itb Institute of Technology
Blanchardstown
Institiúid Teicneolaíochta
Baile Bhlainséir

Honours Degree in Computing

Data Analytics Assessment:
Analyse a dataset

Submitted by: Derek McCarthy, B00007439

Submission date: 18/12/2018

Declaration

I/We declare that this material, which I/We now submit for assessment, is entirely my/our own work and has not been taken from the work of others, except where otherwise stated. I/We have identified and included the source of all facts, ideas, opinions, and viewpoints of others in the assignment references. Direct quotations from books, journal articles, internet sources, module text, or any other source whatsoever are acknowledged, and the source cited are identified in the assignment references.

I/We understand that plagiarism, collusion, and copying are grave and serious offences and accept the penalties that would be imposed should I/we engage in plagiarism, collusion or copying. I acknowledge that copying someone else's assignment, or part of it, is wrong, and that submitting identical work to others constitutes a form of plagiarism. I/We have read and understood the colleges plagiarism policy 3AS08 (available here).

This material, or any part of it, has not been previously submitted for assessment for an academic purpose at this or any other academic institution.

I have not allowed anyone to copy my work with the intention of passing it off as their own work.

Name: Derek McCarthy Dated: 06/12/2018

Contents

Business Understanding

The student dataset provides information on students who have passed or failed their course. The dataset contains information on Intrinsic/Extrinsic Motivation, Self-Efficiency, Self-Regulation, the effort put into studying, number of hours spent studying, Openness, Conscientiousness, high school average grade. It also contains several different attributes that help to identify whether a student is likely to pass or fail their course.

Business objective

- Identify the likelihood of a student Passing or Failing their course.

Data Mining objective

- Create a model that can identify if a student is going to pass or fail a course.
- Identify the attribute that play a role in the likelihood of a student passing or failing a course.

Data Understanding

Describe the data

In this section of the report we will look at the three of the main characteristics of Data Understanding,

- Describing the Data
- Do an initial survey of the data Quality
- Report on findings

As we were provided the dataset we will not need to carry out the Collection of data phase of Data Understanding. The student dataset consists of a variety of different attributes both numeric and categorical which provide information on the student. To get a better understanding of the data we will separate the data into Numeric and Categorical data.

Numeric Data

We will first look at the numeric data see table 1. Each numerical attribute has a,

- Description which describes what the attribute is used for
- Data type of numeric
- Minimum and Maximum values
- Mean (Average) and Standard Deviation (St Dev)

Table 1 - Numeric Data

Numeric Attributes						
Attribute	**Description**	**Data Type**	**Min**	**Max**	**Mean**	**St Dev**
Academic Year	Year student enrolled in first year of college	Numeric	2015	2017	2016.049	0.792
Extrinsic Motivation	Motivated to get a good grade	Numeric	0.200	45	6.699	2.000
Group Work	Prefers working in groups to working alone	Numeric	0	10	6.479	2.841
Intrinsic Motivation	Motivation by a desire to learn new skills	Numeric	0	10	5.540	1.626
Self-Efficacy	Believe in ability	Numeric	1.100	45	5.789	1.932
Self-Regulation	Transform mental abilities	Numeric	0.200	8.500	4.684	1.421

	into academic skills					
Study Effort	Effort put into study	Numeric	0.500	10	5.391	1.624
Study Time	Time spent studying	Numeric	0	10	6.101	1.998
Openness	Likes to learn new things	Numeric	0	9.800	4.593	1.607
Conscientiousness	Being organised and vigilant when doing a task	Numeric	0	10	5.399	1.478
High School Average	Average grade in secondary school	Numeric	40	79.800	60.762	8.727
High School English	English grade from secondary school	Numeric	40	80	57.936	8.645
High School Maths	Maths grade from secondary school	Numeric	40	73	51.377	7.255
Age	The age of student	Numeric	18	51	23.450	5.883
ID		Numeric	1	2459	121.096	703.283

Categorical Data

For the categorical data each attribute has a,

- Description which describes what the attribute is for
- Data type of Nominal
- Value of 1 - 9

Categorical Attributes

Attribute	Description	Data Type	Value 1	Value 2	Value 3	Value 4
Sex	Male of Female	Nominal	Male (658)	Female (577)	N/A	N/A
Discipline	Area of Study	Nominal	Humanities (449)	Computing (344)	Business (330)	Engineering (112)
Modality	How you like to process information	Nominal	Visual (881)	Kinaesthetic (146)	Auditory (107)	N/A
Learning Style		Nominal	Deep (669)	Strategic (344)	Shallow (227)	N/A
Label		Nominal	Pass (721)	Fail (519)	N/A	N/A
Course	Course of Study	Nominal	Computing (219)	Applied Social Care (187)	Business (166)	Community Dev (138)
			Value 5	**Value 6**	**Value 7**	**Value 8**
			Creative Digital Media (125)	Early Child Care (124)	Engineering (112)	International Business (94)
			Value 9			
			Business with IT (70)	N/A	N/A	N/A

Explore the data

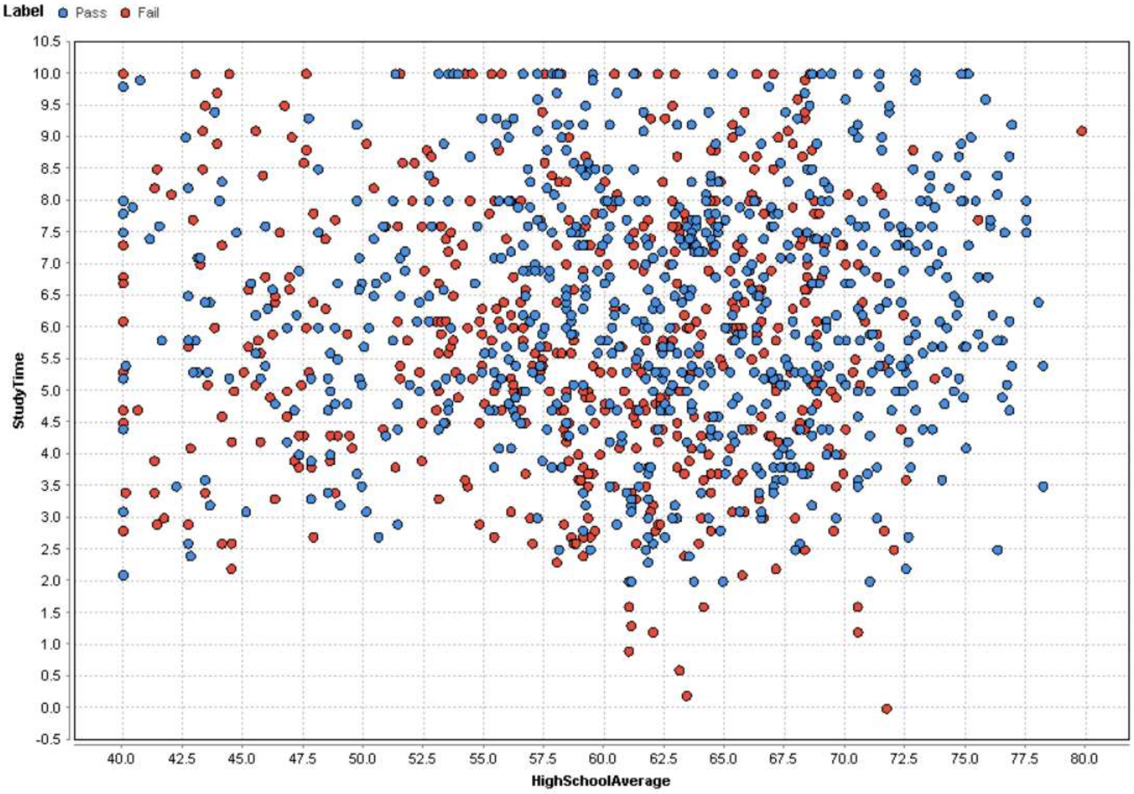


Figure 1 – Study time vs High School Average

In figure 1 we can see a scatter plot of the student's high school average and the time studied. With the student who passed in blue and those who failed in red. As you can see most students with a high school average of 72.5% with a mid-high study time passed with the exception of seven who failed.

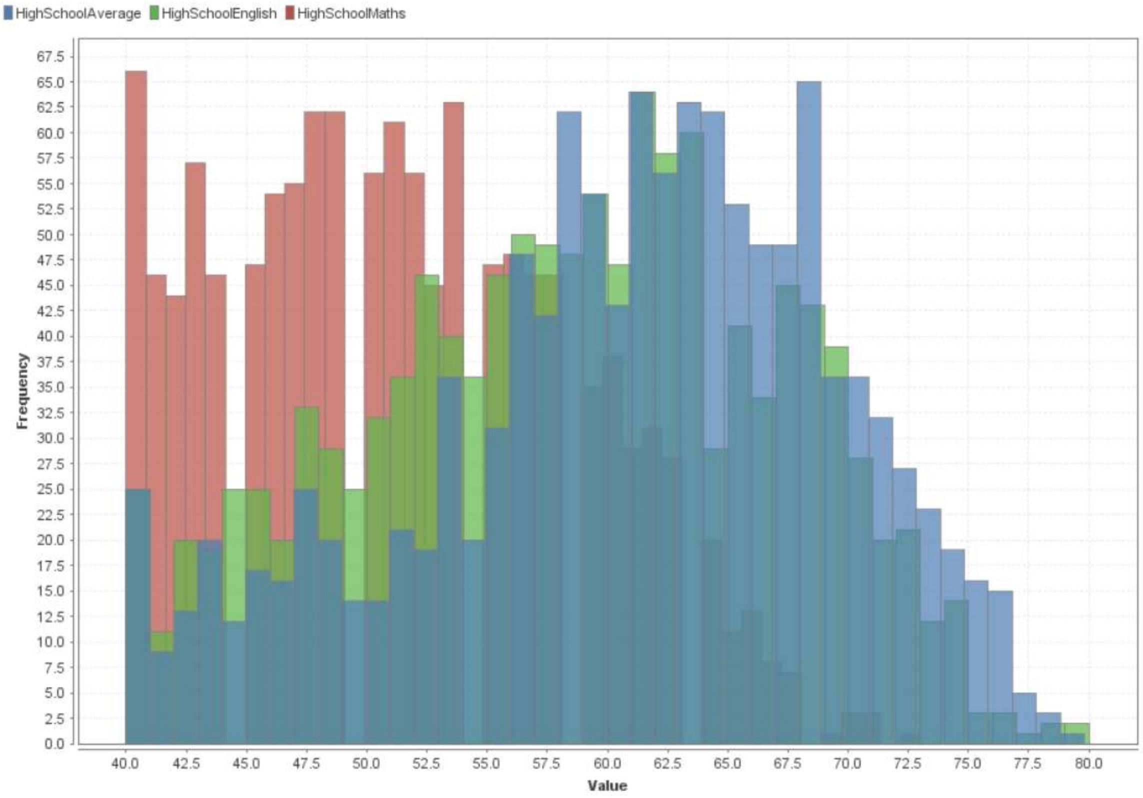


Figure 2 - Grouped Data

In figure 2 we can see a histogram of High School Average, High School English and High School Maths. From this we can establish that these three attributes appear to be a group. All three of their values fall in the range of 40-80. Also, high school maths appears to be skewed while high school English and high school average appear to have a normal distribution.

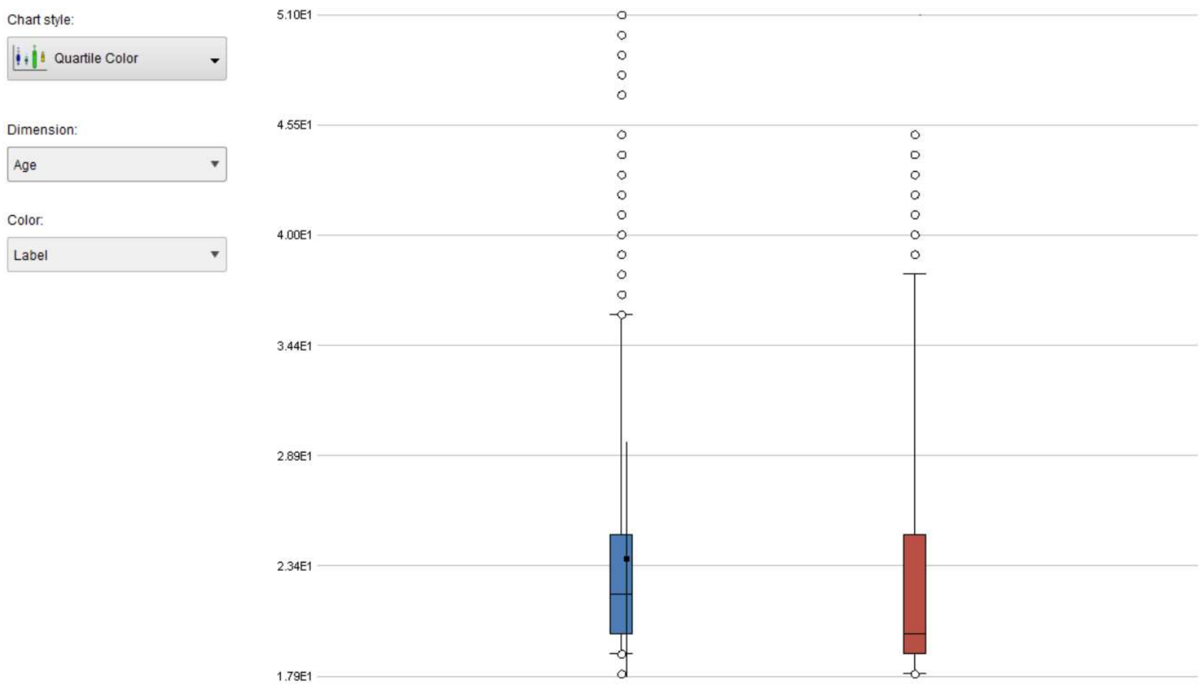


Chart style: Quartile Color

Dimension: Age

Color: Label

Figure 3 – Box Plot of Age Attribute

In figure 3 we can see a box plot of the attribute age. The blue box plot is that of the students who passed while the red is of those who failed. As you can see there is a lot of outliers in both box plots from the 25th percentiles. Also, in the box plot of the students who passed the mean (black circle) is way off the standard deviation line, which indicates that it's not a normal distribution.

Verify data quality

Missing Values

In the dataset there are several attributes that have missing values. The Self-Regulation attribute is the attribute with the most missing values with 936 (75%) in total. Additionally, Modality has 106 (8.5%) missing values while there are 5 attributes that don't have any missing values while the rest have between 1-5 missing values.

Noise, Bias or Outliers

The dataset contains several outliers for example the attribute Self-Efficacy has an outlier value of 45 at row 1189 which is 35 more than then next biggest value in the attribute. Additionally, Extrinsic Motivation also has an outlier with a difference of 35 to the next value on row 1221. With such a difference between values this could cause the data to be skewed. Which could affect the accuracy of prediction of the model.

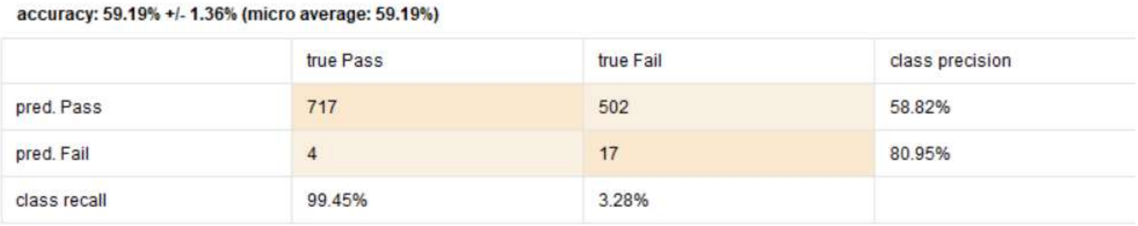
Sufficient Attributes

The dataset contains one attribute that could be identified as sufficient. This attribute being discipline. As the discipline values are a more general form of the attribute course's values this suggest we can remove discipline from the dataset.

Data Preparation

Delete Missing Attribute

As Self-Regulation has more than 40% of its values missing this attribute will need to be removed to improve the accuracy of the model.

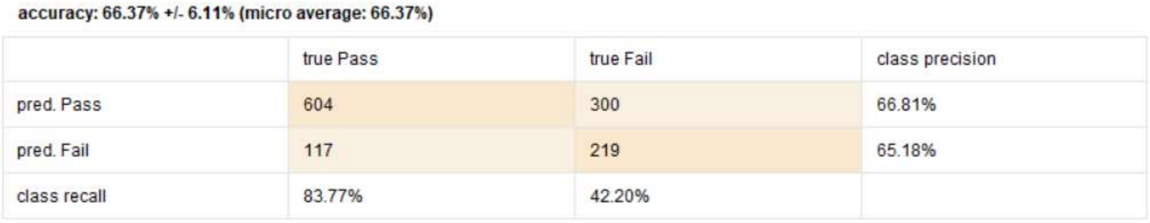


accuracy: 59.19% +/- 1.36% (micro average: 59.19%)

	true Pass	true Fail	class precision
pred. Pass	717	502	58.82%
pred. Fail	4	17	80.95%
class recall	99.45%	3.28%	

Figure 4 - Cross Validation Decision Tree

Before the pre-processing of the dataset, the model was tested using cross validation with decision tree. The accuracy of the model was 59.19% see figure 4. As there are many rows with missing values a filter was applied to ignore these values. The filter used to ignore these values was the no_missing_values filter which brought the accuracy of the model up to 66.37% see figure 5.



accuracy: 66.37% +/- 6.11% (micro average: 66.37%)

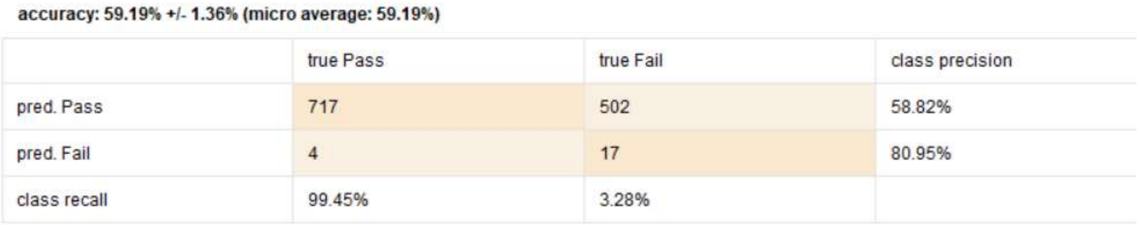
	true Pass	true Fail	class precision
pred. Pass	604	300	66.81%
pred. Fail	117	219	65.18%
class recall	83.77%	42.20%	

Figure 5 - Results of No Missing Values Filter

From just applting a filter to remove missing values not only did the accuracy of the model go up, but also the number of students predicted to pass and fail. Additionaly the true pass recall accuracy went from 99.45% pre-filter to 83.77% post-filter. While the true fail recall accuracy gained significintly from 3.28% pre-filter to 42.20% post-filter.

Select Data

As the attribute Self-Regulation has 936 missing values 75.4% of the total. This indicates that this attribute needs to be deleted as it over the 40% threshold.



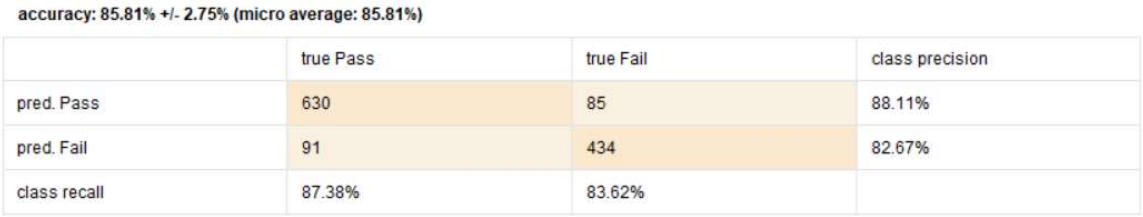
accuracy: 59.19% +/- 1.36% (micro average: 59.19%)

	true Pass	true Fail	class precision
pred. Pass	717	502	58.82%
pred. Fail	4	17	80.95%
class recall	99.45%	3.28%	

Figure 6 - Remove Self-Regulation with Decision Tree

As you can see from figure 7 removing Self-Regulation from the model didn't change the accuracy of the model. The model in figure 7 is using stratified sampling with decision tree.

To try to improve the accuracy we will use k-NN. Which works by comparing the test example with several rows which are similar or close to it.



accuracy: 85.81% +/- 2.75% (micro average: 85.81%)

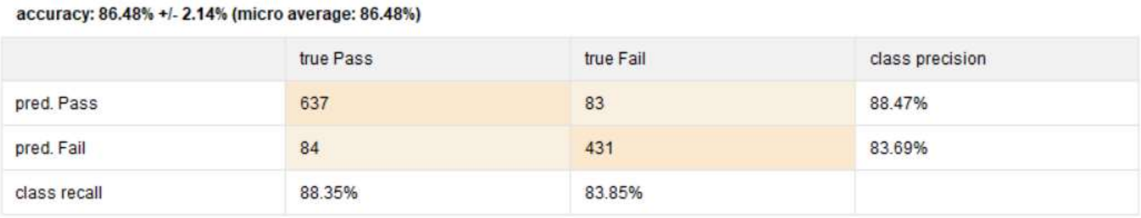
	true Pass	true Fail	class precision
pred. Pass	630	85	88.11%
pred. Fail	91	434	82.67%
class recall	87.38%	83.62%	

Figure 7 - Remove Self-Regulation with k-NN

In figure 8 we can see the results of using cross validation with k-NN. In this instance K (the number of clusters) is set to 1. Using this method, the accuracy of the model has risen significantly from 59.19% to 85.81% using k-NN. Although setting the value of K to 1 or 2 doesn't affect the accuracy in this instance if you increase K to 3 or more the accuracy drops significantly. Also using k-NN the class precision and class recall accuracy rises significantly.

Clean Data

As there are still several attributes with missing values we will have to clean this data. The two techniques we will be using are replace missing values and filter examples. As Modality has just over 8.5% of its values missing we will need to replace the missing values. As for the missing values below 5% we will also filter out these rows using the filter examples operator. As these rows make up less than 5% of the dataset deleting them will not significantly affect the final result.



accuracy: 86.48% +/- 2.14% (micro average: 86.48%)

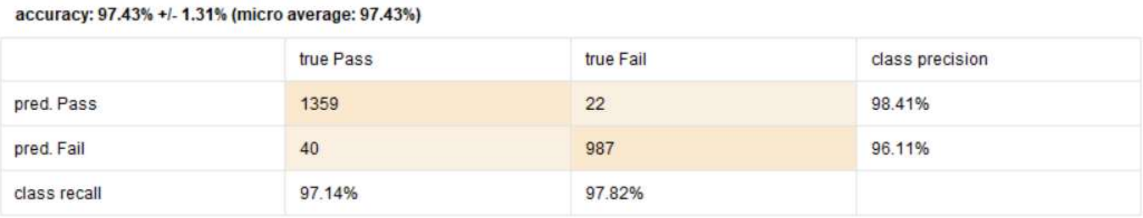
	true Pass	true Fail	class precision
pred. Pass	637	83	88.47%
pred. Fail	84	431	83.69%
class recall	88.35%	83.85%	

Figure 8 - All Missing Values Removed/Replaced

In figure 9 we can see the results of having replaced the values in modality and removing the rows with missing data. As you can see there has being a slight increase of accuracy from 85.81% to 86.48% now. Also, class recall, and class precision has also seeing an increase in accuracy.

Construct Data

To determine the optimal sample size required to model the student dataset several sampling techniques where used. Initially progressive sampling was used to increase the dataset's size. We increased the sampling size until the models accuracy started to level off which was at 2353 rows. At this sample size we used bootstrapping sampling and the accuracy of the model increased from 86.48% to 97.43% with class precision for pass at 98.41% and fail 96.11%. As for recall this also increased significantly to 97.14% for pass and 97.82% for fail (see figure 10). Although this accuracy may appear higher while using bootstrapping, it should be noted that when using bootstrap, it over estimates accuracy.



accuracy: 97.43% +/- 1.31% (micro average: 97.43%)

	true Pass	true Fail	class precision
pred. Pass	1359	22	98.41%
pred. Fail	40	987	96.11%
class recall	97.14%	97.82%	

Figure 9 - Progressive Sampling using Bootstrapping

The second sampling approach taking was to use stratified sampling. With this technique groups are specified in advance and the rows are selected randomly from each group and are typically grouped by a class variable. With this technique there was no increase in performance. The third sampling technique used was Kennard-Stone with a nominal to numerical blending attribute. Kennard-Stone is a sequential technique which selects two rows that are the furthest from one another and the subsequent rows are added by adding new rows that are furthers from objects currently being sampled. Again, using this technique there was no increase in performance. Accuracy was still 97.43% and both recall, and precision were also the same.

To address the attributes that had a large variation in the values range we used normalize. This method of scaling converts all numeric data to fall within a particular range. Doing this makes some algorithms work better such as k-NN or clustering. In the student dataset there were three attributes with a large variation between ranges they were, Age, Extrinsic Motivation and Self Efficacy. By normalizing these three attributes the accuracy of the model increased slightly to 97.72%. Precision increased to 97.43% for pass and 98.12% for fail, and for recall this increased to, 98.63% for pass and 96.49% for fail.

Modelling

For this mining objective an algorithm that can accept both categorical and numeric data is required. Also, as the dataset contains nominal data an algorithm that classifies data rather than predicts data is required. The algorithms chosen for the mining object are Naïve Bayes, Artificial Neural Networks and k-NN. Naïve Bayes was chosen as it classifies data and can assign a row of data to more than one class and also isolated noise does not affect its performance like other algorithms. Artificial Neural Networks was chosen as it can predict complex datasets. But unlike Naïve Bayes it can be sensitive to isolated noise. And, it can be used for the prediction or classification of data. k-NN was chosen as new data is easy to implement and interpret. It gives good accuracy and it doesn't make assumptions about the data. And also it can be used for classification or prediction.

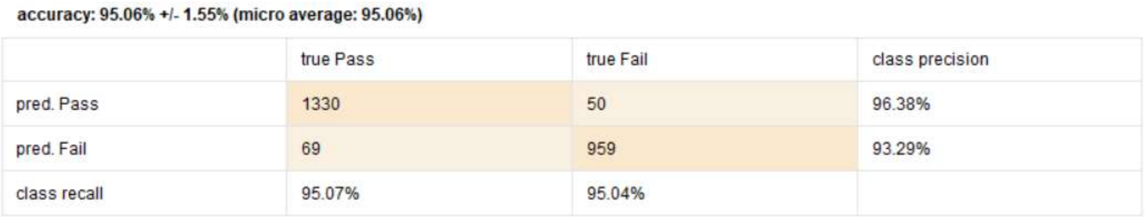
Generate Test Design

To generate a test design cross validation will be used for the creation of the test and training data. With the number of folds (K) set to 10 as this gives the best accuracy. The sample type the model will use is shuffled sampling. By using this method, it allows the entire dataset to be used for both testing and training.

Build and Assess the model

Artificial Neural Networks

The artificial neural network was setup using select attributes, replace missing, filter examples, normalize and bootstrap sampling outlined in the pre-processing section. The cross-validation operator used shuffled sampling with 10 folds. The deep learning operator was chosen as it gave good accuracy. The algorithm used Tanh activation with 17 epochs. Again 17 was chosen for epochs as it was the value which gave best accuracy.



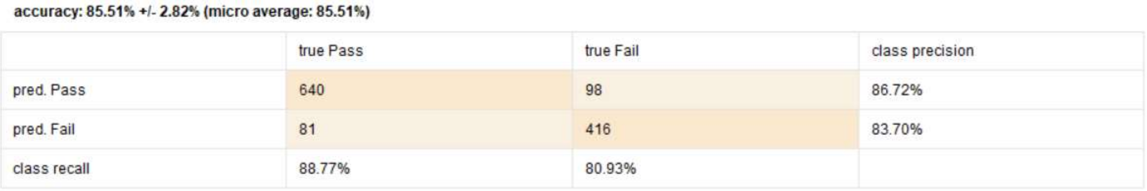
accuracy: 95.06% +/- 1.55% (micro average: 95.06%)

	true Pass	true Fail	class precision
pred. Pass	1330	50	96.38%
pred. Fail	69	959	93.29%
class recall	95.07%	95.04%	

Figure 10 - ANN Model

As you can see from figure 10 the overall accuracy of this model was 95.06%. With class recall for both true pass and true fail being just over 95%. Class precision was 96.38% for the students predicted to pass and 93.29% for the students predicted to fail.

The model was also tested without using bootstrap sampling. With this the accuracy of the model dropped to 85.51% see figure 11. Also, class precision and class recall also dropped significantly.



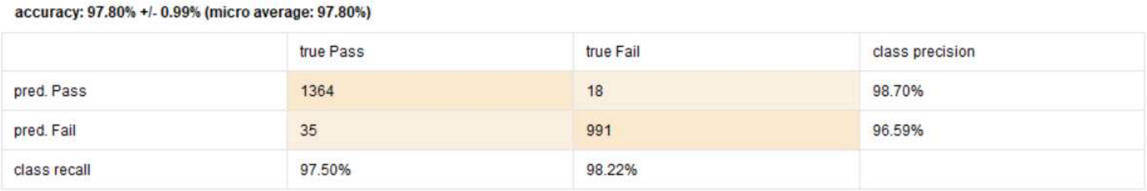
accuracy: 85.51% +/- 2.82% (micro average: 85.51%)

	true Pass	true Fail	class precision
pred. Pass	640	98	86.72%
pred. Fail	81	416	83.70%
class recall	88.77%	80.93%	

Figure 11 - ANN Model Without Bootstrap Sampling

k-Nearest Neighbor (k-NN)

The k-Nearest Neighbour model was setup using select attributes, replace missing, filter examples, normalize and bootstrap sampling outlined in the pre-processing section. The cross-validation operator used stratified sampling with 15 folds. In this model 15 folds produced the best accuracy. The algorithm's k value was set to 1 which gave the best accuracy. The mixed measure and measure types parameters were left at their default values.



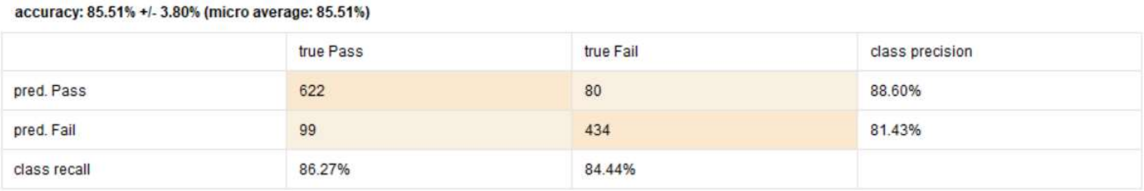
accuracy: 97.80% +/- 0.99% (micro average: 97.80%)

	true Pass	true Fail	class precision
pred. Pass	1364	18	98.70%
pred. Fail	35	991	96.59%
class recall	97.50%	98.22%	

Figure 12 - k-NN Model

As you can see from figure 12 the overall accuracy of this model was 97.80%. With class recall for both true pass and true fail being 97.50% and 98.22% respectively. Class precision was 98.70% for the students predicted to pass and 96.59% for the students predicted to fail.

The model was also tested without using bootstrap sampling. With this the accuracy of the model dropped to 85.51% see figure 13. Also, class precision and class recall also dropped significantly.



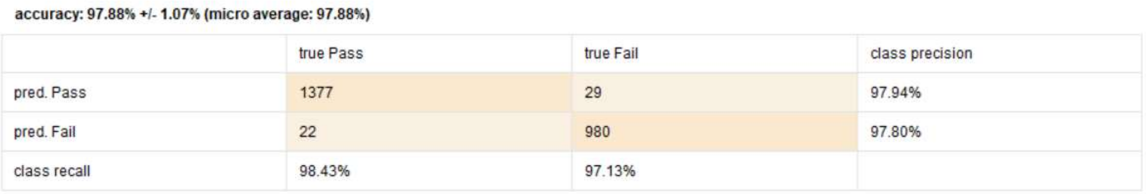
accuracy: 85.51% +/- 3.80% (micro average: 85.51%)

	true Pass	true Fail	class precision
pred. Pass	622	80	88.60%
pred. Fail	99	434	81.43%
class recall	86.27%	84.44%	

Figure 13 - k-NN Model Without Bootstrap Sampling

Naïve Bayes

The Naïve Bayes model was setup using select attributes, replace missing, filter examples, normalize and bootstrap sampling outlined in the pre-processing section. The cross-validation operator used shuffled sampling with 15 folds. In this model 15 folds produced the best accuracy just like with k-NN. The algorithm's estimation was set to full, with bandwidth selection set to fixed. Bandwidth was set to 0.4 as again this gave the best accuracy.



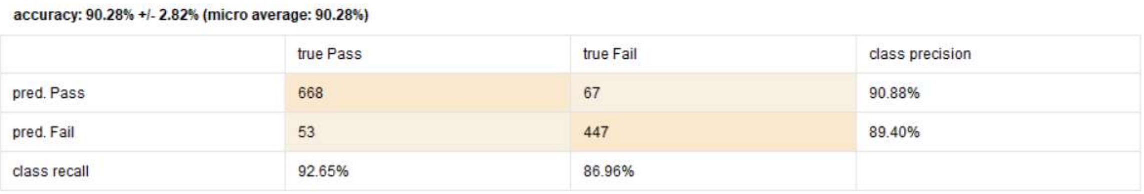
accuracy: 97.88% +/- 1.07% (micro average: 97.88%)

	true Pass	true Fail	class precision
pred. Pass	1377	29	97.94%
pred. Fail	22	980	97.80%
class recall	98.43%	97.13%	

Figure 14 - Naïve Model

As you can see from figure 14 the overall accuracy of this model was 97.88%. With class recall for both true pass and true fail being 98.43% and 97.13% respectively. Class precision was 97.94% for the students predicted to pass and 97.80% for the students predicted to fail.

Again, this model was also tested without using bootstrap sampling. With this the accuracy of the model dropped to 90.28% see figure 15. Also, class precision and class recall also dropped significantly.



accuracy: 90.28% +/- 2.82% (micro average: 90.28%)

	true Pass	true Fail	class precision
pred. Pass	668	67	90.88%
pred. Fail	53	447	89.40%
class recall	92.65%	86.96%	

Figure 15 - Naïve Model Without Bootstrap Sampling

Without bootstrap sampling there wasn't as big of a drop off in the overall accuracy, recall or precision as there was with the other two models.

Evaluation

To conclude the data mining object set out in business understanding has being reached. The model can accurately predict the students who are going to pass or fail their course. Each model implemented had over 95% accuracy in their results. The model which was the most accurate was Naïve Bayes while the Artificial Neural Network model was least accurate.

The model which was least preferred was Artificial Neural Network. As its computation time performance was very slow in comparison to the other two models. The model that was most preferred was the Naïve Bayes algorithm. Not only did this model produce the best results but also produced easy to read visual graphs of each attribute see figure 16.

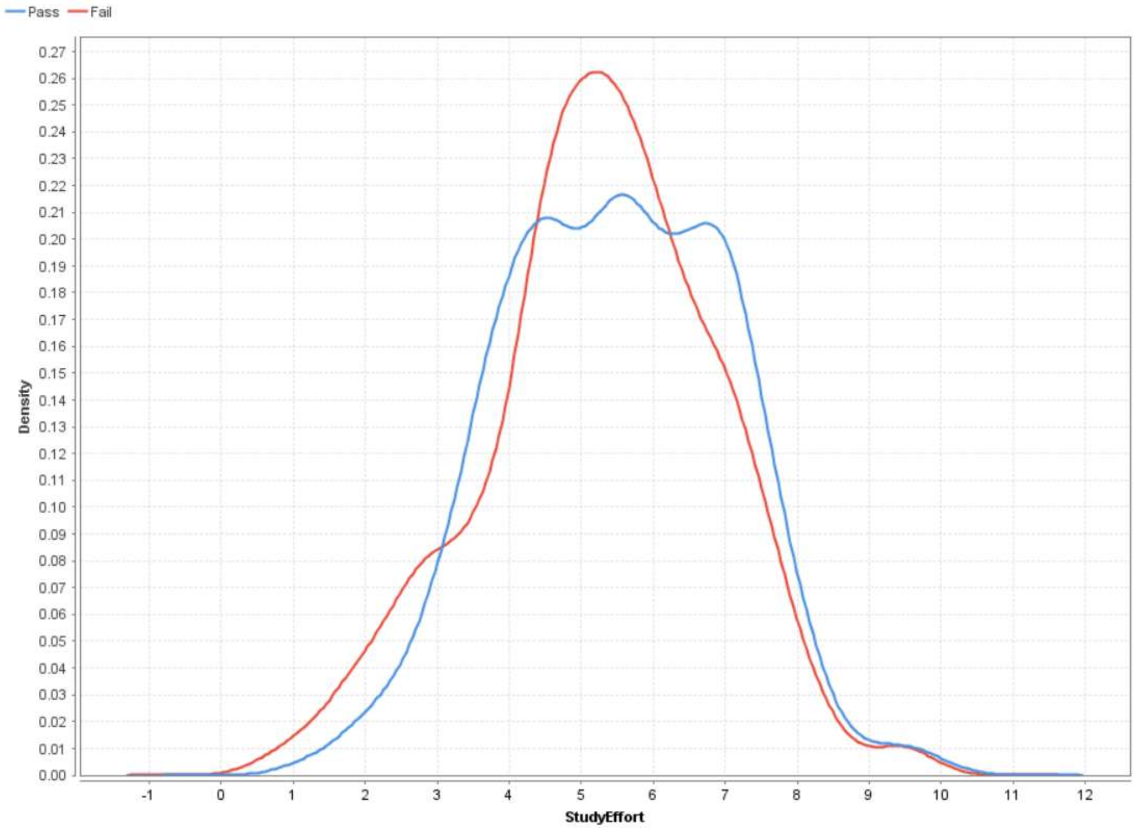


Figure 16 - Stud Effort Graph from Naïve Bayes Model