Students' Performance Prediction by Using Institutional Internal and External Open Data Sources

Farhana Sarker, Thanassis Tiropanis and Hugh C Davis
ECS, University of Southampton, Southampton, United Kingdom
{fs5g09, tt2, hcd}@ecs.soton.ac.uk

Keywords: academic performance, prediction, decision tree, external data sources, internal data sources.

Abstract: The ability to predict students' mark could be useful in a great number of different ways associated with university-level learning. In this study, student's mark prediction models have been developed using institutional internal databases and external open data sources. The results of empirical study for undergraduate students' first year mark prediction show that prediction models based on institutional internal and external data sources provide better performance with more accurate models compared to the models based on only institutional internal data sources. Moreover, this study explores the external data sources (such as National Student Survey result) as one of the best predictors in students' mark prediction. Also, we found that students' first semester performance is the most informative for their first year performance. We envisage that results such as the ones described in this study may increasingly improve the design of future students' predictive models to support students to perform better in their study.

1 INTRODUCTION

The topic of explanation and prediction of academic performance is widely researched. The ability to predict student performance is very important in educational environments. Increasing student success is a long-term goal in all academic institutions. If educational institutions can predict students' academic performance early before their final examination, then extra effort can be taken to arrange proper support for the lower performing students to improve their studies and help them to success.

Students' academic performance is based upon diverse factors like personal, social, psychological and other environmental variables. Various experiments have been carried out in this area to predict students' academic performance. M.N. Quadri and Kalyankar *(2010)* showed that students' performance can be predicted using students' gender, students' parental education, financial background and so on. Al-Radaideh *et al.* (2006) used classification trees to predict the final grade among undergraduate students at Yarmouk University in Jordan. In their study they found high school grade contributed the most in predicting students' final grades. Bharadwaj and Pal (2011) conducted study on the student performance by selecting 300 students from 5 different degree colleges in India. In their study, it was found that students' grade in senior secondary exam, living location, medium of teaching, mother's qualification, family annual income, and student's family status were highly correlated with the student academic performance. Bharadwaj and Pal (2011) in their another study they used students' previous semester marks, class test grade, seminar performance, assignment performance, general proficiency, attendance in class and lab work to predict students' mark in their end semester. Kovacic (2010) used enrollment data to predict successful and unsuccessful student in New Zealand and he found 59.4% and 60.5% of classification accuracy while using decision tree algorithms CHAID and CART respectively. In his study of academic performance prediction, Sajadin Sembiring *et al.* (2011) found Interest, Study Behaviour, Engage Time and Family Support are significantly correlated with the student academic performance. Yadav and Pal (2012) conducted a study using classification tree to predict student academic performance using students' gender, admission type, previous schools marks, medium of teaching, location of living, accommodation type, father's qualification, mother's qualification, father's occupation, mother's occupation, family annual income and so on. In their study, they achieved around 62.22%, 62.22% and 67.77% overall prediction accuracy using ID3, CART and C4.5 decision tree algorithms respectively. In another study Yavev *et al.* (2011) used students' attendance, class test grade, seminar and assignment marks, lab works to predict students' performance at the end of the semester with the help of three decision tree algorithms ID3, CART and C4.5. In their study they achieved 52.08%, 56.25% and 45.83% classification accuracy respectively. Vandamme *et al.* (2007) used decision trees, neural networks and linear discriminant analysis for the early identification of three categories of students: low, medium and high-risk students. Some of the background information such as previous education, number of hours of mathematics, financial independence, and age of the first-year students in Belgian French-speaking universities were significantly related to academic success, while gender, parent's education and occupation, and marital status were not significantly related to the academic success. The Overall correct classification rate they found was 40.63% using decision trees, 51.88% using neural networks and the best result was obtained with discriminant analysis with overall classification accuracy of 57.35%. Cortez & Silva (2008) predicted the secondary student grades of two core classes using past school grades, demographics, social and other school related data. The results were obtained using decision trees, random forests, neural networks and support vector machines. They achieved high level of predictive accuracy when the past grades were included.

The prediction of student performance with high accuracy is beneficial for identify the students with low academic achievements initially. It is required that the teacher can assist the identified students more so that their performance is improved in future. Researchers used various classification methods in their studies to predict students' academic performance, such as decision trees, classification and regression trees, logistic regression, bayesian classification, support vector machine, neural network. Among these decision trees gain popularity in predicting students' performance (Al-Radaideh et al., 2006; B.K. Bharadwaj and Pal., 2011; S. K. Yadav et al., 2011; S. K. Yadev and Pal., 2012). A decision tree is a tree in which each branch node represents a choice between a number of alternatives, and each leaf node represents a decision. Decision tree starts with a root node on which it is for users to take actions. From this node, users split each node recursively according to decision tree learning algorithm (e.g. ID3, C4.5 etc.). The final result is a tree in which each branch represents a possible scenario of decision and its outcome. Among the decision tree algorithms C4.5 gains popularity in terms of its higher performance in classification accuracy (S. K. Yadav et al., 2011; S. K. Yadev and Pal., 2012).

This study aims to explore ways to improve the prediction of students marks by evaluating new predictive models based on the combination of internal higher education institution data sources and external datasets in the linked data cloud. The combination of datasets from internal institutional databases and external data sources presents certain challenges such as data are frequently maintained in different locations, in different formats and often with different identifiers. Data aggregation also presents organizational challenges related to the ownership and use of the data (Arnold, 2010).

Linked data technologies are considered to be well suited for data integration. Linked data is interlinked RDF (Resource Description Framework) data that enables users to retrieve quality information from different data sources[1]. In this study, we examine the sufficiency of existing linked data sources to predict students' first year mark.

In section 2, we define the methodology of this study, section 3 provides the experimentation and results of this study, in section 4 we discuss the findings of this study and section 5 provides conclusion.

2 METHODOLOGY

The purpose of this study is to predict students' mark based on institutional internal datasets and data commonly available in the external open data sources. We used the same variables (as many as available in our internal and external datasets) used by Yadav and Pal (2012) in their studies of predicting students' academic performance. In this study, we considered as institutional internal variables, which are commonly available in the institutional internal databases and external variables are those which can be derived or can be used from institutional external open data sources. At the first step, we developed two models (Model1 and Model2) that based a) on only institutional internal variables and b) on institutional internal variables and external open data sources. Subsequently we extended the above two predictive models (Model3 and Model4) adding students' first semester mark to observe the effect of adding current academic performance on the prediction performance of the models. Moreover, this will help us to analyze the effect of external data sources on the both predictive models before and after adding current academic performance (first semester mark).

[1] http://www.w3.org/DesignIssues/LinkedData.html

Table 1: List of all variables with their description and sources.

Variables name	Description and possible values	Source of the variables
Study_field	Students field of study. Applied (engineering, physics, chemistry etc), Non-applied (Languages etc)	IDS
Gender	Students gender/sex. Male, Female	IDS
Residence	Students Residence/Domicile. UK, Other-EU, Non-EU	IDS
A_level_point	Students result in A level or any other equivalent entry qualifications. A*=140, A=120, B= 100, C=80, D=60 Example, if a student's A level grade is AAA then his A level point counted as AAA=120+120+120=360.	IDS
Adm_Type	Students' admission type. Direct, Clearing	IDS
Accom_Type	Students' accommodation type. University halls, Others	IDS
P_HE	Parents' higher education qualification. Yes, No	IDS
M_Occu_cat	Mother's occupation. Service, House-wife, NA	IDS
F_Occu_cat	Father's occupation. Service, Business, NA	IDS
FirstYr_1stSem_mark	Percentage of mark in first year's first semester. 71%-100%, 61%-70%, 51%-60%, <=50%	IDS
FirstYrMarkrange	Percentage of mark in first year. 71%-100%, 61%-70%, 51%-60%, <=50%	IDS
Part_neighborhood	Students categorized according to their postcode. Lower participation neighborhood, Other neighborhood, NA	EDS (HEFCE)
ONS_soc_eco_class	Students' socio economic class based on parents' occupations. MP-occupations, I-occupations, RM-occupations	EDS (ONS)
P_annual_income	Parents' annual income.	EDS (ONS)
NSS_Q1	Staffs are good at explaining things.	EDS (Unistates)
NSS_Q2	Staffs have made the subject interesting.	EDS (Unistates)
NSS_Q3	Staffs are enthusiastic about what they are teaching.	EDS (Unistates)
NSS_Q4	The course is intellectually stimulating.	EDS (Unistates)
NSS_Q5	The criteria used in marking have been clear in advance.	EDS (Unistates)
NSS_Q6	Assessment arrangements and marking have been fair.	EDS (Unistates)
NSS_Q7	Feedback on my work has been prompt.	EDS (Unistates)
NSS_Q8	I have received detailed comments on my work.	EDS (Unistates)
NSS_Q9	Feedback on my work has helped me clarify things I did not understand.	EDS (Unistates)
NSS_Q10	I have received sufficient advice and support with my studies.	EDS (Unistates)
NSS_Q11	I have been able to contact staff when I needed to.	EDS (Unistates)
NSS_Q12	Good advice was available when I needed to make study choices.	EDS (Unistates)
NSS_Q19	The course has helped me present myself with confidence.	EDS (Unistates)
NSS_Q20	My communication skills have improved.	EDS (Unistates)
NSS_Q21	As a result of the course, I feel confident in tackling unfamiliar problems.	EDS (Unistates)
NSS_Q22	Overall, I am satisfied with the quality of the course.	EDS (Unistates)

In this study we used WEKA for data analysis. The Weka Knowledge Explorer is an easy to use graphical user interface that harnesses the power of the Weka software (G. Holmes *et al.*, 1994). It is an open source software that implements a large collection of machine learning algorithm and is widely used in data mining application (Al-Radaideh et al., 2006; S. K. Yadav et al., 2011; S. K. Yadev and Pal., 2012). In this experimentation we used decision tree classification technique to build the models. The classification tree models have some advantages over traditional statistical models such as logistic regression and discriminant analysis traditionally used in retention studies. First, they can handle a large number of predictor variables, far more than the logistic regression and discriminant analysis would allow. Secondly, the classification tree models are non-parametric and can capture nonlinear relationships and complex interactions between predictors and dependent variable. We used J48 decision tree algorithm to develop the mark prediction model. J48 algorithm is the Weka implementation of the C4.5 top-down decision tree learner proposed by Quinlan (1993). The 10-fold cross validation method was used to validate/evaluate the model in WEKA.

2.1 Data and Data sources

In this study we considered two types of variables, a) variables from institutional internal data sources (IDS) and b) variables from institutional external (open) data sources (EDS).

In this study we motivate to use NSS result published in Unistates website as an external data source to predict students' first year mark. Every year the NSS conducted to measure students' satisfaction in different dimensions of their study subjects in their institutions such as satisfaction in teaching and learning, assessment and feedback, academic support, organization and management, learning resources and personal development. Unistats does not publish individual student data. NSS measures students' satisfaction on their program of study in a 5 points scale (Definitely Disagree, Moderately Disagree, Neither Agree nor Disagree, Moderately Agree, Definitely Agree). The website publishes the percentages of respondents in each scale for an individual course. We considered the actual value (for % Agree) for all the questions for 2010-2011 academic year's published result for the university of Southampton to include in our study to develop the predictive model. Also, Office for National Statistics (ONS[2]) published data has been used in this study to derive parents' annual income (based on ONS published gross annual salary based on SOC2010) and students' socio economic class (based on Standard Occupational Classification 2010). Moreover, we derived participations neighborhood group using Higher Education Funding Council for England (HEFCE[3]) published dataset as some studies found students' participation neighborhood has an impact on students' outcome. Therefore, we considered to include this variable in our prediction model. Table 1 provides the list of all the variables used in this study with their sources.

2.2 Experimentation

The objective of this study is to

a) examine the capability of institutional external open data sources to predict students' mark while combining with only students' enrollment data, and
b) examine the capability of the institutional external data sources while combing with students' enrolment data and current academic performance (students' first semester mark).

Therefore, for the above objectives an analysis of the importance of the variables of the predictive model was necessary. Therefore, we use "select attribute" option from WEKA explorer to select the significant variables/attributes. Considering the variables/attributes with score value greater than "0", we developed four predictive models (Model1, Model2, Model3 and Model4) as described in the methodology. The total number of participants in our study is 149 of which 60.4% is male and 39.4% is female.

3 RESULT AND DISCUSSION

For the first model (model1) we considered 9 input variables/attributes and found all of them are significant and scored greater than "0" for the prediction of first year mark. Table 2 provided the list of these ranked attributes according to their relative importance with their score value. The highest scored attributes are more significant compared to other attributes. From table 2 it is found that students' mark prediction is highly dependent on student' A level point

[2] http://www.ons.gov.uk/ons/index.html
[3] http://www.hefce.ac.uk/

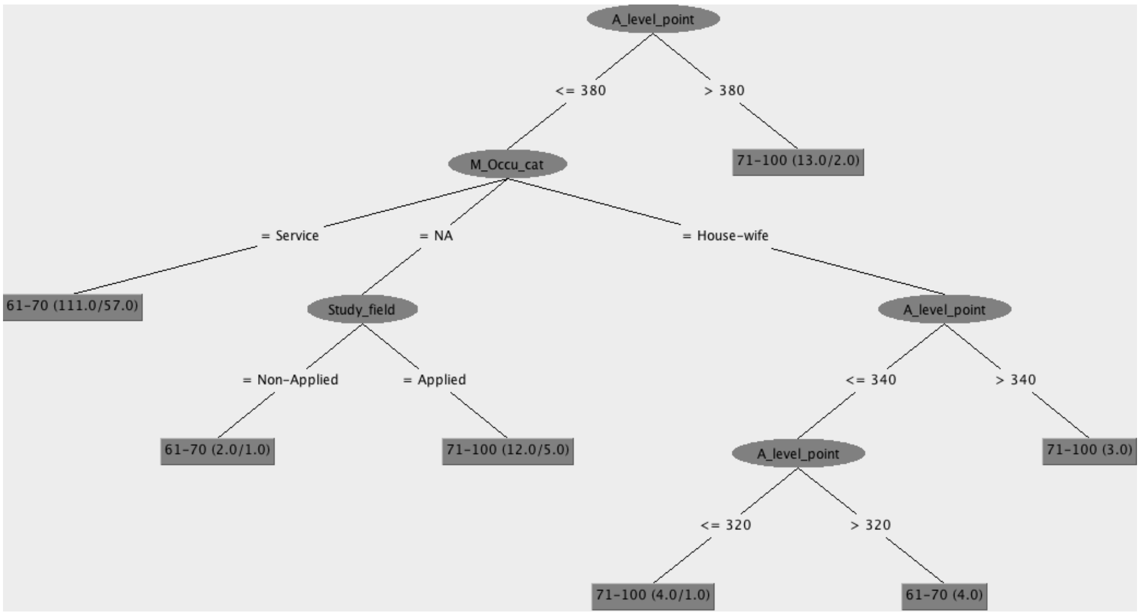

Figure 1: J48 rule for model1

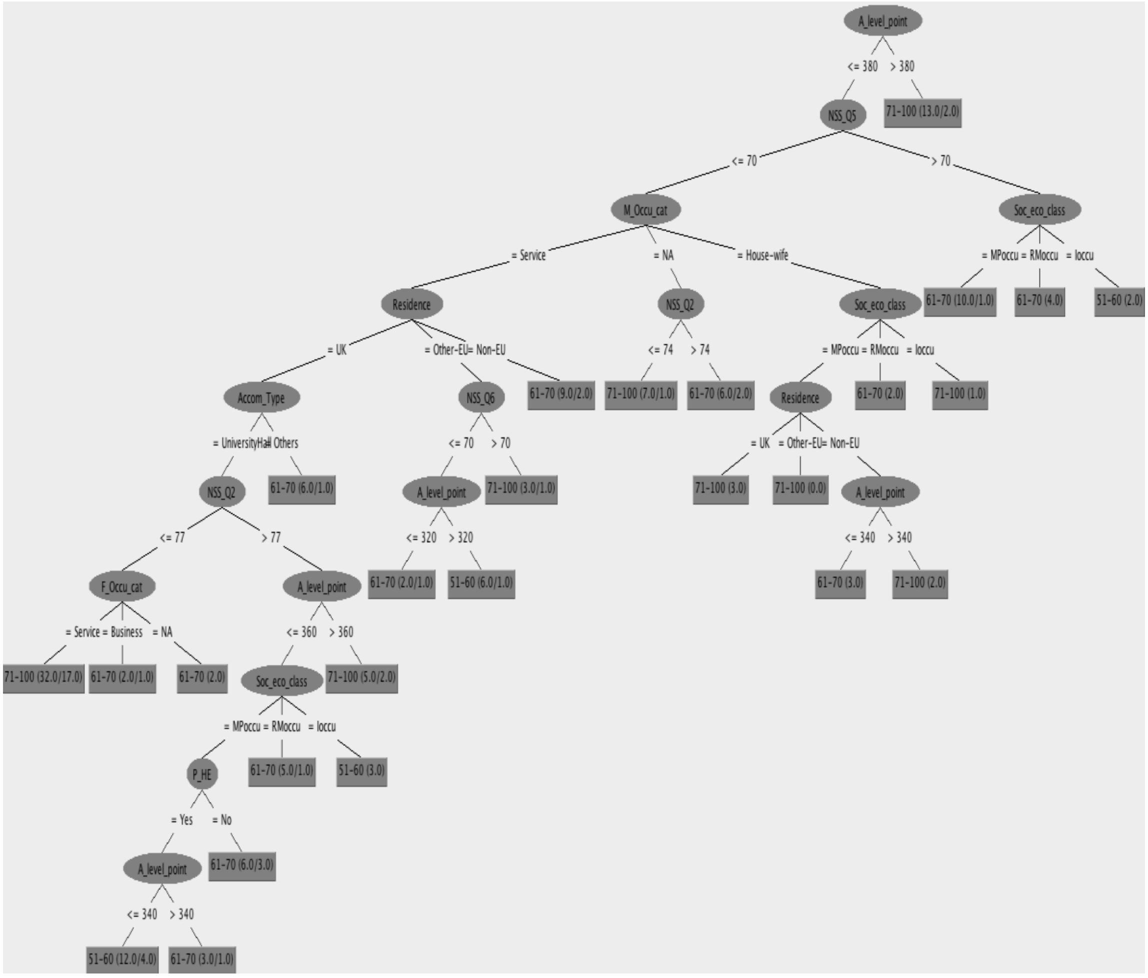

Figure 2: J48 rule for model2

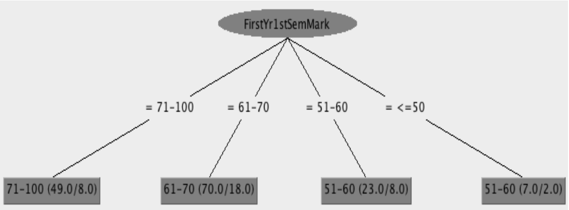

Figure 3: J48 rule for model3

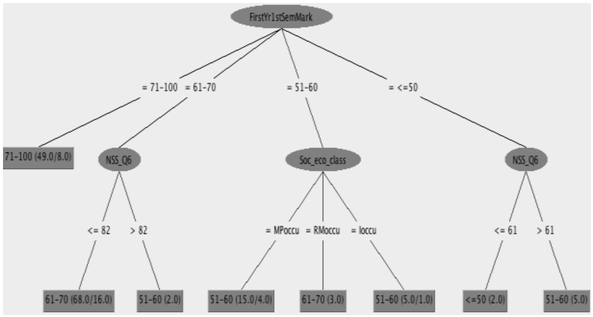

Figure 4: J48 rule for model4

and then mother occupation, field of study, admission type, father occupation and so on. For model2 we considered total 28 internal and external variables of which only 15 variables/attributes are found to be significant and scored greater than "0". Table 3 presents the list of these 15 attributes with their score value. From table 3 it is found that students' mark prediction highly dependent on A level point and NSS five questionnaires (Q2, Q6, Q9, Q5 and Q8). Among other variables mother occupation, study field, admission type, father occupation, socio economic status and so on. For model3 10 out of 10 input variables/attributes are found score value is greater than "0" and considered to build the classification model. Table 4 provides the variables/attributes list with their score values and table 5 provides the list of considered variables for model 4. There are 17 out of 29 internal and external variables/attributes are found to be score value greater than "0" and hence considered for the model development. From table 4 it is found that students' mark prediction is highly dependent on students' first semester mark and then A level point, NSS five questionnaire (Q2, Q6, Q8, Q5, and Q9) and so on. Figure 1, 2, 3 and 4 presents the classification rule generated by J48 decision tree algorithm for model1, model2, model3 and model4 respectively.

The summary of the classification on the datasets using 10-fold cross validation for model1 and model2 presents in table 6; and table 7 presents the accuracy of the classification for model3 and model4. Table 6 shows that model based on institutional internal and external open data sources (model2) performs better in predicting students mark compared to the model using only institutional internal datasets (model1). Compared to the model using only institutional internal databases, model using external data sources achieved around 5.37% more accuracy in the classification. Including students' current academic performance (first semester mark) in the both predictive models (model1 and model2), we get overall accuracy 74.50% for model based on only institutional internal datasets (model3) and 76.51% overall accuracy for model using institutional internal datasets and commonly available external open data sources (model4). It is found that adding students' first semester mark in both models (model1 and model2) the prediction performance increased remarkably from 46.98% to 74.50% and 52.35% to 76.51%. Besides, it is noted that model based on institutional internal and external open data sources performs better among them. Additionally, it can be strongly said that students' first year mark highly dependent on students' current academic performance (first semester mark).

Table 2: Selected variables/attributes with their score for model1

Variables/Attributes name	Score
A_level_point	0.455
M_Occu_cat	0.262
StudyField	0.253
Addm_Type	0.225
F_Occu_cat	0.218
Residence	0.184
Gender	0.17
P_HE	0.126
Accom_Type	0.119

Table 3: Selected variables/attributes with their score for model2

Variables/Attributes name	Score
A_level_point	0.455
NSS_Q2	0.335
NSS_Q6	0.335
NSS_Q9	0.325
NSS_Q5	0.325
NSS_Q8	0.325
M_Occu_cat	0.262
Study_field	0.253
Addm_Type	0.225
F_Occu_cat	0.218
ONS_soc_eco_gp	0.21
Residence	0.184
Gender	0.17
P_HE	0.126
Accom_Type	0.119

Table 4: Selected variables/attributes with their score for model 3

Variables/Attributes name	Score
FirstYr-1stSem_markrange	0.781
A_level_point	0.455
M_Occu_cat	0.262
Study_field	0.253
Addm_Type	0.225
F_Occu_cat	0.218
Residence	0.184
Gender	0.17
P_HE	0.126
Accom_Type	0.119

Table 5: Selected variables/attributes with their score for model 4

Variables/Attributes name	Score
FirstYr-1stSem_markrange	0.781
A_level_point	0.455
NSS_Q2	0.335
NSS_Q6	0.335
NSS_Q8	0.325
NSS_Q5	0.325
NSS_Q9	0.325
M_Occu_cat	0.262
StudyField	0.253
Addm_Type	0.225
F_Occu_cat	0.218
ONS_soc_eco_gp	0.21
Part_neighborhood	0.186
Residence	0.184
Gender	0.17
P_HE	0.126
Accom_Type	0.119

Table 6: Summary of the classification model1 and model2

Model Name	Class	TP Rate	FP Rate	Precision	Recall	Overall accuracy (%)
Model 1	71-100	0.417	0.188	0.513	0.417	46.98
	61-70	0.769	0.69	0.463	0.769	
	51-60	0	0.017	0	0	
	41-50	0	0	0	0	
Model 2	71-100	0.563	0.238	0.529	0.563	52.35
	61-70	0.662	0.393	0.566	0.662	
	51-60	0.276	0.092	0.421	0.276	
	41-50	0	0.021	0	0	

- Model1: based on **only institutional internal database**.
- Model2: based on **institutional internal database and available external data sources**.

Table 7: Summary of the classification model3 and model4

Model Name	Class	TP Rate	FP Rate	Precision	Recall	Overall accuracy (%)
Model 3	71-100	0.854	0.079	0.837	0.854	74.50
	61-70	0.8	0.214	0.743	0.8	
	51-60	0.621	0.083	0.643	0.621	
	41-50	0	0.014	0	0	
Model 4	71-100	0.854	0.079	0.837	0.854	76.51
	61-70	0.815	0.214	0.746	0.815	
	51-60	.69	0.075	0.69	0.69	
	41-50	0	0	0	0	

- Model3: based on **only institutional internal variables** plus **first semester mark**.
- Model4: based on **institutional internal database and available external data sources** plus **first semester mark**.

From table 7 it can also be said that using external data sources in the model improved around 2.01% accuracy of the model compared to the model based on only institutional internal datasets. Table 6 and 7 also presents class wise TP (True positive) rate, FP (False positive) rate, precision and recall value for each model. Our study results also support Kember's study (1995), where the author stated that background characteristics are not good predictors of final outcomes because they are just a starting point. Our study results also strongly support (Zlatko J. Kovacic and Green, 2010), where the authors strongly suggested that the previous academic result plays a major role in predicting students' current academic performance. Also from this study it is evidenced that including external data sources can improve prediction performance. In this study, national student survey (NSS) result contributes significantly in predicting students' mark where five NSS questionnaires (Q2, Q6, Q8, Q5 and Q9) conquered 3rd to 7th significance position while selecting significant variables for the model.

4 CONCLUSION

This study will help to the students and the teachers to improve the performance of the students. This study introduces students mark prediction model development approaches based on institutional internal and external open data sources that can be used in practical settings to predict students' academic performance. The result of this study shows that model based on institutional internal databases and external open data sources performs better than the model based on only institutional internal databases. Furthermore, the result strongly supports that students' current academic performance is the best predictor in predicting students' mark. Among other predictors A level point, NSS results are also highly recommended. This study underlines the importance of linked open data sources in developing predictive models. Therefore, this study suggests more research study using external data sources in this area.

REFERENCES

Al-Radaideh, Q. A., Al-Shawakfa, E. M. & Al-Najjar, M. I. 2006. Mining student data using decision trees, *In the Proceedings of the 2006 International Arab Conference on Information Technology (ACIT'2006)*.

Arnold, K. E. 2010. Signals: Applying Academic Analytics, EDUCAUSE Quarterly (EQ) Magazine, 33(1).

Cortez, P. & Silva, A. 2008. Using data mining to predict secondary school student performance. In the Proceedings of 5th Annual Future Business Technology Conference, Porto, Portugal, 5-12.

B. K. Bharadwaj & Pal, S. 2011. Data Mining: A prediction for performance improvement using classification, *International journal of computer Science and Information security (IJCSIS),* 9(4), 136-140.

B.K. Bharadwaj & Pal., S. 2011. Mining Educational Data to Analyze Students' Performance, *International Journal of Advance Computer Science and Applications (IJACSA),* 2(6), 63-69.

G. Holmes; A. Donkin and I.H. Witten 1994. "Weka: A machine learning workbench". Proc Second Australia and New Zealand Conference on Intelligent Information Systems, Brisbane, Australia. Retrieved 2007-06-25.

Kember, D. 1995. Open learning courses for adults: A model of student progress, *Englewood Cliffs, NJ: Education Technology*.

M.N.Quadri & Kalyankar, D. N. V. 2010. Drop Out Feature of Student Data for Academic Performance Using Decision Tree., *Global Journal of Computer Science and Technology,* 10(2).

Quinlan, R. 1993. *C4.5: Programs for Machine Learning,* Morgan Kaufmann Publishers, San Mateo, CA.

S. K. Yadav, B. K. Bharadwaj & Pal, S. 2011. Data Mining Applications: A comparative study for predicting students' performance, *International journal of Innovative Technology and Creative Engineering (IJITCE),* 1(12).

S. K. Yadev & Pal., S. 2012. Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification, *World of Computer Science and Information Technology (WCSIT),* 2(2), 51-56.

Sajadin Sembiring, M. Zarlis, Dedy Hartama, Ramliana S & Wani, E. 2011. Prediction of student academic performance by an application of data mining techniques *International Proceedings of Economics Development and Research IPEDR,* 6.

Vandamme, J.-P., Meskens, N. & Superby, J.-F. 2007. Predicting academic performance by data mining methods, *Education Economics,* 15(4), 405-419.

Zlatko J. Kovacic & Green, J. S. 2010. Predictive working tool for early identification of 'at risk' students, Open Polytechnic, New Zealand.