DATA MINING FOR BUILDING AN INFORMED DECISION MAKING MODEL FOR CAREER PREDICTION

S. Sathyavathi | N.Niraimathi | K.Priyadarshini

[1](Asst Prof, Dept of IT, Kumaraguru College of Technology, Coimbatore,India, sathyavathi.s.it@kct.ac.in)
[2](Dept of IT, Kumaraguru College of Technology,Coimbatore,India,mathiopn@gmail.com).
[3](Dept of IT, Kumaraguru College of Technology,Coimbatore,India,tomboykpd@gmail.com)

Abstract— *The impact of the credits on the career choice is determined by using mining tools classification (ID3, CHAID). The steps performed for mining the data is classification. The tool used is Rapid miner. The algorithm used for comparing the datasets using classification are ID3, CHAID, Decision tree are used for classification. The performance of the algorithm are compared and it was found that Decision tree provide the maximum accuracy for classifying the fittings prediction based on the skills.*

Keywords— Data mining, classification, Tree induction algorithms, ID3, CHAID, Decision tree.

1. INTRODUCTION

The main objective of higher education institutes is to provide besteducation toits students and to focus the quality of the managerial decision. The Student details can be used to offer a helpful and beneficial recommendations to the academic planners in higher education. And it also helps to understand student's attitude to assist instructors for a better improvised teaching and many other benefits. This paper based on the student's current performance introduce the knowledge of data mining, and select the Classification algorithm using in the analysis of student's performance. Find the Associated factors affecting student's achievement factors, and provide decision support information for educators.Gradepoint average (CGPA) is commonly used to measure the academic potential of a student. But in this paper we add some behavioral attributes that helps to predict student's attitude and performance.

1. Evaluating the personalitytraits and behavioral traits.
2. Prediction based on skills and non-academicparticipation.

This is done by monitoring their performance by considering their 10^{th} mark, 12^{th} mark, CGPA, co-curricular, extra-curricular activities,and participation in paper presentation, workshop, certification course and implant training. Credits are been allocated for each attributes and based upon the credits the students are been categorized and their grades are generated. Ideally early indicators of student's progress can be used to provide appropriate prediction of student success. In this paper, we use Treeinduction algorithms such as Decision Tree,ID3 and CHAID to predict the performance of the students in their academic careers and compare the performance of the Tree induction algorithms.

2. LITERATURE SURVEY

[1]"Data Mining application on student's database"-

The relationship between student's exam result and their success is studied using k-means andcluster analysis techniques. The use of the data mining techniques in education may lead to increase the standard of the education.

[2]"Mining student data to analyze learning behavior"-

In this paper classification algorithm is used to improve the method of teaching by implementing the e-learning system. Certain changes have been made to the existing syllabus. The record includes student's personal or academic details.

[3]"Data mining for education decision support"-

Institution standard is improved by analyzing the curriculum and effectiveness of the course. Data mining algorithm is used for choosing the curriculum amongst the existing system.

[4] "A survey and a data-mining research of recent works"-

By using data mining,learning method is improved by comparing the results of the students periodically by the institution. By analyzing where the student is lagging behind, special coaching is been provided.

[5] "Evaluating student's performance using k-means"-

By considering the academic criteria of the student their performance is evaluated using k-means. Through monitoring and frequent evaluation of the student's performance is done in order to improve their strength and identify their weakness.

[6] "Identification of potential Student academic ability using comparison algorithm K-means"-

In this paper continuous evaluation of student's performance is done and the highly qualified students in academics are been selected. Their high performance in that field is analyzed and they are trained based on their performance by using k-means.

[7] "Predicting student's academic performance using educational data mining"-

In this paper special guidance is given to improve the institute's result by identifying the under performers. More reliable ways are implemented to improve academic results in an efficient way of teaching to achieve the expected result.

[8] "Comparative analysis of decision tree classification algorithms"-

In this paper, Student's performance is analyzed through their cumulative grade point. By using all the classification algorithm the weakness of the student is identified and additional mentoring is provided for the weak student's and their performance is improvised by continuous tracking of the student's details.

[9] "Mining educational data to improve student's performance"-

By applying data mining, student's performance is improved by encountering the failure rate. From the collected data of the student they are been grouped based upon their grades. Data mining is used to improve the grade of the student who have low grades. Rapidminer is used for applying data mining methods. Student's grade is improved by applying mining methods.

[10]"Academic analytics and data mining in higher education" –

This paper helps to restructure the course based upon the performance of the student. Various meetings are conducted to improve the communication between the student and faculty is made mandatory to improve performance of the student in order to evaluate the course impact on the student.

[11] "Performance analysis of UG student's placement selection using decision tree"-

By continuous monitoring of the student's skills, his ability to solve the problem and academics we can identify whether the student will be suitable for placement and what job suits for the student.

[12] "Predicting student's academic performance at degree level"-

The Student'scareer is predicted at the beginning of the course and also at the end of their degree.It helps us to identify the bright student's to guide them in a better way and also to identify the under performers and give special coaching for them. By considering both high performers and under performers details and train them according to their abilities expected results can be achieved.

[13] "A survey of educational data-mining research"-

In this paper the data mining algorithms are been used to enhance the learning abilities of the students. Data mining techniques and tools are been used for improving the learning process of the student. Better understanding of the curriculum made possible by using data mining techniques. Student performance is evaluated based upon the criteria that are mentioned in the course design.

[14]"Educational data mining and learning Analytics"-

In this paper using random forest and clustering algorithm we can understand the relationship between their behavior and learning. Based upon the student's interest learning method is adopted. Each and every student will posses a different learning perspective. Using data mining the standard of the strategic learning method is predicted.

3. PROBLEM DESCRIPTION

The existing system is used to track and analyze the performance of the student's based upon academic marks. Nowadays it alone doesn't help to predict the student's career in Engineering. So in addition to that we add some behavioral attributes that helps to predict student's attitude and performance.

1. Evaluating the personality traits and behavioral traits.
2. Prediction based on skills and non-academic participation.

Appropriate classification algorithm is used to predict the student's performance in their academic careers and compare the performance of Tree induction algorithm namely Decision tree,ID3 and CHAID.

4. MODEL SPECIFICATION

INPUT DATA

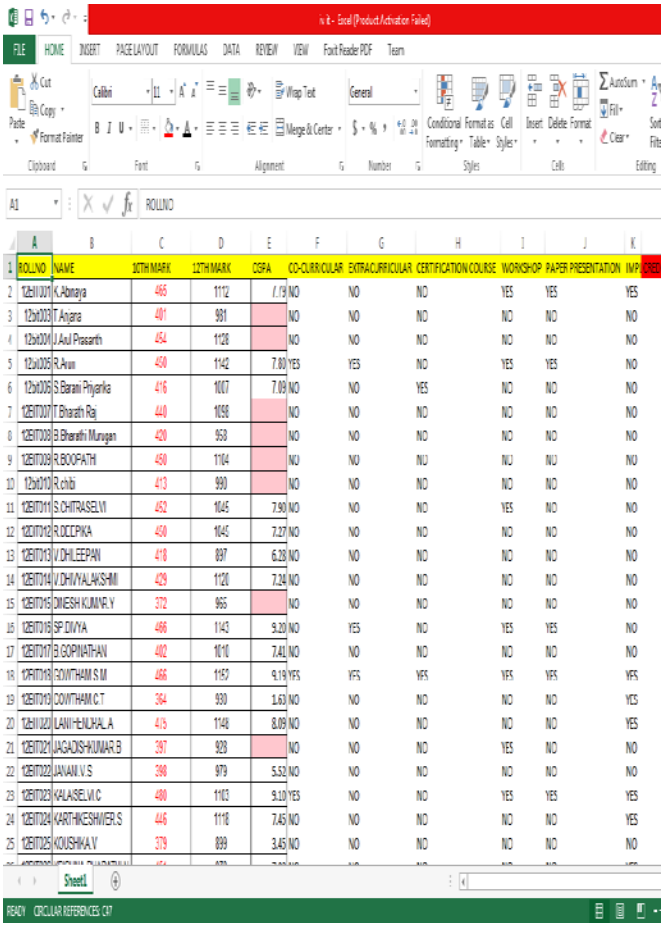
Create a predictive model based on the use of historical data(Past nine years student record(591 records) collected from Information Technology department).The data includes Student's

- Name
- Roll Number
- 10^{th} mark,12^{th} mark, CGPA
- Co-curricular and Extracurricular Activities
- Workshop,paper presentation
- Implant training

This paper reviews Tree induction algorithms and the performance of the algorithms are presented and compared.

MODEL IMPLEMENTATION

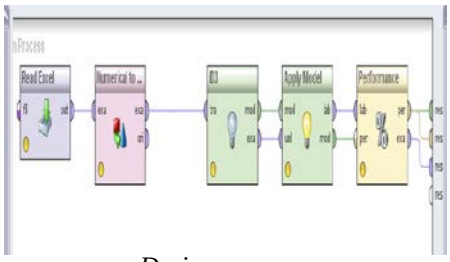
We chose Rapid Miner (RM) for implementation of our model. In our model Input data are imported from Excel format, therefore we need to use the RM import operator Read Excel. For classification, tree induction operator is used.



ROLLNO	NAME	10TH MARK	12TH MARK	CGPA	CO-CURRICULAR	EXTRACURRICULAR	CERTIFICATION COURSE	WORKSHOP	PAPER PRESENTATION	IMPLANT	CREDIT
12bIT001	K.Abinaya	465	1112	7.79	NO	NO	NO	YES	YES	YES	
12bit003	T.Anjana	481	981		NO	NO	NO	NO	NO	NO	
12bit004	U.Asul Praseanth	454	1128		NO	NO	NO	NO	NO	NO	
12bit005	R.Avu	450	1142	7.80	YES	YES	NO	YES	YES	NO	
12bit006	S.Barani Priyanka	416	1007	7.09	NO	NO	YES	NO	NO	NO	
12BIT007	T.Bharath Raj	440	1098		NO	NO	NO	NO	NO	NO	
12BIT008	B.Bharathi Murugan	420	953		NO	NO	NO	NO	NO	NO	
12BIT009	R.BOOFATH	450	1104		NO	NO	NO	NO	NO	NO	
12bit010	R.chibi	413	990		NO	NO	NO	NO	NO	NO	
12BIT011	S.CHITRASELVI	452	1045	7.90	NO	NO	NO	YES	NO	NO	
12BIT012	R.DEEPIKA	450	1045	7.27	NO	NO	NO	NO	NO	NO	
12BIT013	V.DHILEEPAN	418	897	6.23	NO	NO	NO	NO	NO	NO	
12BIT014	V.DHIVYALAKSHMI	429	1128	7.24	NO	NO	NO	NO	NO	NO	
12BIT015	DINESH KUMAR.Y	372	965		NO	NO	NO	NO	NO	NO	
12BIT016	SP.DIVYA	466	1143	9.20	NO	YES	NO	YES	YES	NO	
12BIT017	B.GOPINATHAN	402	1016	7.41	NO	NO	NO	NO	NO	NO	
12BIT018	GOWTHAM S.M	466	1152	9.19	YES	YES	YES	YES	YES	YES	
12BIT019	GOWTHAM C.T	364	930	1.63	NO	NO	NO	NO	NO	YES	
12bIT020	LAVI+ENJALA	415	1148	8.09	NO	NO	NO	NO	NO	YES	
12BIT021	JAGADISH+KUMAR.B	397	928		NO	NO	NO	YES	NO	NO	
12BIT022	JANANI.V.S	398	979	5.52	NO	NO	NO	NO	NO	NO	
12BIT023	KALASELVI.C	480	1103	9.10	YES	NO	NO	YES	YES	YES	
12BIT024	KARTHIKESHWER.S	446	1118	7.45	NO	NO	NO	NO	NO	YES	
12BIT025	KOUSHIKA.V	379	899	3.45	NO	NO	NO	NO	NO	NO	

Training Dataset in Excel Format

In Rapid Miner READ-EXCEL operator is used to read an input data. This operator can be used to load data from Microsoft Excel spreadsheets.

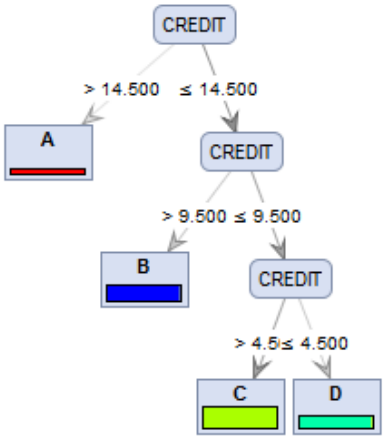


Design process

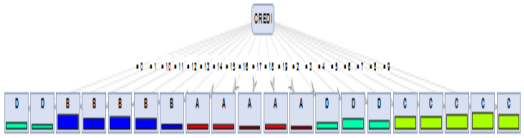
ExampleSet (591 examples, 0 special attributes, 1 regular attribute) Filter (591 / 591 e

Row No.	ROLLNO	GRADE	prediction(...	confidence(...	confidence(...	confidence(...	confidence(A)	IMPLANT	CREDIT
1	12BIT001	B	B	0.995	0	0.005	0	YES	14
2	12bit003	D	D	0	0.982	0.018	0	NO	4
3	12bit004	C	C	0	0.007	0.993	0	NO	8
4	12bit005	A	A	0	0	0	1	NO	15
5	12bit006	C	C	0	0.007	0.993	0	NO	8
6	12BIT007	C	C	0	0.007	0.993	0	NO	5
7	12BIT008	D	D	0	0.982	0.018	0	NO	4
8	12BIT009	C	C	0	0.007	0.993	0	NO	8
9	12bit010	D	D	0	0.982	0.018	0	NO	4
10	12BIT011	B	B	0.995	0	0.005	0	NO	10
11	12BIT012	C	C	0	0.007	0.993	0	NO	8
12	12BIT013	D	D	0	0.982	0.018	0	NO	3
13	12BIT014	C	C	0	0.007	0.993	0	NO	9
14	12BIT015	D	D	0	0.982	0.018	0	NO	3
15	12BIT016	A	A	0	0	0	1	NO	16

Output of the training Dataset in decision tree



Graph view of decision tree



Graph view of both ID3 & CHAID

accuracy: 100.00%

	true A	true C	true B	true D	class precision
pred. A	5	0	0	0	100.00%
pred. C	0	4	0	0	100.00%
pred. B	0	0	5	0	100.00%
pred. D	0	0	0	1	100.00%
class recall	100.00%	100.00%	100.00%	100.00%	

Confusion matrix for decision tree

accuracy: 90.00%

	true A	true C	true B	true D	class precision
pred. A	5	0	0	0	100.00%
pred. C	0	4	0	0	100.00%
pred. B	2	0	7	0	77.78%
pred. D	0	0	0	2	100.00%
class recall	71.43%	100.00%	100.00%	100.00%	

Confusion matrix for ID3

accuracy: 90.00%

	true A	true C	true B	true D	class precision
pred. A	5	0	0	0	100.00%
pred. C	0	4	0	0	100.00%
pred. B	2	0	7	0	77.78%
pred. D	0	0	0	2	100.00%
class recall	71.43%	100.00%	100.00%	100.00%	

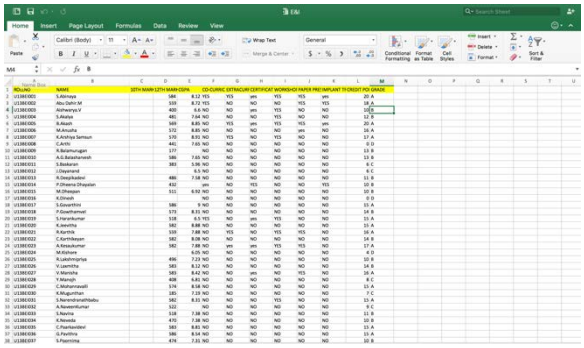
Confusion matrix for CHAID

Training data set gives same accuracy for all three algorithms

accuracy: 99.15%

	true B	true D	true C	true A	class precision
pred. B	182	0	1	0	99.45%
pred. D	0	110	2	0	98.21%
pred. C	0	2	271	0	99.27%
pred. A	0	0	0	23	100.00%
class recall	100.00%	98.21%	98.91%	100.00%	

Electronics and Instrumentation department student data is applied as a test data



Test Dataset in Excel Format

Decision tree gives same accuracy for n records.

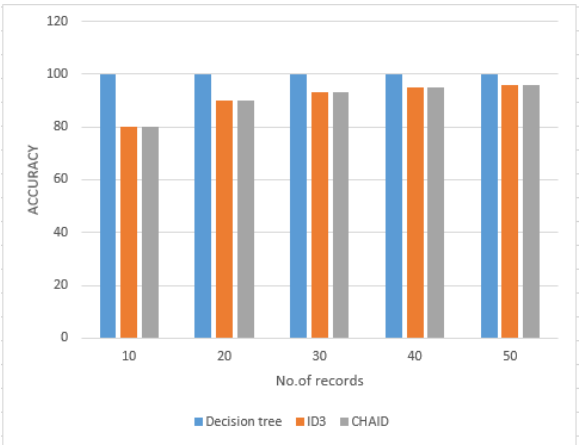
accuracy: 100.00%

	true A	true C	true B	true D	class precision
pred. A	5	0	0	0	100.00%
pred. C	0	4	0	0	100.00%
pred. B	0	0	5	0	100.00%
pred. D	0	0	0	1	100.00%
class recall	100.00%	100.00%	100.00%	100.00%	

According to the number of test data applied ID3 & CHAID algorithm accuracy is varied**.**

No. of test data	ID3	CHAID
10	80%	80%
20	90%	90%
30	93.33%	93.33%

Comparison of Tree Induction Algorithms



5. CONCLUSION

This system is used to predict the career of the student after analyzing all the attributes like CGPA, extra-curricular, behavioral aspect, participation in workshops, paper presentation. For each and every aspect some Credentials are been allocated. Finally based up on the credentials a particular student is been categorized in to Level-A, B, C, D. Analyzing the entire credentials of the student's the success is been predicted for the student. At first prepare the training set for predict the accuracy of Tree induction algorithm using Information Technology student's records. After that Electronics and Instrumentation student's records used as a test data to predict the accuracy of the algorithms. From the obtained results we found that decision tree algorithm gives an accurate result for student's career prediction.

REFERENCES

[1]. Ali Buldu Technical Education of Marmara University, Istanbul, 34722, Turkey "Data mining application on students' data".

[2]. An Approach of Improving Student's Academic Performance by using K-means clustering algorithm and Decision tree by Md. Hedayetul Islam Shovon,MahfuzaHaque.

[3]. C.Anuradha Research Scholar, Bharathiyar University, "A Data Mining based Survey on Student Performance Evaluation System".

[4]. Educational data mining: A survey and a data mining-based analysis of recent works Alejandro Peña-Ayala.

[5]. Evaluating Student's Performance Using k-Means Clustering by Rakesh Kumar Arora, Dr. Dharmendra Badal.

[6]. Identification of Potential Student Academic Ability using Comparison Algorithm K-Means and Farthest First by Athanasies O. P. DewiWiranto H. Utomo Sri Yulianto J. P.

[7]. International Journal of Computer Science and Mobile Computing IJCSMC, Vol. 2, Issue. 7, July 2013, Predicting Students Academic Performance Using Education Data Mining".

[8]. International Journal of Current Engineering and Technology Comparative Analysis of Decision Tree Classification Algorithms.

[9]. International Journal of Information and Communication Technology Research Mohammed M. Abu Tair, Alaa M. El-Halees" Mining Educational Data to Improve Students' Performance".

[10]. Islamic University of Gaza "MINING STUDENTS DATA TO ANALYZE LEARNING BEHAVIOR".

[11]. Paul Baeple "Academic Analytics and Data Mining in Higher Education".

[12]. Performance Analysis of Undergraduate Students Placement Selection using Decision Tree Algorithms by T. Jeevalatha, N.Ananthi, D. Shraavana Kumar.

[13]. Predicting Student Academic Performance at Degree Level: A Case Study by Raheela Asif.

[14]. Richard A. Huebner "A survey of educational data-mining research".

[15]. Ryan S.J.D. Baker, Teachers College, Columbia University "Educational Data Mining and Learning Analytics".