# Early Prediction of students' grades and appropriate Recommendations

Pujitha Reddy Surapareddy, Hsing-Hung Wan

**Abstract**—Educational data mining techniques are widely used in academic prediction on student performance in classroom education. In this paper, we performed analysis to identify the significant and impact of student background, student social activities and student coursework achievement in predicting student academic performance. By using the combination with all the attributes and comparing with different classification tools Decision Tree, Random Forest and SVM, student coursework achievement is the most valuable attribute to predict the student's performance. Moreover, SVM can have the highest accuracy as the prediction model(0.913). Here, we also proposed a recommendation system that recommends courses for students based on similarities of students' course history. The proposed system employs data mining techniques to identify patterns in course enrollment procedure. We also have noticed that clustering the students based on their history of enrollment made the association rules generated using apriori most efficient than the rules generated with out clustering. The frequent items that were found from the rules are depicted as a network. The efficiency of this network is verified by distributing its degree. The result from this is a scale free network.

## 1 INTRODUCTION

The prediction has gained most prominence in every domain like education, health Care, E-Commerce with data mining[1]. There are many practical applications of these prediction models. In this paper we high light more on Educational Data Mining. For instance, every semester, many departments in the university introduce new courses for students. There can be chances that this process may result in the recording of large amounts of data in databases and student files in the registries. However, in most cases, it was found that 80 percent of this data is not being used for analysis[4] and this may result in wastage of what would otherwise be one of the most precious assets of these institutions[12]. Additionally, data mining techniques can help in predicting the students' enrollment percentages so that the department can anticipate the enrollment percentage for any course. Besides, this model will also predict the percentage of enrollment for each course[3]. Furthermore, for many students, it could be a challenge to find an internship or a job after graduation. So, with the help of predictive data analysis and student data set as input, we can provide recommendations to students regarding the internships, job opportunities and contacts of alumni as part of canvas platform. The main aim of this model is to predict the unknown values(final grade in a course and time required for graduation), given the current status of the students. Although many models of predictions like regression, clustering have been studied in the literature, it was in different contexts such as Massive Open Online Courses and Tutoring Systems. However, this model varies from earlier platforms and has many challenges. First there can be different students when seen in terms of their areas(majors, specializations), resulting in different courses and also their enrollment order. So, here the key challenge is effective predictor that can handle heterogeneous data. Second, the previous predictive models have focused mainly on statistical modeling and data mining techniques and have their own limitations. Third, predicting a student's performance is not a one time task in long term degree programs. It requires lot of monitoring and updating as the student enrolls in a new course over time. So, keeping all these above-mentioned challenges, we propose a novel algorithm to predict the student grades and enrollment percentages in a degree program. Our main contributions are threefold.

- We develop a novel algorithm for making predictions based on students' progressive performance states. First, we use a centrality metric to identify the most important features in the given data set. The clustering coefficient will be used to make local predictions given the current progress of students'.Later by synthesizing these predictions, we predict the results throughout the academic year. This keeps the complexity low. We will also do the performance analysis to benchmark the results
- We develop a data-driven course clustering method based on probabilistic matrix factorization, which automatically outputs course clusters based on large, heterogeneous and sparse student course grade data. Base predictors are trained using machine learning algorithms. The input to these algorithms will be the relevant courses from the same cluster.This reduces the complexity of training and removes the noise in the data set.

The rest of this paper is organized as follows. Section II discusses the related work. Section III proposes a novel progressive prediction algorithm to predict the grades and enrollment percentages. Section IV shows the analysis in the form of visuals. Section V briefs some expected results.

- *Pujitha Reddy Surapareddy, HSING-HUNG WAN with the Department of Computer Science, California State University, Fresno .*
*E-mail: pujithareddys@mail.fresnostate.edu, b6i0l5l@mail.fresnostate.edu*
- 

## 2 Literature Review

The content of collecting data and learning analytic by organizations or academic institutions are prepared for predicting student's performance and classifying the main factor which will influence student's performance in the academic. The related work can be separated to two parts, classification data mining and centrality metrics. The former is built to apply the learning algorithm and training the data to predict the student's performance, the latter could be used to determine the key value to affect the student's performance.

### 2.1 Classification Data Mining

In the classification, there are multiple methods to use to predict the student's performance and the widely known are Linear regression[1], Decision Tree[2][3][4][36][37], Random Forest][2][31], Multilayer Perceptron[1][2], Support Vector Machine[3][5][32][38] and Bayesian Networks[4][34][35]. Based on the data collected, those methods can be used to build the predict model for the student's performance.

Febrianti et al. [1] provided a comparison between linear regression and multilayer perceptron as a prediction model to predict the student's performance in final their examination. The data was collected from 50 undergraduate and used posting and attendance as two variables in this model. The result shows multilayer perceptron has a smaller mean absolute error and root mean squared error than linear regression which represented multilayer perceptron had a better accuracy.

Ching-Chieh[2] surveyed the backgrounds and social activities of 395 students to identify at-risk students. The prediction models were four supervised educational data mining techniques, namely Naiive Bayesian, Multilayer Perceptron, Decision Tree and Random Forest. And the result showed that Decision Trees could have the best accuracy (0.924) to predict the risk's student as pass or fail. Mustafa[3] used four different classifier models which are Decision Tree, Support Vector Machines, Artificial Neural Networks, and discriminant analysis to predict instructors' performance through a real course evaluation questionnaire. The result showed that finding the most important factor in the data which can improve the accuracy, also it proved that these classification data mining techniques could not only help to predict the student's performance but also help the instructors to prepare their courses.

Guarín et al. [4] presented a prediction for loss of academic status in Universidad Nacional de Colombia, Bogota campus. The author used naive Bayes and decision tree as data mining techniques to be the classifier model. The result showed that added more data could help the classifier have a better accuracy, however, it doesn't mean it more data presented the better accuracy. It was important to find the most influent factor to add to the data.

Elvira et al. [6] mentioned not only the performance in the class could influence student's grade, but also the social learning environment could be one of the factors to predict student's performance. The data was collected from six consecutive online course installments within six years. Based on 343 student's activity on wiki, blog and microblogging tool, the author used an innovative regression algorithm (Large Margin Nearest Neighbor Regression) to illustrate a higher engagement with social media tools correlates with a higher grade. Instead of referring the extraneous variables to predict the student's performance.

Iti et al. [5] used Support Vector Machine (SVM) to build the predict model of a Psychometric analyses of students' behavior in respect of learning help in enhancing their academic performance. There are two kernels to calculate the accuracy. One is Linear Kernel, the other one is radial basis function kernel (RBF). The result showed that RBF had more accuracy than Linear Kernel which is about 90%.

Based on the those researches which could help us to find the best prediction model to help to predict the student's performance in the future. Moreover, it could also help the teachers to pay more attention on those high-risk students to prevent they fail to the classes.

### 2.2 Centrality Metrics

In the social network, the centrality plays an import role in the graph. Finding the centrality could help to improve the accuracy in classification data mining techniques. Hui[8]studied the correlation of degrees and betweenness centrality to investigate BBS reply networks. And the result showed central nodes with high degree or high betweenness centrality do have high influence and power in online social networks. On the other hand, when central nodes with highest degree and highest betweenness centrality are removed network centralizations decreased typically.

Natarajan et al. [9] provided an idea to use centrality metric such as degree, eigenvector, betweenness and closeness to define the most central state within a country and use this information to design the road/rail transportation networks. Ruchi et al. [10] presented to use a new metric which namely Cross closeness centrality for measuring the multiplex social network and simple network. The datasets were the families which were from two different areas(Danio Rerio and Florentine). After analyzing, the data showed multiplex networks offers much valuable and concrete information compare to the simple network. Prantik et al. [11] observed the influence users from Twitter and use Degree Centrality and Eigenvector Centrality to collect the data. The result showed that indegree and eigenvector centrality should both be considered when finding users who are influential.

Yanping et al. [7] presented a new centrality which combined betweenness centrality and Katz centrality to measure the importance of node. This new centrality not only reduced the problem of betweenness centrality which only focused on the shortest path but also solve the problem of Katz centrality which focused on the adjacent nodes. Lingjie[12] gave a new centrality which depended on the betweenness changes caused by the removal of the largest node in the network. This method was useful to identify the functional and structural of importance of the nodes in a network.

## 3 Data Set

There are two data sets included in this paper. The first one included 649 instances and 33 attributes. The attributes are separated into Student Backgrounds, Student Social Activities and Student Course Result as Table 1 to 4.

| Student Background | | | |
|---|---|---|---|
| Attribute | Description | Type | Value |
| sex | gender of student | binary | male \| female |
| school | school of student | | Mousinho da Silverira \| Gabriel Pereira |
| address | type of students's home address | | rural \| urban |
| Pstatus | cohabitation status of parent | | living together \|apart |
| famsize | size of family | | less 3 than \| greater than 3 |
| schoolsup | extra educational school support | | yes \| no |
| famsup | educational support from family | | yes \| no |
| Mjob | job of mother | nominal | - at home |
| Fjob | job of father | | - civil services |
| reason | reson to choose this school | | - close to home<br>- school reputation<br>- course preference<br>- other |
| guardian | guardian of student | | - father<br>- mother<br>- other |

TABLE 1: Student Background-1. Includes the sex, school, address, parents cohabitation status, family size, extra education support, family education support, mother's jobs, father's jobs, reason to choose this school and student's guardian.

| Student Background | | | |
|---|---|---|---|
| Attribute | Description | Type | Value |
| Medu | education of mother | numeric | 0 # none |
| Fedu | education of father | | 1 # primary education |
| famrel | quality of family relationships | | very bad (1) to excellent (5) |
| age | age of student | | 15 - 22 |
| traveltime | travel time from home to school | | 1 # <15 min<br>2 # 15 to 30 min<br>3 # 30 min to 1 hour<br>4 # > 1hour |
| studytime | weekly study time | | 1 # < 2hours<br>2 # 2 to 5 hours<br>3 # 5 to 10 hours<br>4 # > 10 hours |
| failures | number of failures in pass calss | | n if 1 <= n < 3, else 4 |

TABLE 2: Student Background-2. Includes mother's educations, father's educations, quality of family relationships, age of students, travel time, study time, and failures.

The two classification lists below will be used to predict the final grade G3. They will be divided into two parts for the prediction.

- 2 Level classification(fail/pass)
- 5 Level classification(A/B/C/D/E)(Table 5)

The second data set includes 325199 instances and – attributes from the courses enrolled through canvas at Harvard and MIT.

The Data Set is acquired from Canvas Network Person-Course (1/2014 - 9/2015). This data set consists of one row per-person, per-course which means if an individual is enrolled in three different courses, there will be three different records with same student id and three course id's. So, In Order to generate frequent item set, redundancy should be removed. Data set should be refined in a way that there will be a unique row for each student.

| Student Social Activities | | | |
|---|---|---|---|
| Attribute | Description | Type | Value |
| activies | extra-curricular | binary | yes \| no |
| higher | plans for higher education | | |
| internet | home internet access | | |
| nursery | hursery school attended | | |
| paidclass | extra paid classes | | |
| romantic | in romantic relationship | | |
| absences | absences from school | numeric | very low(1) to very high (5) |
| health | status of current health | | |
| freetime | free time after school | | |
| goout | outing with friends | | |
| alc | consume alcohol in weekday | | |
| Wald | consume alcohol in weekend | | 0 - 93 |

TABLE 3: Student Social Activities. Includes activities, plan to higher education, internet at home, attend nursey school, free time after school, extra paid classes within the course subject...etc.

| Student Course Work | | | |
|---|---|---|---|
| Attribute | Description | Type | Value |
| G1 | 1st grade period | numeric | 0 - 20 |
| G2 | 2nd grade period | | |

TABLE 4: Student Course Work. Determines the grades of the first exam and the second exam.

| Mark | 16 - 20 | 14 - 15 | 12 - 13 | 10 - 11 | 0 - 9 |
|---|---|---|---|---|---|
| Grade | A | B | C | D | E |

TABLE 5: 5 Level Classification. Separates the score of the final exam into five levels as target.

| Student ID | CSCI 41 | CSCI 115 | CSCI 174 | CSCI 274 |
|---|---|---|---|---|
| 1 | A | B | C | N |
| 2 | B | C | C | F |
| 3 | N | F | F | N |
| 4 | A | B | N | B |

TABLE 6: students data

This data set is made de-identified and consists of over 325,000 aggregate records from 238 Canvas Network courses. Data Set has grades in range of 0-1. The course names are changed in a way that matches the institutional data set codes.To make it more meaningful we filtered as follows.

## 4 Methodology

Based on Ching-Chieh[2], Decision Tree could have the best accuracy between Naiive Bayesian, Multi layer Perceptron, Decision Tree and Random Forest. However, SVM could also have a higher accuracy prediction for student's performance[5]. In this experiment, Decision Tree, Random Forest and SVM would used to build the prediction model to help to predict the student's performance.

| Grade Range | Grade Scale |
|---|---|
| [0.7-1.0] | A |
| (0.7-0.5] | B |
| (0.5-0.1] | C |
| (0.1-0.0] | F |
| Not taken course | N |

TABLE 7: Grades Table

### 4.1 Grade Prediction

- Decision Trees
  Decision Trees [2]helps to break the data into two groups and repeated this process several time until the data has no sub-data to break to make any sub-groups.By this method[3], the complex data could be converted to a single variable in order to finding the pattern in the model. Fig. 1 shows a sample representation for Decision Tree. In this experiment, student's information could be built into this model to help to predict the student's performance.

**Algorithm 1:** Training a Decision Tree[39]

Input: training data E and attribute set F
TreeGrowth(E,F)
if stopping_cond(E,F) = $true$ then
$leaf$ = createNode()
$leaf.label$ = Classify(E)
return $leaf$
else
$root$ = createNode().
$root.test\_cond$= find_best_split(E,F)
let V = {v|v is a possible outcome of root.test_cond}
for each v $\epsilon V do$
$E_v = \{$e|root.test_cond(e)=v $and$ e $\epsilon E\}$
$child$ = TreeGrowth($E_v, F)$
add $child$ as descendent of $root$ and label the edge ($root \rightarrow child$ as $v$)
end for
endif
return root

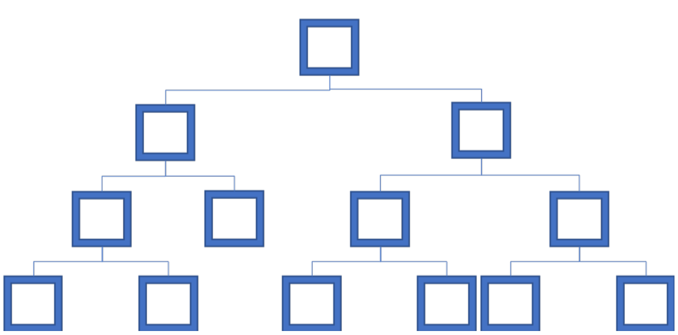

Fig. 1: Decision Trees. The attributes would be split in every level from top to down, until it reaches the targets.

- Random Forest
  As this data has multiple attributes, decision tree would be too simple to predict the student's performance. Therefore, Random Forest could help to improve the prediction model. RF is a classification which includes multiple Decision Trees to [2][31] help to increase the accuracy when new data is added.Each tree in the forest gives a classification and "vote" for that class. The forest chooses the classification having the most votes. Fig 2 shows the sample representation for Random Forest.

**Algorithm 2:** Training a Random Forest[41]

Input: Data set D = ($x_1$,y1),(x2,y2),...,(xm,ym,);
Feature subset size K.
Process:
N ← create a tree node based on D;
If all instances in the same class return N
F ← the set of features that can be split further;
If F is empty then return N
F ← select K features from F randomly;
N.f ← feature has the best split point in F;
N.p← the best split on N.f;
Dl ← subset of D with values on N.f smaller than N.p;
Dr ← subset of D with values on N.f no smaller than N.p;
Nl ← call the process with parameters (Dl, K);
Nr ← call the process with parameters (Dr, K);
return N
output: a Random Forest

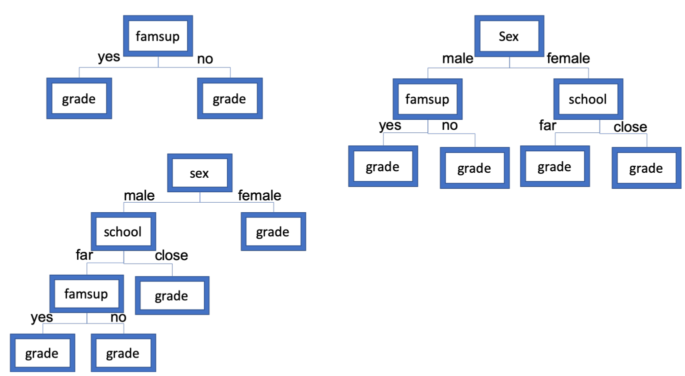


Fig. 2: Random Forest. As Decision Tree, attributes are split in each level and instead of one tree, multiple trees could help to increase the accuracy.

- Support Vector Machine(SVM) Instead of putting all the attributes to build the predict model, SVM takes two attributes each time to create the prediction model which can help the student course work result has the same levels to compared with other attributes.SVM is a supervised learning technique that aims to classify the data. It takes a hyperplane for splitting the dataset into two groups with a gap which is namely margin. And in order to increasing the accuracy, the largest margin is selected. SVM could have a better accuracy when the variable is important in the data[3]. Fig. 3 shows a sample representation for SVM. In this experiment, 2-Dimension graph will be used to predict the student's performances to compare with other classifications.

**Algorithm 3:** Training a SVM [40]

Input: x and y loaded with training labeled data, p ←0 or p ← partially trained SVM
C ←some value (10 for example)
repeat:
  for all and grades $\{x_i, y_i\}$ , do
    Optimize $p_i$ and $p_j$
  end for
until no changes in p
Ensure: Retain only the support vectors($p_i > 0$)

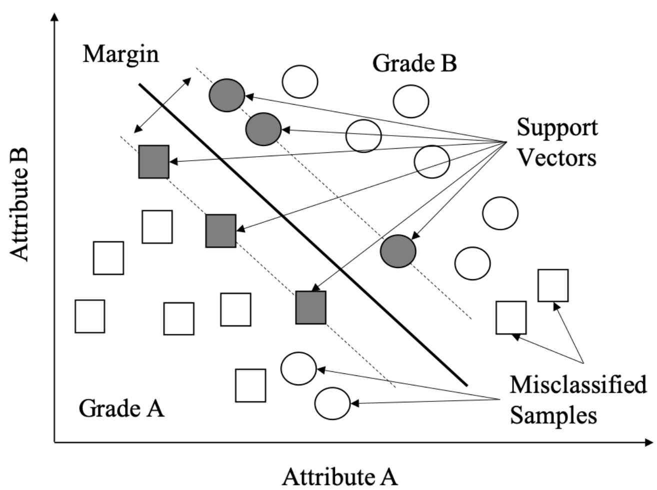


Fig. 3: Support Vector Machine(SVM)[3]. This paper is using a 2-dimension SVM graph to predict the student's grades.

Figure 4 explains the flow of the classification processes. The student's information will be separated into different groups which will be used as the attributes of these three classifications( Decision Tree, Random Forest and SVM) to predict the student's final grade with 2 levels and 5 levels.

### 4.2 Course Recommendation

Decision Making plays a crucial role in students' life while planning career. The first step involves choosing their majors and related courses during their graduation. To help students in decision making we proposed a course recommendation algorithm that takes the similarity among courses, students and recommend a course with predicted grade. First we experiment the existing algorithms(Frequent Item Set Mining) and find the courses that are taken more frequently.Later, we modify the existing algorithm by using clustering before data mining. The latter gives high accuracy when compared to former.

Each Student's enrollment is made as a transaction, which involves the list of courses, student is enrolled in.

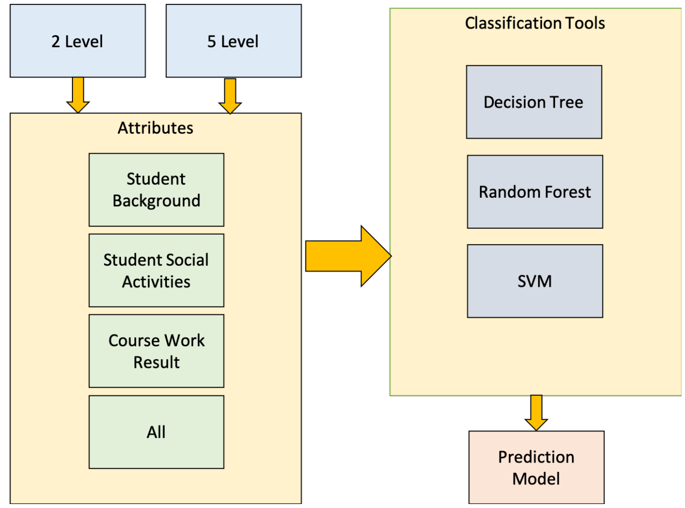


Fig. 4: Classification Data Mining Process. Using Decision Tree, Random Forest and SVM to predict different levels of attributes.

Each student is considered as basket and items are the courses.Support is calculated for each transaction and association rules are generated to predict students enrollment with expected grade and generate frequent item sets.

**Algorithm 4:** Apriori Algorithm [20]

Input: Set of Transactions
Apriori(T,E)
PROCEDURE GetFrequentItemSets freqSets[] = null
for all Itemsets i in S do
if support ¿ min$_{s}upportthen$
freqSets[] ¡- i
end if
end for
end procedure

## 5 Data Visualization

### 5.1 Data Set and Analysis

Descriptive analysis is an important first step for conducting statistical analyses. It gives you an idea of the distribution of your data, helps you detect outliers and typos, and enable you identify associations among variables, thus preparing you for conducting further statistical analyses.

Figure 5 shows the show the Precision , Recall and F-Measure of 2 Level in different attributes of using Decision Tree, Random Forest and SVM classification.

Figure 6 shows the show the Precision , Recall and F-Measure of 5 Level in different attributes of using Decision Tree, Random Forest and SVM classification .

Figure 7 shows the clusters that are generated through k-means clustering. Clustering is done in order to improve the performance of the algorithm as well as the patterns in frequent item sets. Feature Extraction is done using Principal Component Analysis. Features such as discipline,

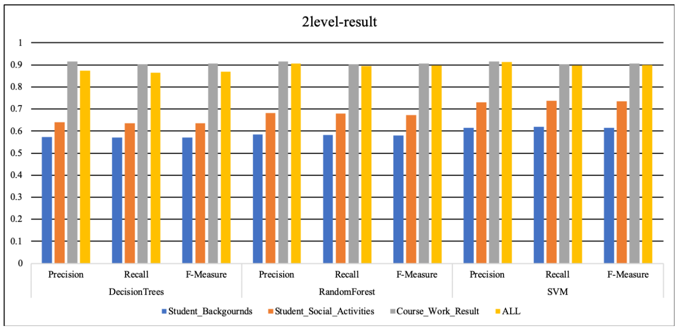


Fig. 5: Precision, Recall and F-Measure 2 Level Classification Result. Decision Tree, Random Forest and SVM as different groups to compare the predictions

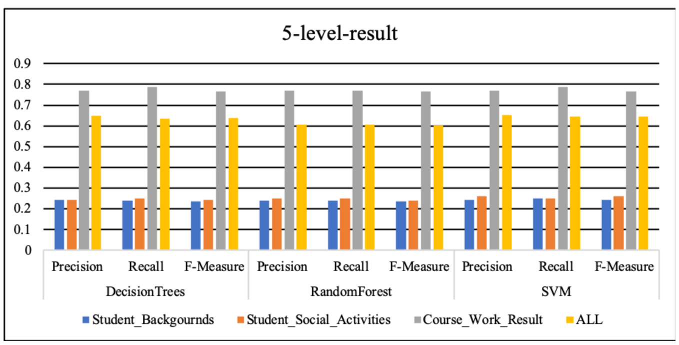


Fig. 6: Precision, Recall and F-Measure 5 Level Classification Result. Decision Tree, Random Forest and SVM as different groups to compare the predictions

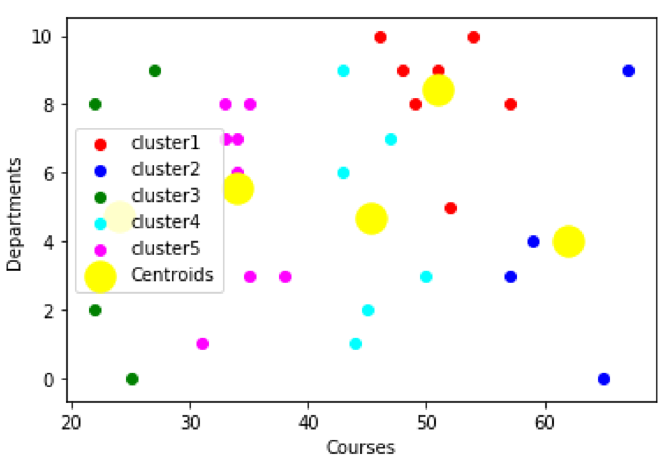


Fig. 7: Clusters among coruses and disciplines

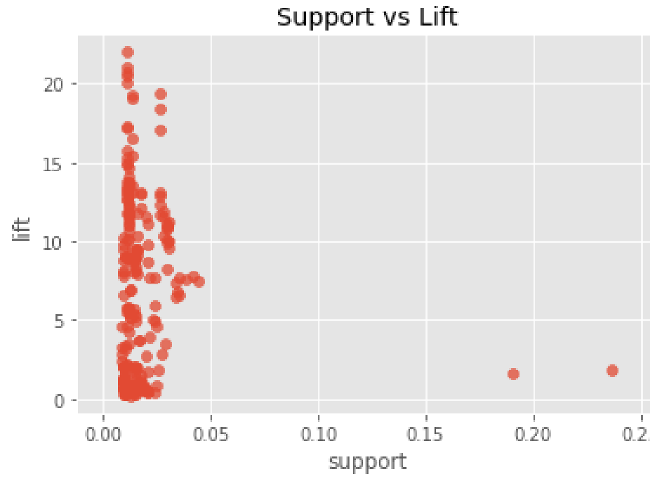


Fig. 8: Support VS Lift

number of hours they have spent and their grade are taken to cluster similar items.

Figure 8 shows the distribution of support versus lift for the rules those are generated from apriori algorithm. Support gives us a value that determines how often the course was taken by the students. Lift is a parameter that measures the importance of a rule. So this graph shows that most of the rules whose support is between 0.00 and 0.005 have their importance varying between 0 and 20.

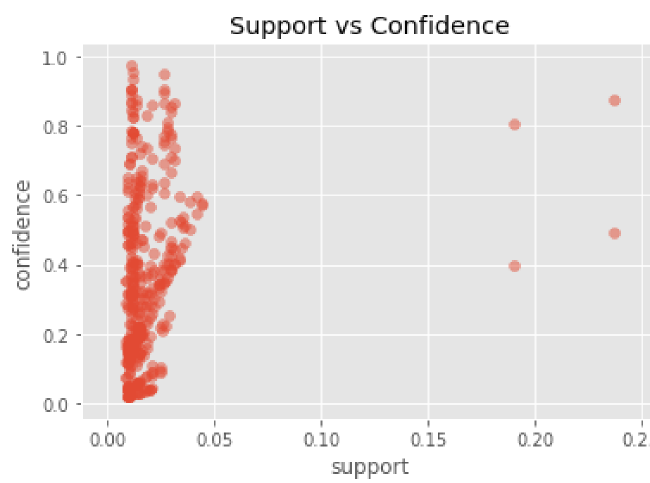


Fig. 9: Support VS Confidence

Figure 9 shows the distribution of support versus confidence for the rules those are generated from apriori algorithm. Support gives us a value that determines how often the course was taken by the students.Confidence tells us than how often the predicted rule is true. From the graph it is clear that most of the predictions are true and most of their confidence is between 0 and 1.

Figure 10 shows the distribution of lift versus confidence for the rules those are generated from apriori algorithm. From the graph it is clear that as lift increases confidence

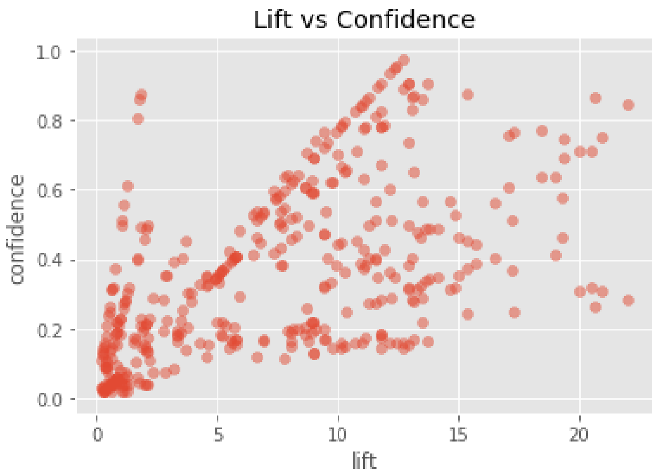


Fig. 10: Lift VS Confidence

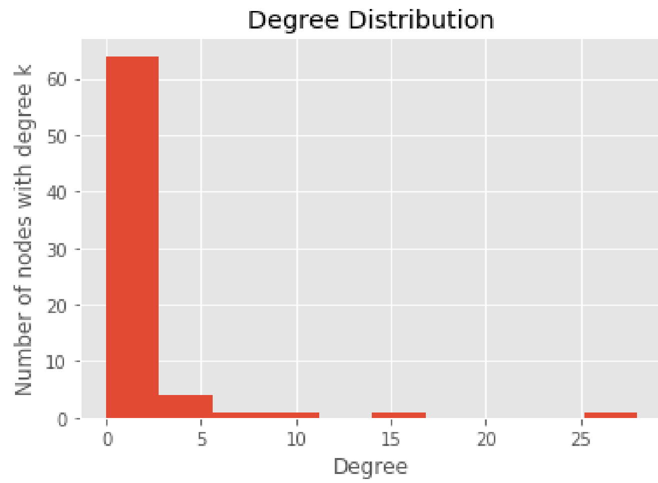


Fig. 12: Degree Distribution of the network Generated from Rules

also increases. This means that the rules with higher importance has more confidence.

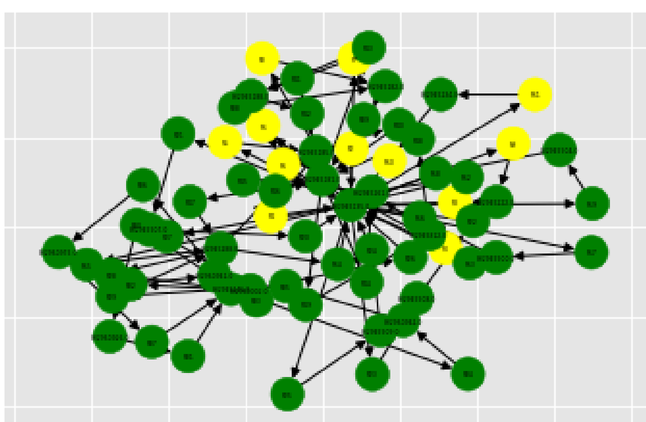

Fig. 11: DiGraph Representing connections among Rules Generated

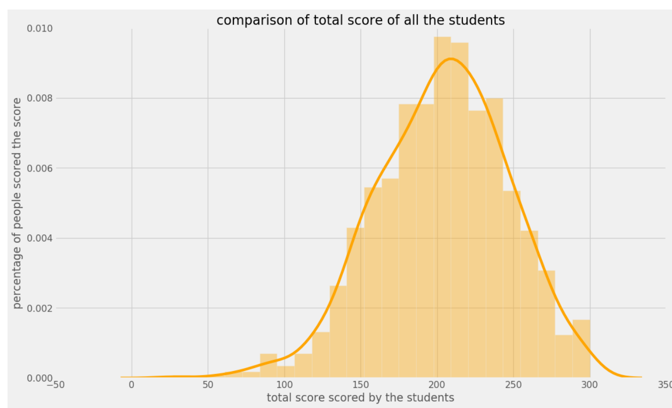


Fig. 13: Distribution of parents degree

Figure 11 shows a directed graph that represents the networks of courses and their connections. Nodes here are the courses and edges are the frequent patterns derived from association rule mining. The main application of this would be enrollment percentage prediction for the courses offereds.

Figure 12 shows the degree distribution of the network of courses created. When we observe tha pattern it is a scale free network which justifies that this is a real world network.

Figure 13 shows the distribution of score students' score in mathematics. This is a normal distribution. But most of the curve is towards the right side. It shows that majority of class has score more than half of the grade.

## 6 Results

This data approach student achievement in secondary education of two Portuguese schools. The data attributes include student grades, demographic, social and school related features) and it was collected by using school reports and questionnaires. Two data sets are provided regarding the performance in two distinct subjects: Mathematics (mat) and Portuguese language (por). The two datasets were modeled under binary/five-level classification and regression tasks. Important note: the target attribute G3 has a strong correlation with attributes G2 and G1. This occurs because G3 is the final year grade (issued at the 3rd period), while G1 and G2 correspond to the 1st and 2nd period grades. It is more difficult to predict G3 without G2 and G1, but such prediction is much more useful.

As the result shows in Table 8 of 2 Level prediction. The highest accuracy is 0.916 when the attributes are course work result. The student's background and student's social activities attributes can also have the accuracy above 0.5. However, the total attributes have a higher accuracy than student's background and student's social activities as 0.88 which means the course work result can improve the accuracy to predict the student's performance. Moreover, when the attributes are student's backgrounds, student's social activities and all, SVM can have the highest average accuracy among these three classifications(Decision Tree, Random Forest and SVM). On the other hand, the result shows in Table 9 of 5 Level prediction. The highest accuracy is lower than 2 Level as 0.771. It represents to predict the target as more levels the accuracy will get lower. After all, the Course Work Result can help the classifications have the best accuracy in both 2 level and 5 level prediction.

| Classification | Metrics | Student_Backgourdns | Student_Social_Activities | Course_Work_Result | ALL |
|---|---|---|---|---|---|
| DecisionTrees | Precision | 0.5735 | 0.641 | 0.916 | 0.874 |
| | Recall | 0.571 | 0.636 | 0.902 | 0.865 |
| | F-Measure | 0.5705 | 0.635 | 0.906 | 0.868 |
| RandomForest | Precision | 0.584 | 0.682 | 0.916 | 0.907 |
| | Recall | 0.5815 | 0.68 | 0.902 | 0.894 |
| | F-Measure | 0.5795 | 0.672 | 0.906 | 0.8975 |
| SVM | Precision | 0.614 | 0.731 | 0.916 | 0.913 |
| | Recall | 0.619 | 0.738 | 0.902 | 0.897 |
| | F-Measure | 0.615 | 0.734 | 0.906 | 0.9 |

TABLE 8: Precision, Recall and F-Measure 2 Level Classification Result. SVM has the highest prediction of all the classifications (0.913). Course Work Result is more important attribute than other attribute which could improve the accuracy in all attributes.

| Classification | Metrics | Student_Backgourdns | Student_Social_Activities | Course_Work_Result | ALL |
|---|---|---|---|---|---|
| DecisionTrees | Precision | 0.2425 | 0.245 | 0.771 | 0.6495 |
| | Recall | 0.2415 | 0.2495 | 0.7885 | 0.634 |
| | F-Measure | 0.235 | 0.242 | 0.7665 | 0.637 |
| RandomForest | Precision | 0.2405 | 0.25 | 0.77 | 0.605 |
| | Recall | 0.2405 | 0.25 | 0.77 | 0.605 |
| | F-Measure | 0.2365 | 0.2415 | 0.765 | 0.6035 |
| SVM | Precision | 0.242 | 0.26 | 0.771 | 0.651 |
| | Recall | 0.249 | 0.25 | 0.7885 | 0.644 |
| | F-Measure | 0.242 | 0.261 | 0.7665 | 0.644 |

TABLE 9: Precision, Recall and F-Measure 5 Level Classification Result. SVM has the highest prediction of all the classifications (0.651). Comparing to lower levels of targets, the levels of target increase, the accuracy will decrease.

Table 10 below represents the item set that is generated after data pre processing and clustering.

| | userid_DI | course_id_DI_1 | course_id_DI_10 | course_id_DI_11 |
|---|---|---|---|---|
| 34001 | 832455231 | 832945135.0 | 832945504.0 | 832945222.0 |

TABLE 10: Item Set, Representing transactions of students. User Id represents the student id and all other columns represent the courses they are enrolled in the order of enrollment date.

Table 11 below explains the frequent item sets, their support values generated using Market Basket Analysis. Here Each Student is like a basket and items are the frequent patterns that appeared the most

Table 12 below shows the rules that are generated with antecedent, precedent, support, confidence and lift values for a given data set.

## 7 Conclusion

In the current research, Decision Tree and Random Forest are used widely to predict the student's performance. However, these two methods have the same problem which the prediction model is too simple to predict the target. If there is a new attribute, the structure would have a big change and the result will be inaccurate. Therefore,

| | support | itemsets |
|---|---|---|
| 0 | 0.235145 | (832945135.0) |
| 1 | 0.076880 | (832945145.0) |
| 2 | 0.076880 | (832945181.0) |
| 3 | 0.066179 | (832945182.0) |
| 4 | 0.070121 | (832945185.0) |
| 5 | 0.077161 | (832945188.0) |
| 6 | 0.054351 | (832945222.0) |
| 7 | 0.033512 | (832945234.0) |

TABLE 11: Frequent Item sets and their support values

| antecedents | consequents | antecedent support | consequent support | support | confidence | lift |
|---|---|---|---|---|---|---|
| (832945181.0) | (832945135.0) | 0.076880 | 0.235145 | 0.014644 | 0.190476 | 0.810037 |
| (832945135.0) | (832945181.0) | 0.235145 | 0.076880 | 0.014644 | 0.062275 | 0.810037 |
| (832945182.0) | (832945135.0) | 0.066179 | 0.235145 | 0.009575 | 0.144681 | 0.615283 |
| (832945135.0) | (832945182.0) | 0.235145 | 0.066179 | 0.009575 | 0.040719 | 0.615283 |
| (832945185.0) | (832945135.0) | 0.070121 | 0.235145 | 0.011264 | 0.160643 | 0.683164 |
| (832945135.0) | (832945185.0) | 0.235145 | 0.070121 | 0.011264 | 0.047904 | 0.683164 |
| (832945188.0) | (832945135.0) | 0.077161 | 0.235145 | 0.014644 | 0.189781 | 0.807081 |
| (832945135.0) | (832945188.0) | 0.235145 | 0.077161 | 0.014644 | 0.062275 | 0.807081 |
| (832945222.0) | (832945135.0) | 0.054351 | 0.235145 | 0.010138 | 0.186528 | 0.793249 |

TABLE 12: Association Rules generated from rule mining using Apriori algorithm

we use the SVM to cover all the attributes to predict the student's performance to avoid this problem. And the result proves that the accuracy is higher than Decision Tree and Random Forest. Moreover, based on the result of this research, the attributes could be an import factor to improve the accuracy. Moreover, less attributes can also reduce the time consuming in the prediction models. Therefore, how to select the attributes of student's background will be the most important factor in order to improve the accuracy of the prediction model.Future work involves generating rules using Park Chen and Yu, Multi Stage and Multi Hash Algorithm. We aim to improve efficiency of the rule generation and clustering using the above mentioned algorithms. This makes the recommendations appropriate for the user.

## References

[1] Widyahastuti, Febrianti, and Viany Utami Tjhin. "Predicting students performance in final examination using linear regression and multilayer perceptron." 2017 10th International Conference on Human System Interactions (HSI). IEEE, 2017.

[2] Kiu, Ching-Chieh. "Data Mining Analysis on Student's Academic Performance through Exploration of Student's Background and Social Activities." 2018 Fourth International Conference on Advances in Computing, Communication & Automation (ICACCA). IEEE, 2018.

[3] Agaoglu, Mustafa. "Predicting instructor performance using data mining techniques in higher education." IEEE Access 4 (2016): 2379-2387.

[4] Guarín, Camilo Ernesto López, Elizabeth León Guzmán, and Fabio A. González. "A model to predict low academic performance at a specific enrollment using data mining." IEEE Revista Iberoamericana de tecnologias del Aprendizaje 10.3 (2015): 119-125.

[5] Burman, Iti, and Subhranil Som. "Predicting Students Academic Performance Using Support Vector Machine." 2019 Amity International Conference on Artificial Intelligence (AICAI). IEEE, 2019.

[6] Popescu, Elvira, and Florin Leon. "Predicting academic performance based on learner traces in a social learning environment." IEEE Access 6 (2018): 72774-72785.
[7] Zhang, Yanping, et al. "Identifying node importance by combining betweenness centrality and katz centrality." 2015 International Conference on Cloud Computing and Big Data (CCBD). IEEE, 2015.
[8] Li, Hui. "Centrality analysis of online social network big data." 2018 IEEE 3rd International Conference on Big Data Analysis (ICBDA). IEEE, 2018.
[9] Meghanathan, Natarajan, and Raven Lawrence. "Centrality analysis of the united states network graph." (2016).
[10] Mittal, Ruchi, and M. P. S. Bhatia. "Cross-Layer Closeness Centrality in Multiplex Social Networks." 2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT). IEEE, 2018.
[11] Howlader, Prantik, and K. S. Sudeep. "Degree centrality, eigenvector centrality and the relation between them in Twitter." 2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT). IEEE, 2016.
[12] Zhou, Lingjie, et al. "Multi-hop Based Centrality of a Path in Complex Network." 2017 13th International Conference on Computational Intelligence and Security (CIS). IEEE, 2017.
[13] Song, Zhichao, et al. "A novel measure of centrality based on betweenness." 2015 Chinese Automation Congress (CAC). IEEE, 2015.
[14] Ariely, Dan, and Klaus Wertenbroch. "Procrastination, deadlines, and performance: Self-control by precommitment." Psychological science 13.3 (2002): 219-224.
[15] HarvardX-MITx Person-Course Academic Year 2013 De-Identified dataset, version 2.0 - MITx and HarvardX Dataverse", Dx.doi.org.
[16] L. Geng and H. Hamilton, "Interestingness measures for data mining", ACM Computing Surveys, vol. 38, no. 3, p. 9-es, 2006. 10.1145/1132960.1132963.
[17] M. J. Zaki. SPADE: An efficient algorithm for mining frequent sequences. Machine Learning, 42(1/2):31)60, 2001.
[18] Kuncel, Nathan R., Sarah A. Hezlett, and Deniz S. Ones. "A comprehensive meta-analysis of the predictive validity of the graduate record examinations: implications for graduate student selection and performance." Psychological bulletin 127.1 (2001): 162.
[19] A survey", Knowledge-Based Systems, vol. 123, pp. 154-162, 2017. Available: 10.1016/j.knosys.2017.02.009.
[20] G. Adomavicius and A. Tuzhilin, "Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions", IEEE Transactions on Knowledge and Data Engineering, vol. 17, no. 6, pp. 734-749, 2005. Available: 10.1109/tkde.2005.99.
[21] J. Bobadilla, F. Ortega, A. Hernando and A. Gutiérrez, "Recommender systems survey", Knowledge-Based Systems, vol. 46, pp. 109-132, 2013. Available: 10.1016/j.knosys.2013.03.012.
[22] R. Farzan and P. Brusilovsky, "Social navigation support in a course recommendation system", in International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems, Berlin, 2006, pp. 91-100.
[23] H. Thanh-Nhan, H. Nguyen and N. Thai-Nghe, "Methods for building course recommendation systems", in 2016 Eighth International Conference on Knowledge and Systems Engineering (KSE), 2016, pp. 163-168.
[24] A. Al-Badarenah and J. Alsakran, "An Automated Recommender System for Course Selection", International Journal of Advanced Computer Science and Applications, vol. 7, no. 3, 2016. Available: 10.14569/ijacsa.2016.070323.
[25] C. Romero and S. Ventura, "Educational data mining: A survey from 1995 to 2005", Expert Systems with Applications, vol. 33, no. 1, pp. 135-146, 2007. Available: 10.1016/j.eswa.2006.04.005.
[26] D. Yang, M. Piergallini, I. Howley and C. Rose, "Forum thread recommendation for massive open online courses", in Educational Data Mining 2014., 2014.
[27] D. Upendran, S. Chatterjee, S. Sindhumol and K. Bijlani, "Application of Predictive Analytics in Intelligent Course Recommendation", Procedia Computer Science, vol. 93, pp. 917-923, 2016. Available: 10.1016/j.procs.2016.07.267.
[28] A. Tewari, A. Kumar and A. Barman, "Book recommendation system based on combine features of content based filtering, collaborative filtering and association rule mining", in 2014 IEEE International Advance Computing Conference (IACC), 2014, pp. 500-503.
[29] S. Parvatikar and B. Joshi, "Online book recommendation system by using collaborative filtering and association mining", in 2015 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), 2015, pp. 1-4.
[30] F. Wang and H. Shao, "Effective personalized recommendation based on time-framed navigation clustering and association mining", Expert Systems with Applications, vol. 27, no. 3, pp. 365-377, 2004. Available: 10.1016/j.eswa.2004.05.005.
[31] Sokkhey, P., Okazaki, T. (2019, June). Comparative Study of Prediction Models on High School Student Performance in Mathematics. In 2019 34th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC) (pp. 1-4). IEEE.
[32] Roslina, A. H., Noraziah, A. (2010, August). Prediction of hepatitis prognosis using Support Vector Machines and Wrapper Method. In 2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery (Vol. 5, pp. 2209-2211). IEEE.
[33] Liu, J., Danait, N., Hu, S., Sengupta, S. (2013, December). A leave-one-feature-out wrapper method for feature selection in data classification. In 2013 6th International Conference on Biomedical Engineering and Informatics (pp. 656-660). IEEE.
[34] Itoh, K., Itoh, H., Funahashi, K. (2012, August). Forecasting Students' Grades Using a Bayesian Network Model and an Evaluation of Its Usefulensss. In 2012 13th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (pp. 331-336). IEEE.
[35] Shiratori, N. (2017, July). Modeling dropout behavior patterns using Bayesian networks in small-scale private university. In 2017 6th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI) (pp. 170-173). IEEE.
[36] Long, X., Wu, Y. (2012, March). Application of decision tree in student achievement evaluation. In 2012 International Conference on Computer Science and Electronics Engineering (Vol. 2, pp. 243-247). IEEE.
[37] Supianto, A. A., Dwitama, A. J., Hafis, M. (2018, November). Decision Tree Usage for Student Graduation Classification: A Comparative Case Study in Faculty of Computer Science Brawijaya University. In 2018 International Conference on Sustainable Information Engineering and Technology (SIET) (pp. 308-311). IEEE.
[38] Guo, Q., Chen, M., An, D., Ye, L. (2019, June). Prediction of Students' Course Failure Based on Campus Card Data. In 2019 International Conference on Robots Intelligent System (ICRIS) (pp. 361-364). IEEE.
[39] Tan, Pang-Ning, et al. "Introduction to Data Mining (Second Edition)." Introduction to Data Mining, 2018, www-users.cs.umn.edu/ kumar001/dmbook/index.php.
[40] Pedersen, R., Schoeberl, M. (2006, June). An embedded support vector machine. In 2006 International Workshop on Intelligent Solutions in Embedded Systems (pp. 1-11). IEEE.
[41] Zhou, Zhi-Hua. Ensemble methods: foundations and algorithms. Chapman and Hall/CRC, 2012.