RESEARCH ARTICLE OPEN ACCESS

Integrating Weka Into Web Application: Predicting Student's Performance

D.Fatima[1], Dr.Sameen Fatima[2], Dr.A.V.Krishna Prasad[3]
[1]*Assistant Professor, CSE Department, MVSR Engg College, Osmania University, Hyderabad*
[2]*Professor & Principal, University College of Engineering, Osmania University, Hyderabad*
[3]*Associate Professor, CSE Department, MVSR Engg College, Osmania University, Hyderabad*

ABSTRACT
In today's world, there are many stand alone data mining tools that can be used by the academicians to carry out data mining tasks. Any Educational Institute can show the curiosity to know the future performance of recently joined students. To address this, We have analyzed the data set containing information about students, and results in first year of the previous batch of students. By applying the ID3 (Iterative Dichotomiser 3), C4.5, Naive Bayes, Multilayer Perceptron and K-Nearest Neighbour classification algorithms on this data, we have predicted the general and individual performance of freshly admitted students in future examinations and made the entire implementation dynamic to train the prediction parameters itself when new training sets are fed into the web application.
Key Words: Educational Data Mining, Prediction Techniques, Student Performance, PHP-Java-Bridge etc.

Date of Submission: 20-12-2017 Date of acceptance: 03-01-2018

I. INTRODUCTION

In recent years, Data mining has emerged as an important field where a huge data is to be analyzed and important information that makes sense needs to be extracted.

1.1 Problem Statement

Data Mining has wide spread applications across various domains which include education, finance, marketing etc. Various data mining tools have been developed using which information is extracted from large collections of data. However most of the tools developed are stand-alone and only computer professionals can understand and know how to use these tools. Hence it is not easy for someone who is not aware of data mining concepts and tools has to perform processing on data.

1.2 Objective

Here comes the scenario where we need to develop a mechanism that hides different issues in using a tool. If we integrate the tool into the web and hide the complexities, a faculty from any department can use it. Mostly faculties use data mining tools for making predictions for their students about who will become the promising students and who may fail. When the faculty has this information, he can show extra focus on those students and change their teaching styles. This prediction making is a part of the data mining tool. Analyzing the past performance of admitted students would provide a better perspective of the probable academic performance of students in the future. This can very well be achieved using the concepts of data mining. One of the data mining tools is Weka that is written in Java programming language. Integrating Weka into web application and then creating a user friendly interface makes the tool accessible and even a layman can understand its working and can make use of it. And the users can focus on only output.

The Under Graduate Engineering Educational Institutes take the student admissions for the Engineering branches like Mechanical, Civil, CSE, IT, EEE, ECE courses based on Entrance exams and their merit score. The newly joined students are from different boards like (SSC, CBSE, ICSE, Intermediate Board, International etc), different locations and with different background. In this paper, we narrated the data collection, pre-processing, mining process and how we drawn some conclusions on Engineering first year students data of our institute.

1.3 Scope

This Technical paper discusses about the integration of open source data mining tool developed by University of Wakaito called Weka into a web application by taking classification and regression problem to predict student's performance as an example. It can be used to make the classification and regression model

programmatically using classification and regression algorithms provided by Weka. The built classification model is tested for accuracy before being applied to realistic situations like singular and bulk evaluation. This generates arff file format which is the native file format of weka on the fly. This is capable of taking inputs either from an excel file or an open source database like MySQL.

II. LITERATURE SURVEY

The increase of instrumental educational software, the use of the Internet in education, and the establishment of state databases of student information has created large repositories of data. All this information provides a goldmine of educational data that can be explored and exploited to understand how students learn.

Today, almost all the educational institutes facing the problem of handling and managing the exponential growth of educational data for effective use. The data generated by any type of information systems supporting learning or education (in schools, colleges, universities, and other academic or professional learning institutions providing traditional and modern forms and methods of teaching, as well as informal learning) can be analyzed by EDM. These data are not restricted to interactions of individual students with an educational system (e.g., navigation behaviour, input in quizzes and interactive exercises) but might also include data from collaborating students (e.g., text chat), administrative data (e.g., school, school district, teacher), demographic data (e.g., gender, age, school grades), student affectivity (e.g., motivation, emotional states), and so forth. Data generated by educational institutes have unique characteristics such as multiple levels of hierarchy (subject, assignment, question levels), context (a particular student in a particular class encountering a particular question at a particular time on a particular date), fine grained (recording of data at different resolutions to facilitate different analyses), and longitudinal (much data recorded over many sessions for a long period of time, e.g., spanning semester and yearlong courses).

III. IMPLEMENTATION OF A WEB APPLICATION

The architecture is as follows.

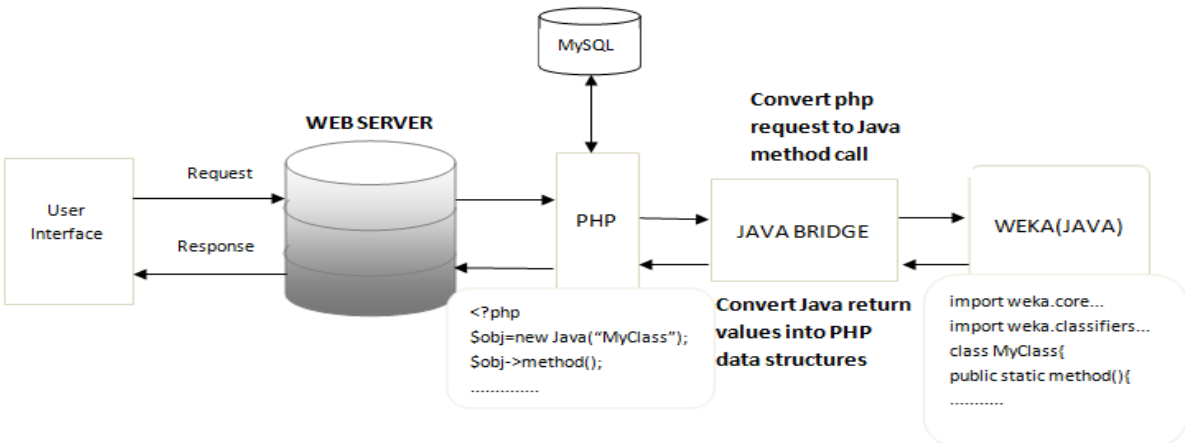


Figure 1: Architectural Design

3.1 Pre-Processing and creation of a model

Firstly, information about students who have been admitted to the second year was collected and saved into the database. This included the details submitted to the college at the time of enrolment. From this data, extraneous information was removed and the relevant information was taken and then it is checked whether it contains any numerical data. If it contains numerical data, we converted that data into discrete data and then saved into another table in database. Once we have this pre-processed data, we then apply the classification and regression algorithms like ID3, C4.5, Naive Bayes, Multilayer Perceptron, K-Nearest Neighbour and Linear Regression algorithm; and create a model which can be used for prediction. This can be elaborated as

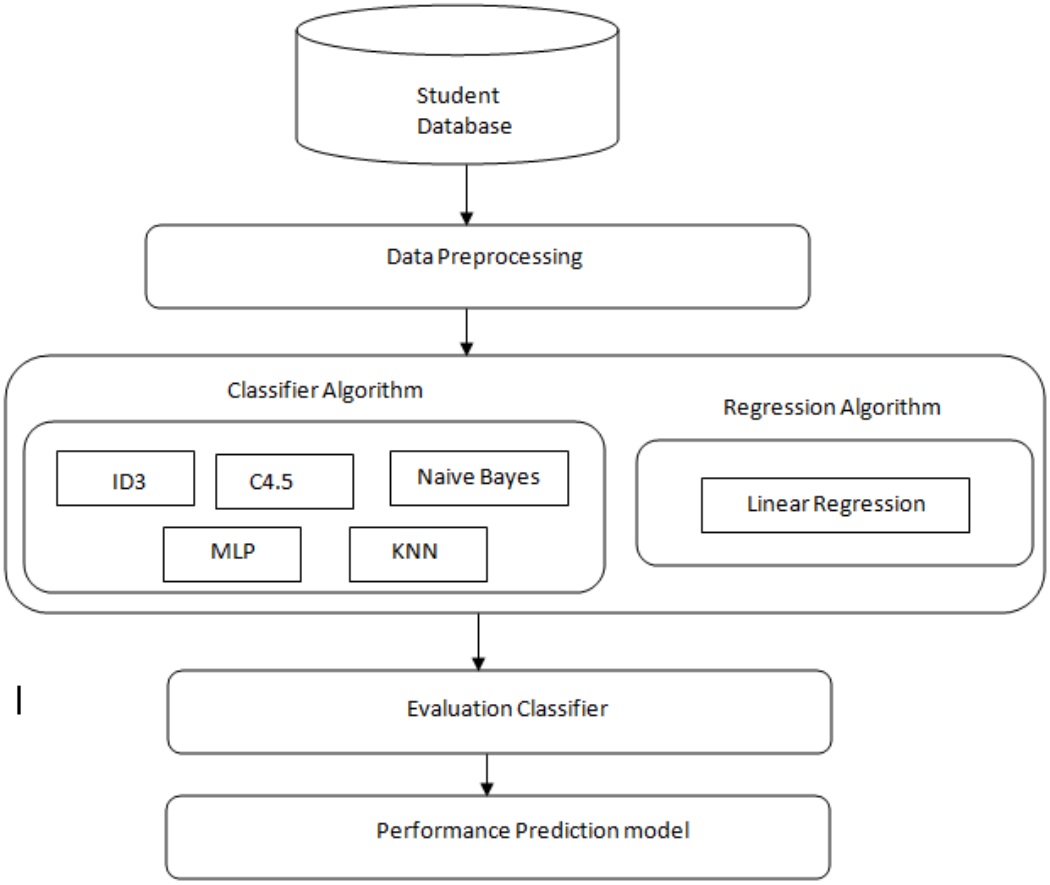


Figure 2: Pre-processing and creation of a model

3.1.1 Student Database

We collected the data from academic branch and created a database for the first year students. We used Microsoft Excel with attributes as Student_full_name, Application_id, Gender, Caste, Percentage_of_marks etc[16]. In board examinations of classes X and XII, percentage of marks obtained in Physics, Chemistry and Mathematics in class XII, marks obtained in the entrance examination, admission type, etc. For ease of performing data mining operations, the data was filled into a MySQL database directly using PHP Excel i.e. the contents of the excel sheet is directly saved as it is into the database.

3.11 Data Pre-Processing

Once we gathered details of all the students, we then partitioned the training dataset further, considering various feasible splitting attributes, i.e. the attributes which would have a higher impact on the performance of a student. For example, 'location' is one such splitting attribute, and then segmented the data according to students' locality.

Here, attributes which are not relevant such as students residential address, name, application ID, etc. had been removed. For example, the admission date of the student was not influential in predicting the future performance of the student. The relevant attributes that had been retained are those for merit score or marks scored in entrance examination, gender, percentage of marks scored in Physics, Chemistry and Mathematics in the board examination of class XII and admission type. Finally, the "class" attribute was added and it held the predicted result, which can be either "Pass" or "Fail" for classification process and the "passpercentage" attribute was added that holds the predicted percentage for regression process.

Since the attributes for marks would are numeric in nature, we discretized it to produce better results, and specific classes were defined. Thus, the "merit" attribute had a value "good" if the merit score of the student was 120 or above out of a maximum score of 200, and was classified as "bad" if the merit score was below 120. Also, the value that can be held by the "percentage" attribute of the student are three - "distinction" if the percentage of marks scored by the student in Physics, Chemistry and Mathematics was 70 or above, class value of "first_class" if the percentage was less than 70 and greater than or equal to 60, then it was classified as "second_class" if the percentage was less than 60. The attribute for admission type is labelled as

"type" and the value held by it can be either "AI" (short for All-India), if the student was admitted to a seat available for All-India candidates, or "OTHER" if the student was admitted to another seat.

1.3.1 Attribute Selection

In this step, we find the attributes that influence the classification more i.e. the influencing attributes are ranked according to the order of influence they have. Attribute selection is performed on the pre-processed data. On the pre-processed data, Chi-Squared Attribute Evaluation and the Ranker algorithm is applied with respect to the class.

IV. RESULTS

The Results are explained with Screen shots. Attributes that influences the classification (ranking of attributes). This page shows the ranking of the attributes that influences the classification more. It also provides the list of algorithms that can be applied on the pre-processed training set.

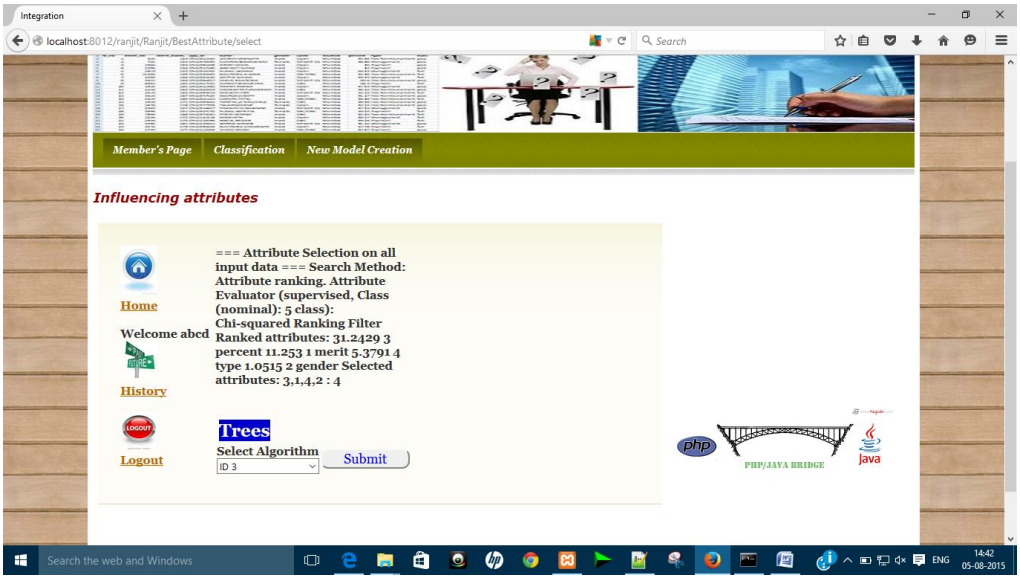


Figure 3: Influencing attributes

Verification of the model created for ID3 algorithm. This page shows the verification of the model that is being created when we selected the ID3 algorithm.

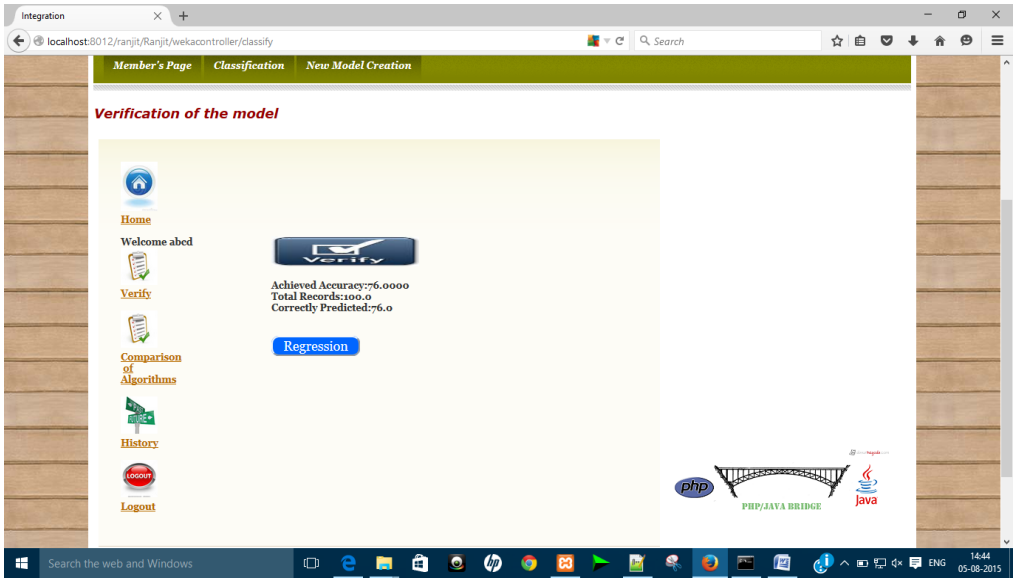
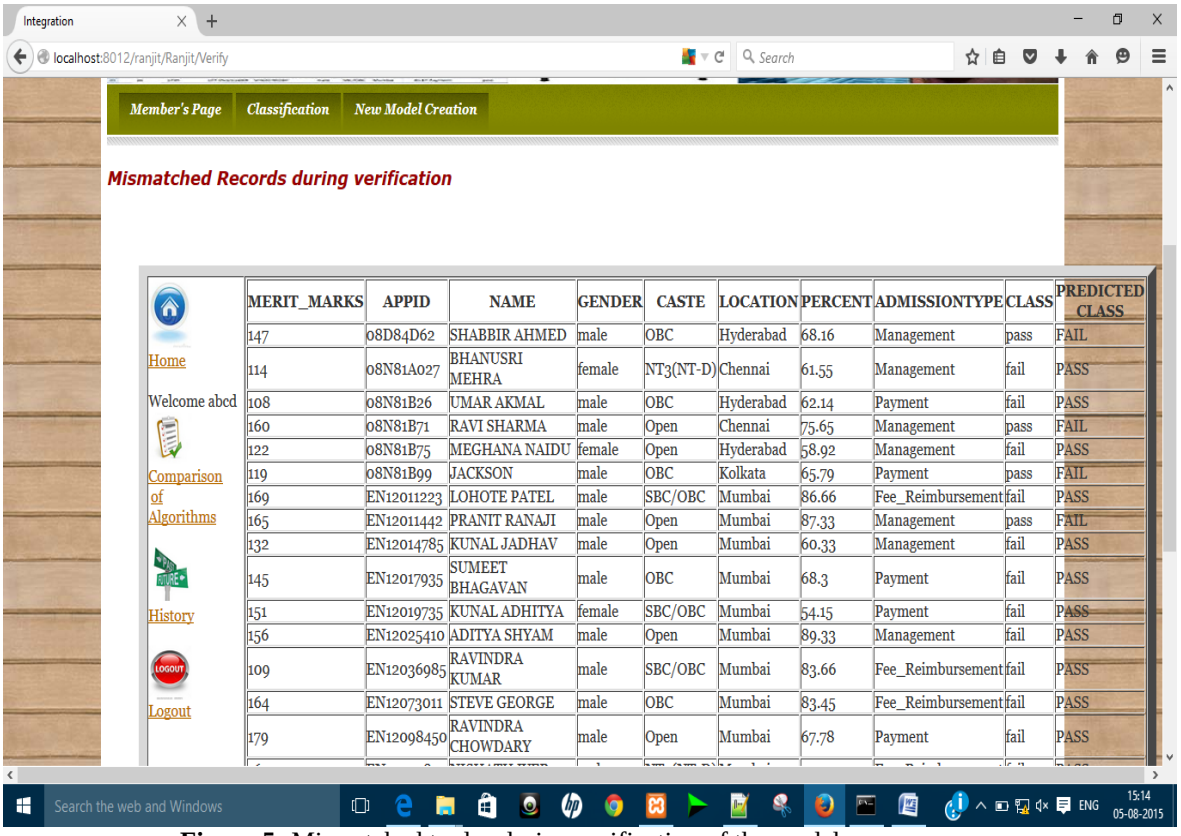


Figure 4: Verification of the model created for ID3 algorithm



MERIT_MARKS	APPID	NAME	GENDER	CASTE	LOCATION	PERCENT	ADMISSIONTYPE	CLASS	PREDICTED CLASS
147	08D84D62	SHABBIR AHMED	male	OBC	Hyderabad	68.16	Management	pass	FAIL
114	08N81A027	BHANUSRI MEHRA	female	NT3(NT-D)	Chennai	61.55	Management	fail	PASS
108	08N81B26	UMAR AKMAL	male	OBC	Hyderabad	62.14	Payment	fail	PASS
160	08N81B71	RAVI SHARMA	male	Open	Chennai	75.65	Management	pass	FAIL
122	08N81B75	MEGHANA NAIDU	female	Open	Hyderabad	58.92	Management	fail	PASS
119	08N81B99	JACKSON	male	OBC	Kolkata	65.79	Payment	pass	FAIL
169	EN12011223	LOHOTE PATEL	male	SBC/OBC	Mumbai	86.66	Fee_Reimbursement	fail	PASS
165	EN12011442	PRANIT RANAJI	male	Open	Mumbai	87.33	Management	pass	FAIL
132	EN12014785	KUNAL JADHAV	male	Open	Mumbai	60.33	Management	fail	PASS
145	EN12017935	SUMEET BHAGAVAN	male	OBC	Mumbai	68.3	Payment	fail	PASS
151	EN12019735	KUNAL ADHITYA	female	SBC/OBC	Mumbai	54.15	Payment	fail	PASS
156	EN12025410	ADITYA SHYAM	male	Open	Mumbai	89.33	Management	fail	PASS
109	EN12036985	RAVINDRA KUMAR	male	SBC/OBC	Mumbai	83.66	Fee_Reimbursement	fail	PASS
164	EN12073011	STEVE GEORGE	male	OBC	Mumbai	83.45	Fee_Reimbursement	fail	PASS
179	EN12098450	RAVINDRA CHOWDARY	male	Open	Mumbai	67.78	Payment	fail	PASS

Figure 5: Mismatched tuples during verification of the model

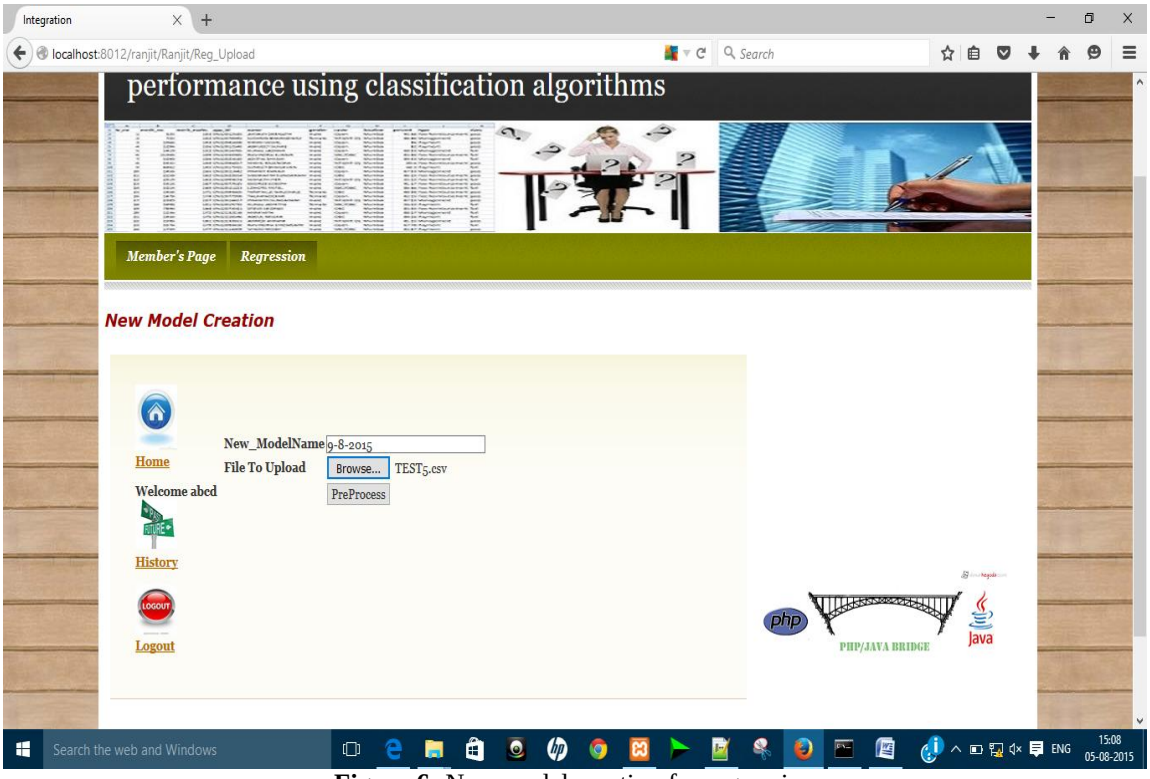


Figure 6: New model creation for regression

This page shows the singular evaluation page where we provide different inputs and the calculated predicted class and pass percentage are shown when we click on submit.

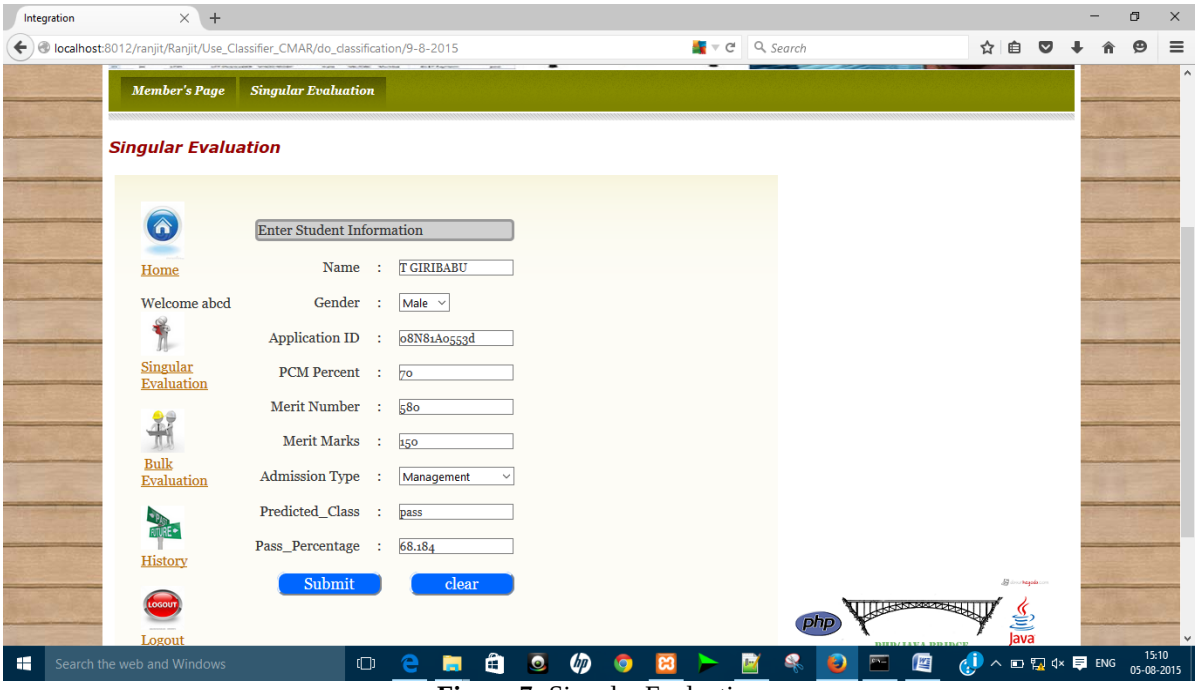
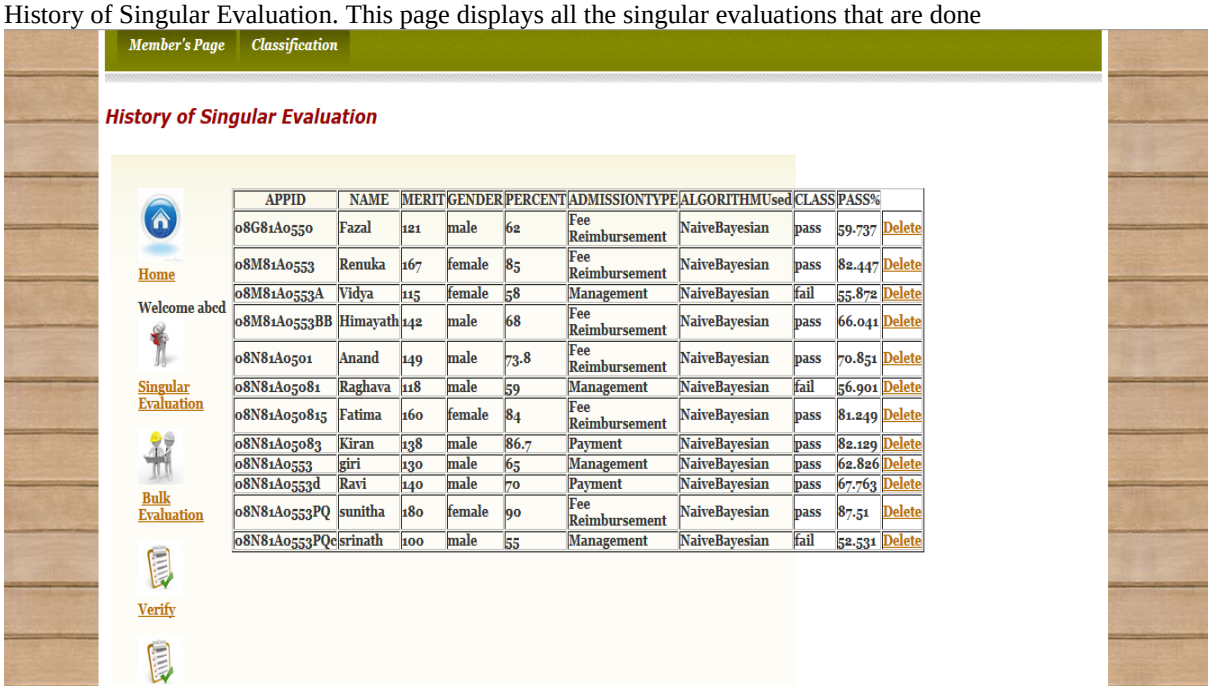


Figure 7: Singular Evaluation

History of Singular Evaluation. This page displays all the singular evaluations that are done



APPID	NAME	MERIT	GENDER	PERCENT	ADMISSIONTYPE	ALGORITHMUsed	CLASS	PASS%	
08G81A0550	Fazal	121	male	62	Fee Reimbursement	NaiveBayesian	pass	59.737	Delete
08M81A0553	Renuka	167	female	85	Fee Reimbursement	NaiveBayesian	pass	82.447	Delete
08M81A0553A	Vidya	115	female	58	Management	NaiveBayesian	fail	55.872	Delete
08M81A0553BB	Himayath	142	male	68	Fee Reimbursement	NaiveBayesian	pass	66.041	Delete
08N81A0501	Anand	149	male	73.8	Fee Reimbursement	NaiveBayesian	pass	70.851	Delete
08N81A05081	Raghava	118	male	59	Management	NaiveBayesian	fail	56.901	Delete
08N81A050815	Fatima	160	female	84	Fee Reimbursement	NaiveBayesian	pass	81.249	Delete
08N81A05083	Kiran	138	male	86.7	Payment	NaiveBayesian	pass	82.129	Delete
08N81A0553	giri	130	male	65	Management	NaiveBayesian	pass	62.826	Delete
08N81A0553d	Ravi	140	male	70	Payment	NaiveBayesian	pass	67.763	Delete
08N81A0553PQ	sunitha	180	female	90	Fee Reimbursement	NaiveBayesian	pass	87.51	Delete
08N81A0553PQc	srinath	100	male	55	Management	NaiveBayesian	fail	52.531	Delete

Figure 8: History of Singular Evaluation

Bulk Evaluation. This page shows the bulk evaluation where we provide a test file that contains many records, batch and the branch.

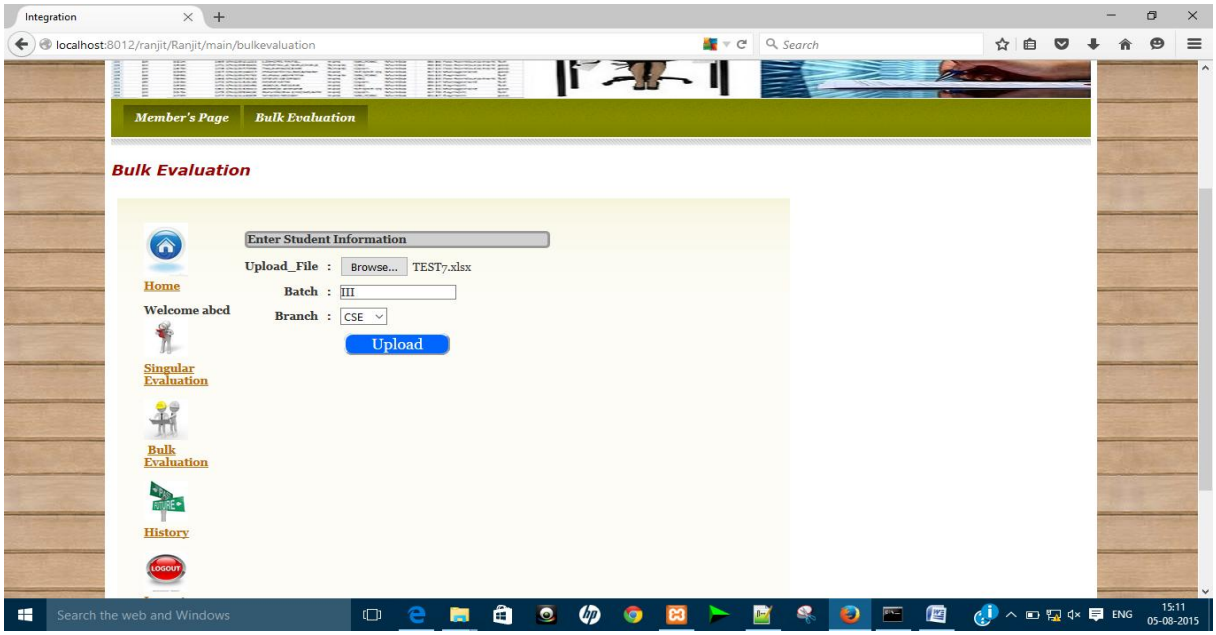
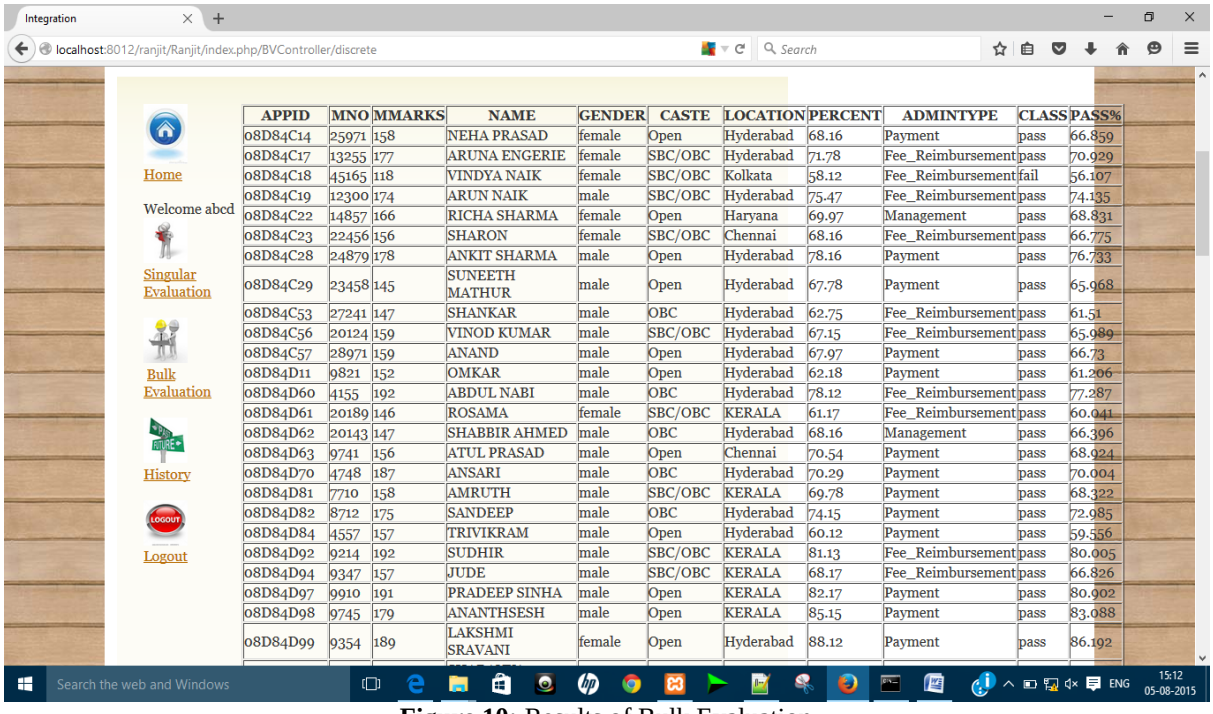


Figure 9: Bulk Evaluation

Results of Bulk Evaluation. This displays the result of bulk evaluation i.e. for the entire input train set, predictions are performed and displayed here.



APPID	MNO	MMARKS	NAME	GENDER	CASTE	LOCATION	PERCENT	ADMINTYPE	CLASS	PASS%
08D84C14	25971	158	NEHA PRASAD	female	Open	Hyderabad	68.16	Payment	pass	66.859
08D84C17	13255	177	ARUNA ENGERIE	female	SBC/OBC	Hyderabad	71.78	Fee_Reimbursement	pass	70.929
08D84C18	45165	118	VINDYA NAIK	female	SBC/OBC	Kolkata	58.12	Fee_Reimbursement	fail	56.107
08D84C19	12300	174	ARUN NAIK	male	SBC/OBC	Hyderabad	75.47	Fee_Reimbursement	pass	74.135
08D84C22	14857	166	RICHA SHARMA	female	Open	Haryana	69.97	Management	pass	68.831
08D84C23	22456	156	SHARON	female	SBC/OBC	Chennai	68.16	Fee_Reimbursement	pass	66.775
08D84C28	24879	178	ANKIT SHARMA	male	Open	Hyderabad	78.16	Payment	pass	76.733
08D84C29	23458	145	SUNEETH MATHUR	male	Open	Hyderabad	67.78	Payment	pass	65.968
08D84C53	27241	147	SHANKAR	male	OBC	Hyderabad	62.75	Fee_Reimbursement	pass	61.51
08D84C56	20124	159	VINOD KUMAR	male	SBC/OBC	Hyderabad	67.15	Fee_Reimbursement	pass	65.989
08D84C57	28971	159	ANAND	male	Open	Hyderabad	67.97	Payment	pass	66.73
08D84D11	9821	152	OMKAR	male	Open	Hyderabad	62.18	Payment	pass	61.206
08D84D60	4155	192	ABDUL NABI	male	OBC	Hyderabad	78.12	Fee_Reimbursement	pass	77.287
08D84D61	20189	146	ROSAMA	female	SBC/OBC	KERALA	61.17	Fee_Reimbursement	pass	60.041
08D84D62	20143	147	SHABBIR AHMED	male	OBC	Hyderabad	68.16	Management	pass	66.396
08D84D63	9741	156	ATUL PRASAD	male	Open	Chennai	70.54	Payment	pass	68.924
08D84D70	4748	187	ANSARI	male	OBC	Hyderabad	70.29	Payment	pass	70.004
08D84D81	7710	158	AMRUTH	male	SBC/OBC	KERALA	69.78	Payment	pass	68.322
08D84D82	8712	175	SANDEEP	male	OBC	Hyderabad	74.15	Payment	pass	72.985
08D84D84	4557	157	TRIVIKRAM	male	Open	Hyderabad	60.12	Payment	pass	59.556
08D84D92	9214	192	SUDHIR	male	SBC/OBC	KERALA	81.13	Fee_Reimbursement	pass	80.005
08D84D94	9347	157	JUDE	male	SBC/OBC	KERALA	68.17	Fee_Reimbursement	pass	66.826
08D84D97	9910	191	PRADEEP SINHA	male	Open	KERALA	82.17	Payment	pass	80.902
08D84D98	9745	179	ANANTHSESH	male	Open	KERALA	85.15	Payment	pass	83.088
08D84D99	9354	189	LAKSHMI SRAVANI	female	Open	Hyderabad	88.12	Payment	pass	86.192

Figure 10: Results of Bulk Evaluation

4.1 Comparison of Classifiers

The algorithm that has the highest accuracy is considered to be the best algorithm and it is dynamically called to carry out the singular and bulk evaluation tasks. I.e. if two or more models are available with same model name but created with different classification algorithms, the model that has the highest accuracy is being selected for carrying out the singular and bulk evaluation. In this table, we will show the comparative study for different algorithms that are being applied. We will compare different aspects like build time required for creating models, true positive rate for class 'pass' and 'fail' of different algorithms, and the accuracy rate obtained.

Execution time to build the model

	ID3	C4.5	NAIVE BAYES	MLP	KNN
Time	0.326sec	0.328sec	0.391sec	4.094sec	0.329sec

Table 1: Execution time to build the model

The above table shows the execution time to build the model when different algorithms are selected.

Comparison of algorithms

	ID3	C4.5	NAIVE BAYES	MLP	KNN
Pass	0.949	0.955	0.955	0.940	0.940
Fail	0.394	0.394	0.424	0.424	0.303

Table 2: Comparison of algorithms

The above table shows the true positive rates for the 'pass' class and 'fail' class when different algorithms are used.

Prediction performance of classification algorithms

	ID3	C4.5	NAIVE BAYES	MLP	KNN
Accuracy	76.0%	77.0%	78.0%	75.0%	73.0%

Table 3: Prediction performance of the classification algorithms

The above table shows the accuracy percentages for different algorithms.

V. CONCLUSION AND FUTURE WORK

In this paper, we used only classification and regression modules of weka. In classification, we used five decision tree algorithms and in regression, we used linear regression and did predictions on the student data set that is collected from academic section. As of now, we predicted whether the student passes or fails and what percentage he may get in his/her final examination. We tested our application with a moderate data set having fields like application id, merit marks, merit number, name, gender, location, admission type, caste and generated a model with an accuracy of 80%. This model also displayed all the mismatched records. Both singular and bulk evaluation are evaluated with equal ease.

There is a scope of predicting whether what class a student gets like first class, second class, distinction i.e. logistical regression can be done, what attendance percentage can a student have etc. Extra-curricular activities can also have significant impact for predicting student's performance. This can be extended to other weka modules like clustering, association rule mining, selecting featured attributes, visualization.

REFERENCES

[1]. Han, J. and Kamber, M., (2006) Data Mining: Concepts and Techniques, Elsevier.
[2]. Dunham, M.H., (2003) Data Mining: Introductory and Advanced Topics, Pearson Education Inc.
[3]. Kantardzic, M., (2011) Data Mining: Concepts, Models, Methods and Algorithms, Wiley-IEEE Press.
[4]. Ming, H., Wenying, N. and Xu, L., (2009) "An improved decision tree classification algorithm based on ID3 and the application in score analysis", Chinese Control and Decision Conference (CCDC), pp1876-1879.
[5]. Xiaoliang, Z., Jian, W., Hongcan Y., and Shangzhuo, W., (2009) "Research and Application of the improved Algorithm C4.5 on Decision Tree", International Conference on Test and Measurement (ICTM), Vol. 2, pp184-187.
[6]. Data mining in Education: ARTICLE in WILEY INTERDISCIPLINARY REVIEWS: DATA MINING AND KNOWLEDGE DISCOVERY · JANUARY 2013 by Cristobal Romero and Sebastian Ventura.
[7]. Educational data mining: A survey from 1995 to 2005 by C. Romero and S. Ventura.
[8]. A Survey on the Classification Techniques In Educational Data Mining by Nitya Upadhyay and Vinodini Katiyar
[9]. RMatlab-app2web: Web Deployment of R/MATLAB Applications by Armin Varmaz, Christian Fieberg and Andreas Varwig.
[10]. CodeIgnitor User Guide Version 2.14, http://ellislab.com/codeigniter/user-guide/toc.html
[11]. Weka programmatic use, https://weka.wikispaces.com/Programmatic+Use

[12]. MySQL – The world's most popular open source database, http://www.mysql.com/
[13]. PHP- http://www.w3schools.com/php/www.wiki.pentaho.com
[14]. Predicting students' performance using id3 and c4.5 classification algorithms. Kalpesh Adhatrao, Aditya Gaykar, Amiraj Dhawan, Rohit Jha and Vipul Honrao Department of Computer Engineering,Fr. C.R.I.T., Navi Mumbai, Maharashtra, India
[15]. A Survey on Predicting Student Performance. A.Dinesh Kumar ,Dr.V.Radhika Sri Krishna Arts and Science College Coimbatore, India

International Journal of Engineering Research and Applications (IJERA) is **UGC approved** Journal with Sl. No. 4525, Journal no. 47088. Indexed in Cross Ref, Index Copernicus (ICV 80.82), NASA, Ads, Researcher Id Thomson Reuters, DOAJ.

D.Fatima "Integrating Weka Into Web Application: Predicting Student's Performance." International Journal of Engineering Research and Applications (IJERA) , vol. 7, no. 12, 2017, pp. 77-85.