IJARSCT

ISSN (Online) 2581-9429

IJARSCT
Impact Factor: 6.252

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

Volume 2, Issue 4, May 2022

Student Performance Prediction using Support Vector Machine

Prathmesh Jadhav, Abhishek Toshniwal, Yash Golawar, Pratik Khomhane
Sinhgad College of Engineering, Pune, Maharashtra, India

Abstract: *Predicting students' academic progress is critical in a university setting for early diagnosis of at-risk pupils. The goal of this work is to offer data mining models that use classification methods based on Support Vector Machine (SVM) algorithms to forecast students' academic accomplishment after a preparatory year, as well as to select the optimal approach. Based on graduation CGPA, students' academic achievement is classified as High, Average, or Below Average, and these classifications are applied to a newly produced dataset.*

Keywords: Support Vector Machine (SVM), computer vision, neural networks, Data Mining, Error Measurement, Accuracy.

I. Introduction

Education is a field that generates and accumulates a lot of data. During the typical educational process, thousands of hours are spent at school and performing countless chores at home. Students' interactions with instructional materials generate a large amount of data. Education management systems and online educational platforms collect data on student participation in the online system. Statistics on students' progress and the outcomes of assignments and exercises, as well as involvement in group projects and dialogues, are examples. The university has accumulated data on its candidates throughout time, including their gender, age, and grades in numerous subjects upon high school graduation. Following that, data is collected on the same individuals, but time as students. Attendance, grades in various subjects, the success of scientific activities, the types of tasks assigned, the teachers who taught the course, and so forth.

Data mining, also known as knowledge discovery, is growing more popular since it aids in the study of data from multiple perspectives and the summary of that data into useful information. In educational data mining, decision trees, neural networks, k-nearest neighbour, naive bayes, support vector machines, and other methods are used.

On a csv dataset, the purpose of this study is to use the Support Vector Machine approach to predict and analyse student performance.

II. Related Work

To predict student achievement, numerous research have been undertaken. The Decision Tree (DT) approach was used to predict engineering student achievement in [8]. The researchers gathered information on 340 students in order to forecast how well they will fare on their first-year exam. In their training set, the model they constructed was 60% correct. [9] estimated final student grades using parameters from two different datasets using WEKA data mining methods. Each dataset includes statistics on a group of students from a specific college course during the last four semesters. Among the other classifiers, the IBK has the highest accuracy. [10] summarised previous studies that used various data mining tools to forecast, analyse, and evaluate student performance. The authors of [11] employed DT classification methods to measure student performance and created classifier models using artificial neural networks. The work was analysed using a number of criteria in attempt to predict student achievement. Analyzing the learner's strengths and weaknesses in order to improve future performance. This conclusion emphasises the importance of data mining approaches in course evaluation and higher education data mining. The writers of [12] provide a study that will aid students and teachers in improving the grades of at-risk kids. Data from the student's previous database, including attendance, seminar attendance, and assignment grades, was utilised to forecast performance at the end of the semester. When compared to other categorization algorithms, the authors employed the Nave Bayes algorithm, which had the highest accuracy. Multiple decision tree algorithms were tested on an educational dataset to classify students' educational performance, according to the researchers in [13]. The study's main goal

DOI: 10.48175/IJARSCT-3939

is to find the best algorithm among the most extensively used decision tree algorithms and to provide a standard for each one. On the studied dataset, the Classification and Regression Tree strategy outperformed the other methods based on the accuracy and precision obtained by 10-fold cross validations.

[14] An overview of data mining approaches for forecasting student performance was provided, with a focus on how to use the prediction algorithm to uncover the most significant components of a student's data. To predict student achievement, researchers frequently employ neural networks and decision trees as categorization algorithms. [15] Future outcomes, as well as the factors that influence them, were found and assessed using data mining approaches. The analysis was carried out by applying the FP Growth Algorithm to determine the Association rules for the same, which were then ranked by Lift Metric. The Rule Based Induction Method was then used to classify the data.

III. Data Set and Description

Using various classification approaches, the student's The performance of students was predicted using a college database. Each dataset contains 396 instances, each with 12 characteristics. Table 1 lists the characteristics, along with their descriptions and values. The following attributes were taken into account for prediction:

Travel time:
Home to school or college travel time.
Type numeric: 1- <15minutes, 2- 15-30minutes, 3- 30-1hour, 4- >1hour.

Study time:
Weekly avg. study time of student.
Type numeric: 1- <2hour, 2- 2-5hour, 3- 5-10hour,
4- >10hour

Attended Extra-class:
Whether student has attended any extra class.
Type Boolean: Yes or No

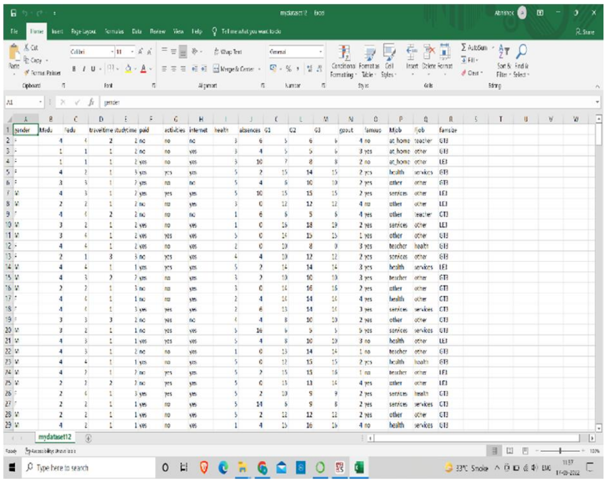
Extra-curricular:
Whether student has taken part in any Extra-curricular activities.
Type Boolean: Yes or No

Tuition:
Whether student has taken any tuition other than college.
Type Boolean: Yes or No

Health Condition:
Students' health condition throughout semester.

Absent days:
For how many days student was absent.
Type numeric:0-90 days

Test 1 Mark:
Marks Scored by student in test 1.
Type numeric: 0-100
Test 2 Mark:
Marks Scored by Student in test 2.
Type numeric: 0-100



	gender	Medu	Fedu	traveltime	studytime	paid	activities	internet	health	absences	G1	G2	G3	goout	famsup	Mjob	Fjob	famsize
2	F	4	4	2	2	no	no	no	3	6	5	6	6	4	no	at_home	teacher	GT3
3	F	1	1	1	2	no	no	yes	3	4	5	5	6	3	yes	at_home	other	GT3
4	F	1	1	1	2	yes	no	yes	3	10	7	8	8	2	no	at_home	other	LE3
5	F	4	2	1	3	yes	yes	yes	5	2	15	14	15	2	yes	health	services	GT3
6	F	3	3	1	2	yes	no	no	5	4	6	10	10	2	yes	other	other	GT3
7	M	4	3	1	2	yes	yes	yes	5	10	15	15	15	2	yes	services	other	LE3
8	M	2	2	1	2	no	no	yes	3	0	12	12	12	4	no	other	other	LE3
9	F	4	4	2	2	no	no	no	1	6	6	5	6	4	yes	other	teacher	GT3
10	M	3	2	1	2	yes	no	yes	1	0	16	18	19	2	yes	services	other	LE3
11	M	3	4	1	2	yes	yes	yes	5	0	14	15	15	1	yes	other	other	GT3
12	F	4	4	1	2	yes	no	yes	2	0	10	8	9	3	yes	teacher	health	GT3
13	F	2	1	3	3	no	yes	yes	4	4	10	12	12	2	yes	services	other	GT3
14	M	4	4	1	1	yes	yes	yes	5	2	14	14	14	3	yes	health	services	LE3
15	M	4	3	2	2	yes	no	yes	3	2	10	10	10	3	yes	teacher	other	GT3
16	M	2	2	1	3	no	no	yes	3	0	14	16	16	2	yes	other	other	GT3
17	F	4	4	1	1	no	no	yes	2	4	14	14	14	4	yes	health	other	GT3
18	F	4	4	1	3	yes	yes	yes	2	6	13	14	14	3	yes	services	services	GT3
19	F	3	3	3	2	no	yes	no	4	4	8	10	10	2	yes	other	other	GT3
20	M	3	2	1	1	no	yes	yes	5	16	6	5	5	5	yes	services	services	GT3
21	M	4	3	1	1	yes	yes	yes	5	4	8	10	10	3	no	health	other	LE3
22	M	4	3	1	2	no	no	yes	1	0	13	14	14	1	no	teacher	other	GT3
23	M	4	4	1	1	yes	no	yes	5	0	12	15	15	2	yes	health	health	GT3
24	M	4	2	1	2	no	yes	yes	5	2	15	15	16	1	no	teacher	other	LE3
25	M	2	2	2	2	no	yes	yes	5	0	13	13	14	4	yes	other	other	LE3
26	F	2	4	1	3	yes	yes	yes	5	2	10	9	9	2	yes	services	health	GT3
27	F	2	2	1	1	yes	no	yes	5	14	6	9	8	2	yes	services	services	GT3
28	M	2	2	1	1	yes	no	yes	5	2	12	12	12	2	yes	other	other	GT3
29	M	4	2	1	1	yes	no	yes	1	4	15	16	16	4	no	health	services	GT3

Figure 1: Data description and values

IV. SVM CLASSIFICATION

In this step, the data was classified using the classification method (SVM), and the students' status was then supplied. The preceding phase's test and training data records were collected and used to classify the data using the SVM algorithm. A database was used to compile the data. Several of the training record sets were tested. After applying the SVM classifier function to the data sets, the classification information were obtained. The accuracy of the prognosis for overall functioning was established after these classification details were analysed to acquire the results. Finally, based on the conditions, the students were divided into groups. Internal Marks for certain theoretical courses, Internal Lab Marks, whether he or she engages in sports, department activities, or is experiencing any psychological stress as a result of the college or family environment, and so on.

It's a promising data classification approach for both linear and non-linear data. It uses non-linear mapping to translate the original training data into higher dimensions. Students are referred to as multidimensional things since they are defined by a variety of aspects. Within this additional dimension, it finds the linear optimal separation hyperplane (that is, a "decision boundary" separating pupils from one class from another). A hyperplane can be used to split data into two groups at any time (H1 and H2). The SVM uses support vectors ("critical" preliminary information samples) and edges to locate this hyperplane (Large edge and small edge, which is characterised by help vectors).

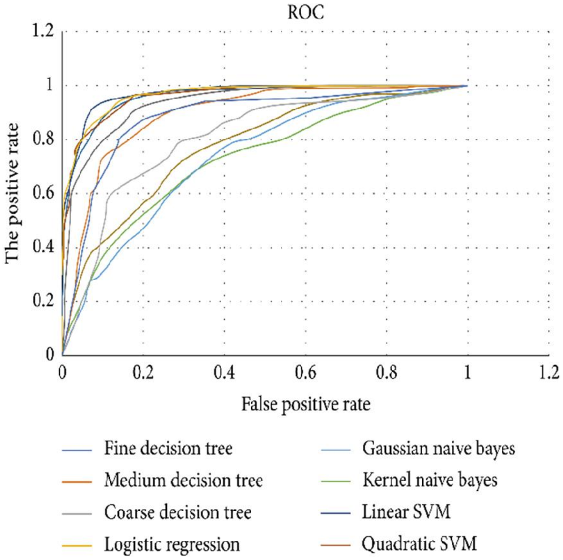


Figure 2: FP rate

When compared to other classifiers, SVM has a higher True positive rate, as shown in the graph above.

V. Experimental Results and Discussion

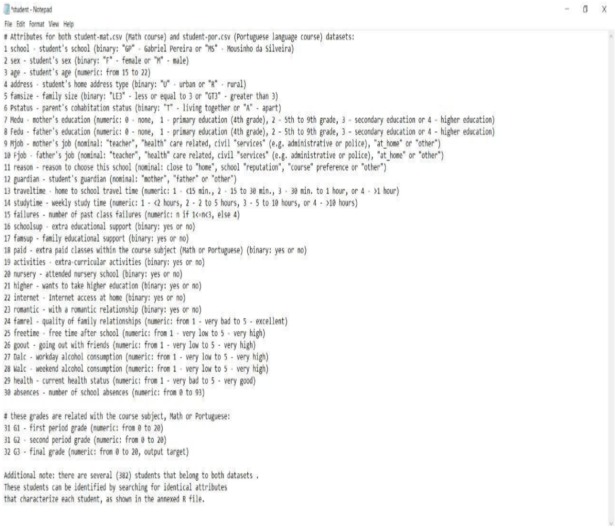
The 396 instances were the focus of the experiment and debate. Our mission is to discover students who are at the top of their class in any graduating programme. For this, we obtained data from both first-year and second-year or final-year students. Because SVM gives the best results for small numbers of vectors, the dataset was likewise constrained to 396 students. The students' personal information was taken from their database. The students' attendance status was also retrieved from the same database. Because it has an impact on a student's grade

Every institution/university requires that each of its students complete the bare minimum of tutoring hours for each topic. Assume a student does not regularly attend classes due to a high level of involvement in other academic endeavours. The poor attendance rate of a student will almost likely affect his or her academic progress. As a result, their professors should issue a warning to keep their attendance within the "safe" range. Questionnaires were utilised to collect data from students. The following is the format of the questionnaire. This information is fed into the system.

Model of a Questionnaire:

- Your Name:
- Father’s Name:
- Father’s Occupation:
- Mother’s Name:
- Mother’s Occupation:
- ADDRESS:
- DOB:
- Email id:
- Gender:
- Age:
- Course:
- Course Completion:
- Joined any Tuition:
- Attended Extra classes:
- Travel Time:
- Study Time(weekly):
- Health Condition:
- Co-Curricular Activities:
- Absent days:
- Marks Scored in Test1:
- Marks Scored in Test 2:

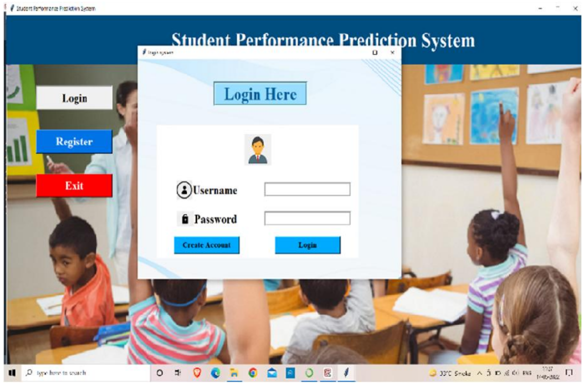
Figure 3 shows a number of other similar questions that were posed.

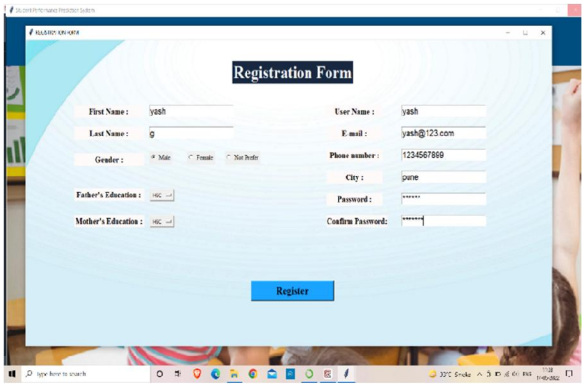


```
*student - Notepad
File  Edit  Format  View  Help
# Attributes for both student-mat.csv (Math course) and student-por.csv (Portuguese language course) datasets:
1 school - student's school (binary: "GP" - Gabriel Pereira or "MS" - Mousinho da Silveira)
2 sex - student's sex (binary: "F" - female or "M" - male)
3 age - student's age (numeric: from 15 to 22)
4 address - student's home address type (binary: "U" - urban or "R" - rural)
5 famsize - family size (binary: "LE3" - less or equal to 3 or "GT3" - greater than 3)
6 Pstatus - parent's cohabitation status (binary: "T" - living together or "A" - apart)
7 Medu - mother's education (numeric: 0 - none,  1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)
8 Fedu - father's education (numeric: 0 - none,  1 - primary education (4th grade), 2 - 5th to 9th grade, 3 - secondary education or 4 - higher education)
9 Mjob - mother's job (nominal: "teacher", "health" care related, civil "services" (e.g. administrative or police), "at_home" or "other")
10 Fjob - father's job (nominal: "teacher", "health" care related, civil "services" (e.g. administrative or police), "at_home" or "other")
11 reason - reason to choose this school (nominal: close to "home", school "reputation", "course" preference or "other")
12 guardian - student's guardian (nominal: "mother", "father" or "other")
13 traveltime - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
14 studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
15 failures - number of past class failures (numeric: n if 1<=n<3, else 4)
16 schoolsup - extra educational support (binary: yes or no)
17 famsup - family educational support (binary: yes or no)
18 paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
19 activities - extra-curricular activities (binary: yes or no)
20 nursery - attended nursery school (binary: yes or no)
21 higher - wants to take higher education (binary: yes or no)
22 internet - Internet access at home (binary: yes or no)
23 romantic - with a romantic relationship (binary: yes or no)
24 famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
25 freetime - free time after school (numeric: from 1 - very low to 5 - very high)
26 goout - going out with friends (numeric: from 1 - very low to 5 - very high)
27 Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
28 Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
29 health - current health status (numeric: from 1 - very bad to 5 - very good)
30 absences - number of school absences (numeric: from 0 to 93)

# these grades are related with the course subject, Math or Portuguese:
31 G1 - first period grade (numeric: from 0 to 20)
31 G2 - second period grade (numeric: from 0 to 20)
32 G3 - final grade (numeric: from 0 to 20, output target)

Additional note: there are several (382) students that belong to both datasets .
These students can be identified by searching for identical attributes
that characterize each student, as shown in the annexed R file.
```

Figure 3: Questionnaire Model with Description

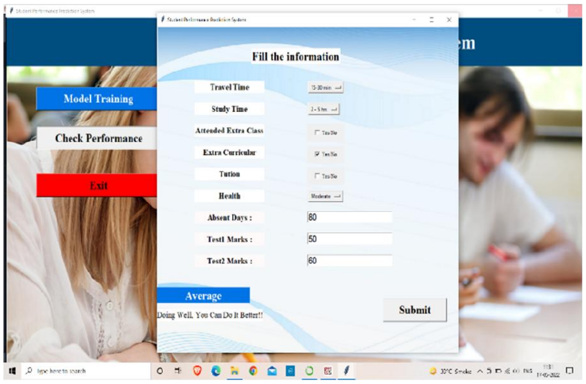
During the pre-processing step, only a few samples were chosen from the data input. In most cases, variables that are unrelated to the forecast are ignored. DOB and E-mail ID, for example, are not included in any partitioning or categorization algorithms. Occasionally, the data values obtained are insufficient or unavailable. These kinds of samples should all be ignored. A range of factors, such as parents' education and career, may indirectly improve a student's performance. They do, however, play a minor role in the film. As a result, they, too, omitted.





Registration Form

First Name : yash
Last Name : g
Gender : ● Male ○ Female ○ Not Prefer
Father's Education : HSC
Mother's Education : HSC
User Name : yash
E-mail : yash@123.com
Phone number : 1234567899
City : pune
Password : ******
Confirm Password: *******



Fill the information

Travel Time: 15-30 min
Study Time: 2 - 5 hrs
Attended Extra Class: ☐ Yes/No
Extra Curricular: ☑ Yes/No
Tution: ☐ Yes/No
Health: Moderate
Absent Days : 80
Test1 Marks : 50
Test2 Marks : 60

Average
Doing Well, You Can Do It Better!!

VI. Conclusion

The effectiveness of DM approaches to predict students' development after a year of preparation in higher education was investigated in this model. A student's academic success is measured using the Grade Point Average (GPA) (excellent, average, or poor). Throughout the experiment, we used SVM classifiers on the student dataset to predict the student's academic year achievement. The SVM classifier outperformed other strategies in predicting student achievement, according to the findings. Additionally, grades on the first and second prelim tests, as well as the final exam, Skills course, and extracurricular skills course, all had a part in predicting students' academic progress. By giving out timely warnings to students, the findings will assist predict students' final grades early enough for suitable interventions to be undertaken. With competent counselling, the number of low-achieving pupils can be reduced.

References

[1]. M. Goyal and R. Vohra, IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 2, No. 1, March 2012, "Applications of Data Mining in Higher Education."

[2]. R. Huebner, Studies in Higher Education Journal, "A study of educational data mining research," 2012.

[3]. M.S. Mythili and A.R. Mohamed Shanavas, "An Analysis of Students' Performance Using Classification Algorithms," IOSR, Journal of Computer Engineering, Volume 16, Number 1, January 2014.

[4]. S. Lakshmi Prabha and A.R.Mohamed Shanavas, "Educational data mining applications," ORAJ, Vol. 1, No. 1, August 2014.

[5]. "Data mining in course management systems: Moodle case study and tutorial," Computers & Education, vol. 51, no. 1, 2008, pp. 368-384.

[6]. S. Ayesha, T. Mustafa, A. Sattar, and M. Khan, "Data mining model for higher education system," European Journal of Scientific Research, vol.43, no.1, pp.24-29, 2010.

[7]. Z. J. Kovacic, "Early prediction of student success: Mining student enrollment data," InSITE 2010 Proceedings.

[8]. I. Milos, S. Petar, V. Mladen, and A. Wejdan, INFOTEH- JAHORINA Vol. 15, March 2016, 684.

[9]. P.Kavipriya, A Review on Predicting Students' Academic Performance Using Data Mining Techniques Earlier, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 6, Issue 12, December 2016 ISSN: 2277 128X.

[10]. N. Ankita and R. Anjali, Student Performance Analysis Using Data Mining Technique, International Journal of Innovative Research in Computer and Communication Engineering, Vol. 5, Issue 1, January 2017.

[11]. P. Shruthi and B. Chaitra, Student Performance Prediction in the Education Sector Using Data Mining, Volume 6, Issue 3, March 2016, International Journal of Advanced Research in Computer Science and Software Engineering.

[12]. 2012, S.K. Yadav, B. Bharadwaj, and S. Pal. A Comparative Study of Data Mining Applications for Predicting Student Performance Vol. 1, No. 12 of the International Journal of Innovative Technology and Creative Engineering (ISSN: 2045-711),A. Mohamed Shahiria, W. Husaina, N. Abdul Rashida, "A Review on Predicting Student's Performance using Data Mining Techniques" Procedia Computer Science 72 (2015) 414 – 422, ELSEVIER. December.

[13]. K. Kohli and S. Birla, " Data Mining on Student Database to Improve Future Performance", International Journal of Computer Applications (0975 – 8887), Volume 146 – No.15, July 2016.

[14]. Hilal Almarabeh, "Analysis of Students' Performance by Using Different Data Mining Classifiers", International Journal of Modern Education and Computer Science (IJMECS), Vol.9, No.8, pp.9-15, 2017.DOI: 10.5815/ijmecs.2017.08.02

[15]. K. B. Eashwar, R. Venkatesan and D. Ganesh, Student Performance Prediction Using Svm, International Journal of Mechanical Engineering and Technology 8(11), 2017, pp. 649–662.

DOI: 10.48175/IJARSCT-3939