WELLINGTON RENTAL PRICES AND AFFORDABILITY FOR STUDENTS AT VICTORIA UNIVERSITY

Business Understanding

The business understanding I had of the problem had a large impact on the selection of data. I selected the datasets based on a criteria of wanting to model the rental markets of wellington, Otago and Auckland in relation to student affordability. The model needed to include a dataset which had the rent of the areas based on property types and number of bedrooms, giving an idea of both the availability of those properties based on student numbers and on the affordability of the rooms based on the average student income. For the income of the students I used a dataset which had figures back to 2010 with average student allowances and accommodation benefits, along with a dataset which contained the number of students at the universities per year. This was also compared with a general model of the university hostels, which was a dataset I created based on statistics found online based on the cost and number of rooms relevant to each university, which was also used as a general idea of affordability. I also had a plan to take the CBD areas of the cities or the suburbs closest to the universities in order to get a more specific idea, which was where the detailed rents and bonds came in again.

Data Manipulation and Pre-processing

The datasets were manipulated and data pre-processed in the following ways:
Student income, numbers and average rent and number of bonds were all culled down to the relevant years used in this assignment (2010 to present) through the use of R and WEKA allowing the removal of columns and rows that were irrelevant to the current dataset. I also used some manual manipulation to transpose rows to columns in order to take advantage of removing the irrelevant data (i.e. this making it easy to take out Manawatu as it was not a relevant region, removing all students barring just the bachelor students as they were not what the model was attempting to explain).
This helped with data understanding as it provided a clearer image of the data before attempting preparation. These details were all combined into one csv which would later be matched with another one, as they were all complete details barring the student numbers for 2018 which were imputed using WEKA to fill the missing values (screenshot below of that process, as well as sample code from R used to remove useless columns).
Detailed rent and bond numbers were combined using R merge with the SAU Table in order to get regional names for the data instead of the SAU numbers. From there, the columns which were irrelevant were removed (such as water, TA, and the years not looked at by this model), and then the data for central areas was merged together to get a total number for Auckland Central and Wellington CBD, with the Otago University option taken as was.
The data was then extrapolated into a Year and Month format in Excel, and from there I merged the two datasets together after filling in missing values.

The important aspects of the data used to create the ML tools and having the biggest impact on data preparation were the student income, which was taken as is, along with the affordability class which was derived based on the general price of the rental / hostel in comparison to the student income (which was taken as the average values of each student allowance, accommodation benefit and living cost help given by Studylink).

The number of bonds was manipulated with the students attending each university to give a general availability ratio for each area, a feature that could be used to determine affordability as it would impact on the competition for the students.
The features of year, month, rent and student numbers were taken as is from the folders and not manipulated, but the number of students was imputed for the 2018 year. The only feature selection that occurred in the initial pre processing stage was the removal of the years prior to 2010 and the removal of student numbers relative to that as well, and some feature manipulation came in from the missing values.

The initial datasets for this part came from:
Student allowance: https://www.msd.govt.nz/about-msd-and-our-work/publications-resources/statistics/studylink/student-allowance-ytd-12.html
Student numbers: https://www.educationcounts.govt.nz/statistics/tertiary-education/participation
Hostel statistics are included in the datasets portion of the assignment, showing how they were collated.
Regional bond numbers, rent figures: https://catalogue.data.govt.nz/dataset/rental-bond-data-by-quarter-detailed
Detailed bond numbers and rent figures: https://catalogue.data.govt.nz/dataset/rental-bond-data-by-region

Snapshots of datasets during processing (data taken while in excel format):

Student numbers / allowances / Regional data:

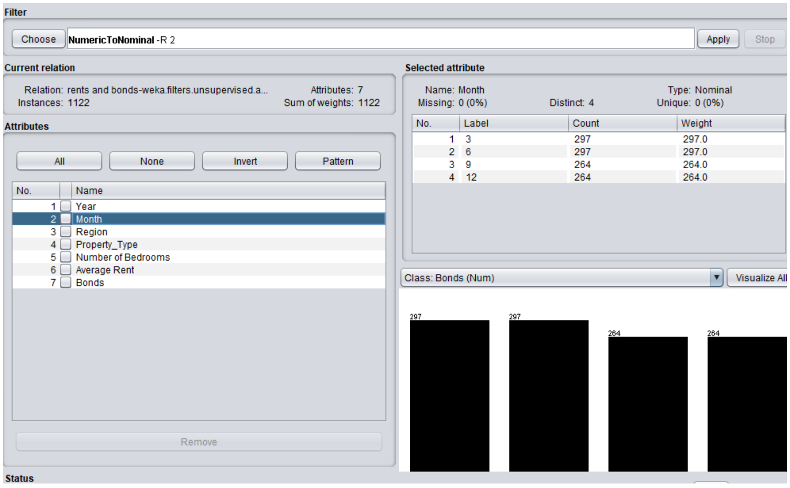
Year	Month	City	Average R	Number o	Number	Student A	Student
2010	1	Auckland	383	3991	26,580	$170.68	$35.03
2010	2	Auckland	389	4367	26,580	$170.68	$35.03
2010	3	Auckland	382	7091	26,580	$170.68	$35.03
2010	4	Auckland	394	5242	26,580	$170.68	$35.03
2010	5	Auckland	396	4690	26,580	$170.68	$35.03
2010	6	Auckland	399	4426	26,580	$170.68	$35.03
2010	7	Auckland	397	5098	26,580	$170.68	$35.03
2010	8	Auckland	395	5194	26,580	$170.68	$35.03
2010	9	Auckland	396	5199	26,580	$170.68	$35.03
2010	10	Auckland	396	4724	26,580	$170.68	$35.03
2010	11	Auckland	400	4811	26,580	$170.68	$35.03
2010	12	Auckland	403	4646	26,580	$170.68	$35.03

SAU merged Rent data:

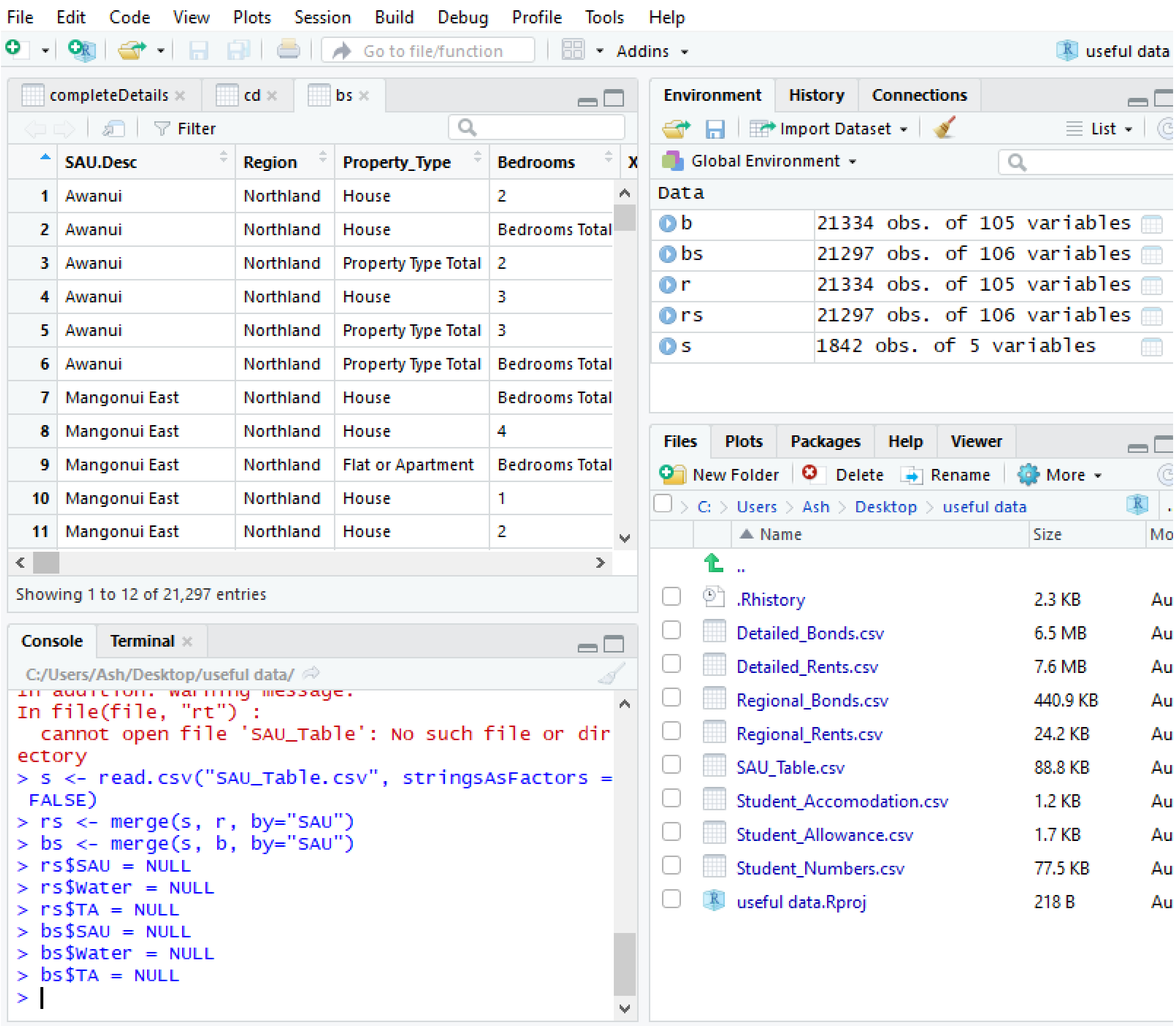
Suburb	Region	Property_	Bedrooms		X1.03.201	X1.06.201	X1.09.201	X1.12.201	X1.03.201	X1.06.201	X1.09.201	X1.12.201	X1.03.201	X1.06.201	X1.09.201	X1.12.201	X1.03.201	X1.06.2
Auckland	Auckland	Flat or Ap	1	304	310	314	313	313	332	344	337	336	346	342	341	346	351	3
Auckland	Auckland	Flat or Ap	2	370	372	381	387	386	416	422	409	424	438	434	418	422	445	4
Auckland	Auckland	Flat or Ap	3	500	459	497	661	498	564	565	642	537	635	665	595	595	605	7
Auckland	Auckland	Flat or Ap	4	514	NA	420	524	NA	504	NA	NA	734	NA	NA	NA	NA	NA	NA
Auckland	Auckland	House	5+	768	NA	723	652	518	682	NA	NA	790	670	NA	NA	NA	600	NA
Auckland	Auckland	House	Bedrooms	336	334	344	350	348	374	382	372	378	393	381	376	383	392	3
Auckland	Auckland	House	1	294	346	296	280	287	354	317	NA	400	375	NA	NA	361	335	3
Auckland	Auckland	House	2	NA	392	393	NA	657	390	438	NA	900	413	459	432	360	388	NA
Auckland	Auckland	House	3	404	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Auckland	Auckland	House	Bedrooms	358	372	382	371	349	436	446	409	393	473	426	394	373	414	4
Auckland	Auckland	Room or E	1	130	136	119	141	130	170	170	136	81	148	152	150	141	161	1
Auckland	Auckland	Room or E	Bedrooms	101	191	162	224	257	242	293	193	216	222	255	262	249	275	2
Parnell	Auckland	Flat or Apartment		1	341	344	339	357	339	364	354	361	354	383	368	374	374	3
Parnell	Auckland	Flat or Apartment		2	368	462	447	457	480	485	473	516	475	499	526	484	489	5
Parnell	Auckland	Flat or Apartment		3	596	627	634	624	685	771	811	788	764	903	705	925	648	7
Parnell	Auckland	Flat or Apartment		4	NA	NA	NA	828	NA	NA	974	NA	NA	NA	NA	NA	NA	NA
Parnell	Auckland	Flat or Apartment		5+	390	404	404	422	407	417	485	513	422	536	423	390	472	5
Parnell	Auckland	Flat or Apartment		Bedrooms	450	466	452	465	490	534	536	535	509	521	506	553	496	5
Parnell	Auckland	House		1	468	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
Parnell	Auckland	House		2	699	615	646	623	723	804	695	714	703	647	676	660	784	7
Parnell	Auckland	House		3	712	675	651	696	692	724	744	818	734	726	743	745	725	9
Parnell	Auckland	House		4	NA	966	959	966	NA	995	NA	NA	NA	NA	NA	839	902	NA
Parnell	Auckland	House		5+	665	742	751	846	1003	1010	719	666	791	731	744	775	734	7

Sheet1 Sheet3 Sheet2

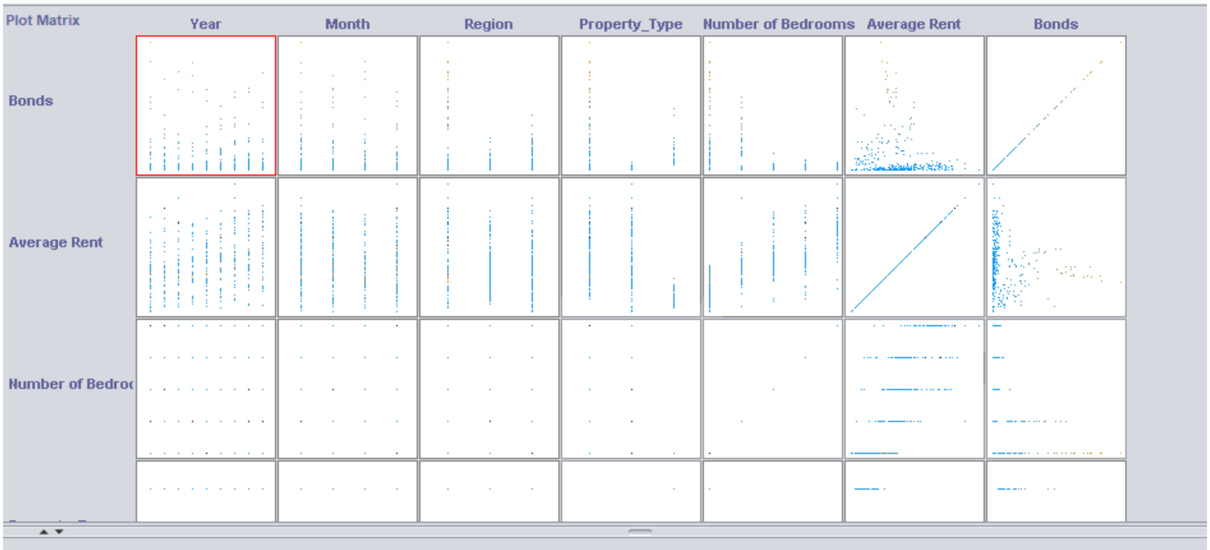
WEKA Processing (conversion of numeric to nominal example):



Example of R code used to remove the columns that weren’t useful:



Visualisation of data set (interim rent and bond data):



R data filling in missing values for rent and bond:

```
downloaded_packages
> rb <- knnImputation(rb, k = 10, scale = T, meth =
 "weighAvg", distData = NULL)
Error in knnImputation(rb, k = 10, scale = T, meth
= "weighAvg", distData = NULL) :
  could not find function "knnImputation"
> rb <- DMwR::knnImputation(rb, k = 10, scale = T,
meth = "weighAvg", distData = NULL)
> View(rb)
> View(rb)
>
```

Final merged data set:

2013	9	Wellingto	370	4770	15,335	$173.29	35.18	Flat or Apa	2	448	138	0.028930818	0	0
2013	9	Wellingto	370	4770	15,335	$173.29	35.18	Flat or Apa	4	476	8	0.001677149	0	0
2013	9	Otago	354	2680	16,060	$173.29	35.18	House	4	490	11	0.004104478	0	0
2013	9	Auckland	449	14837	26,905	$173.29	35.18	House	3	490	14	0.000943587	1	0
2013	9	Wellingto	370	4770	15,335	$173.29	35.18	Flat or Apa	5	497	10	0.002096436	0	0
2013	9	Otago	354	2680	16,060	$173.29	35.18	Flat or Apa	4	505	8	0.002985075	0	0
2013	9	Auckland	449	14837	26,905	$173.29	35.18	House	4	522	13	0.000876188	1	0
2013	9	Auckland	449	14837	26,905	$173.29	35.18	Flat or Apa	4	588	20	0.001347981	1	0
2013	9	Auckland	449	14837	26,905	$173.29	35.18	Flat or Apa	5	615	10	0.000673991	1	0
2013	9	Auckland	449	14837	26,905	$173.29	35.18	House	5	622	8	0.000539193	1	0
2013	9	Otago	354	2680	16,060	$173.29	35.18	House	5	648	17	0.006343284	0	0
2013	9	Otago	354	2680	16,060	$173.29	35.18	Flat or Apa	5	652	16	0.005970149	0	0
2013	9	Auckland	449	14837	26,905	$173.29	35.18	Flat or Apa	3	710	30	0.002021972	1	0
2013	12	Auckland	456	15307	26,905	$173.29	35.18	Room or E	1	149	49	0.00320115	1	1
2013	12	Wellingto	413	6456	15,335	$173.29	35.18	House	3	191	8	0.001239157	1	1
2013	12	Wellingto	413	6456	15,335	$173.29	35.18	Room or E	1	194	49	0.007589839	1	1
2013	12	Wellingto	413	6456	15,335	$173.29	35.18	House	1	232	40	0.006195787	1	1
2013	12	Otago	335	3204	16,060	$173.29	35.18	Flat or Apa	1	240	15	0.004681648	0	0
2013	12	Otago	335	3204	16,060	$173.29	35.18	House	1	250	7	0.002184769	0	1
2013	12	Wellingto	413	6456	15,335	$173.29	35.18	House	2	284	16	0.002478315	1	1
2013	12	Wellingto	413	6456	15,335	$173.29	35.18	Flat or Apa	1	294	169	0.0261772	1	1
2013	12	Otago	335	3204	16,060	$173.29	35.18	House	2	303	7	0.002184769	0	0
2013	12	Otago	335	3204	16,060	$173.29	35.18	Flat or Apa	2	327	5	0.001560549	0	0
2013	12	Otago	335	3204	16,060	$173.29	35.18	Room or E	1	335	109	0.034019975	0	1

finishedData

Data Preparation

In order to take care of the data preparation -> modelling phase of the CRISP-DM model I first modelled the raw, processed data using a J48 Decision Tree in order to see how accurately the classifier would be able to classify the data. On the raw data I achieved an accuracy of 93.955% with only 68 of the 1057 instances contained within my data set misclassified. This was useful as I had a baseline to compare what the removed attributes could do to the data i.e. an accuracy increase or decrease.

From here I used the visualise tab to see how the attributes reacted with each other and to see if there were any with high correlation meaning one could be removed without impacting on the other factors of the data. This allowed a general idea of the impact of a feature such as month, which could have a low impact, as against average rent, which could be a very high impact feature on deciding which class the instance had within the classifier.

Model Creation and Evaluation

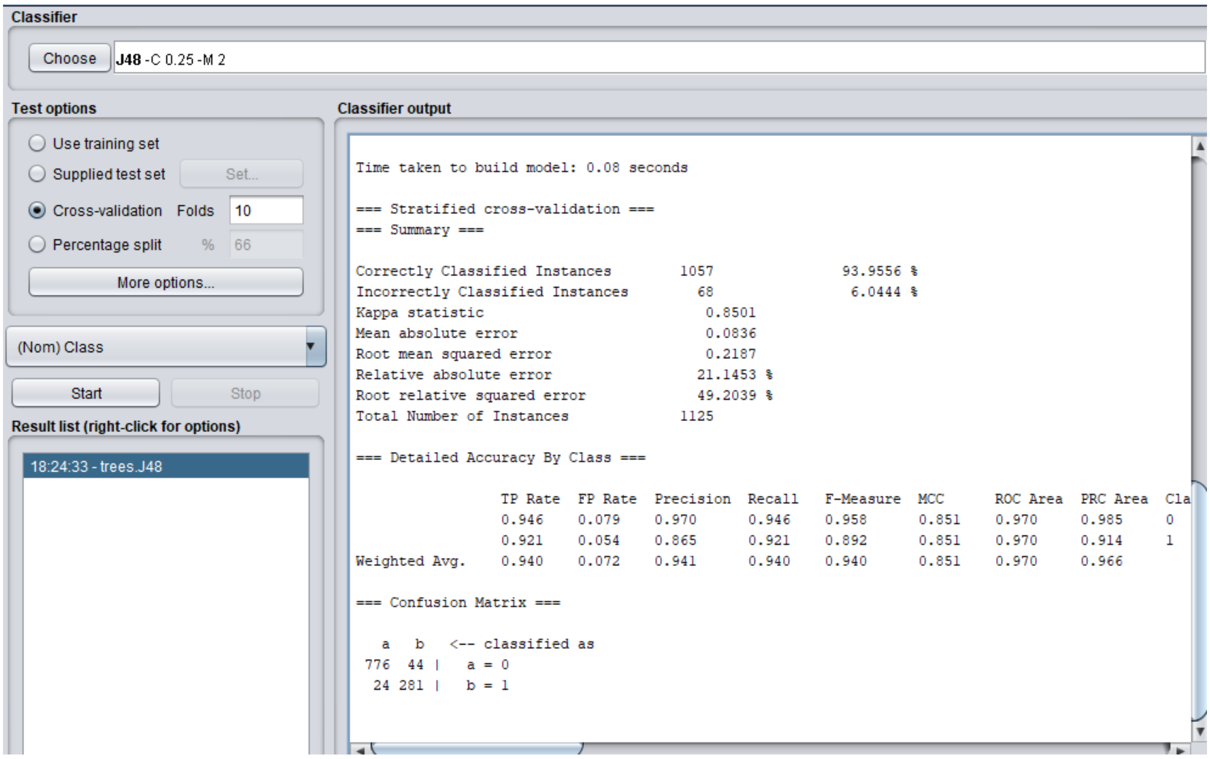
After this I moved on to processing the data with the select attributes option within WEKA. As my main method of modelling and classifying the data for this assignment was through the use of the Decision Trees, the first technique that was modelled on the data was the information gain measure. I used this as it is a measure of entropy and this relates to the purity/impurity measure used when deciding the next node to add to the tree based on the attributes. Through this, the output achieved showed that average rent and number of bedrooms had the highest impact on the data, while the features of above national average and month had the smallest impact. In order to check the impact of different evaluators, I also tested the data set with the Pearson correlation to see if the relation between the attributes and the class gave a better result than the information gain measure. This also resulted in the attribute of month having the least impact on the dataset, as well as number of bonds, but the relation between student/bond ratio and class is more obvious in this method, showing that it is more important than the information gain might suggest. Student Accommodation Benefit Average had the second least impact on the data overall, and would be a prime attribute for removal to test the impact on the dataset, while rent and ratio when removed could result in a wildly different accuracy and testing quality on the data which could be worth testing in the dimension reduction. Year also had a relatively low impact.

The first test to see how the removed attributes impacts on the tool output was by simply removing month from the data, using the WEKA explorer tools before re running the J48 classifier. This was done first and solo as it had the smallest impact on the data overall, and provided a good base point to start with while ensuring that the data understanding was still prevalent for anyone interpreting the data. The factors within the J48 tree were kept the same (10 cross validation, default batch size, pruning etc), to keep the integrity of the changes caused by the manipulation of the data and removal of attributes. The removal of this attribute decreased the overall accuracy of the classifier, though it decreased the number of instances that were incorrectly classified as affordable when they were not. It also decreased the number of correct unaffordable classifications.

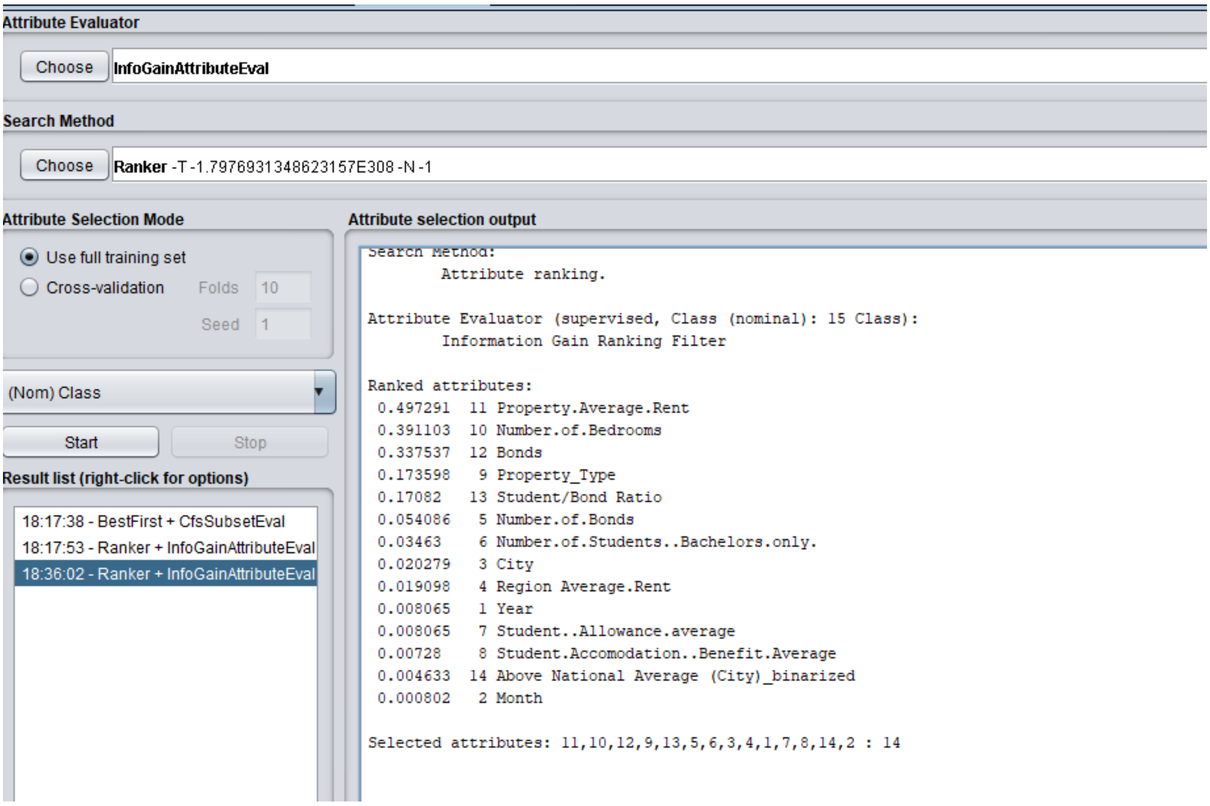
The second test also removed year and month, along with student accommodation benefit average. This reduced the total number of attributes by three by 12, and removed the three attributes with the lowest overall information gain in the j48 decision tree. The factors were kept the same as above, and the classification on the j48 was run again. Again the removal of these attributes decreased the number of instances correctly classified as affordable though the percentage accuracy was not changed from simply removing the month factor. The high importance attributes remain the same, and the overall structure of the tree remains the same.

In order to see the dimension reduction effect when the most important factor is removed from the data set, I then ran the same test with the rent removed but the rest of the attributes remaining within the data set. Interestingly this only led to a 3% decrease in the overall accuracy of the tree, though the structure was very different. I believe that this shows for this dataset that the attributes are all necessary in their own way to provide the most accurate outcome from the decision tree learning tool. The more the dimensionality of this dataset is reduced, the less accurate it becomes.

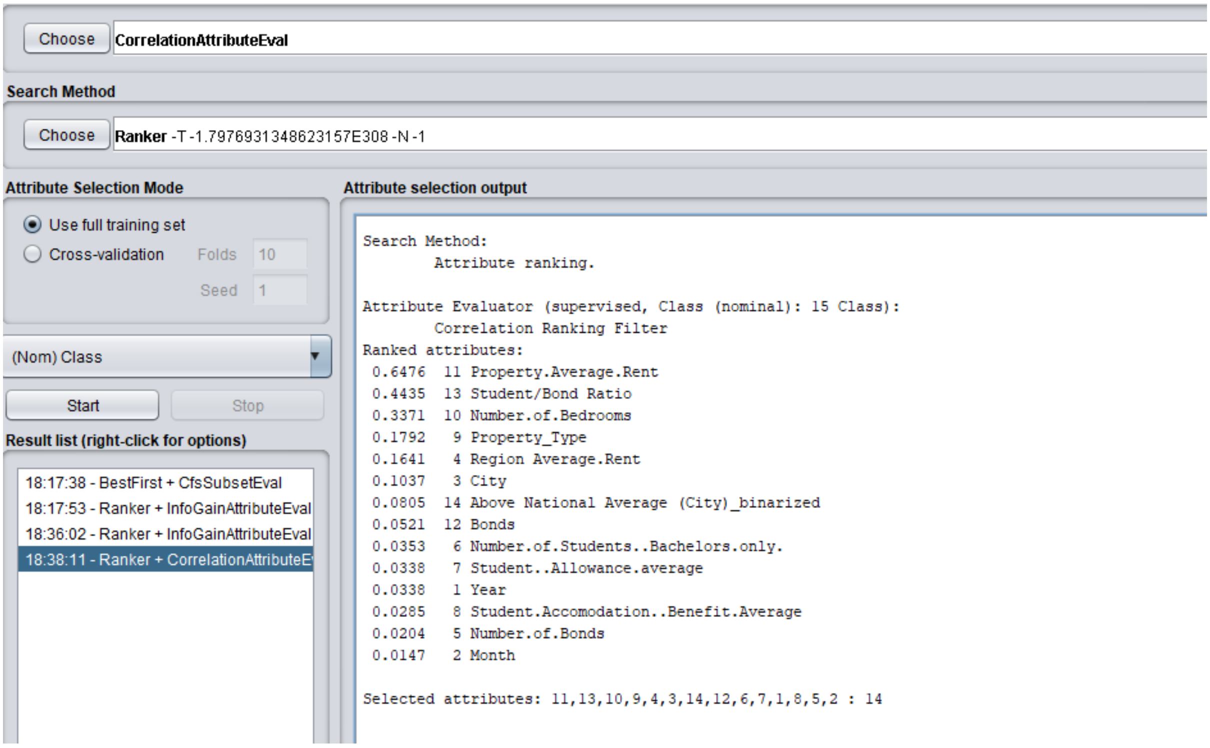
Data through the J48 tree before attribute selection:



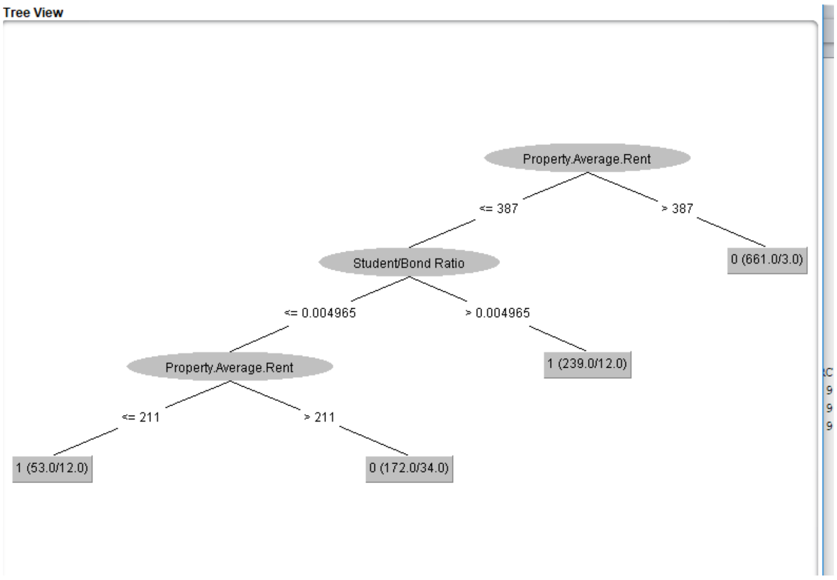
Information Gain on full dataset



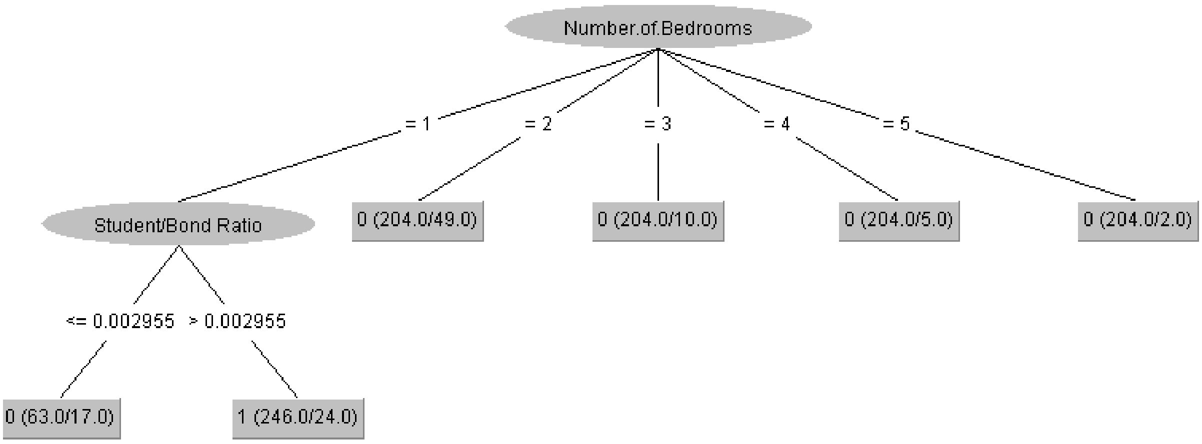
Pearson correlation on full dataset:



Decision tree with month/year/average accomodation removed



Decision tree with rent removed



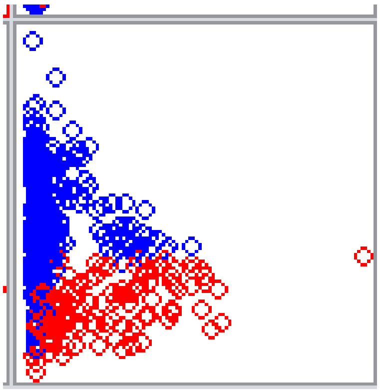
Rent removed statistics:

```
Time taken to build model: 0.01 seconds

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

Correctly Classified Instances        1016               90.3111 %
Incorrectly Classified Instances       109                9.6889 %
Kappa statistic                          0.7473
Mean absolute error                      0.156
Root mean squared error                  0.2847
Relative absolute error                 39.4425 %
Root relative squared error             64.0377 %
Total Number of Instances             1125

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Cla
                 0.951    0.226    0.919      0.951    0.935      0.749    0.917     0.952     0
                 0.774    0.049    0.855      0.774    0.812      0.749    0.917     0.790     1
Weighted Avg.    0.903    0.178    0.901      0.903    0.902      0.749    0.917     0.908

=== Confusion Matrix ===

   a   b   <-- classified as
 780  40 |   a = 0
  69 236 |   b = 1
```

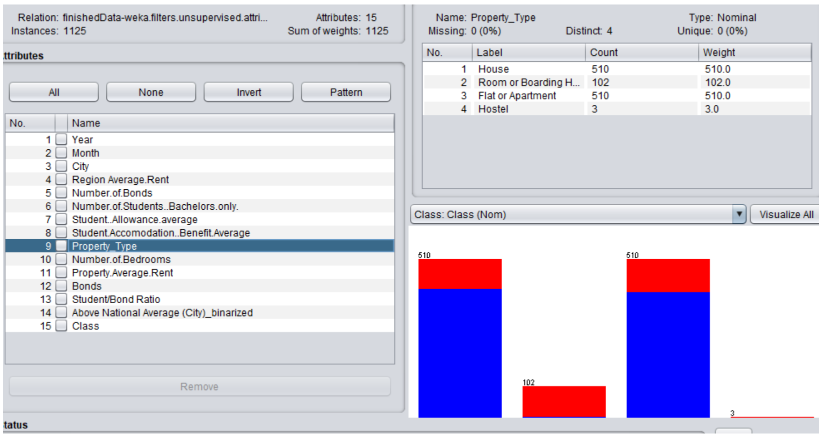
Evaluation and Deployment

The hypotheses of this assignment was that the Wellington rental market is more affordable for students than either the Auckland or Otago markets. This hypothesis was proven through evidence such as the number of properties deemed affordable in each region, but the hypothesis was also impacted by the low student/accommodation ratio. The property type deemed most affordable was the hostel, followed by the room or boarding situation, then the house or flat. The affordability is most affected by the price of the rental, though there were also significant impacts of the student ratio and the number of bedrooms available to rent. These factors are seen through the dimensionality reduction details, and the visualisation of data both in the WEKA explorer looking at attributes and in the graphs of the comparisons also shown in WEKA:

Student/bond ratio against property average rent:



Property type:

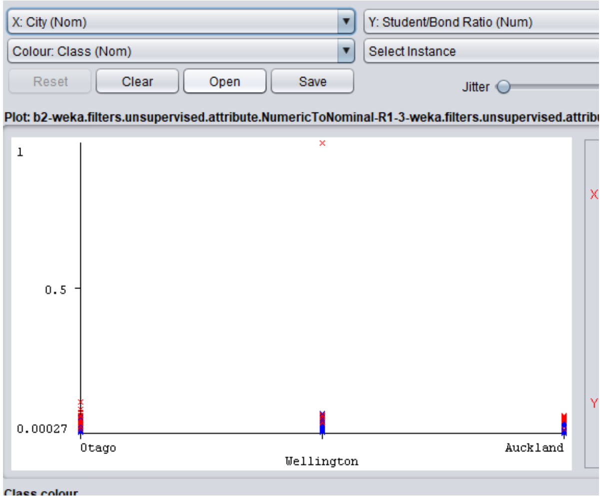


The consequences of these hypotheses are that students may see that the rental market in Wellington is not as dire as reports may claim, and that it has a good, in many ways affordable market for students to rent. However the number of properties available is far less than that of Auckland, though more than Otago, which may mean that as student numbers increase the competition will go up, the student/bond ratio will go down and it will become less and less affordable for students to rent in Wellington.

The pattern that has been established of affordability could be protected with increases in the number of one bedroom properties up for rent, including increasing the number of hostel rooms available by increasing the number of students which could be housed. This however brings up a space and means issue, as the hypotheses of Wellington rental market being better than Auckland and Otago will require a large number of new buildings, and the unrealistic space and cost of building larger hostels making that a less affordable option for Universities as well. There is also an option for students to rent a multiple bedroom house together, breaking up the price into a fraction of what it is for an individual person and thus making it more affordable for students to rent in those regions.

There are two types of fake data fed into the pipeline in order to test the hypotheses of adding more hostel rooms and also breaking up rent for affordability. In order to test the affordability the hypotheses change comes from each student paying for one of the bedrooms in the house i.e. the rent of a 5 bedroom house now becomes 1/5th of the original amount to compensate for breaking it up.

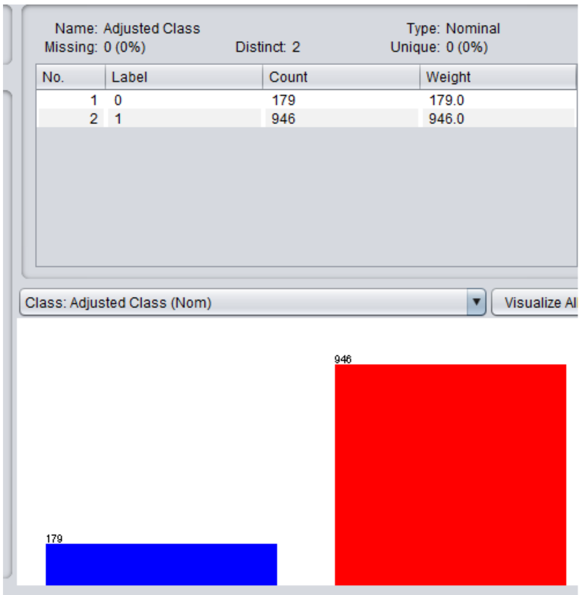
With the extra hostel data, the classification of the instances does not change a large amount. The ratio changes for the relevant data point, becoming a seeming outlier and splitting off from the rest of the student/bond ratios for the other instances. If this was to be a reality though the university would need to build hostels to accommodate the number of students. In the dataset used here that would require building a 12,000 room hostel to just cater for every bachelors student, not including the large number of post grads or an increase in the student numbers which will come with the current political stance of free university fees. The major changes in data points for the data set occurs in the following image:



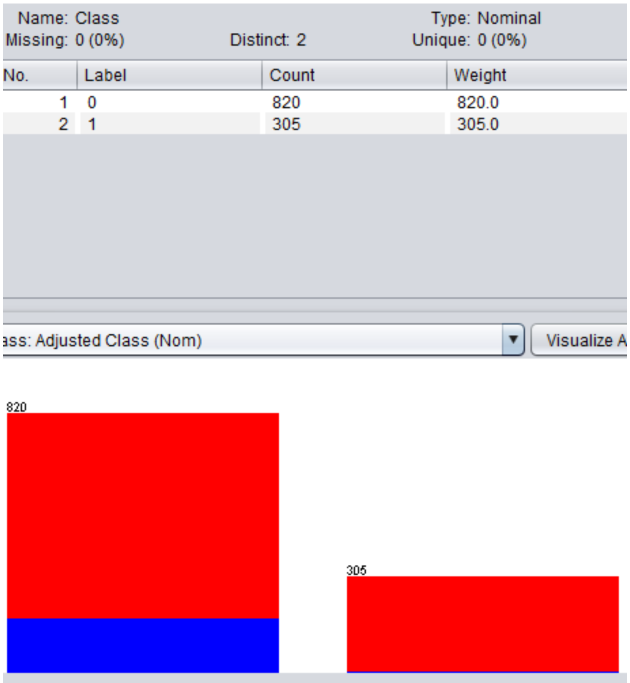
Which shows the major change in the student ratio.

The other way to split the dataset, in terms of rent/number of rooms, is shown below in a few snapshots from the data:

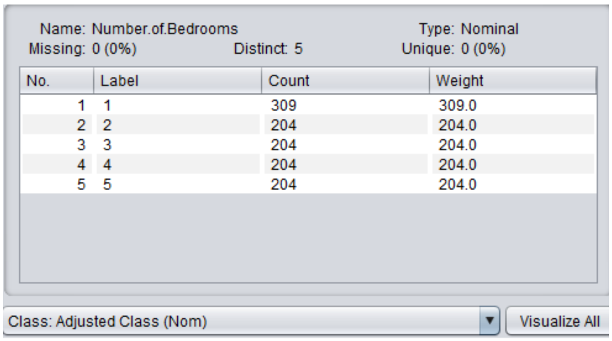
Overall affordability:



Original class against adjusted class:



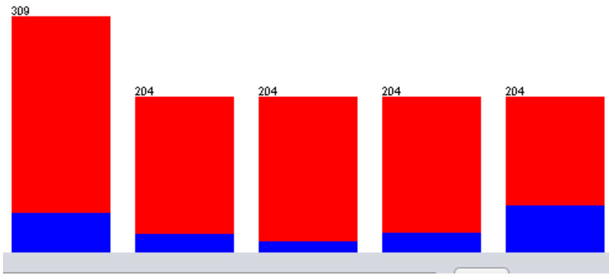
Number of bedrooms:



Name: Number.of.Bedrooms Type: Nominal
Missing: 0 (0%) Distinct: 5 Unique: 0 (0%)

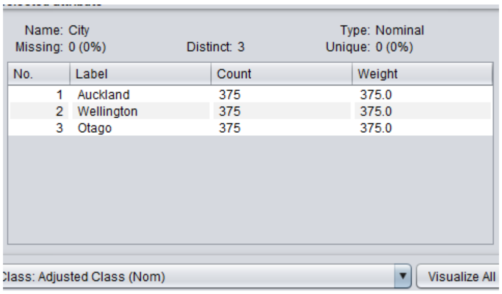
No.	Label	Count	Weight
1	1	309	309.0
2	2	204	204.0
3	3	204	204.0
4	4	204	204.0
5	5	204	204.0

Class: Adjusted Class (Nom) Visualize All



Adjusted rent vs original rent:

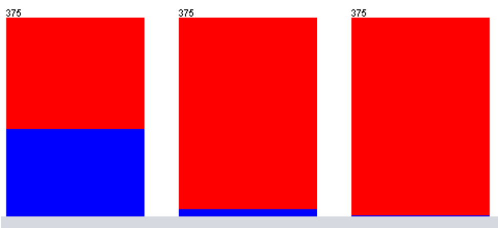
And city affordability:



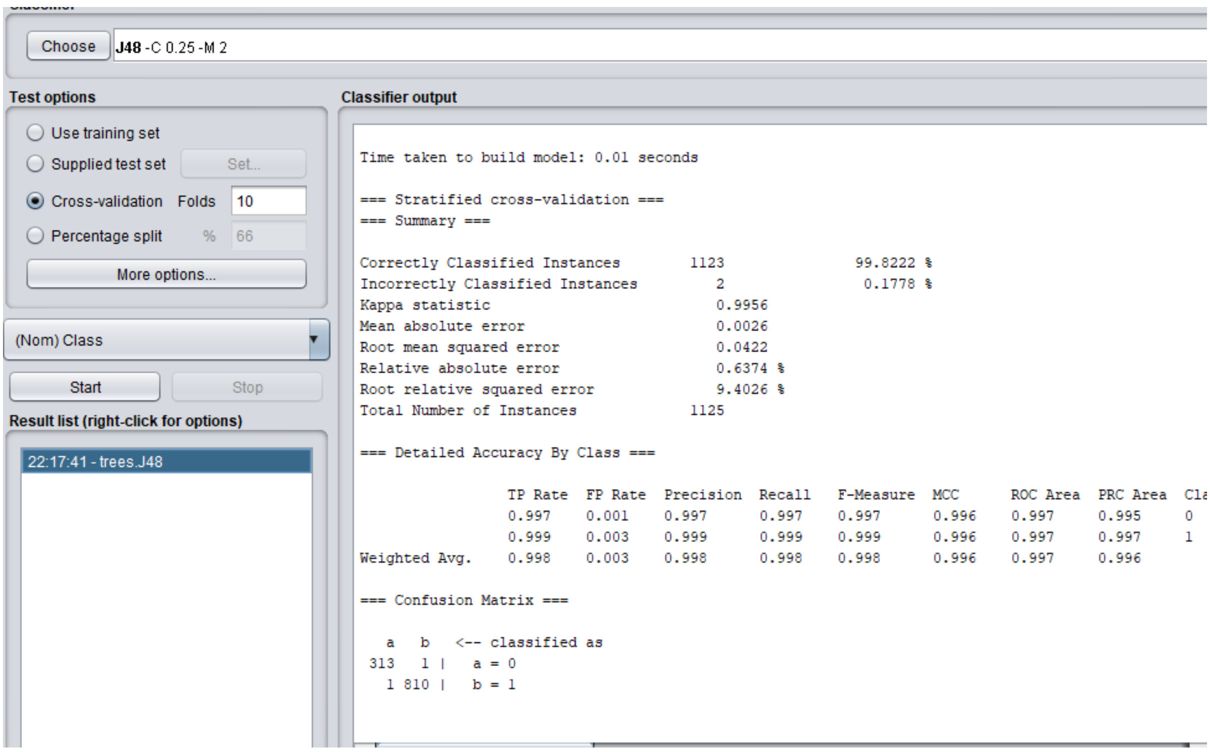
Name: City Type: Nominal
Missing: 0 (0%) Distinct: 3 Unique: 0 (0%)

No.	Label	Count	Weight
1	Auckland	375	375.0
2	Wellington	375	375.0
3	Otago	375	375.0

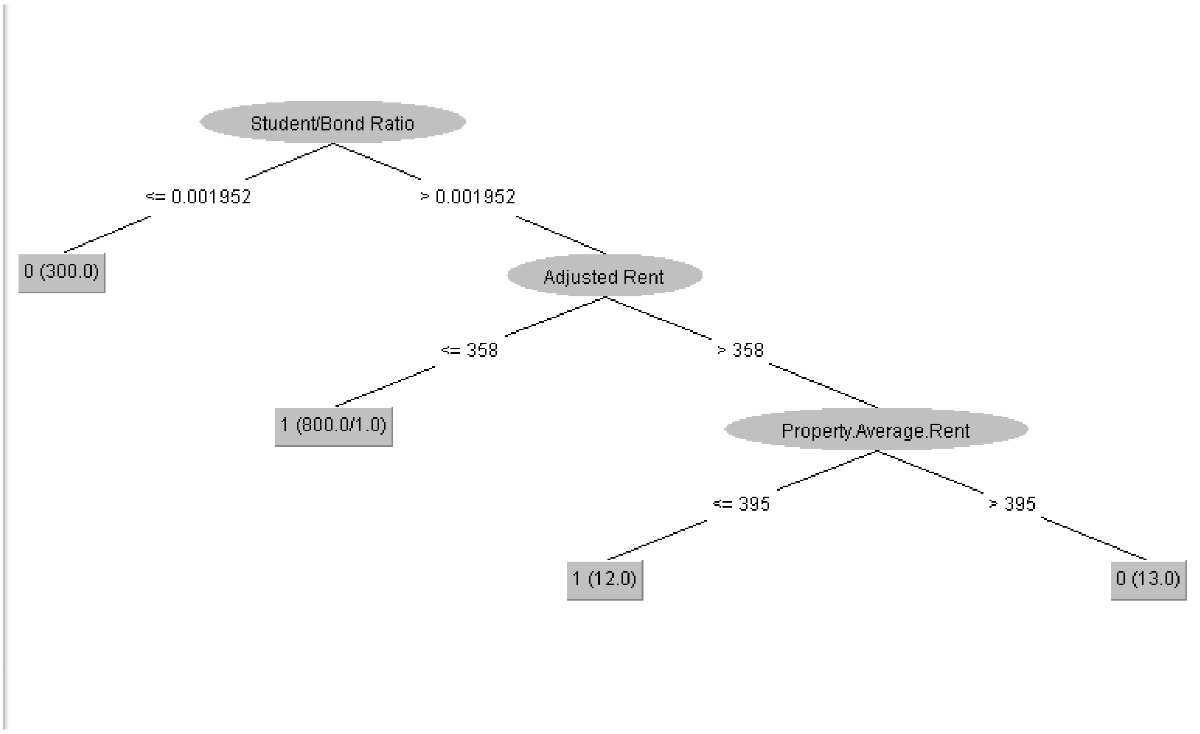
Class: Adjusted Class (Nom) Visualize All



And the classification output:



With new tree:



This shows that by splitting up the rent the affordability in all areas becomes much higher, though there are still some places which are beyond what a student could afford. The reality of this however is that it is very difficult for students to find groups of people they trust that they can rent with, and also in that as the number of bedrooms in a place increase the competition becomes much higher as the students become competitive in the market against more families and pairs of income, while the student income is capped much lower. This competition gives the landlords control too, allowing them to drive up the price of rent beyond what most students can afford. The most affordable accommodation would remain the hostels, and there would still be a low ratio of students to affordable housing even with this option of splitting the rental cost.