



Ordinary Degree in Computing: Data Mining.

Data Mining Assessment: Mine a dataset

Submitted by: Nicky Randles, B00058026

Submission date

12/12/14

Declaration

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Ordinary Degree in Computing in the Institute of Technology Blanchardstown, is entirely my own work except where otherwise stated.

Author: Nicholas Randles

Dated: 12/12/14

Table of Contents

Business

Understanding.....	1
--------------------	---

- Introduction.....1
- Business Objective.....1
- Data mining objective.....1

Data Understanding.....	2
-------------------------	---

- Describe the data.....2
- Explore the data.....3
- Verify the data.....5

Data Preparation.....	7
-----------------------	---

1. Replace missing values.....7
2. Filter Examples.....7
3. Normalization.....8
4. Detect Outliers.....9
5. Select Attributes.....10

Modeling.....	11
---------------	----

Evaluation.....	14
-----------------	----

Business Understanding

Introduction

The Australian open dataset contains the statistics for each match that has taken place in the tournaments between two players. It includes players of both genders. There are 40 attributes in the dataset which help to determine who the best player in the match was. The class label for the dataset is result. There are 254 rows of data in the data set. Each row contains information about player 1 and 2 in the match such as first serve percentage (FSP1 + 2), first serve won (FSW1+ 2), second serve percentage (SSP1 + 2), second serve won(SSW 1 + 2), aces won(ACE1 + 2), double faults committed (DBF1 + 2), winners earned (WNR1 + 2), unforced errors committed(UFE1 + 2), break points created(BPC1 + 2), break points won(BPW1 + 2), net points attempted(NPA1 + 2), net points won(NPW1 + 2), total points won(TPW1 + 2), set 1 results(ST1 1+2), set 2 results(ST2 1+2), set 3 results(ST3 1+2) , set 4 results(ST4 1+2) , set 5 results(ST5 1+2).

Business Objective

- My objective is to find out who the best players are and how they perform in their matches.

Data mining objective

- To predict how well each player performs during the matches.
- To predict who the best player was in the match.

Data Understanding

Describe the data

There is 1 special attribute and 40 regular attributes in the dataset. The class label is result. Result is always either zero or one. One means that player 1 won the match. Of the 40 regular attributes, 29 are numerical and 11 are nominal.

Numerical attributes

Name	Description	Mean	Min	Max	Standard deviation
Round	Round of the tournament at which game is played	1.992	+/- 1.511	1.000	15.000
FSP1	First Serve Percentage for player 1	61.768	+/- 8.047	40.000	86.000
FSW1	First Serve Won by player 1	38.236	+/- 16.978	3.000	109.000
SSP1	Second Serve Percentage for player 1	38.232	+/- 8.047	14.000	60.000
SSW1	Second Serve Won by player 1	16.744	+/- 8.682	1.000	47.000
ACE1	Aces won by player 1	6.948	+/- 6.849	0.000	41.000
DBF1	Double Faults committed by player 1	4.210	+/- 4.142	0.000	50.000
WNR1	Winners earned by player 1	25.228	+/- 17.622	0.000	111.000
UFE1	Unforced Errors committed by player 1	28.295	+/- 17.247	0.000	81.000
BPC1	Break Points Created by player 1	3.984	+/- 3.438	0.000	45.000
BPW1	Break Points Won by player 1	8.933	+/- 5.224	0.000	28.000
NPA1	Net Points Attempted by player 1	10.598	+/- 7.095	0.000	37.000
NPW1	Net Points Won by player 1	16.282	+/- 11.206	1.000	61.000
TPW1	Total Points Won by player 1	91.051	+/- 35.781	5.000	231.000
ST11	Set 1 result for Player 1	4.854	+/- 1.923	0.000	7.000
FSP2	First Serve Percentage for player 2	61.332	+/- 7.859	39.000	86.000
FSW2	First Serve Won by player 2	38.177	+/- 17.445	0.000	114.000
SSP2	Second Serve Percentage for player 2	38.909	+/- 8.737	14.000	100.000
SSW2	Second Serve Won by player 2	16.768	+/- 8.915	1.000	57.000
ACE2	Aces won by player 2	5.916	+/- 5.722	0.000	32.000
DBF2	Double Faults committed by player 2	4.433	+/- 3.137	0.000	18.000
WNR2	Winners earned by player 2	25.094	+/- 16.410	0.000	82.000
UFE2	Unforced Errors committed by player 2	29.886	+/- 18.700	0.000	96.000
BPC2	Break Points Created by player 2	3.585	+/- 2.340	0.000	10.000
BPW2	Break Points Won by player 2	8.324	+/- 4.864	0.000	22.000
NPA2	Net Points Attempted by player 2	11.809	+/- 8.832	0.000	49.000
NPW2	Net Points Won by player 2	17.517	+/- 12.747	0.000	66.000
TPW2	Total Points Won by player 2	90	+/- 36.305	1.000	230.000
ST12	Set 1 result for Player 2	4.771	+/- 1.957	0.000	7.000

Nominal attributes

Name	Description	Mode	Least
Player 1	Name of Player 1	Rafael Nadal (7)	Lucas Lako (1)
Player 2	Name of Player 2	Roger Federer (6)	Albert Montanes (1)
ST21	Set 2 Result for Player 1	6 (128),	NA (1)
ST31	Set 3 Result for Player 1	NA (89)	8 (2)
ST41	Set 4 Result for Player 1	NA (193)	5 (1)
ST51	Set 5 Result for Player 1	NA (231)	8 (1)
ST22	Set 2 Result for Player 2	6 (99),	NA (1)
ST32	Set 3 Result for Player 2	NA (89)	9 (1)
ST42	Set 4 Result for Player 2	NA (193)	5 (1)
ST52	Set 5 Result for Player 2	NA (231)	1 (1)
Gender	MALE is the men's competition matches; FEMALE is the Women's competition matches.	Male (127)	Male (127)

Explore the data

I noticed from looking at the box plots that most of the attributes are quite normally distributed (Figure 1). The box plots also helped me identify predictive attributes (Figure 2). By looking at the scatter plots I was able to spot outliers in many of the attributes (Figure 3). I was also able to find correlation between different attributes in the scatter plots (Figure 4). By looking at the histograms I was able to see many outliers and skewness of attributes (Figure 5). I was able to notice all of the attribute values ranges by looking at the parallel plots (Figure 6).



Figure 1: Box plot for Player 1

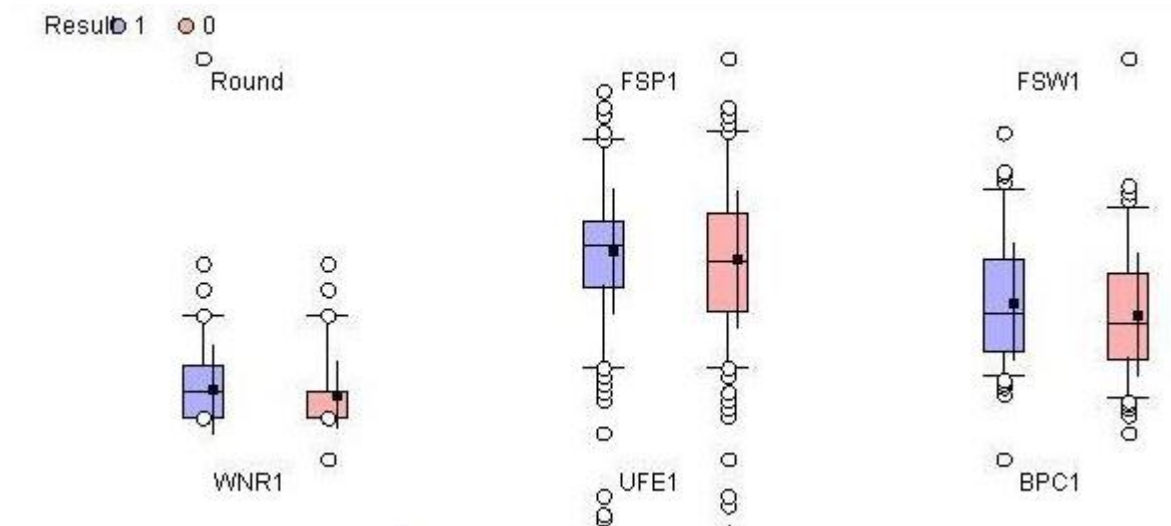


Figure 2: Box plots for Result

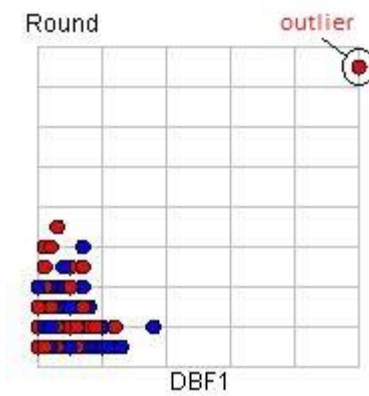


Figure 3: Scatter plot of attribute round with outlier



Figure 4: Negative correlation between attributes FSP1 and SSP1

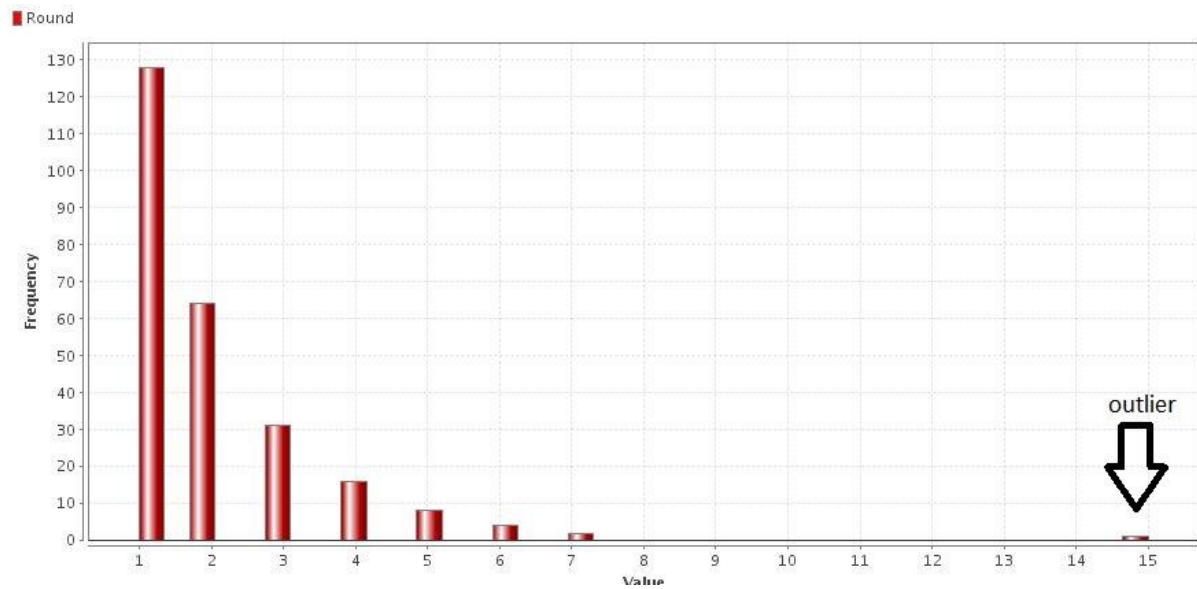


Figure 5: Histogram of attribute round with positive skew and outlier

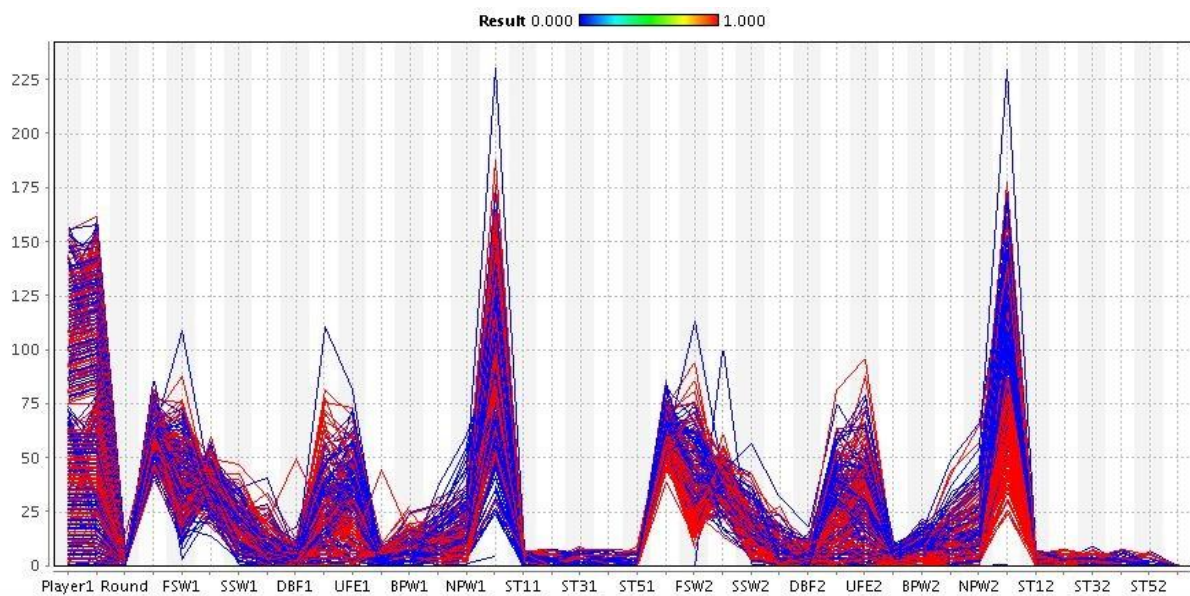


Figure 6: Parallel plot for attribute result

Verify data quality

The following table represents all of the attributes in the dataset. It shows whether or not they have outliers and what percentage of missing values they have. It also shows their individual data quality.

Name	Type	Outliers	Missing Data	Data Quality
Result	Integer	No	0%	Good – No missing values or outliers.
Player1	Polynomial	No	0%	Good – No missing values or outliers.
Player2	Polynomial	No	0%	Good – No missing values or outliers.
Round	Integer	Yes	0%	Good – No missing values but has outliers. Data is skewed.
FSP1	Integer	No	0%	Good – No missing values or outliers.
FSW1	Integer	Yes	0%	Good – No missing values but has outliers.
SSP1	Integer	No	0%	Good – No missing values or outliers.
SSW1	Integer	No	0%	Good – No missing values or outliers.
ACE1	Integer	Yes	2%	Ok – Small amount of missing values and outliers. Data is skewed.
DBF1	Integer	Yes	0.8%	Ok – Small amount of missing values and outliers.
WNR1	Integer	Yes	0%	Good – No missing values but has outliers. Data is skewed.
UFE1	Integer	No	0%	Good – No missing values or outliers.
BPC1	Integer	Yes	0.4%	Ok – Small amount of missing values and outliers.
BPW1	Integer	Yes	0.4%	Ok – Small amount of missing values and outliers.
NPA1	Integer	Yes	17.7%	Bad – Quite large amount of missing data missing and has outliers.
NPW1	Integer	No	17.7%	Ok – Good bit of data missing but has no outliers.
TPW1	Integer	Yes	0%	Good – No missing values but has outliers.
ST11	Integer	No	0.4%	Good – No outliers but has a small amount of missing data.
ST21	Polynomial	No	0.4%	Good – No outliers but has a small amount of missing data.
ST31	Polynomial	No	0.4%	Good – No outliers but has a small amount of missing data.
ST41	Polynomial	No	0.4%	Good – No outliers but has a small amount of missing data.
ST51	Polynomial	No	0.4%	Good – No outliers but has a small amount of missing data.
FSP2	Integer	No	0.4%	Good – No outliers but has a small amount of missing data.
FSW2	Integer	Yes	0%	Good – No missing values but has outliers.
SSP2	Integer	Yes	0%	Good – No missing values but has outliers.
SSW2	Integer	Yes	0%	Good – No missing values but has outliers.
ACE2	Integer	Yes	2%	Ok – Small amount of missing values and outliers. Data is skewed.
DBF2	Integer	Yes	0.8%	Ok – Small amount of missing values and outliers.
WNR2	Integer	Yes	0%	Ok – Small amount of missing values and outliers.
UFE2	Integer	Yes	0%	Ok – Small amount of missing values and outliers.
BPC2	Integer	No	0.4%	Ok – Small amount of missing values and outliers.
BPW2	Integer	No	0.4%	Ok – Small amount of missing values and outliers.
NPA2	Integer	Yes	17.7%	Bad – Quite large amount of missing data missing and has outliers.
NPW2	Integer	Yes	17.7%	Bad – Quite large amount of missing data missing and has outliers.
TPW2	Integer	Yes	0%	Good – No missing values but has outliers.
ST12	Integer	No	0.4%	Good – No outliers but has a small amount of missing data.
ST22	Polynomial	No	0.4%	Good – No outliers but has a small amount of missing data.
ST32	Polynomial	No	0.4%	Good – No outliers but has a small amount of missing data.
ST42	Polynomial	No	0.4%	Good – No outliers but has a small amount of missing data.
ST52	Polynomial	No	0.4%	Good – No outliers but has a small amount of missing data.
Gender	Binominal	No	0%	Good – No missing values or outliers.

In the Australian open dataset there are quite a few outliers present and there are missing values in the dataset. Attributes such as NPA1, NPW1, NPW1, and NPW2 have a good bit of missing values. Some of the attributes are skewed. The dataset will definitely need to be cleaned up. I will have to remove some of the outliers and I will have to handle the missing values. Overall the data quality is quite good. It has sufficient attributes that will help me to accomplish my data mining objectives.

Data Preparation

The Australian open dataset is quite small. It only has 254 rows. It will not need a percentage of the rows removed. The dataset has some outliers and missing values. 17.7% is the highest amount of missing values in a column. Although this is quite high, it is not high enough to delete the entire column. These missing values will need to be replaced. There is quite a large variation of attributes so the data will have to be normalised. Some attributes such as player 1, player 2 and gender will be deleted as they are not helping the data mining objective.

accuracy: 81.45% +/- 6.31% (mikro: 81.50%)			
	true 1	true 0	class precision
pred. 1	110	23	82.71%
pred. 0	24	97	80.17%
class recall	82.09%	80.83%	

Figure 7: Baseline accuracy with k-NN

Techniques Used

There were a number of different techniques I used to fix up the dataset. I had to use different approaches to remove unnecessary data, and clean up the data.

1. Replace missing values

For the attributes that had greater than 5% and less than 40% missing values I used the replace missing values operator. There was only 4 missing attributes with this percentage of missing values, they were NPA 1, NPA 2, NPW1 and NPW 2.

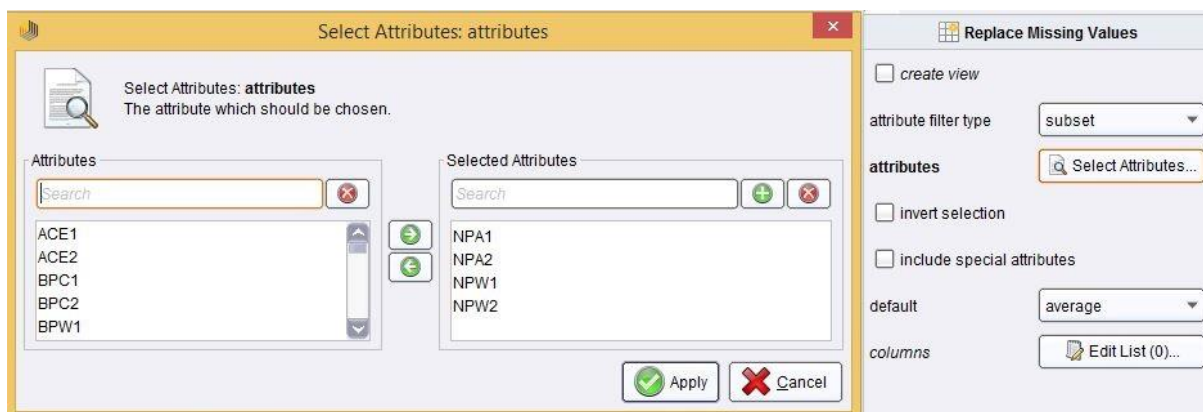


Figure 8: Selected attributes and parameters for Replace missing values operator

2. Filter Examples

The rest of the attributes with missing values were below 5%. For these, I removed the rows with missing values by using the Filter examples operator. I set the condition class to

‘no-missing-attributes’. This removed all the remaining attributes from the dataset that had missing values. By doing this the accuracy increased to 84.20%.

accuracy: 84.20% +/- 3.84% (mikro: 84.21%)			
	true 1	true 0	class precision
pred. 1	114	19	85.71%
pred. 0	20	94	82.46%
class recall	85.07%	83.19%	

Figure 9: Accuracy after removing remaining attributes with missing values

3. Normalisation

By looking at the meta data of the dataset I could tell there was a large variation in the range in the attributes. This is mainly caused by the outliers in the dataset. I normalised the attributes to reduce the number of outliers and put the attributes in the same range.

Role	Name	Type	Statistics	Range	Missings
label	Result	binominal	mode = 1 (134), least = 0 (113)	1 (134), 0 (113)	0
regular	NPA1	integer	avg = 10.838 +/- 6.437	[0.000 ; 37.000]	0
regular	NPW1	integer	avg = 16.462 +/- 10.184	[1.000 ; 61.000]	0
regular	NPA2	integer	avg = 11.931 +/- 8.073	[0.000 ; 49.000]	0
regular	NPW2	integer	avg = 17.765 +/- 11.645	[1.000 ; 66.000]	0
regular	Player1	polynomial	mode = Rafael Nadal (7), least	Rafael Nadal (7), Na Li (6), Torr	0
regular	Player2	polynomial	mode = Roger Federer (6), leas	Roger Federer (6), David Ferrer	0
regular	Round	integer	avg = 2 +/- 1.525	[1.000 ; 15.000]	0
regular	FSP1	integer	avg = 61.688 +/- 8.049	[40.000 ; 86.000]	0
regular	FSW1	integer	avg = 38.854 +/- 16.732	[11.000 ; 109.000]	0
regular	SSP1	integer	avg = 38.312 +/- 8.049	[14.000 ; 60.000]	0
regular	SSW1	integer	avg = 16.976 +/- 8.663	[1.000 ; 47.000]	0
regular	ACE1	integer	avg = 6.988 +/- 6.862	[0.000 ; 41.000]	0
regular	DBF1	integer	avg = 4.231 +/- 4.175	[0.000 ; 50.000]	0
regular	WNR1	integer	avg = 25.591 +/- 17.663	[0.000 ; 111.000]	0
regular	UFE1	integer	avg = 28.304 +/- 17.306	[0.000 ; 81.000]	0
regular	BPC1	integer	avg = 4.008 +/- 3.465	[0.000 ; 45.000]	0
regular	BPW1	integer	avg = 9.024 +/- 5.238	[0.000 ; 28.000]	0

Figure 10: Data before being normalised

Role	Name	Type	Statistics	Range	Missings
label	Result	binominal	mode = 1 (134), least = 0 (113)	1 (134), 0 (113)	0
regular	NPA1	real	avg = 0.293 +/- 0.174	[0.000 ; 1.000]	0
regular	NPW1	real	avg = 0.258 +/- 0.170	[0.000 ; 1.000]	0
regular	NPA2	real	avg = 0.243 +/- 0.165	[0.000 ; 1.000]	0
regular	NPW2	real	avg = 0.258 +/- 0.179	[0.000 ; 1.000]	0
regular	Round	real	avg = 0.071 +/- 0.109	[0.000 ; 1.000]	0
regular	FSP1	real	avg = 0.471 +/- 0.175	[0.000 ; 1.000]	0
regular	FSW1	real	avg = 0.284 +/- 0.171	[0.000 ; 1.000]	0
regular	SSP1	real	avg = 0.529 +/- 0.175	[0.000 ; 1.000]	0
regular	SSW1	real	avg = 0.347 +/- 0.188	[0.000 ; 1.000]	0
regular	ACE1	real	avg = 0.170 +/- 0.167	[0.000 ; 1.000]	0
regular	DBF1	real	avg = 0.085 +/- 0.083	[0.000 ; 1.000]	0
regular	WNR1	real	avg = 0.231 +/- 0.159	[0.000 ; 1.000]	0
regular	UFE1	real	avg = 0.349 +/- 0.214	[0.000 ; 1.000]	0
regular	BPC1	real	avg = 0.089 +/- 0.077	[0.000 ; 1.000]	0
regular	BPW1	real	avg = 0.322 +/- 0.187	[0.000 ; 1.000]	0
regular	TPW1	real	avg = 0.330 +/- 0.170	[0.000 ; 1.000]	0
regular	ST11	real	avg = 0.701 +/- 0.269	[0.000 ; 1.000]	0
regular	FSP2	real	avg = 0.474 +/- 0.167	[0.000 ; 1.000]	0

Figure 11: Data after being normalised

When I set the method to Z-transformation my accuracy increased to 87.83%.

accuracy: 87.83% +/- 3.19% (mikro: 87.85%)			
	true 1	true 0	class precision
pred. 1	120	16	88.24%
pred. 0	14	97	87.39%
class recall	89.55%	85.84%	

Figure 12: Accuracy with method Z-transformation

When I set the method to range transformation my accuracy increased to 93.12%.

Normalize

☐ create view

attribute filter type

all

☐ invert selection

☐ include special attributes

method

range transforma...

min

0.0

max

1.0

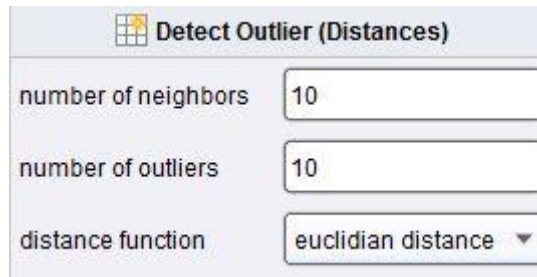
Figure 13: Normalise parameters

accuracy: 93.12% +/- 3.18% (mikro: 93.12%)			
	true 1	true 0	class precision
pred. 1	126	9	93.33%
pred. 0	8	104	92.86%
class recall	94.03%	92.04%	

Figure 14: Accuracy with method range transformation

4. Detect Outliers (Distances)

By looking at the histograms and scatter plots of the attributes, I was able to see that there still was outliers in the dataset. I used the detect outliers (distances) operator to detect them based on their distance to their neighbours.



Detect Outlier (Distances)

number of neighbors: 10


number of outliers: 10

distance function: euclidian distance

Figure 15: Parameters I set for Detect Outliers (Distances)

5. Select Attributes

I used the select attributes operator to remove the attributes I detected with the detect outliers class. I also used it to get rid of attributes that were not necessary to help achieve the data mining object such as gender, player 1, and player 2.



Select Attributes: attributes

Select Attributes: **attributes**
The attribute which should be chosen.

Attributes: ACE1, ACE2, BPC1, BPC2, BPW1, BPW2

Selected Attributes: Gender, Player1, Player2, outlier

attribute filter type: subset

☒ Invert selection

☒ include special attributes

Apply Cancel

Figure 16: Selected attributes and parameters for select attributes

accuracy: 91.12% +/- 3.49% (mikro: 91.09%)			
	true 1	true 0	class precision
pred. 1	125	13	90.58%
pred. 0	9	100	91.74%
class recall	93.28%	88.50%	

Figure 17: Accuracy after removing attributes with select attributes operator

Modelling

The data algorithms that I thought well suited the Australian open dataset were the decision tree and k-nearest neighbour algorithm. K-nearest neighbour suited the dataset because it is able to take all different types of attributes. This is very important as the dataset contains many different types of attributes such as binomial, polynomial and integer attributes. The decision tree is able to take categorical, and some numerical attributes so it can also be used.

I will experiment with both the decision tree and k-nearest neighbour algorithms to see which gives the best results. I will add the model to the training set and then apply the model to the testing set. I will then enter a performance operator so I can then evaluate the performance of both algorithms. I will then connect an x-validation method up to the Australian open dataset.

The accuracy for the dataset using the decision tree algorithm with the default parameters was quite high at 87.83%.

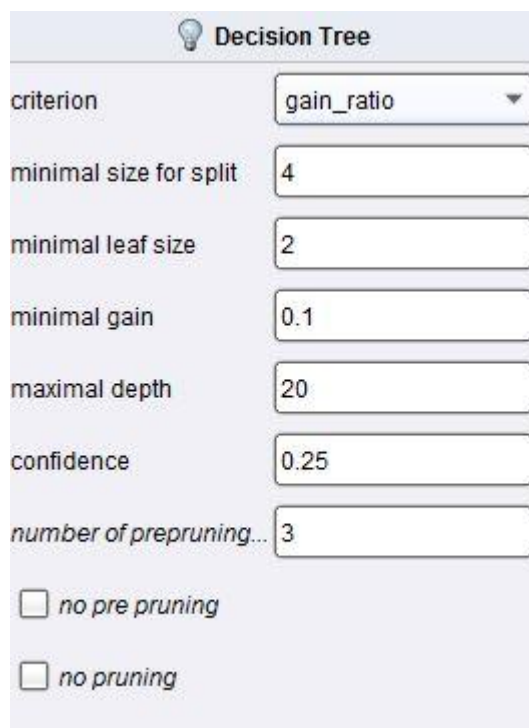


Figure 18: Default parameter of decision tree

accuracy: 87.83% +/- 7.97% (mikro: 87.80%)			
	true 1	true 0	class precision
pred. 1	120	17	87.59%
pred. 0	14	103	88.03%
class recall	89.55%	85.83%	

Figure 19: Accuracy of dataset using decision tree with default parameters

I then increased the 'minimal size for split' to 5 and the accuracy increased to 89.42% (Fig 20). I increased the 'minimal leaf size' to 4 and the accuracy went up to 90.17% (Fig 21). I tried 5 but the accuracy went down to 88.57% (Fig 22). I changed the 'minimal gain' to 0.2 and the accuracy went up to 91.31% (Fig 23).

accuracy: 89.42% +/- 6.56% (mikro: 89.37%)			
	true 1	true 0	class precision
pred. 1	123	16	88.49%
pred. 0	11	104	90.43%
class recall	91.79%	86.67%	

Figure 20: Accuracy when 'minimal size for split' changed to 5

accuracy: 90.17% +/- 3.60% (mikro: 90.16%)			
	true 1	true 0	class precision
pred. 1	120	11	91.60%
pred. 0	14	109	88.62%
class recall	89.55%	90.83%	

Figure 21: Accuracy when 'minimal leaf size' changed to 4

accuracy: 88.57% +/- 3.31% (mikro: 88.58%)			
	true 1	true 0	class precision
pred. 1	117	12	90.70%
pred. 0	17	108	86.40%
class recall	87.31%	90.00%	

Figure 22: Accuracy when 'minimal leaf size' changed to 5

accuracy: 91.32% +/- 4.25% (mikro: 91.34%)			
	true 1	true 0	class precision
pred. 1	123	11	91.79%
pred. 0	11	109	90.83%
class recall	91.79%	90.83%	

Figure 23: Accuracy when 'minimal gain' changed to 0.2

accuracy: 88.57% +/- 5.49% (mikro: 88.58%)			
	true 1	true 0	class precision
pred. 1	115	10	92.00%
pred. 0	19	110	85.27%
class recall	85.82%	91.67%	

Figure 23: Accuracy when 'minimal gain' changed to 0.3

The accuracy for the dataset using the k-NN algorithm with the default parameters was 81.45%.



k: 1
☐ weighted vote
 measure types: MixedMeasures
 mixed measure: MixedEuclideanD...

Figure 24: Default parameter for k-NN

accuracy: 81.45% +/- 6.31% (mikro: 81.50%)			
	true 1	true 0	class precision
pred. 1	110	23	82.71%
pred. 0	24	97	80.17%
class recall	82.09%	80.83%	

Figure 25: Accuracy of dataset using k-NN with default parameters

I then increase k to 5 and the accuracy went up to 85.43%. I then increased it to ten and the accuracy remained the same. I then increased it to 15 and the accuracy went up to 87.02%. I then increase k to 20 and it the accuracy went up to 89.00%.

accuracy: 85.45% +/- 3.00% (mikro: 85.43%)			
	true 1	true 0	class precision
pred. 1	115	18	86.47%
pred. 0	19	102	84.30%
class recall	85.82%	85.00%	

Figure 26: Accuracy when k = 5

accuracy: 87.02% +/- 4.94% (mikro: 87.01%)			
	true 1	true 0	class precision
pred. 1	116	15	88.55%
pred. 0	18	105	85.37%
class recall	86.57%	87.50%	

Figure 27: Accuracy when k = 15

accuracy: 89.00% +/- 3.79% (mikro: 88.98%)			
	true 1	true 0	class precision
pred. 1	121	15	88.97%
pred. 0	13	105	88.98%
class recall	90.30%	87.50%	

Figure 28: Accuracy when k = 20

Evaluation

The dataset was quite hard to clean up as it had missing values and had a good few outliers. I was able to manage it by using a number of different operators. I was able to deal with a lot of the outliers and the large variation in ranges by using the Normalisation operator. This operator had a very positive effect on my data and increased my accuracy substantially. I handled the missing values two ways. For the attributes with over 5% and less than 40% I used the replace attribute operator to replace the attributes and for the ones with less than 5% I removed them using the Filter examples operator. This helped me to get closer to completing my business and data mining goals.

I used two data algorithms, k-NN and decision trees. I found k-NN more effective probably because it is more suitable for this type of data compared to decision trees. It was more efficient and it gave me a higher accuracy than the decision tree did. Both algorithms were not that scalable but k-NN was definitely more scalable. Although, it was quite hard to read the test data in k-NN compared to the decision tree which was relatively easy to understand.

All of the above operators and algorithms helped me to achieve my business and data mining objectives.

accuracy: 93.12% +/- 3.18% (mikro: 93.12%)			
	true 1	true 0	class precision
pred. 1	126	9	93.33%
pred. 0	8	104	92.86%
class recall	94.03%	92.04%	

Figure 29: Final accuracy of the Australian dataset using k-NN

From doing this assignment I have learned a lot about data mining. I have learned how to analyse a dataset and all of its attributes. I have learned how to clean up a dataset. I have a lot about all of the different operators on rapidminer and how to implement them to improve my dataset. I have learned a lot about decision trees and k-NN. I discovered how to alter their parameters to get the best accuracy. I have also learned how important data mining is in the real world from doing this assignment and how it can benefit all companies.