must take into account the diverse academic backgrounds and varied performance of students in relevant courses because each student will have a different reaction to them. For example, a study conducted by Palincsar and Brown (1984) showed that implicit instructions could help average students to achieve greater

understanding and success in class, whilst the same teaching method would hinder the learning process of lowerperformance students.

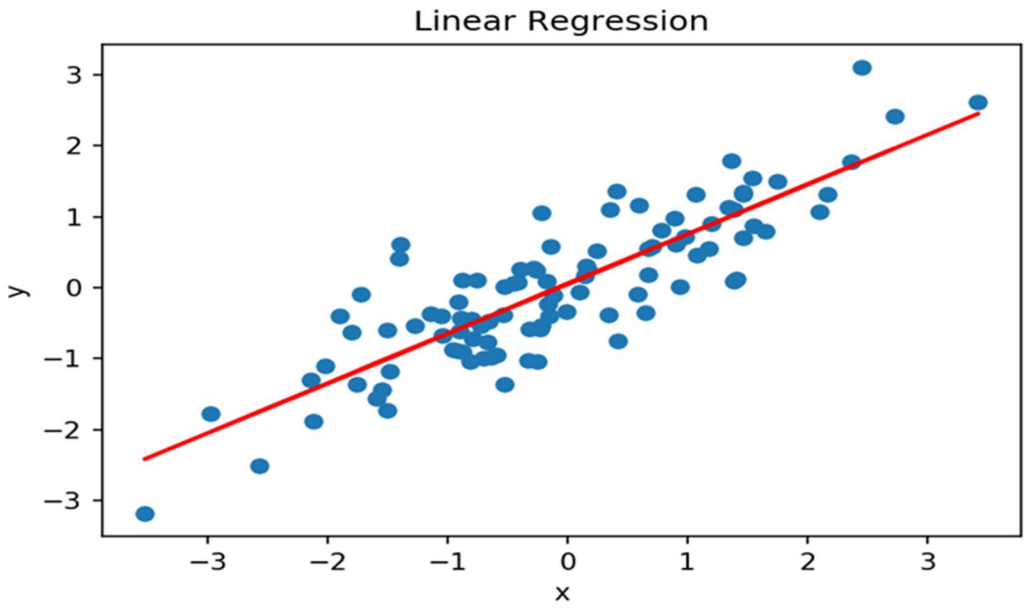
Many education researchers and instructors have made extensive efforts in constructing effective models to predict student academic performance in a class (Emerson & Taylor, 2004; Holland, James, & Richards, 1966; Kotsiantis, Pierrakeas, & Pintelas, 2003; Lowis & Castley, 2008; Pittman, 2008). The results of these predictive models can help the instructor determine whether or not a pedagogical and instructional intervention is needed. For example, the instructor can determine how well, or how poorly, students may perform in the class. Then, appropriate pedagogical and instructional interventions (for example, designing an innovative and effective teaching and learning plan) can be developed and implemented to help these academically at-risk students.

Variables such as students' prior knowledge and prior achievement contribute significantly to the prediction accuracy of the model that predicts student academic 3 performance (Fletcher, 1998). Thompson and Zamboanga (2003) concluded that prior knowledge and prior achievement (such as GPA) are significant predictors of student academic performance in a class and represented 40% to 60% of variance in learning new information (Dochy, 1992; Tobias, 1994).

However, if prior knowledge is insufficient or even incorrect, learning and understanding of new information will be hindered (Dochy, Segers, & Buehl, 1999). Psychological variables, such as goals, are controversial predictors for academic achievement. Some studies found that psychological variables were significant predictors (Cassidy & Eachus, 2000) and increased the amount of variance explained for academic achievement (Allen, Robbins, & Sawyer, 2010). However, other studies discovered that the change in explained variance was not significant when psychological variables were included (French, Immekus, & Oakes, 2005). It has been suggested that the variables have different effects on different learning subjects (Marsh, Vandehey, & Diekhoff, 2008). Identifying and choosing effective modeling approaches is also vital in developing predictive models. Various mathematical techniques, such as regression and neural networks, have been employed in constructing predictive models. These mathematical techniques all have advantages and disadvantages. For example regression, one of the most commonly used approaches to constructing predictive models, is easy to understand and provides explicit

mathematical equations. However, regression should not be used to estimate complex relationships and is susceptible to outliers because the mean is included in regression formulas.

On the other hand, neural networks can fit any linear or nonlinear function without specifying an explicit mathematical model for the 4 relationship between inputs and output; thereby, it is relatively difficult to interpret the results. In a recent work by Fang and Lu (2010), a decision-tree approach was employed to predict student academic achievement in an engineering dynamics course. Their model (Fang & Lu, 2010) only generates a set of "if-then" rules regarding a student's overall performance in engineering dynamics. This research focused on developing a set of mathematical models that may predict the numerical scores that a student will achieve on the dynamics final comprehensive exam.



Chapter 2

OBJECTIVE

The Main Objectives of this Study:

- Predict the student's success or failure
- Predict the final grade

The problem of the student final grade prediction in a particular course has recently been addressed using data mining techniques. Researchers usually examine study-related records, e.g. the age, gender and the field of study because of their easy availability in university information systems. The most typical way how to obtain such data is to conduct questionnaires but it tends to have a lower response rate. Therefore, only the data originated from the college are considered for our experiments.

As stated previously, student low academic performance in the engineering dynamics course has been a long-standing problem. Before designing and implementing any pedagogical and instructional interventions to improve student learning in engineering dynamics, it is important to develop an effective model to predict student academic performance in this course so the instructor can know how well or how poorly the students in the class will perform. This study focused on developing and validating mathematical models that can be employed to predict student academic performance in engineering dynamics.

The goal of this study is to develop a validated set of mathematical models to predict student academic performance in engineering dynamics, which will be used to identify the academically-at-risk students. The predicted results were compared to the actual values to evaluate the accuracy of the models. The three objectives of the proposed research are as follows:

1. Identify and select appropriate mathematical (i.e., statistical and data mining) techniques for developing predictive models.

2. Identify and select appropriate predictor variables/independent variables that can be used as the inputs of predictive models.

3. Validate the developed models using the data collected in four semesters and identify academically-at-risk students.

Three research questions have been designed to address each research objective of the study. These three research questions include:

1. How accurate will predictions be if different statistical/data mining techniques such as multiple linear regression (MLR), multilayer perceptron (MLP) networks, radial basis function (RBF) networks, and support vector machine (SVM) are used?

2. What combination of predictor/independent variables yields the highest prediction accuracy?

3. What is the percentage of academically at-risk students that can be correctly identified by the model?

Student academic performance is affected by numerous factors. The scope of the research is limited to the investigation of the effects of a student's prior achievement, domain-specific prior knowledge, and learning progression on their academic performance in the engineering dynamics course. Psychological factors, such as selfefficacy, achievement goals, and interest, were not included in constructing predictive models. In the future study, psychological factors will be considered for developing the predictive models and further interviews will be conducted to confirm the identified academically at-risk students and diagnose if those students have psychology-related issues and problems in addition to having academic problems. How to effectively apply the predictive models will also be examined in the future study.

A variety of commonly used literature databases were examined, including the Education Resources Information Center, Science Citation Index, Social Science Citation Index, Engineering Citation Index, Academic Search Premier, the ASEE annual conference proceedings (1995-2011), and the ASEE/IEEE Frontier in Education conference proceedings (1995-2011). The only paper on predictive modeling of student academic performance in the engineering dynamics course is done by Fang and Lu (2010). However, not only did their work use only one modeling approach (a decision tree approach), but their work took into account only student prior domain knowledge.

Chapter 3

LITERATURE REVIEW

Samrat Singh, Dr. Vikesh Kumar [1] .Data Mining is a powerful tool for academic performance. Educational Data Mining is concerned with developing new methods to discover knowledge from educational database and can used for decision making in educational system.

M. Goyal and R. Vohra [2] .Data analysis plays an important role for decision support irrespective of type of industry like any manufacturing unit and educations system. If data mining techniques such as clustering, decision tree and association are applied to higher education processes, it would help to improve students performance, their life cycle management, selection of courses, to measure their retention rate and the grant fund management of an institution.

Jason Brownlee [3]. After you have found a well performing machine learning model and tuned Sample output to test PDF Combine only P a g e | 7 it, you must finalize your model so that you can make predictions on new data.

Neelam Naik & Seema Purohit [4] . The quality higher education is required for growth and development of country. Professional education is one of the pillars of higher education. Data mining techniques aim to discover hidden knowledge in existing educational data, predict future trends and use it for betterment of higher educational institutes as well as students.

Alaa M.El-Halees, Mohammed M. Abu Tair. [5] Educational data mining concerns with developing methods for discovering knowledge from data that come from educational domain. In this paper we used educational data mining to improve graduate students' performance, and overcome the problem of low grades of graduate students.

B.K. Bharadwaj and S. Pal [6] .Now-a-days the amount of data stored in educational database increasing rapidly. These databases contain hidden information for improvement of students' performance. The performance in higher education in India is a turning point in the academics for all students. This academic performance is influenced by many factors, therefore it is essential to develop predictive data mining model for students' performance so as to identify the difference

between high learners and slow learners student. In the present investigation, an experimental methodology was adopted to generate a database.

Suchita Borkar, K. Rajeswari [7] .Education Data Mining is a promising discipline which has an imperative impact on predicting students' academic performance. In this paper, student's performance is evaluated using association rule mining algorithm. Research has been done on assessing student's performance based on various attributes. In our study important rules are generated to measure the correlation among various attributes which will help to improve the student's academic performance.

Randhir Singh, M.Tiwari, Neeraj Vimal [8]. Educational institutions are important parts of our society and playing a vital role for growth and development of nation and prediction of student's performance in educational environments is also important as well. Student's academic performance is based upon various factors like personal, social, psychological etc.

D.Magdalene Delighta Angeline [9].The objective of the educational institution that is producing good results in their academic exams can be achieved by using the data mining techniques which can be applied to predict the performance of the students and to impart the quality of education in the educational institutions. Data mining is used to extract meaningful information and to develop relationships among variables stored in large data set.

Mrs. M.S. Mythili, Dr. A.R.Mohamed Shanavas [10] . In recent years, the analysis and evaluation of students‟ performance and retaining the standard of education is a very important problem in all the educational institutions. The most important goal of the paper is to analyze and evaluate the school students‟ performance by applying data mining classification algorithms in WEKA tool.

S. Anupama Kumar and Dr. Vijayalakshmi M.N [11] .Educational data mining is used to study the data available in the educational field and bring out the hidden knowledge from it. Classification methods like decision trees, rule mining, Bayesian network etc can be applied on the educational data for predicting the students behavior, performance in examination etc.

Prediction in single course :-

Title : Predicting student academic performance in an engineering dynamics course :

A comparison of four types of predictive mathematical models.

Author : S. Huang, N. Fang Year : 2013

Description: The work presented in this paper does not predict performance at degree level but at a course level. However, it is interesting as it suggests a kind of upper bound for the accuracy that can be achieved when predicting performance at the end of a degree. They employed four mathematical models namely multiple linear regression, multilayer perception networks, radial basis functions and support vector machines to predict student's academic performance in an engineering dynamics course.It has worked on the data of 323 undergraduate students who took dynamics course at Utah State University in four semesters. Their predictor variables were the students' cumulative GPA; grades earned in four pre-requisite courses i.e. statistics, calculus I, calculus II and physics; and scores on three dynamics midterm examinations. The paper used six combinations of predictor variables to develop a total of 24 predictive mathematical models. For all the four models, they achieved an average prediction accuracy of 81%– 91%. This work shows that previous marks can predict the grade of a course with high accuracy.

Prediction by Using Data Mining Classification :-

Title : Student Performance Prediction by Using Data Mining Classification Algorithms, International Journal of Computer Science and Management Research.

Author : DorinaKabakchieva\\ Year : 4 November 2012\\

Description: The research is focused on the development of data mining models for predicting student performance, based on their personal, higher secondary and university International Journal of Pure and Applied Mathematics Special Issue 231 performance by using One Rule Learner, Decision Tree, Neural Network and K-Nearest Neighbour classification methods and found that Neural Network model predicts with high accuracy than other three models. The present study differs from other works in various aspects. First, by using the predicting values that effect the programming skills most are only considered, no socioeconomic data is considered. It also gives some limit to what can be achieved when predicting graduation performance. Indeed the predictors include midterm examinations that can be expected to correlate well with the final exam of the course, more than marks of single courses with the graduation mark.

Factors Affecting Student's Performance

Title: Factors Affecting Student Performance : A Case of Private Colleges Bangladesh e-Journal of Sociology

Author : Hijazi, S. T., and Naqvi, R.S.M.M Year : 2006.

Description: A study on the student's performance by selecting a sample of 300 students (225males, 75 females) from a group of colleges under Punjab University in Pakistan. The conclusion that was stated as Student's attitude towards attendance in class, hours spent in study on daily basis after college, students' fathers income, mother's age and mother's educational qualification are significantly related to student performance by means of simple linear regression analysis, it was found that the factors like mother's educational qualification and student's father's income were highly interconnected with the student academic performance. The present study differs from other works in various aspects. First, by using the predicting values that effect the programming skills most are only considered, no socio-economic data is considered.

Analyzing University Data for Determining Student Profiles.

Title : Analyzing University Data for Determining Student Profiles and Predicting Performance, Proceedings of the 4th International Conference on Educational Data Mining Scheduling Independent Tasks.

Author : D. Kabakchieva , K. Stefanova, V.Kisimov Year : July 6-8, 2011.

Description: The study predicts student's university performance by using student's International Journal of Pure and Applied Mathematics Special Issue 232 personal and pre-university characteristics. They experimented on the data of 10330 students of a Bulgarian educational society, each student being described by 20 attributes. (e.g., gender, date of birth, living area, and country, place and total score from previous classes, current semester scores, total university grades, etc.). In the papers they have applied different data mining algorithms such as the decision tree C4.5, Naive Bayes, Bayesian networks, K-nearest neighbors (KNN) performance at the end of the degree.It is easier to gather marks of students than their socioeconomic data. Therefore if a reasonable prediction can be reached without socio-economic data, it makes the implementation of a performance support system in a university easier. If courses can be identified with a major impact on graduation performance, then measures can be taken at the level of those courses, making also the implementation of a performance support system easier. In this study, the

performance of a student at the end of the degree will be a class A, B, C, D or E, which represents the interval in which her/his final mark lies. This allows to differentiate between strong and weak students.

The various techniques for prediction using different data mining techniques in the discussed below. There are many applications and area where prediction can be applied to predict some useful information. One such application/ area is Healthcare prediction. In the healthcare environment, there are many diseases which can be predicted before the analysis.

Regression Model for Predicting Engineering Students Academic Performance

Published By: Blue Eyes Intelligence Engineering & Sciences Publication

Retrieval Number: F1015376S19/19©BEIESP

Some of the researchers focused on heart disease and the discovery of rules in medical data to predict the existence of disease [2] by using a data mining technique called association rule mining which maps the medical data to the transaction and forms association rules. Association rule mining can be used to extract the relationship between the item sets. Some authors presented a study which compares the data mining techniques used by different researchers for prediction of heart disease [1] and tabulated the accuracy of using different data mining techniques such as Decision tree, NaiveBayes, JRip, CART etc and finally concluded with the best data mining technique with number of correctly classified samples for further understanding. Our present study is mainly focused on the predicting the student's academic performance by using data mining algorithms. Here are some researchers who conveyed on the usage of data mining algorithms for the student's academic performance. Oloruntoba et al(2017) proposed a study based on the prediction of student's academic performance using data mining techniques. The study identifies the relationship between the student's academic performance and their final scores. The model is built using the support vector machine technique and it was compared with other classification algorithms. The final result has shown that the accuracy obtained through SVM classification is much greater than the other algorithms.

A research project done by Dorina Kabakchieva(2012)Bulgarian University mainly focused on the usage of data mining techniques for university management. The results achieved by selected data mining algorithms for classification doesn't reveal any worthy outcomes.

Zlatko(2010) explored the student's demographic attributes along with their corresponding study environment which is used for the analysis of these factors affecting their success rates in their course of the study. The results show that the important factors to distinguish between successful and unsuccessful students and for predicting the category of students, the CART algorithm is used which produce an overall percentage of 60.5%. It does not contain adequate information for distinguishing between successful and unsuccessful students. Some authors explored the difference between data mining techniques [5] [6][7] and explored the comparison of the methods for educational learners and provided a better predictive model among all the data mining techniques. Some researchers focused on the use of classification algorithms and provided the comparison of all the classifiers in their paper [10]. In all these works, the authors concentrated on the students' performance prediction using different data mining techniques to carry out the analysis but the work mainly focuses on building and validating the regression technique of undergraduate students of Engineering stream who use the internet for various purposes. The rest of the paper is unfolded as follows, Section 3 discusses the objective and techniques in detail followed by the data collection and how research method is applied for the above-proposed problem and in last Section, the detailed explanations of the performance metrics are discussed. Finally, it ends with the conclusion along with the future work of the paper.

Chapter 4

METHODOLOGY

Student Characteristics Regression is the most often used technique for student performance prediction. Researchers usually examined study-related (SR) data . Our study-related data contained attributes such as higher secondary school background, the medium of study, syllabus, mathematics and English marks. We built a multi linear regression for programming course based on the training set and evaluated the results using various models (SVM, RF, Decision tree).The method that achieved best results was subsequently validated on the test set. Grade prediction The regression is a commonly used technique for student grade prediction. And also SVM Reg., Random Forest, decision Tree are used to validate the models each other. The baseline model predicts the programming marks of the training set of a given instances of attributes. In addition to accurately predicting student‟s performance, the multi-regression model can be used to analyze how the different features contribute to the predicted grades and thus gain some insights about the student‟s performance. For a proper analysis of the estimated model parameters, it is more convenient that all the attributes have non-negative values which will make all the model‟s components to contribute additively to the predicted grades.

6.1 IMPLEMENTATION:

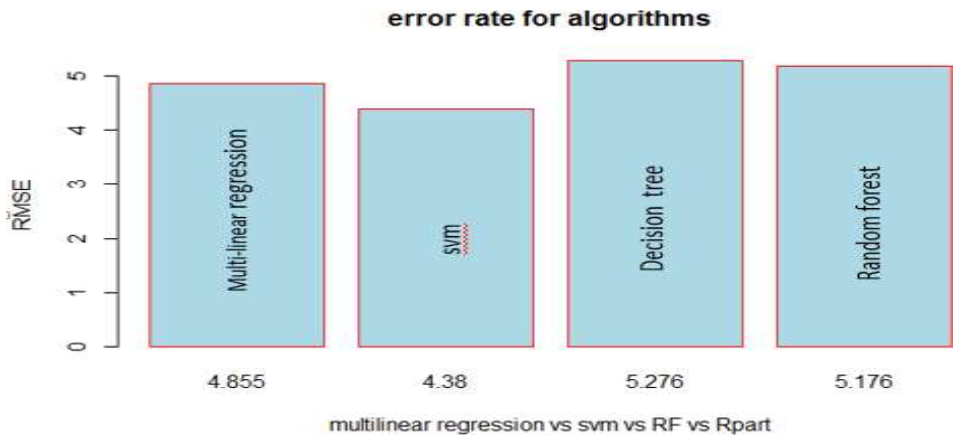
Multi-linear regression model:

Y = 0.0267(m)+0.00444(e) + (-0.25824(sicse)/0.59825(sstate)) - 0.49231(llocal)+0.56192(bckmaths) – 0.59242(csyes)+1.79781(int) lm(formula = c ~ m + e + s + l + bck + cs + int, data = trai)

6.2 COMPARING APPROACHES:

In comparison with the method using all grades, both approaches had positive effects on the number of calculations. RMSE gives the standard deviation of the model prediction error. A smaller value indicates better model performance. In this case, residuals of models are considered and RMSE is calculated for both multi linear regression and SVM has almost same error rate so

any of the models can be used to test the data and predict the values. Fig represents the RMSE rate for algorithms



6.3 ROOT MEAN SQUARE ERROR

Root mean square is calculated by identifying residuals and squaring residuals and finding mean of the squared residuals and finally calculating the root of the mean squared gives RMSE. By using RMSE we can identify that which algorithm gives high accuracy results for the given datasets. In this case we observe that SVM, linear model has least RMSE. Experimental results represents the graphical output for four algorithms in a test data.

The goal of this study was to develop a validated set of statistical and data mining models to predict student academic performance in an engineering dynamics course.

The three objectives of this research were as follows:

1. Identify and select appropriate mathematical (i.e., statistical and data mining) techniques for constructing predictive models.
2. Identify and select appropriate predictor variables (i.e., independent variables) that can be used as inputs for predictive models.
3. Validate the developed models using the data collected during multiple semesters to identify academically-at-risk students.

Three research questions were designed to address each research objective:

1. How accurate will predictions be if different statistical and data mining modeling techniques such as traditional multiple linear regression, MLP networks, RBF networks, and SVM are used?
2. What particular combination of predictor variables will yield the highest prediction accuracy?
3. What is the percentage of academically-at-risk students that can be correctly identified by the models?

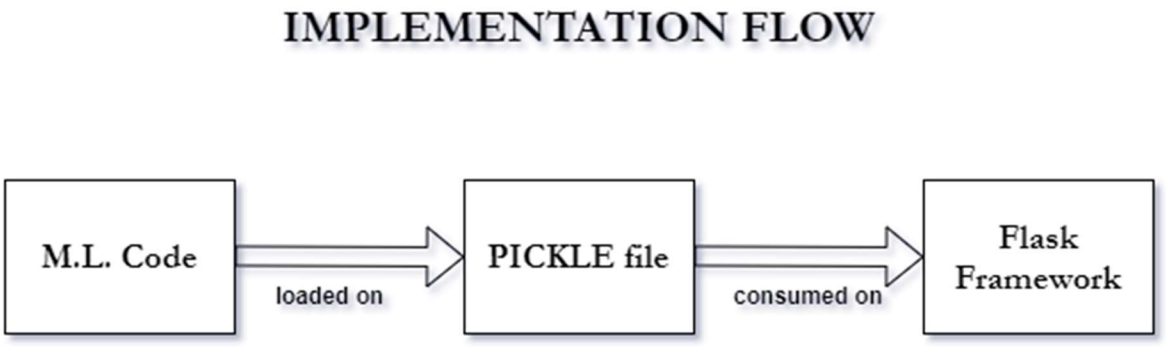
Overall Framework : Cabena, Hadjinian, Stadler, Verhees, and Zanasi (1997) created a five-stage model for data mining processes, including the determination of business objectives, data preparation, data mining, results analysis, and knowledge assimilation. Feelders, Daniels, and Holsheimer (2000) illustrated six stages of the data mining process, including defining the problem definition, acquiring background information, selection and preprocessing of data, analyzing and interpreting, as well as reporting acquired data. Pittman (2008) proposed a data mining process model for education, which includes determining a dataset based on student retention rates, domain knowledge, and data availability.

Based on extensive literature review and the experience in teaching engineering dynamics, data regarding students' prior achievement, domain-specific prior knowledge, and learning progression were collected. Eight variables (X1, X2, …, X8) were selected as the candidate predictor/independent variables of the predictive models. X1 (cumulative GPA) indicates prior achievement. X2~ X5 (grades earned in the prerequisite courses for engineering dynamics) indicate prior domain knowledge. X6~X8 (grades earned from three engineering dynamics mid-term exams) indicate learning progression in this particular course. Data collected from four semesters in Fall 2008-Spring 2011 were used to develop and validate the models.

The reasons for selecting these particular variables are discussed below.

- X1 (cumulative GPA) was included because it is a comprehensive measurement of a student's overall cognitive level.

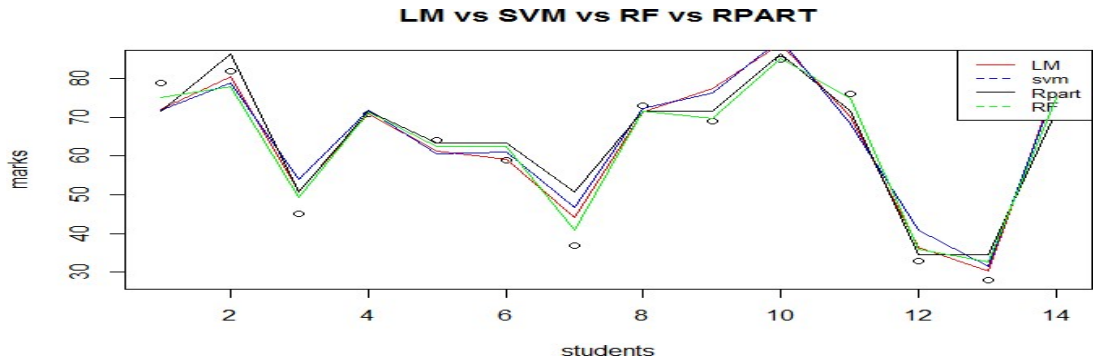
- X2 (statics grade) was included because numerous concepts of statics (such as free-body diagram, force equilibrium, and moment equilibrium) are employed throughout the dynamics course.
- X3 and X4 (calculus I and II grades) are an accurate measurement of a student's mathematical skills needed to solve calculus-based dynamics problems.
- X5 (physics grade) was used to measure a student's basic understanding of physical concepts and principles behind various dynamics phenomena.
- X6 (score of dynamics mid-term exam #1) measures student problem-solving #1 #2 #3 #4 0 25 50 75 100 125 The Number of Students Semester No. MAE CEE Other 17.2% 26.5% 56.3% 27.6% 34.5% 37.9% 15.1% 37.7% 47.2% 11.8% 23.5% 64.7% 44 skills concerning "kinematics of a particle" and "kinetics of a particle: force and acceleration."
- X7 (score of dynamics mid-term exam #2) measured student problem-solving skills concerning "kinetics of a particle: work and energy" and "kinetics of a particle: impulse and momentum." • X8 (score of dynamics mid-term exam #3) is a measurement of student problem-solving skills on "planar kinetics of a rigid body" and "planar kinetics of a rigid body: force and acceleration."



Chapter 5

RESULT

EXPERIMENTAL RESULTS :



Applying linear model and Linear model coefficients

```
In [15]: # y = m * x + c
         from sklearn.linear_model import LinearRegression
         lr = LinearRegression()
         lr.fit(X_train,y_train)#linear regression

Out[15]: LinearRegression(copy_X=True, fit_intercept=True, n_jobs=None, normalize=False)

In [16]: lr.coef_

Out[16]: array([[3.93571802]])

In [17]: y_pred  = lr.predict(X_test)
```

y_pred

```
In [18]: lr.intercept_

Out[18]: array([50.44735504])

In [19]: pd.DataFrame(np.c_[X_test, y_test, y_pred], columns = ["study_hours", "student_marks_original","student_marks_predicted"])
```

Out[19]:

	study_hours	student_marks_original	student_marks_predicted
0	8.300000	82.02	83.113815
1	7.230000	77.55	78.902596
2	8.670000	84.19	84.570030
3	8.990000	85.46	85.829460
4	8.710000	84.03	84.727459
5	7.700000	80.81	80.752384
6	5.690000	73.61	72.841591
7	5.390000	70.90	71.660875
8	5.790000	73.14	73.235162

SVM model coefficient

```
> predsvm

Call:
svm(formula = c ~ m + e + s + l + bck + cs + int, data = trai)


Parameters:
   SVM-Type:  eps-regression
 SVM-Kernel:  radial
       cost:  1
      gamma:  0.1111111
    epsilon:  0.1


Number of Support Vectors:  414
```

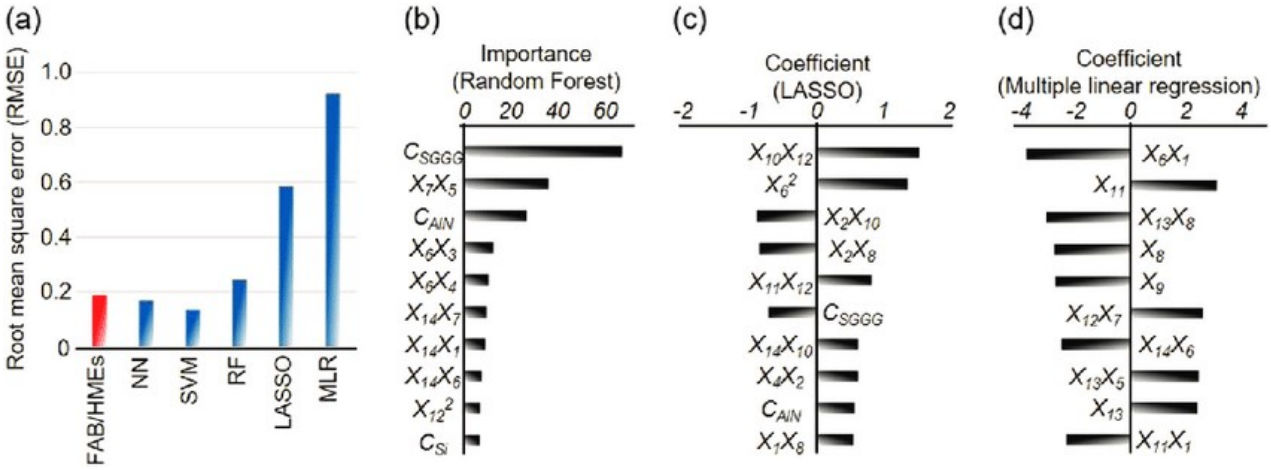
The very first phase in any system developing life cycle is preliminary investigation. The feasibility study is a major part of this phase. A measure of how beneficial or practical the development of any information system would be to the organization is the feasibility study. The feasibility of the development software can be studied in terms of the following aspects: 1. Operational Feasibility: The Application will reduce the time consumed to maintain manual records and is not tiresome and cumbersome to maintain the records. Hence operational feasibility is assured. 2. Technical Feasibility: Minimum hardware requirements: 1.66 GHz Pentium Processor or Intel compatible processor. 1 GB RAM. 80 MB hard disk space. 3. Economic feasibility: Once the hardware and software requirements get fulfilled, there is no need for the user of our system to spend for any additional overhead. For the user, the Application will be economically feasible in the following aspects: 15 The Application will find out the more efficient algorithm to predict the student performance. Hence reducing the extra cost used on the less efficient algorithm. Our Application will reduce the time that is wasted in manual processes.

Research Question #1 : How accurate will predictions be if different statistical and data mining modeling techniques such as traditional MLR, MLP networks, RBF networks, and SVM are used?

A total of 24 predictive models have been developed by using MLR, MLP, RBF, 97 and SVM techniques. The prediction accuracy of MLP models remains nearly unchanged in spite of the change in relevant parameters, such as the maximum training epochs. The initial value of these parameters does not significantly affect the prediction accuracy of MLP and RBF models. The prediction accuracy of SVM models is affected by changing the penalty factor C and the width of kernel 2 σ . In cases in which all above-mentioned parameters are optimized, and based on the

average prediction accuracy and the percentage of accurate predictions, the order of the overall prediction accuracy of the four types of models is:

RBF MLP < MLR < SVM



Research Question #2 : What combination of predictor/independent variables yields the highest prediction accuracy?

According to the combinations of predictors, the 24 models are grouped into the following six sets:

- Models using X1 as predictors
- Models using X1, X2, X3, X4, and X5 as predictors
- Models using X6 as the only predictor
- Models using X1, X2, X3, X4, X5, and X6 as predictors
- Models using X1, X2, X3, X4, X5, X6, and X7 as predictors
- Models using X1, X2, X3, X4, X5, X6, X7, and X8 as predictors

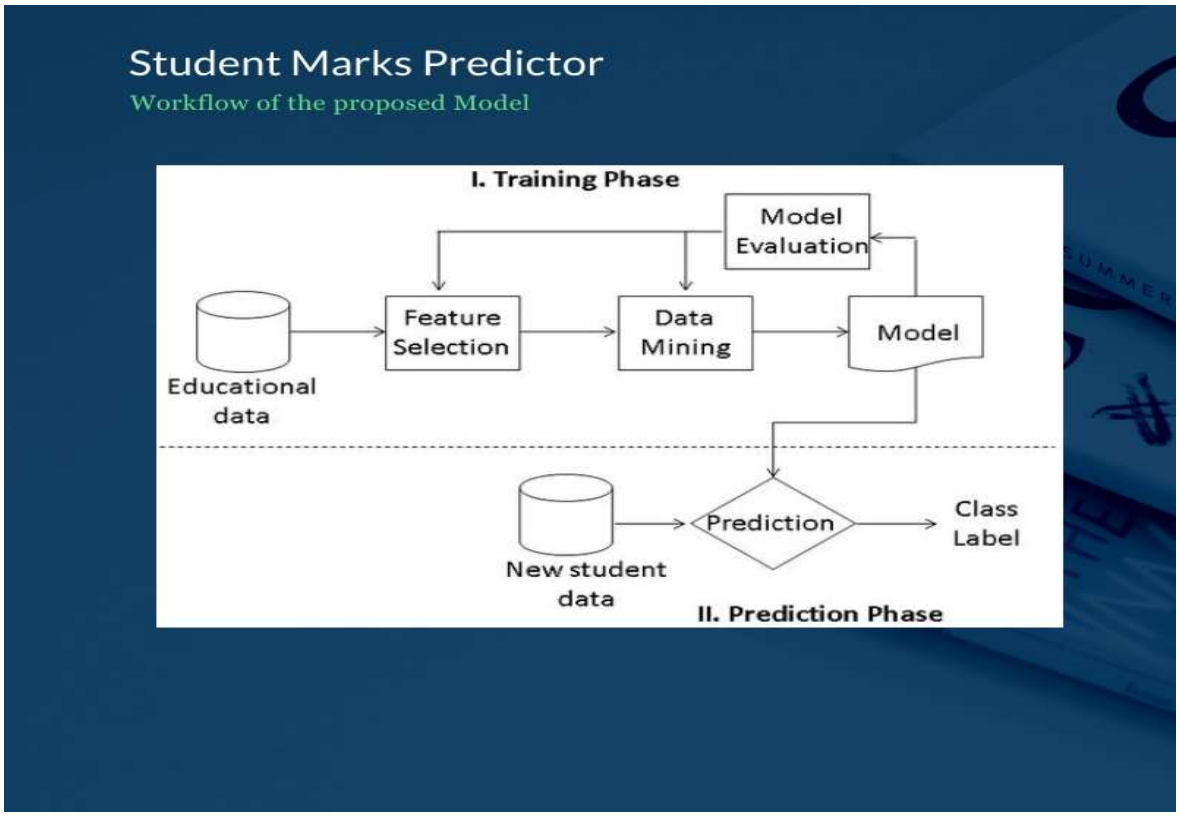
Chapter 6

CONCLUSIONS AND FUTURE SCOPE

6.1 CONCLUSIONS

Predicting student‟s performance would boost the results of student‟s grades and gives teachers a better approach for teaching the students who are at risk of failure. Regression models,tree based models and created to make the best predictions with high accuracy. The basic idea is to increase the efficiency of the prediction results using various algorithms . Thus by finding the RMSE, **we observed that SV, linear model gives optimum results.**

Modular Description



Working on Data :

This module is used for accessing the raw data. This not only accessed the data, but also filled many flaws that existed in the raw data. This is also used in extracting the needed data and exempting the less needed ones.

Finding the best algorithm :

Using the data on various machine learning models to find out the much suited ones. This checked the errors and accuracy of each algorithms so to compare them, to find the best suited one.

Testing our model :

This module trains and test our data, by passing some sample data so to calculate their accuracy in every conditions.

Implementing our Model :

After the model is converted in .pkl file, we implemented our model using flask.

Comparison of Different Modeling Techniques :

The following observations are made:

- In internal validation, SVM models have relatively low APA, but relatively high PAP.
- RBF models yield the lowest average PAP among the four types of models in internal validation.
- Although MLP models generate good APA in external validation, RBF and SVM models outperform MLP models in terms of PAP.
- RBF and SVM 91 models have the nearly the same level of performance in terms of APA and PAP.
- The MLP models have the lowest performance among the four types of models based on the data collected in this study.

Identifying Academically At-Risk Students :

One of the purposes of this study is to identify academically at-risk students. Tables 18-21 show the percentage of academically at-risk students that have been correctly identified by the four types of predictive models. A cell in the table is called a "good cell" if the value in it is larger than 50, which means that more than 50% of academically at-risk students are correctly identified by the model. In Tables 18- 21, there are a total of 19 "good cells" which are highlighted in bold.

Comparison of different combinations of predictors: The models with X1~X8 as predictors yield nine good cells. The models with X1~X7 and X1~X6 as predictors have four good cells. The average percentage of academically at-risk students correctly identified in Semesters #2-#4 (external validation) is 58.8% for models using X1~X8 as predictors, 41.2% for models using X1~X7 as predictors, and 40.9% for models using X1~X6 as predictors. Comparison of different modeling techniques: Both RBF and SVM models generate seven good cells. However, SVM Model #19 fails to correctly identify any academically at-risk student in Semester #4. On average, RBF models correctly identify 64.1% of 92 academically at-risk students in Semester #2, 46.7% of those students in Semester #3, and 28.1% in Semester #4. SVM models identify 64.1% of those students in Semester #2, 44.7% in Semester #3, and 10.5% in Semester #4. Table 23 shows an example of identifying academically at-risk students

6.1 FUTURE SCOPE OF THE WORK

At this point of time we have made this by seeing the pandemic situation. But it can be used on wider perspective.

- **a)** For finding out weaker students out of the class so that their upcoming performance could get increased.
- **b)** For various coaching and college institution to increase their performance of students every year.

References

1. S. Huang, N. Fang, “Predicting student academic performance in an engineering dynamics course: A comparison of four types of predictive mathematical models,” Computer and Education, 2013.
2. DorinaKabakchieva,” Student Performance Prediction by Using Data Mining Classification Algorithms”, International Journal of Computer Science and Management Research Vol 1 Issue 4 November 2012.
3. D. Kabakchieva , K. Stefanova, V. Kisimov, Analyzing University Data for Determining Student Profiles and Predicting Performance, Proceedings of the 4th International Conference on Educational Data Mining, Eindhoven, the Netherlands. July 6-8, 2011.
4. D. Kabakchieva, Predicting Student Performance by Using Data Mining Methods for Classification, Cybernetics and Information Technologies, vol. 13, No. 1, pp. 61- 72, 2013.
5. Hijazi, S. T., and Naqvi, R.S.M.M.,” Factors Affecting Student‟s Performance: A Case of Private Colleges”, Bangladesh eJournal of Sociology,2006.
6. S. Haykin, Neural Networks: A comprehensive Foundation. 2nd ed. Prentice Hall, Upper Saddle River, New Jersey, 1999, p.157, 171, 184.
7. P. Golding, S. McNamarah, “Predicting Academic Performance in the School of Computing & Information Technology (SCIT),” Proceedings of 35th ASEE /IEEE Frontiers in Education Conference, 2005.
8. P. Golding, O. Donaldson, “Predicting Academic Performance”, Proceedings of 36th ASEE /IEEE Frontiers in Education Conference, 2006.
9. Ventura, S., Romero, C., López, M.-I., and Luna, J.-M. 2013.Predicting students' final performance from participation in on International Journal of Pure and Applied Mathematics Special Issue 236 line discussion forums. Computers & Education 68, 458-472.