Prediction of Work Integrated Learning Placement Using Data Mining Algorithms

Mosima Anna Masethe, Hlaudi Daniel Masethe

Abstract— Data mining in education is used to study data and discover new patterns that can be used for decision making. The classification algorithms are applied on educational data set for predicting work integrated learning placement based on student performance. J48, Bayes Net, Naive Bayes, Simple Cart, and REPTREE algorithms are applied to student data set to predict their performance for placement in the work place. The decision tree from the prediction shows likely students ready for placement in the work environment. The research compares different data mining techniques for classifying student's based on data set for the semester before final examinations.

Index Terms—Algorithms, Classification, Data-Mining, Work-Placement, WIL

I. Introduction

The Placement of students at industry is challenging and depends on a number of factors, such as student academic performance, communication skills, capabilities, problem solving and disability factors, etc. Work Integrated Learning (WIL) is a mechanism to enhance professional practice and development of work readiness skills in graduates [1]. The ability to predict performance of students using data mining techniques will help assist employability practitioners in placement of students for work integrated learning. The research applies data mining techniques to predict if the student in software development will pass or fail the course [2]. WIL in the Universities of Technology in South Africa refers to work based learning activities at an approved industry workstation and the experience that integrate theory and practice [1].

Application of data mining algorithms are helpful to determine information that can be used to establish pedagogical basis for taking educational decisions [3]. The research compiles data set from student data in their final semester of the National Diploma in Software Development. The problem exists in predicting if the student will pass or fail the course in the last semester of their studies, while they also have to be placed for WIL. Predictions are important for identifying and assisting in student's performance for referral [4]. Classification algorithms using a data mining tool are then applied to the data set for analysis and student placement. The patterns to be discovered in the application of educational data mining techniques will be used to enhance decision making in student placement in the work place [5]. The prediction of student placement in their final semester will help employability practitioners and student for proper progress. The research applies data mining techniques such as J48, Bayes Net, Simple Cart, Naïve Bayes, and RepTree classification algorithms to interpret potential and useful knowledge. The predictions of student performance in a university help in identifying outstanding students for industry placement and bursary allocations [4].

The marks acquired by the student during the semester decide his future for placement in the work environment [6]. It is important for the WIL practitioners or employability practitioners to predict whether the student will pass or fail the final examination in the last semester of their studies.

Manuscript received July25, 2014; revised August 15, 2014. This work was supported in part by Tshwane University of Technology.
M.A. Masethe is with the Department of Software Engineering at Tshwane University of Technology, eMalahleni, 1035, South Africa (phone: +2713-653-3187; (e-mail: masethema@tut.ac.za).
H.D. Masethe is with the Department of Software Engineering at Tshwane University of Technology, Pretoria 0001, South Africa (phone: +27 12-382-9714; fax: +27 866-214-011; (e-mail: masethehd@tut.ac.za).

II. Literature

The researchers [7] developed a model to predict students' performance and make specific recommendations that can influence the final examination successfully using decision tree classifier. The data was acquired from Moodle e-learning system and proper transformation and discretization techniques were applied to the data [7]. The researchers [7] experimented with J48 decision tree classifier algorithms and successfully predicted the teaching unit that needs an extra attention to improve student's performance.

The researchers [8] used two approaches, non-parametric regression and ReliefF-based algorithms to predict grades that university students will obtain in the final examinations. The researchers [8] used 10-fold stratified cross-validation with traditional algorithms and compared performance to the proposed fuzzy nonparametric regression and ReliefF-based algorithm. The researchers [8] argue that the proposed technique overcomes the weakness of the traditional algorithm used and has accurate prediction.

The application of the C4.5 decision tree algorithm was used to predict student's academic performance on student's internal assessment data [6]. The researchers [6] used J48

ISBN: 978-988-19252-0-6
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

algorithm in WEKA for prediction analysis and the technique was able to assist the academics to identify weak students and to improve on their performance.

The researchers [9] used Cross Industry for Standard Model for Data Mining (CRISP-DM) and other techniques to study socio-economic relationship on student's performance. The researchers [10] apply Educational Data Mining (EDM) and regression analysis to analyze online learning behaviors and student performance in the course. The researchers [11] applied the kernel method data mining technique to analyze association between behavior and success of students, and established a model for predicting student performance

III. DATA MINING ALGORITHMS

The Data Mining Algorithms defines a method of discovering important patterns or knowledge from a pre-processed data source with an attempt to describe connections between the data and generate a predictive model [7]. Educational Data Mining (EDM) is a science to discover knowledge and techniques to explore data gathered from educational environments [8]. EDM can predict student performance accurately in all institutions of higher learning for categorizing strong and weak students [8].

The acceptance of EDM by institutions of higher learning as an analytical and decision making tool offers prospects to explore unexploited data generated by learning management systems [10]. EDM has the prospective to support universities recognize the dynamics and patterns of a variety of learning atmosphere and to progress student education experience [10]. Adoption of data mining techniques in universities has the prospective to advance quality of education with foundation for operational understanding of the learning process [10]. Data Mining (DM) is a method to determine useful information from stored data [5].

The objective of the prediction methods is to develop a model that can infer characteristic of data or predicted parameter from combination of other data [5]. The task of data mining in this research is to build models for prediction of the class based on selected attributes.

A. *J48 Algorithm*

The J48 classification algorithm is WEKA's version of the implementation of the C4.5 decision tree algorithm, which uses a greedy technique to induce decision trees and make use of reduced- error pruning. The algorithm was developed from ID3 algorithm for handling missing data, continuous data, pruning, splitting and generating rules [12]. The technique uses Gain Ratio instead of Information Gain for splitting purpose [12]:

$$\text{Gain Ratio (D, S)} = \text{Gain (D, S)/ Split INFO}$$

$$\text{Where, Split INFO} = -\left(\sum_{i=1}^{s} \frac{D_i}{D} \log_2 \frac{D_i}{D}\right)$$

In order to categorize a given set, Information Gain as a metric is compulsory, with a function to deliver a balance in the splitting [13]. Providing a data set that contains attributes, we can measure the entropy as a degree of impurity

$$Entropy = \sum_{j} -P_j \log_2 P_j$$

And determining the best attribute for a node in the tree, we use the Information Gain as a measure, such that [13] Information Gain, Gain (S,A) attributes are defines as:

$$Gain(S, A) = Entropy(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

B. *REPTREE Algorithm*

The REPTREE classification algorithm is a technique that builds trees using entropy as impurity measure and also makes use of reduced-error pruning.

C. *NAÏVE BAYES Algorithm*

The algorithm is based on Bayes rule of provisional possibility and adopts independence between attributes values in a data set [14]. The algorithm requires small amount of training data to predict a classification model. The technique signifies a method to probabilistic discovery of knowledge and gives efficient algorithm for data classification [15]. The algorithm makes use of the Bayesian theorem with naïve independent assumptions as in the formula [3]

$$P(Ci \mid X) = \frac{P(X \mid Ci)P(Ci)}{P(X)}$$

D. *CART Algorithm*

The CART algorithm is a binary recursive partitioning technique proficient of processing nominal attributes as predictors [16]. The algorithm offers no interior performance measures for tree selection but measured on independent test data [16].

IV. RESEARCH METHODOLOGY

A. *Data Transformation*

The research follows three basics steps in data mining: pre-processing the collected data to become suitable for data mining, application of data mining algorithms to the data, and post-processing to evaluate and visualize to make the right decisions [7]. The data transformation phase is carried out to improve the data quality input for educational data mining [5]. The dataset parameters and values are extracted from the academic data at the university. The data cleaning

ISBN: 978-988-19252-0-6
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

practice removed and filled the omitted or lacking data

B. Classification Model

Classification, is one task in data mining, a type of machine learning in human learning from past experience to discover new knowledge or prediction method [7]. Classification is a type of prediction that construct a pattern based on training set and uses the pattern to classify a new training set [5].

The research applies J4.8 algorithm which is a revised version 8 of C4.5 decision tree algorithm developed by Quinlan [17]. C4.5 is a modification to the ID3 algorithm [17] with improvements that include dealing with numeric data, missing data and noisy data.

Waikato Environment for Knowledge Analysis (WEKA) is a machine learning and data mining software tool compiled in Java and dispersed under GNU Public license and mostly used by academic researchers [18]. WEKA includes many standard data mining techniques such as classification, regression, clustering and association techniques [18]. In EDM, WEKA has been used for prediction due to its proficiency in discovering, analysis and predicting student's behavior [5].

V. Experimental results and analysis

The study carried out experiments in order to assess the performance of classification techniques for predicting a class based on the data set. The algorithms are evaluated using stratified 10-fold validation.

For the purpose of the research, the self-explanatory attributes in Table I are regarded important and converted into the ARFF format for the WEKA software. The attribute activities considered in the study are Gender, Attendance, Sponsor, and grades for the following subjects (IDC30AT, DSO34AT, DS034BT, and ISY34BT) and Semester Grades which are the most powerful influencer for the final marks of the students. These activities are considered effective to strength the learning process and determine an assessment for performance. The data collection for the purpose of the research was carried out during an academic year from January to May, just before the final examinations. Table I summarizes attribute activities as input variables done by the student in the course. The data transformation involved discretization of the nominal data into categorical classes or attributes manually.

The mark attributes intervals and labels are categorized as follows:

Fail – If value < = 40
Average – If value > 40 and value <50
Pass –If value > =50
Excellent – If value >=60

Table I: Student related variables for WIL placements

Variables	Description	Possible Values
StudentNo	Student Number	Identification
Gender	Student gender	Male, Female
SUB_ID	Subject Identification	Subject Code: (IDC30AT, DSO34AT, DSO34BT, ISY34BT)
SemGrade	Semester Performance	Poor, Average, Good, Excellent
Citizenship	South African or Non-South African	South African (SA), Non-SA
Sponsor	Educational Sponsor	Parents, Student Loan, Scholarship

A. Confusion Matrix of J48 Algorithm

=== Confusion Matrix ===

```
 a  b  c  d   <-- classified as
32  0  0  3 |  a = Fail
 1  6  1  2 |  b = Pass
 0  0 15  0 |  c = Excellent
 4  4  0  1 |  d = Average
```

B. Confusion Matrix of REPTREE Algorithm

=== Confusion Matrix ===

```
 a  b  c  d   <-- classified as
32  1  1  1 |  a = Fail
 1  3  2  4 |  b = Pass
 1  0 14  0 |  c = Excellent
 0  4  2  3 |  d = Average
```

C. Confusion Matrix of NAÏVE BAYES Algorithm

=== Confusion Matrix ===

```
 a b  c  d   <-- classified as
     34 0  1  0 |  a = Fail
1 2 5 2 | b = Pass
0 0 15 0 | c = Excellent
1 3 2 3 | d = Average
```

D. Confusion Matrix of BAYES NET Algorithm

=== Confusion Matrix ===

```
 a  b  c  d   <-- classified as
34  0  0  1 |  a = Fail
 1 4 3 2 | b = Pass
0   1 14  0 |  c = Excellent
1   2 2 4 | d = Average
```

ISBN: 978-988-19252-0-6
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

E. Confusion Matrix of SIMPLE CART Algorithm

=== Confusion Matrix ===

```
 a  b  c  d   <-- classified as
33  0  0  2 |  a = Fail
 1  1  4  4 |  b = Pass
 1  0 14  0 |  c = Excellent
 2  1  1  5 |  d = Average
```

Observation of the confusion matrix shows that Bayes Net and Naïve Bayes algorithm produce good results, strongly suggesting that educational data mining techniques are able to predict student performance. The confusion matrix clearly categorizes the accuracy of the model to successfully identify students likely to pass or fail the course.

The predicted results are compared to the original examination results for the month of June 2014, for the accuracy of the model. The decision tree using the algorithms indicates clearly that's students with poor attendance do not perform well, where else those that have a higher attendance of the classes perform very well in overall. The employability practitioners can then place students based on the predicted results, as the classified model show student performance. Table II and Table III shows classification accuracy based on different techniques applied on the data set, which shows the best classification technique to be Bayes Net algorithm. J48 and Naïve Bayes perform similar in this data set, while Simple Cart algorithm out-performed the REPTREE technique. Naïve Bayes and REPTREE learn rapidly in time to build a new model in the data set.

Table II: Predictive performance of the classifiers

Evaluation Criteria	Classifiers				
	J48	REPTREE	NAÏVE BAYES	BAYES NET	SIMPLE CART
Timing to build model (in sec)	0.07	0.01	0	0.02	0.13
Correctly Classified instances	54	52	54	56	53
Incorrectly Classified instances	15	17	15	13	16
Predictive Accuracy	78.26	75.36	78.26	81.15	76.811

Table III: Comparison of estimates

Evaluation Criteria	Classifiers				
	J48	REPTREE	NAÏVE BAYES	BAYES NET	SIMPLE CART
Kappa Statistics	0.6631	0.6257	0.6619	0.709	0.6395
Mean Absolute Error	0.1224	0.1355	0.1297	0.1197	0.1423
Root Mean Squared	0.2959	0.2809	0.2546	0.2493	0.3154
Relative Absolute Error	36.8391	40.7842	39.0395	36.0516	42.8304
Root Relative Squared Error	72.8102	69.1145	62.6424	61.3555	77.6079

The research made use of classification algorithms incorporating a number of machine learning methods for automatically analyzing the data set provided. The decision tree classifiers allowed a tree-shaped representation of the learning results. The classifier algorithm characterized the student's performance in terms of attributes that are generated from the data set.

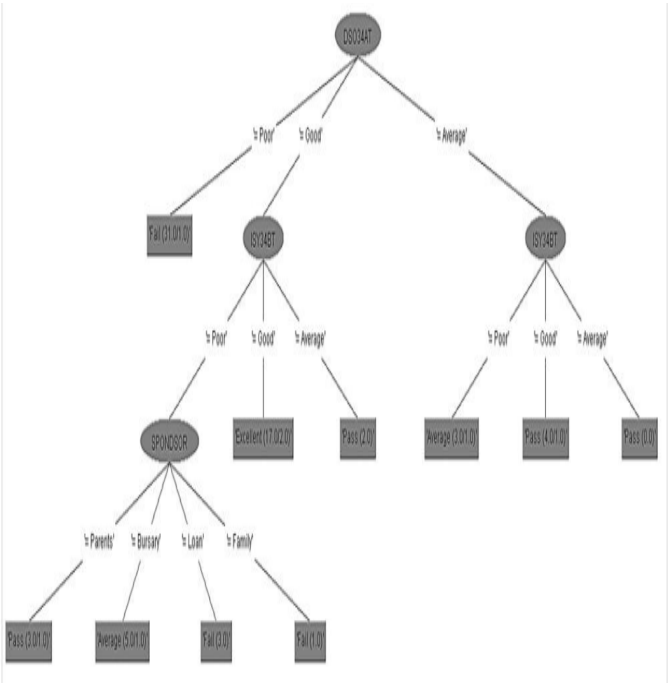

Fig .1. J48 Decision Tree Model

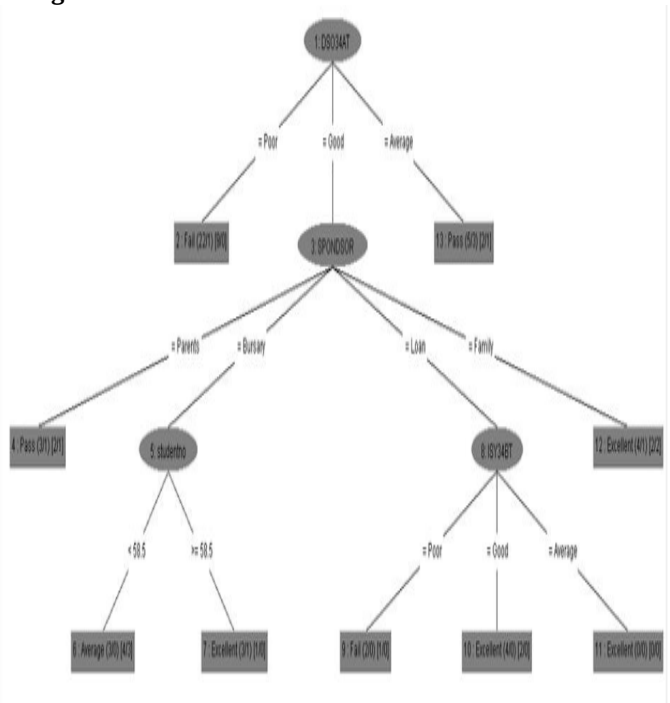

Fig .2. REPTREE Decision Tree Model

ISBN: 978-988-19252-0-6
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

The decision tree model in Figure I, & II provides guidance to WIL coordinators and employability practitioners for work integrated learning placements.

VI. Conclusion

Data mining techniques are effective for implementation on educational data set. The result shows success in implementing educational data mining techniques to classify a decision tree model. The classification techniques have successfully predicted the number of students who are likely to pass their examination for work integrated learning placement. The research findings, shows that Bayes Net algorithm out-performed used data mining techniques in this study.

Acknowledgment

Thanks giving to the institution, Tshwane University of Technology for such great support in everything, starting with finance ending with your words of encouragements to keep on writing, and we further extend our thanks giving to the supportive librarian that we always have Ms. Ruth Segage, from eMalahleni campus**.**

References

[1] D. Jackson, "The contribution of work-integrated learning to undergraduate employability skill outcomes," *Asia-Pacific J. Coop. Educ.*, vol. 14, no. 2, pp. 99–115, 2013.

[2] A. Zafra and S. Ventura, "Predicting Student Grades in Learning Management Systems with Multiple Instance Genetic Programming," no. Mil, pp. 307–314, 2009.

[3] A. K. Pal, "Classification Model of Prediction for Placement of Students," *Int. J. Mod. Educ. Comput. Sci.*, vol. 11, pp. 49–56, 2013.

[4] N. Nghe, P. Janecek, and P. Haddawy, "A comparative analysis of techniques for predicting academic performance," in *37th ASEE/IEEE Frontiers In Education Conference*, 2007, pp. 7–12.

[5] A. AZIZ, N. ISMAIL, and F. AHMAD, "MINING STUDENTS'ACADEMIC PERFORMANCE.," *J. Theor. Appl. Inf. Technol.*, vol. 53, no. 3, 2013.

[6] S. A. Kumar, "EFFICIENCY OF DECISION TREES IN PREDICTING STUDENT ' S ACADEMIC PERFORMANCE," *Comput. Sci. Inf. Technol.*, vol. 2, pp. 335–343, 2011.

[7] S. Milinković and M. Maksimović, "USING DECISION TREE CLASSIFIER FOR ANALYZING STUDENTS ' ACTIVITIES," *JITA*, vol. 3, no. 2, pp. 87–95, 2013.

[8] J. Ghasemian, M. Moallem, and Y. Alipour, "Predicting students ' grades using fuzzy non-parametric regression method and ReliefF-based algorithm," *Adv. Comput. Sci. an Int. J.*, vol. 3, no. 2, pp. 43–51, 2014.

[9] O. State, "Mining Parent Socio- Economic Factors to Predict Students' Academic Performance in Osun State College of Technology, Esa Oke," *Int. J. Eng. Res. Technol.*, vol. 2, no. 12, pp. 1677–1683, 2013.

[10] M'hammed Abdous, W. He, and C.-J. Yen, "Using Data Mining for Predicting Relationships between Online Question Theme and Final Grade," *Educ. Technol. Soc.*, vol. 15, no. 3, pp. 77–88, 2012.

[11] S. Sembiring, M. Zarlis, D. Hartama, S. Ramliana, and E. Wani, "PREDICTION OF STUDENT ACADEMIC PERFORMANCE BY AN APPLICATION OF DATA MINING TECHNIQUES.," in *International Conference on Management and Artificial Intelligence (IPEDR)*, 2011, vol. 6, pp. 110–114.

[12] K. R. Lakshmi, M. V. Krishna, and S. P. Kumar, "Performance Comparison of Data Mining Techniques for Predicting of Heart Disease Survivability," *Int. J. Sci. Res. Publ.*, vol. 3, no. 6, pp. 1–10, 2013.

[13] A. Badr, E. Din, and I. S. Elaraby, "Data Mining : A prediction for Student ' s Performance Using Classification Method," *World J. Comput. Appl. Technol.*, vol. 2, no. 2, pp. 43–47, 2014.

[14] B. K. Bhardwaj, "Data Mining : A prediction for performance improvement using classification," *Int. J. Comput. Sci. Inf. Secur.*, vol. 9, no. 4, 2011.

[15] E. Review, "Data mining approach for predicting student performance," *J. Econ. Bus.*, vol. X, no. 1, pp. 3–12, 2012.

[16] P. Ajith, B. Tejaswi, and M. S. S. Sai, "Rule Mining Framework for Students Performance Evaluation," *Int. J. Soft Comput. Eng.*, vol. 2, no. 6, pp. 201–206, 2013.

[17] K. Daimi and R. Miller, "Analyzing Student Retention with Data Mining.," in *International Conference on Data Mining*, 2009, pp. 55–60.

[18] Baumgartner D and S. G, "Large Experiment and Evaluation Tool for WEKA Classifiers," in *International Conference on Data Mining*, 2009, pp. 340–345.