Analysis of Educational Data Mining using Classification

Team Members:

Bandrevu Mounika - 121014009
Davuluri Sri Haneesha - 121014014

Introduction:

Higher education institutions are often very curious to know about the success rate of the students throughout their study. For this reason, they need to use several methods like physical examination, Statistical methods and currently prevailing data mining techniques for the prediction of student's performance. An upcoming area of research which uses techniques of data mining is known as Educational Data Mining. It involves machine learning algorithms and statistical techniques to help the user for interpretation of student's learning habits, their academic performance and further improvement if required. In this paper we will discuss various techniques of data mining which are useful for predicting performance level of students. For this we used the datasets from kalboard 360 and applied it on weka to analyze the data mining techniques.

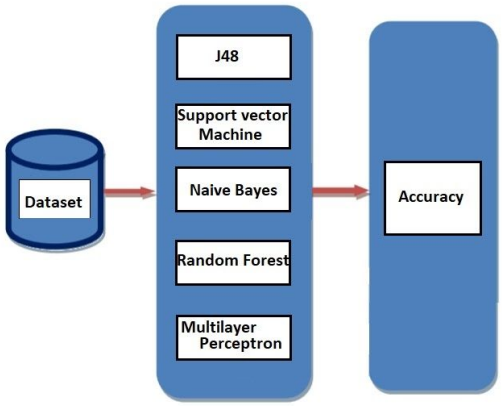

Figure 1: EDM Process

Methodology:

For doing quick analysis on data with the help of data mining techniques, there are many open source softwares like weka, rapid miner, orange, knime, SSDt (SQL Server data Tools) designed for data investigation and to get understandable structure for future use. In this paper, we use WEKA (Waikato Environment for Knowledge Analysis) which is best suited for the analysis of data and to build a model to get predictive outcomes.

Datasets:

In this paper, we are using kalboard 360 dataset which lies in the domain of education and gathered using a learning management system (LMS). This type system always facilitates users with a contemporary use for the resources related to education with the help of an instrument and Internet connection. Collection of data is done through the tool which is called learner activity tracker tool, called experience API (xAPI), a major part of the training and learning architecture (TLA) which authorize to check progress of learning and actions of learner's which may be an article's reading or watching a training video. The experience API helps the learning activity providers to

determine the learner, activity and objects that describe a learning experience. There are 16 features and 480 student records in this dataset.

Data Description:

The dataset is collected through two educational semesters: 245 student records are collected during the first semester and 235 student records are collected during the second semester.
The attributes are as follows:

1. Gender - student's gender (nominal: 'Male' or 'Female')
2. Nationality- student's nationality (nominal:' Kuwait',' Lebanon',' Egypt',' SaudiArabia',' USA',' Jordan','Venezuela',' Iran',' Tunis',' Morocco',' Syria',' Palestine',' Iraq',' Lybia')
3. Place of birth- student's Place of birth (nominal:' Kuwait',' Lebanon',' Egypt',' SaudiArabia',' USA','Jordan','Venezuela',' Iran',' Tunis',' Morocco',' Syria',' Palestine',' Iraq',' Lybia')
4. Educational Stages- educational level student belongs- (nominal: 'lowerlevel', 'MiddleSchool', 'HighSchool')
5. Grade Levels- grade student belongs (nominal: 'G-01', 'G-02', 'G-03', 'G-04', 'G-05', 'G-06', 'G-07','G-08', 'G-09', 'G-10', 'G-11', 'G-12 ')
6. Section ID- classroom student belongs (nominal:'A','B','C')
7. Topic- course topic (nominal:' English',' Spanish', 'French',' Arabic',' IT',' Math',' Chemistry', 'Biology', 'Science',' History',' Quran',' Geology')
8. Semester- school year semester (nominal:' First',' Second')
9. Parent responsible for student (nominal:'mom','father')
10. Raised hand- how many times the student raises his/her hand on classroom (numeric:0-100)
11. Visited resources- how many times the student visits a course content(numeric:0-100)
12. Viewing announcements-how many times the student checks the new announcements(numeric:0-100)
13. Discussion groups- how many times the student participate on discussion groups (numeric:0-100)
14. Parent Answering Survey- parent answered the surveys which are provided from school or not (nominal:'Yes','No')
15. Parent School Satisfaction- the Degree of parent satisfaction from school(nominal:'Yes','No')
16. Student Absence Days-the number of absence days for each student (nominal: above-7, under-7)

The students are classified into three numerical intervals based on their total grade/mark:(*Class*)

1. Low-Level: interval includes values from 0 to 69
2. Middle-Level: interval includes values from 70 to 89,
3. High-Level: interval includes values from 90-100.

Experimentation:

The experimental results and discussion have been done on selecting 163 instances. Five selected classification algorithms were used; Random Forest, Naive Bayse, Multilayer Perceptron, Support Vector machine and J48 each one has its own characteristics to classify the data set. The following are the performance results of all classifiers by using WEKA 3.8.4 version.

❖ **J48 ALGORITHM**

=== Summary ===

Correctly Classified Instances	120	**73.6196 %**
Incorrectly Classified Instances	43	26.3804 %
Kappa statistic	0.6001	
Mean absolute error	0.2215	
Root mean squared error	0.3761	
Relative absolute error	51.0963 %	
Root relative squared error	80.6429 %	
Total Number of Instances	163	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.643	0.172	0.738	0.643	0.687	0.482	0.766	0.645	M
	0.818	0.076	0.800	0.818	0.809	0.737	0.891	0.747	L
	0.796	0.158	0.684	0.796	0.736	0.613	0.854	0.654	H
Weighted Avg.	0.736	0.142	0.738	0.736	0.735	0.590	0.826	0.675	

=== Confusion Matrix ===

```
 a  b  c   <-- classified as
45  8 17 |  a = M
 7 36  1 |  b = L
 9  1 39 |  c = H
```

- **Naive Bayes**

=== Summary ===

Correctly Classified Instances	105	**64.4172 %**
Incorrectly Classified Instances	58	35.5828 %
Kappa statistic	0.4633	
Mean absolute error	0.2347	
Root mean squared error	0.4085	
Relative absolute error	54.1375 %	
Root relative squared error	87.5898 %	
Total Number of Instances	163	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.514	0.237	0.621	0.514	0.563	0.287	0.760	0.633	M
	0.795	0.126	0.700	0.795	0.745	0.644	0.938	0.861	L
	0.694	0.184	0.618	0.694	0.654	0.494	0.881	0.798	H
Weighted Avg.	0.644	0.191	0.641	0.644	0.639	0.446	0.845	0.744	

=== Confusion Matrix ===

```
 a  b  c   <-- classified as
36 14 20 |  a = M
 8 35  1 |  b = L
14  1 34 |  c = H
```

- **Random Forest**

=== Summary ===

Correctly Classified Instances	117	**71.7791 %**
Incorrectly Classified Instances	46	28.2209 %
Kappa statistic	0.5586	
Mean absolute error	0.2613	
Root mean squared error	0.3412	
Relative absolute error	60.2767 %	
Root relative squared error	73.1477 %	
Total Number of Instances	163	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.786	0.333	0.640	0.786	0.705	0.449	0.830	0.806	M
	0.773	0.067	0.810	0.773	0.791	0.716	0.961	0.913	L
	0.571	0.061	0.800	0.571	0.667	0.570	0.927	0.850	H
Weighted Avg	0.718	0.180	0.734	0.718	0.717	0.557	0.895	0.848	

=== Confusion Matrix ===

```
 a  b  c   <-- classified as
55  8  7 |  a = M
10 34  0 |  b = L
21  0 28 |  c = H
```

❖ **Support Vector Machine**

=== Summary ===

Correctly Classified Instances	123	**75.4601 %**
Incorrectly Classified Instances	40	24.5399 %
Kappa statistic	0.6196	
Mean absolute error	0.2822	
Root mean squared error	0.3655	
Relative absolute error	65.0982 %	
Root relative squared error	78.3673 %	
Total Number of Instances	163	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.771	0.258	0.692	0.771	0.730	0.509	0.750	0.625	M
	0.750	0.050	0.846	0.750	0.795	0.728	0.934	0.776	L
	0.735	0.088	0.783	0.735	0.758	0.659	0.885	0.702	H
Weighted Avg.	0.755	0.151	0.761	0.755	0.756	0.613	0.840	0.689	

=== Confusion Matrix ===

a	b	c	<-- classified as
54	6	10	a = M
11	33	0	b = L
13	0	36	c = H

❖ **Multilayer Perceptron**

=== Summary ===

Correctly Classified Instances	126	**77.3006 %**
Incorrectly Classified Instances	37	22.6994 %
Kappa statistic	0.6479	
Mean absolute error	0.1676	
Root mean squared error	0.3655	
Relative absolute error	38.6723 %	
Root relative squared error	78.3633 %	
Total Number of Instances	163	

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.800	0.247	0.709	0.800	0.752	0.547	0.820	0.736	M
0.795	0.050	0.854	0.795	0.824	0.762	0.929	0.834	L
0.714	0.070	0.814	0.714	0.761	0.670	0.919	0.855	
Weighted Avg. 0.773	0.141	0.780	0.773	0.774	0.642	0.879	0.798	

=== Confusion Matrix ===

a	b	c	<-- classified as
56	6	8	a = M
9	35	0	b = L
14	0	35	c = H

Tabulation:

	Classifiers				
Criteria	***Random Forest***	***Naïve Bayes***	***Multilayer Perceptron***	***Support Vector Machine***	***DT - J48***
Accuracy %	67.40%	64.40%	76.07%	75.40%	73.60%
Correctly Classified Instances	110	105	124	123	120
Incorrectly Classified Instances	53	58	39	40	43

Figure 2: Performance Result

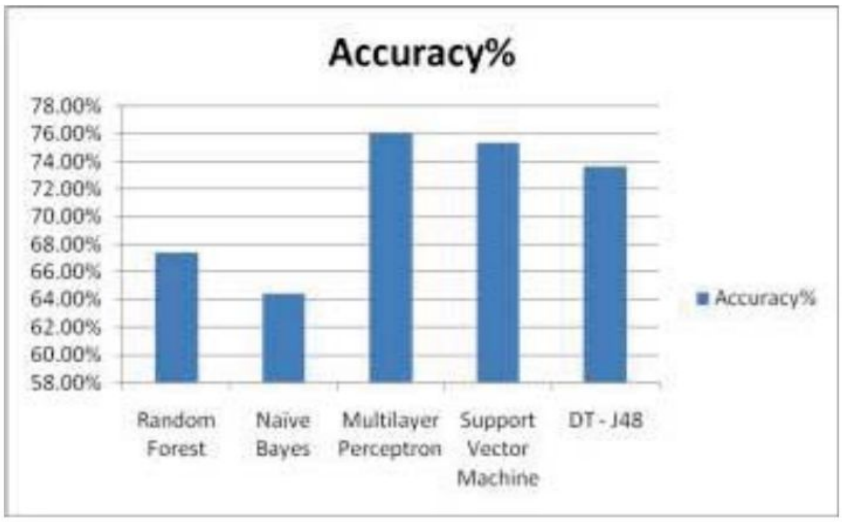

Figure 3: Classifiers Accuracy Performance

Improvement from Paper:

- **Random Tree**

  === Summary ===

Correctly Classified Instances	119	**73.0061 %**
Incorrectly Classified Instances	44	26.9939 %
Kappa statistic	0.5833	
Mean absolute error	0.1921	
Root mean squared error	0.4254	
Relative absolute error	44.3127 %	
Root relative squared error	91.2034 %	
Total Number of Instances	163	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.729	0.258	0.680	0.729	0.703	0.467	0.725	0.603	M
	0.705	0.084	0.756	0.705	0.729	0.635	0.813	0.606	L
	0.755	0.088	0.787	0.755	0.771	0.676	0.830	0.671	H
Weighted Avg.	0.730	0.160	0.733	0.730	0.731	0.575	0.780	0.624	

=== Confusion Matrix ===

a	b	c	<-- classified as
51	9	10	a = M
13	31	0	b = L
11	1	37	c = H

Results:

Data mining has a significant importance in educational institutions. The knowledge acquired by the usage of data mining techniques can be used to make successful and effective decisions that will improve and progress the student's performance in education. Data set contains 163 instances and sixteen attributes. Five classifiers are used under WEKA and the comparisons are made based on the accuracy among these classifiers and different error measures are used to determine the best classifier. Experiments results show that Multilayer Perceptron has the best performance among other classifiers.

Further Improvement Scope:

In future, research related to the same can be done using classification and clustering applications to increase the prediction result in terms of speed and exactness in the field of education.

References:

Base Paper:https://ieeexplore.ieee.org/document/8862214

Dataset: https://www.kaggle.com/aljarah/xAPI-Edu-Data

Software Used: WEKA 3.8.4