CS 773 COURSE PROJECT

AUTHOR: NANDITH REDDY M.

AUGUST 1, 2017

Nmala001@odu.edu

Table of Contents:

Open University Learning Analytics Data

Executive Summary:

The main aim of the project is to analyze and predict the students who are highly at risk of failing the module early in the course and warn them prior so that the students can be wary of their tentative grade in the course and do start doing well later on.

In order to perform the Analysis and develop a system which can correctly predict the students who are at risk of failing the course.

The Types of data that has been used are:

I) Demographic Data(Static) and
II) The Actions or data collected through various actions of the student on Virtual Learning Environment.

Now a days we are aware that most of the Websites and Social media sites have a huge database of the users and every user's activity is tracked and captured using the machine learning and Data Mining techniques so as to develop smart systems that is useful for both users and the businesses mutually.

Hence it becomes much easier than before to analyze the Users actions and create smart systems for predicting and Analyzing the data that is at hand.

In the current system, the data of the student's activity in the Virtual Learning Environment (VLE) is available at hand (for instance number of hits made to a particular resource, the dates of access and also the scores and performance of them in each semester).

For all the weeks we have the different attributes of student's activity apart from their Demographic data that are used to build the predictive models that we aim to develop in the current project.

The Various Models are:

i) Bayesian classifier
ii) K simple means technique using the VLE Data.
iii) J 48 Decision Tree
iv) Decision Table
v) Naive Bayes Classifier

The main of the system is to warn the students of their performance and notify the instructors and management about the students who are in danger of failing if not warned early in the semester.

By using the data that we have at hand for the previous semesters, we aim to build a predictive model to detect the list of students who can pass, fail and who can secure a distinction in the course.

Introduction:

The model that we aim to develop is to help the institutions, managements and also the organizations to implement the timely interventions and help the students to keep their focus on the course and track their performance in periodic intervals.

The timely interventions will for sure help the management in the student retention and also help the students to stay in track with the course.

Data Collection:

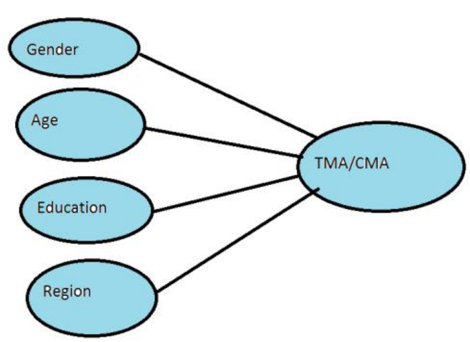
The Data set is obtained from the OU VLE (Virtual Learning Environment) the data collected is then cleaned and also appropriate merge techniques are obtained with the python script to get the proper Testing and Training data

Two Approaches Used:

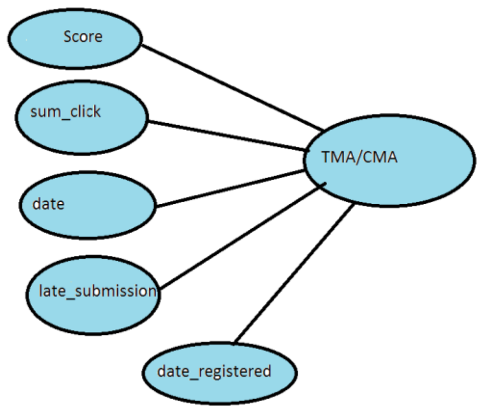
In the predictive Modelling discussed through this paper, we predict the number of instances that are correctly predicted.

The Approaches that we used are:

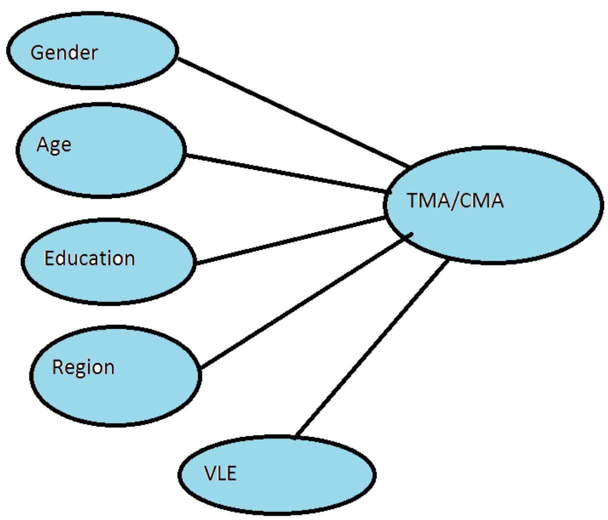
i) **Predictions using only the demographic data:**



ii) **Predictions without using demographic data but VLE data:**



iii) **Predictions Using Both VLE and Demographic Data**



Selection of Attributes for our Analysis:

Demographic Attributes: Gender, Region, Highest Education, band, disability

VLE Attributes: date registered, sum click and score are among the important VLE attributes that we used

Predictive Modelling:

The predictive modelling that we adopted is the total number of clicks that the student has made on a week to week basis and how it is greatly impacting the result of the student.

The sum of clicks for each week till a particular assessment date are calculated. And also the scores obtained by the student in each of the assessments is taken into consideration.

Thus all the important attributes required for building the model are obtained.

After getting the comprehensive dataset, we divided the dataset into two. 70% of the data is considered as the training set and the rest 30% as the testing data.

Naïve Bayes:

This algorithm is used as the overall average error is much lesser than the other algorithms.

J48:

By applying a decision tree like **J48**on that dataset would allow you to predict the target variable of a new dataset record.

Decision Tree:

A decision tree is a graph that uses a branching method to illustrate every possible outcome of decision. Programmatically, they can be used to assign monetary/time or other values to possible outcomes so that decisions can be automated.

Problem Statement:

The model that we aim to develop is to help the institutions, managements and also the organizations to implement the timely interventions and help the students to keep their focus on the course and track their performance in periodic intervals.

The timely interventions will for sure help the management in the student retention and also help the students to stay in track with the course.

Solution Methodology:

The solution for this project is based on the data collected from the VLE. All the files that have been provided have been looked into individually and then decide on the attributes that can be the significant indicators to determine the “final result” of the student.

The tools that we used for the project are Weka , R and Python using Jupyter.

The data sets are merged by performing a series of python scripting using jupyter notebook.

1. The significant attributes for the data analysis can be obtained through the information gain

Experimental Setup and Data Used:

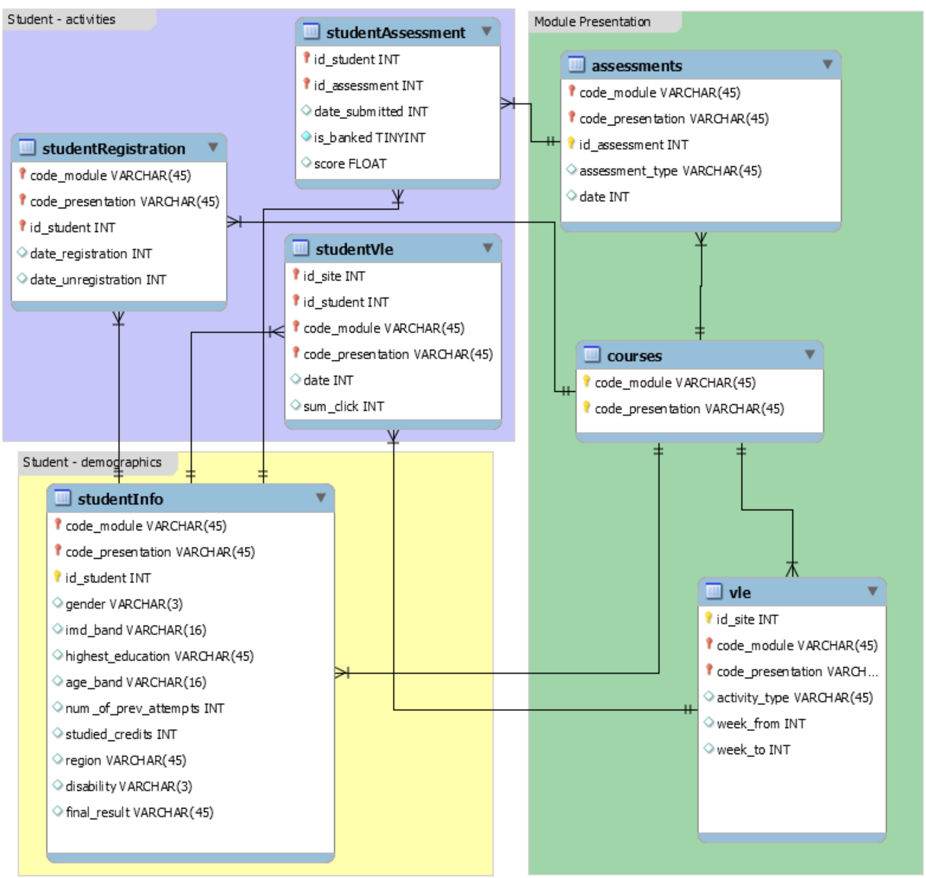


Figure 1: OULAD data set schema [12]

Step1:

In the initial phase the data is cleaned and seen if there are any duplicates or the messy data in our files so as to avoid any redundant data from the system. This will help us create the predictive model with some amount of accuracy.

Step2:

The tables have to be properly merged with appropriate association rules making all the necessary join. So that we have all the attributes that are necessary to build our model.

Step3:

For Merging the Data, We used the python scripting using anaconda and Jupiter Notebook environments. In addition to the Jupiter Notebook, we also used Weka to load the Training and Testing data. 70% of the instances have been taken as the training data and the rest 30 % as the training data out of the 10000 instances taken randomly.

Step4:

First the loaded data is preprocessed and the appropriate classifiers are used to get the predictive modelling and obtain the results.

Step5:

Finally, All the Classifiers are compared according to the amount of accuracy achieved.

Information Gain:

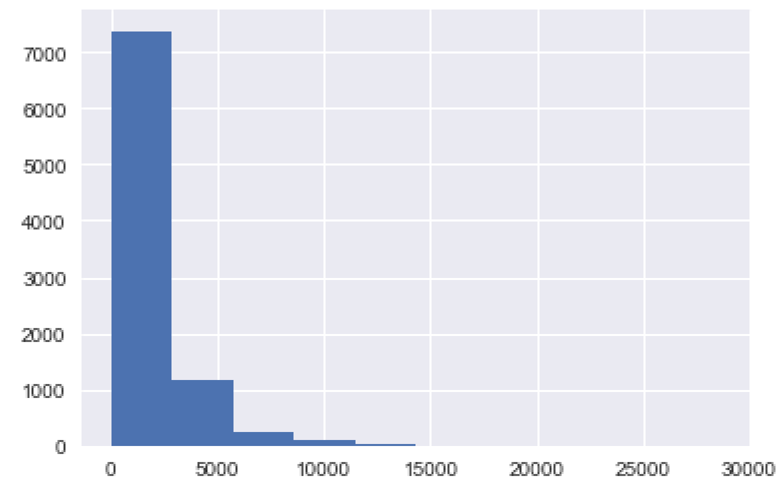
Attribute	Info Gain
Code_module	0.03331
Studied_credits	0.0291
Imd_band	0.0194
Highest_education	0.0220
Date_registration	0.0126
Num_prev_attempts	0.0112
region	0.0099
Code_presentation	0.0090
gender	0.000365
Age_band	0.00477

Table1: Info gain of the merged data set

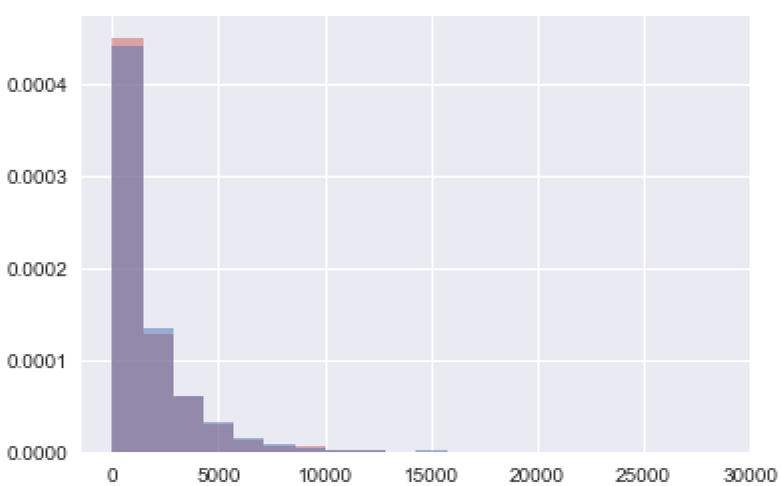
Results:

The Naïve Bayes classifier algorithm is run in the python script by providing 70% Training data and 30% testing data.

Here are the results obtained:

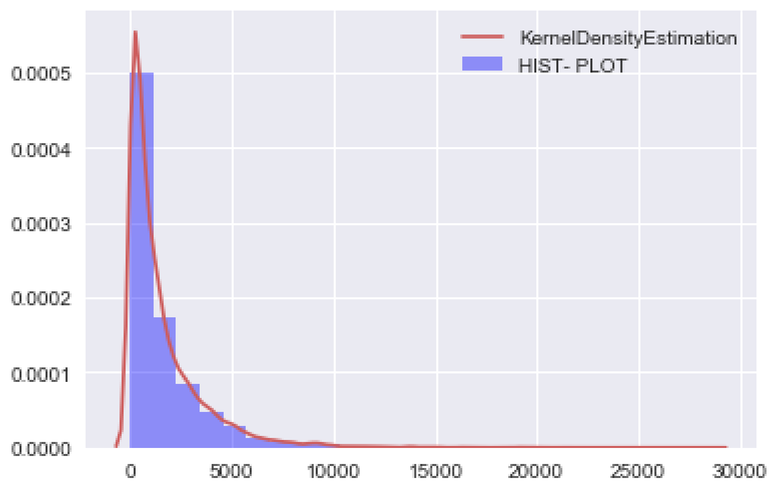


Histogram for test data array between sum_click(x-axis) and total students passed(y-axis)

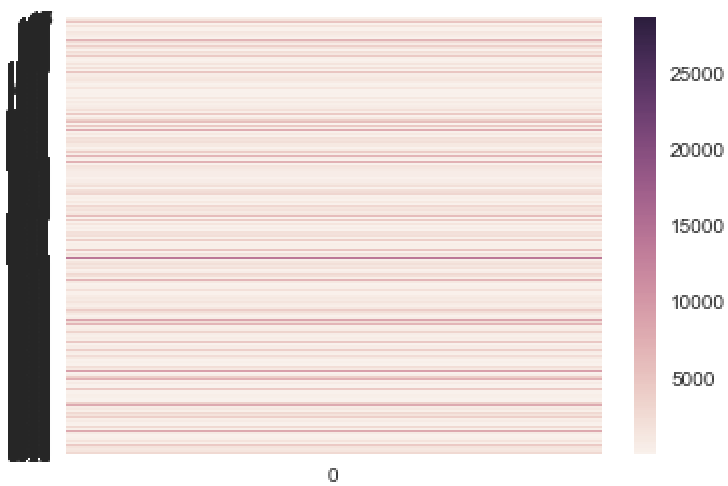


Histogram for test data array between sum_click(x-axis) and total students passed(y-axis)

For training and testing data



Histogram and Kernel Density Estimation for test data array between sum_click(x-axis) and total students passed(y-axis)



Heat Map Obtained for the Test Array shows the distribution of number of sum clicks

For the total of 8000 instances provided, the classifier predicted the results with 47% accuracy.

Running The classifiers on Weka:

By taking the attributes final result and Scores:

Naïve Bayes:

i) **When test set is supplied:**

Correctly Classified Instances	69	57.5 %
Incorrectly Classified Instances	51	42.5 %

ii) **When Cross Validation Is used:**

Correctly Classified Instances	3983	54.4945
Incorrectly Classified Instances	3326	45.5055 %

iii) **When Percentage split is 66%:**

Correctly Classified Instances	1357	54.6076 %
Incorrectly Classified Instances	1128	45.3924 %

J48 Classifier:

i) **When test set is supplied:**

Correctly Classified Instances	67	55.8333 %
Incorrectly Classified Instances	53	44.1667 %

ii) **When Cross Validation Is used:**

Correctly Classified Instances	3993	54.6313 %
Incorrectly Classified Instances	3316	45.3687 %

iii) **When Percentage split is 66%:**

Correctly Classified Instances	1352	54.4064 %
Incorrectly Classified Instances	1133	45.5936 %

Decision Tree:

i) **When test set is supplied:**

Correctly Classified Instances 69 57.5 %
Incorrectly Classified Instances 51 42.5 %

ii) **When Cross Validation Is used:**

Correctly Classified Instances 3965 54.2482 %

Incorrectly Classified Instances 3344 45.7518 %

iii) **When Percentage split is 66%:**

Correctly Classified Instances 1352 54.4064 %

Incorrectly Classified Instances 1133 45.5936 %

By taking the attributes final result and Sum_click:

Decision Tree:

i) **When test set is supplied:**

Correctly Classified Instances 69 57.5 %
Incorrectly Classified Instances 51 42.5 %

ii) **When Cross Validation Is used:**

Correctly Classified Instances 1425 52.9543 %

Incorrectly Classified Instances 1266 47.0457 %

iv) **When Percentage split is 66%:**

Correctly Classified Instances 476 52.0219 %

Incorrectly Classified Instances 439 47.9781 %

J48 Classifier:

iv) **When test set is supplied:**

Correctly Classified Instances 69 57.5 %

Incorrectly Classified Instances 51 42.5 %

v) When Cross Validation Is used:

Correctly Classified Instances 1425 52.9543 %

Incorrectly Classified Instances 1266 47.0457 %

vi) When Percentage split is 66%:

Correctly Classified Instances 1352 54.4064 %

Incorrectly Classified Instances 1133 45.5936 %

Naïve Bayes:

iv) When test set is supplied:

Correctly Classified Instances 69 57.5 %

Incorrectly Classified Instances 51 42.5 %

v) When Cross Validation Is used:

Correctly Classified Instances 1425 52.9543 %

Incorrectly Classified Instances 1266 47.0457 %

vi) When Percentage split is 66%:

Correctly Classified Instances 473 51.694 %

Incorrectly Classified Instances 442 48.306 %

A. With All Demographic Data:

Naive Bayes

Supplied Test set:

=== Summary ===

Correctly Classified Instances 57 47.5 %

Incorrectly Classified Instances 63 52.5 %

Kappa statistic 0.2086

Mean absolute error 0.2814
Root mean squared error 0.3977
Relative absolute error 91.3309 %
Root relative squared error 102.5952 %
Coverage of cases (0.95 level) 94.1667 %
Mean rel. region size (0.95 level) 68.125 %
Total Number of Instances 120

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.000	0.009	0.000	0.000	0.000	-0.028	0.735	0.170	Fail
	0.435	0.275	0.682	0.435	0.531	0.164	0.663	0.712	Pass
	0.724	0.352	0.396	0.724	0.512	0.321	0.802	0.628	Distinction
	0.500	0.148	0.273	0.500	0.353	0.273	0.705	0.348	Withdrawn
Weighted Avg.	0.475	0.258	0.515	0.475	0.464	0.197	0.707	0.610	

=== Confusion Matrix ===

a	b	c	d		<-- classified as
0	5	1	4	|	a = Fail
1	30	28	10	|	b = Pass
0	6	21	2	|	c = Distinction
0	3	3	6	|	d = Withdrawn

Cross Validation:(10 Folds)

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances 3716 50.8414 %
Incorrectly Classified Instances 3593 49.1586 %
Kappa statistic 0.2311
Mean absolute error 0.2781
Root mean squared error 0.4021
Relative absolute error 88.1582 %
Root relative squared error 101.2418 %
Coverage of cases (0.95 level) 92.4066 %
Mean rel. region size (0.95 level) 68.7919 %
Total Number of Instances 7309

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
0.220	0.056	0.331	0.220	0.264	0.197	0.740	0.262	Fail
0.596	0.410	0.635	0.596	0.615	0.184	0.638	0.656	Pass
0.560	0.210	0.386	0.560	0.457	0.308	0.769	0.447	Distinction

	0.344	0.101	0.381	0.344	0.362	0.254	0.724	0.339	Withdrawn
Weighted Avg.	0.508	0.285	0.515	0.508	0.507	0.220	0.687	0.524	

=== Confusion Matrix ===

a	b	c	d	<-- classified as
180	412	64	163	a = Fail
238	2372	968	405	b = Pass
22	533	780	57	c = Distinction
103	420	208	384	d = Withdrawn

Percentage split:(61%):

=== Summary ===

Correctly Classified Instances	1553	54.4721 %
Incorrectly Classified Instances	1298	45.5279 %
Kappa statistic	0.2485	
Mean absolute error	0.2685	
Root mean squared error	0.3886	
Relative absolute error	85.2104 %	
Root relative squared error	98.243 %	
Coverage of cases (0.95 level)	94.0372 %	
Mean rel. region size (0.95 level)	69.0196 %	
Total Number of Instances	2851	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.240	0.053	0.348	0.240	0.284	0.221	0.766	0.267	Fail
	0.685	0.479	0.637	0.685	0.660	0.209	0.649	0.678	Pass
	0.533	0.169	0.427	0.533	0.474	0.336	0.792	0.481	Distinction
	0.263	0.067	0.417	0.263	0.323	0.239	0.741	0.359	Withdrawn
Weighted Avg.	0.545	0.311	0.533	0.545	0.533	0.239	0.703	0.548	

=== Confusion Matrix ===

a	b	c	d	<-- classified as
72	173	15	40	a = Fail
87	1076	301	106	b = Pass
6	233	290	15	c = Distinction
42	207	73	115	d = Withdrawn

J48

Supplied Test Set:

=== Summary ===

Correctly Classified Instances	61	50.8333 %
Incorrectly Classified Instances	59	49.1667 %
Kappa statistic	0.1823	
Mean absolute error	0.2695	
Root mean squared error	0.4297	
Relative absolute error	87.4693 %	
Root relative squared error	110.8295 %	
Coverage of cases (0.95 level)	80.8333 %	
Mean rel. region size (0.95 level)	60.4167 %	
Total Number of Instances	120	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.000	0.073	0.000	0.000	0.000	-0.081	0.468	0.088	Fail
	0.623	0.490	0.632	0.623	0.628	0.133	0.610	0.631	Pass
	0.414	0.154	0.462	0.414	0.436	0.270	0.736	0.443	Distinction
	0.500	0.111	0.333	0.500	0.400	0.327	0.671	0.283	Withdrawn
Weighted Avg.	0.508	0.336	0.508	0.508	0.506	0.168	0.634	0.505	

=== Confusion Matrix ===

```
 a  b  c  d   <-- classified as
 0  6  1  3 |  a = Fail
 7 43 12  7 |  b = Pass
 1 14 12  2 |  c = Distinction
 0  5  1  6 |  d = Withdrawn
```

Cross Validation 10 Folds:

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances	4431	60.6239 %
Incorrectly Classified Instances	2878	39.3761 %
Kappa statistic	0.329	
Mean absolute error	0.2345	
Root mean squared error	0.394	
Relative absolute error	74.3215 %	
Root relative squared error	99.2047 %	

Coverage of cases (0.95 level) 84.2933 %
Mean rel. region size (0.95 level) 59.1736 %
Total Number of Instances 7309

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.276	0.056	0.385	0.276	0.321	0.256	0.699	0.271	Fail
	0.793	0.496	0.657	0.793	0.718	0.311	0.679	0.669	Pass
	0.480	0.079	0.587	0.480	0.528	0.434	0.791	0.516	Distinction
	0.341	0.064	0.488	0.341	0.401	0.322	0.685	0.354	Withdrawn
Weighted Avg.	0.606	0.301	0.587	0.606	0.589	0.330	0.704	0.547	

=== Confusion Matrix ===

```
   a    b    c    d   <-- classified as
 226  470   33   90 |    a = Fail
 230 3157  349  247 |    b = Pass
  22  641  668   61 |    c = Distinction
 109  538   88  380 |    d = Withdrawn
```

Percentage split:(80 %)

=== Evaluation on test split ===

Time taken to test model on training split: 0 seconds

=== Summary ===

Correctly Classified Instances 911 62.3119 %
Incorrectly Classified Instances 551 37.6881 %
Kappa statistic 0.3435
Mean absolute error 0.2358
Root mean squared error 0.3881
Relative absolute error 74.6926 %
Root relative squared error 97.6345 %
Coverage of cases (0.95 level) 87.6197 %
Mean rel. region size (0.95 level) 63.1156 %
Total Number of Instances 1462

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.292	0.046	0.429	0.292	0.347	0.293	0.725	0.268	Fail
	0.834	0.527	0.653	0.834	0.732	0.332	0.683	0.672	Pass
	0.464	0.059	0.646	0.464	0.540	0.463	0.803	0.524	Distinction
	0.319	0.056	0.524	0.319	0.397	0.325	0.693	0.368	Withdrawn
Weighted Avg.	0.623	0.311	0.607	0.623	0.601	0.351	0.712	0.552	

=== Confusion Matrix ===

```
   a   b   c   d   <-- classified as
  45  87   5  17 |   a = Fail
  35 662  53  44 |   b = Pass
   2 138 128   8 |   c = Distinction
  23 127  12  76 |   d = Withdrawn
```

Decision Table:

Supplied Test Set:

=== Summary ===

```
Correctly Classified Instances          65               54.1667 %
Incorrectly Classified Instances        55               45.8333 %
Kappa statistic                          0.1375
Mean absolute error                      0.3125
Root mean squared error                  0.3937
Relative absolute error                101.4328 %
Root relative squared error            101.5602 %
Coverage of cases (0.95 level)         100      %
Mean rel. region size (0.95 level)      93.9583 %
Total Number of Instances              120
```

=== Detailed Accuracy By Class ===

```
                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.100    0.109    0.077      0.100    0.087      -0.008   0.592     0.123     Fail
                 0.768    0.725    0.589      0.768    0.667      0.049    0.497     0.595     Pass
                 0.310    0.033    0.750      0.310    0.439      0.396    0.707     0.515     Distinction
                 0.167    0.028    0.400      0.167    0.235      0.209    0.676     0.289     Withdrawn
Weighted Avg.    0.542    0.437    0.566      0.542    0.520      0.144    0.574     0.506
```

=== Confusion Matrix ===

```
  a  b  c  d   <-- classified as
  1  8  0  1 |  a = Fail
 11 53  3  2 |  b = Pass
  0 20  9  0 |  c = Distinction
  1  9  0  2 |  d = Withdrawn
```

Cross-Validation:(10 fold)

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	4222	57.7644 %
Incorrectly Classified Instances	3087	42.2356 %
Kappa statistic	0.2638	
Mean absolute error	0.2953	
Root mean squared error	0.3776	
Relative absolute error	93.6006 %	
Root relative squared error	95.0716 %	
Coverage of cases (0.95 level)	99.1244 %	
Mean rel. region size (0.95 level)	92.9573 %	
Total Number of Instances	7309	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.258	0.112	0.224	0.258	0.240	0.137	0.666	0.257	Fail
	0.809	0.571	0.629	0.809	0.708	0.258	0.652	0.658	Pass
	0.376	0.046	0.660	0.376	0.479	0.417	0.786	0.533	Distinction
	0.239	0.030	0.586	0.239	0.339	0.310	0.704	0.365	Withdrawn
Weighted Avg.	0.578	0.337	0.583	0.578	0.556	0.283	0.687	0.544	

=== Confusion Matrix ===

```
   a    b   c   d   <-- classified as
 211  535   16   57 |    a = Fail
 448 3221  202  112 |    b = Pass
 126  723  524   19 |    c = Distinction
 156  641   52  266 |    d = Withdrawn
```

Percentage Split:(66%)

=== Summary ===

Correctly Classified Instances	1366	54.9698 %
Incorrectly Classified Instances	1119	45.0302 %
Kappa statistic	0.2845	
Mean absolute error	0.3166	
Root mean squared error	0.3848	
Relative absolute error	100.4075 %	
Root relative squared error	96.9909 %	
Coverage of cases (0.95 level)	100 %	
Mean rel. region size (0.95 level)	100 %	
Total Number of Instances	2485	

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0.451	0.273	0.164	0.451	0.241	0.121	0.542	0.265	Fail
	0.703	0.430	0.663	0.703	0.682	0.275	0.658	0.649	Pass
	0.432	0.009	0.916	0.432	0.587	0.580	0.776	0.587	Distinction
	0.223	0.004	0.905	0.223	0.358	0.413	0.609	0.369	Withdrawn
Weighted Avg.	0.550	0.266	0.696	0.550	0.567	0.339	0.661	0.553	

=== Confusion Matrix ===

```
  a   b   c   d   <-- classified as
 119 144   1   0 |   a = Fail
 385 954  13   5 |   b = Pass
 124 144 207   4 |   c = Distinction
  97 197   5  86 |   d = Withdrawn
```

B. With Demographic data:

(Attributes like Gender, Age, disability, region and highest education have been removed)

Naïve Bayes:

1. **Supplied Test Set**

   Correctly Classified Instances 58 48.3333 %

   Incorrectly Classified Instances 62 51.6667 %

2. **Cross validation for ten folds:**

    Correctly Classified Instances 3597 49.2133 %

    Incorrectly Classified Instances 3712 50.7867 %

3. **Percentage Split:**

    Correctly Classified Instances 1308 52.6358 %
    Incorrectly Classified Instances 1177 47.3642 %

J48:

1. **Supplied Test Set**

    Correctly Classified Instances 61 50.8333 %

    Incorrectly Classified Instances 59 49.1667 %

2. **Cross validation for ten folds:**

    Correctly Classified Instances 4106 56.1773 %

    Incorrectly Classified Instances 3203 43.8227 %

3. **Percentage Split:**

    Correctly Classified Instances 1337 53.8028 %

    Incorrectly Classified Instances 1148 46.1972 %

Decision Tree:

1. **Supplied Test Set**

    Correctly Classified Instances 1420 57.1429 %

    Incorrectly Classified Instances 1065 42.8571 %

2. **Cross validation for ten folds:**

    Correctly Classified Instances 4170 57.0529 %

    Incorrectly Classified Instances 3139 42.9471 %

3. **Percentage Split:**

    Correctly Classified Instances 1420 57.1429 %

    Incorrectly Classified Instances 1065 42.8571 %

From the Results, by applying the three classifiers we observe that the decision tree yields better results when compared to other classifiers at hand.

Conclusion:

With the help of the predictive model built with the help of the student's activities and actions in the VLE, we can accurately predict students at risk and also proper feedback can be provided so as it bring the student back on track. We have consider the sum of clicks that the student has made in so and so resource and also the scores obtained by the student. It is observed that demographic data do not affect the overall accuracy. It is also observed that Decision tree fetches the most accurate results.

References:

[1] https://en.wikipedia.org/wiki/Weka_(machine_learning)

[2] http://jupyter.org/

[3] https://www.continuum.io/downloads

[4] http://www.laceproject.eu/publications/analysing-at-risk-students-at-open-university.pdf

[5] http://www.laceproject.eu/learning-analytics-review/analysing-at-risk-students-at-open-university/

[6] http://oro.open.ac.uk/42529/

[7] https://analyse.kmi.open.ac.uk/

[8] http://analytics.ncsu.edu/sesug/2016/EPO-271_Final_PDF.pdf

[9] https://www.slideshare.net/JakubKuzilek/lak15-ou-analysemaster

[10] http://whatis.techtarget.com/definition/decision-tree

[11] http://data-mining.business-intelligence.uoc.edu/home/j48-decision-tree

[12] Kuzilek, J., Hlosta, M., Herrmannova, D., Zdrahal, Z. and Wolff, A. OU Analyse: Analysing At-Risk Students at The Open University. Learning Analytics Review, no. LAK15-1, March 2015, ISSN: 2057-7494.