Honours Degree in Computing

Data Analytics Assessment:
Analyse a dataset

Submitted by: Stephen Blaney, B00076157

Submission date 10th December 2017

Declaration

I declare that this material, which I now submit for assessment, is entirely my own work and has not been taken from the work of others, except where otherwise stated. I 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 understand that plagiarism, collusion, and copying are grave and serious offences and accept the penalties that would be imposed should I 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 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: Stephen Blaney Dated: 26/10/2017

Table of Contents . . .

Contents

1. Business Understanding

The thyroid dataset is derived from UCI machine Learning Repository and is an example of a classification dataset from the Garavan Institute in Australia supplied by Ross Quinlan. The dataset contains various thyroid related information on patients e.g. age, gender, goitre etc.

1.1 Business objective

To increase efficiency in hospitals by predicating if a patient is negative, (same) decreased binding protein(worst), or increased binding protein (better), to identify whether they need to remain hospitalized.

1.2 Data Mining objective

To build a classification model to predict when the status of patient when they are better worst or the same. This knowledge could then be used to increase efficiency among staff and hospital resources.

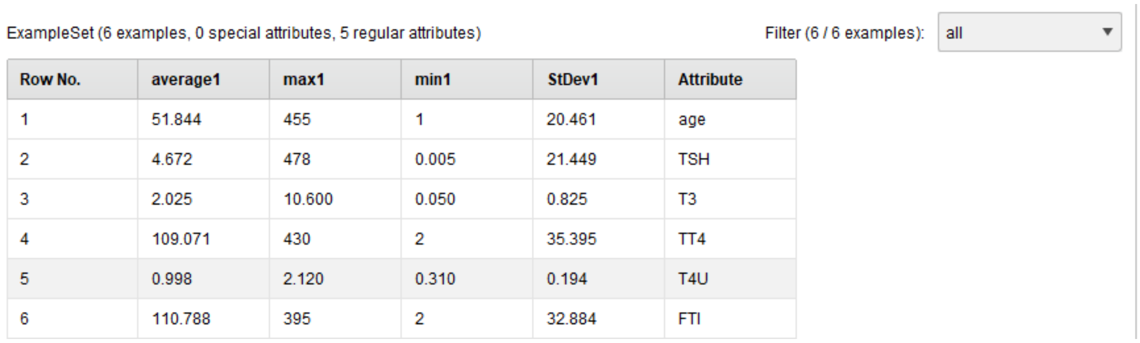
2.Data Understanding.

2.1 Describe the data

The dimensionality of this this database (Number of columns) is 30 which are either Boolean or continuously valued. The number of data instances is 2800 with 972 test instances.

Attributes	**Datatype**	**Description**
Class(label)	Polynomial	Negative, increase binding protein decreased binding protein.
Age	Integer	Age of patients
Gender	Polynomial	Gender of patients
Thyroxine	Polynomial	Main hormone secreted by thyroid gland
QuaryThyroxine	Polynomial	
Antityroid	Polynomial	Medicine to produce less thyroid hormone.
Sick	Polynomial	Patients are sick or not
Pregnant	Polynomial	Patient pregnant or not
TyroidSurgery	Polynomial	Patient required surgery or not.
I131 treatment	Polynomial	thyroid radiation treatment
Hypothyroid	Polynomial	Underactive thyroid

Hyperthyroid	Polynomial	Thyroid gland produces too much hormone thyroxine
Lithium	Polynomial	lithium inhibits thyroid hormone release
Goitre	Polynomial	swelling in the neck resulting from an enlarged thyroid gland
Tumor	Polynomial	Abnormal growth of cells
Hypopituitary	Polynomial	pituitary gland does not produce one or more of its hormones
Psych	Polynomial	Psych evaluation
TSHmeasure	Polynomial	Thyroid stimulating hormone (TSH) in your blood
TSH	Real	Measure of thyroid stimulating hormone (TSH) in your blood
T3measured	Polynomial	T3 test
T3	Real	hormone called triiodothyronine
TT4measured	Polynomial	TT4 test
TT4	Integer	Thyroid hormones bound to proteins
T4Umeasured	Polynomial	T4U test results
T4U	Real	T4U amount of thyroxine that the body absorbs
FTI	Polynomial	Indicates primary hypothyroidism



ExampleSet (6 examples, 0 special attributes, 5 regular attributes)

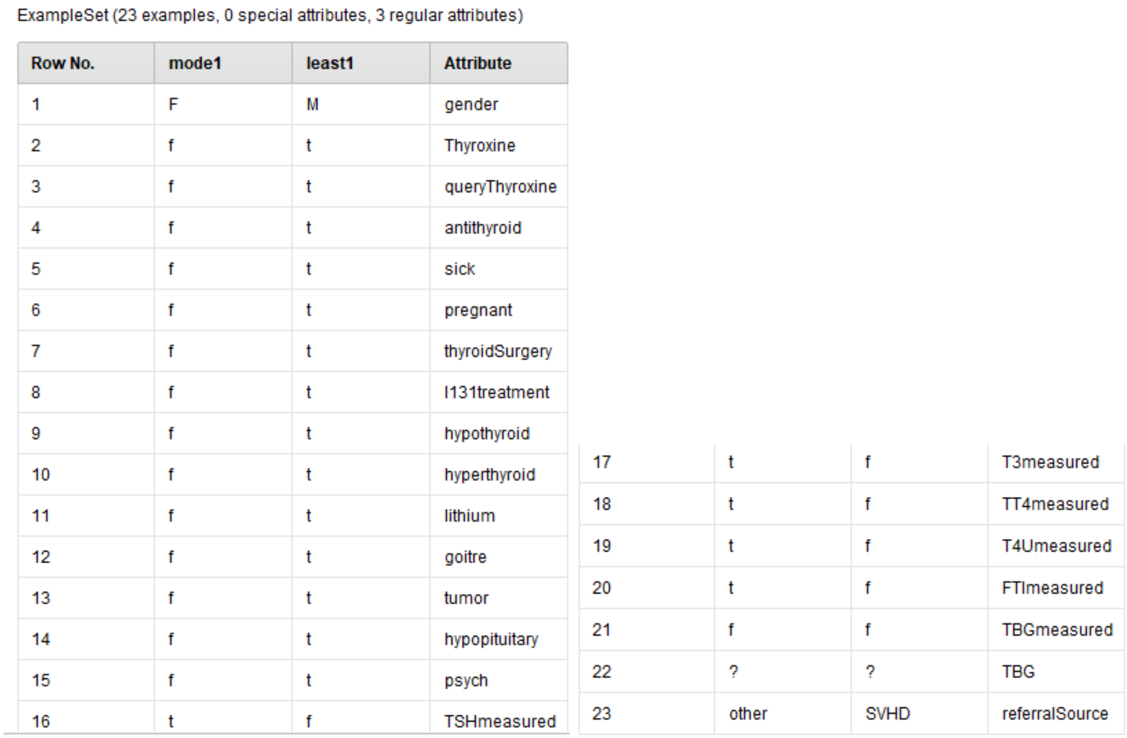
Filter (6 / 6 examples): all

Row No.	average1	max1	min1	StDev1	Attribute
1	51.844	455	1	20.461	age
2	4.672	478	0.005	21.449	TSH
3	2.025	10.600	0.050	0.825	T3
4	109.071	430	2	35.395	TT4
5	0.998	2.120	0.310	0.194	T4U
6	110.788	395	2	32.884	FTI

Figure 1: Min, Max, Average and standard deviation of all numeric attributes.

2.1.1 Meta data understanding figure 1 example1: The range for age seems to be off (Min 1 Max 455) indicating that this could be human error. This needs to be addressed during the data preparation stage.

2.1.2 Meta data understanding figure1 example 2: Thyroid stimulating hormone (TSH) attribute. A range (min and max) 0.5 to 4.5 is considered normal. With average being 4.672. Standard deviation of 21449 T4U is the only attribute that has normal distribution from viewing the histograms.

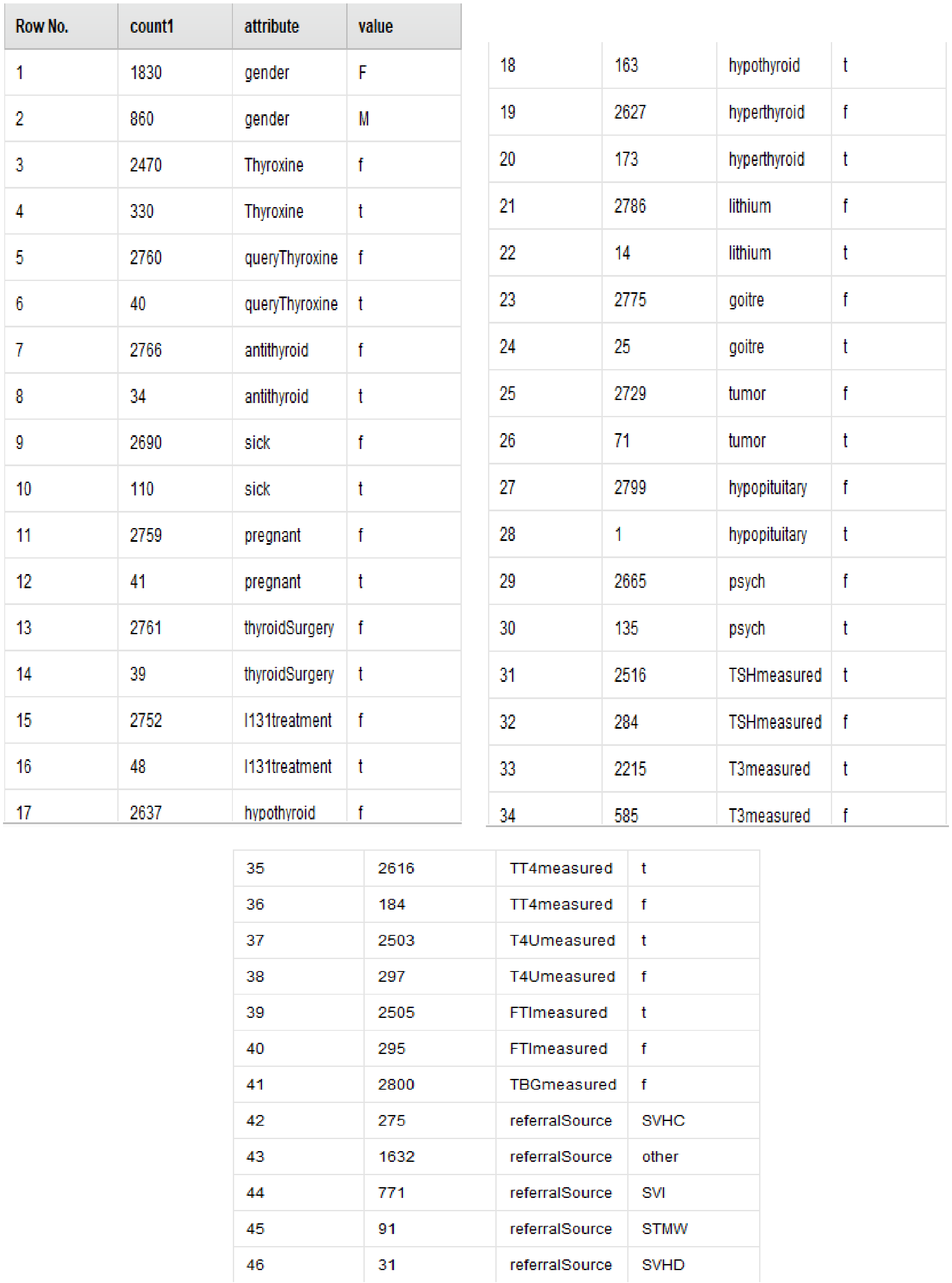


ExampleSet (23 examples, 0 special attributes, 3 regular attributes)

Row No.	mode1	least1	Attribute
1	F	M	gender
2	f	t	Thyroxine
3	f	t	queryThyroxine
4	f	t	antithyroid
5	f	t	sick
6	f	t	pregnant
7	f	t	thyroidSurgery
8	f	t	I131treatment
9	f	t	hypothyroid
10	f	t	hyperthyroid
11	f	t	lithium
12	f	t	goitre
13	f	t	tumor
14	f	t	hypopituitary
15	f	t	psych
16	t	f	TSHmeasured
17	t	f	T3measured
18	t	f	TT4measured
19	t	f	T4Umeasured
20	t	f	FTImeasured
21	f	f	TBGmeasured
22	?	?	TBG
23	other	SVHD	referralSource

Figure 2: The mode and least frequent attribute value for nominal attributes

2.1.3 Meta data understanding figure2 example 1: Interesting the TBG attribute seems to have all of its values missing this attribute can be deleted. Referral Source mode and least don't have the conventional Boolean values of true and false and don't appear to be very intuitive in understanding the data.



Row No.	count1	attribute	value
1	1830	gender	F
2	860	gender	M
3	2470	Thyroxine	f
4	330	Thyroxine	t
5	2760	queryThyroxine	f
6	40	queryThyroxine	t
7	2766	antithyroid	f
8	34	antithyroid	t
9	2690	sick	f
10	110	sick	t
11	2759	pregnant	f
12	41	pregnant	t
13	2761	thyroidSurgery	f
14	39	thyroidSurgery	t
15	2752	I131treatment	f
16	48	I131treatment	t
17	2637	hypothyroid	f
18	163	hypothyroid	t
19	2627	hyperthyroid	f
20	173	hyperthyroid	t
21	2786	lithium	f
22	14	lithium	t
23	2775	goitre	f
24	25	goitre	t
25	2729	tumor	f
26	71	tumor	t
27	2799	hypopituitary	f
28	1	hypopituitary	t
29	2665	psych	f
30	135	psych	t
31	2516	TSHmeasured	t
32	284	TSHmeasured	f
33	2215	T3measured	t
34	585	T3measured	f
35	2616	TT4measured	t
36	184	TT4measured	f
37	2503	T4Umeasured	t
38	297	T4Umeasured	f
39	2505	FTImeasured	t
40	295	FTImeasured	f
41	2800	TBGmeasured	f
42	275	referralSource	SVHC
43	1632	referralSource	other
44	771	referralSource	SVI
45	91	referralSource	STMW
46	31	referralSource	SVHD

Figure 3: Nominal counts showing the attributes that are true and false in the dataset

2.1.4 Meta data understanding figure3 example 1: There seems to be many females in this set, in contrast to the men according to this nominal count

2.2 Explore the data

This dataset is very unbalanced and will cause trouble later. A model could predict everyone as Negative and be correct for most of the rows in the dataset.

By using a scatterplot correlated attributes could be discovered A good way of identifying this is this see if the attributes produce a diagonal sequence.

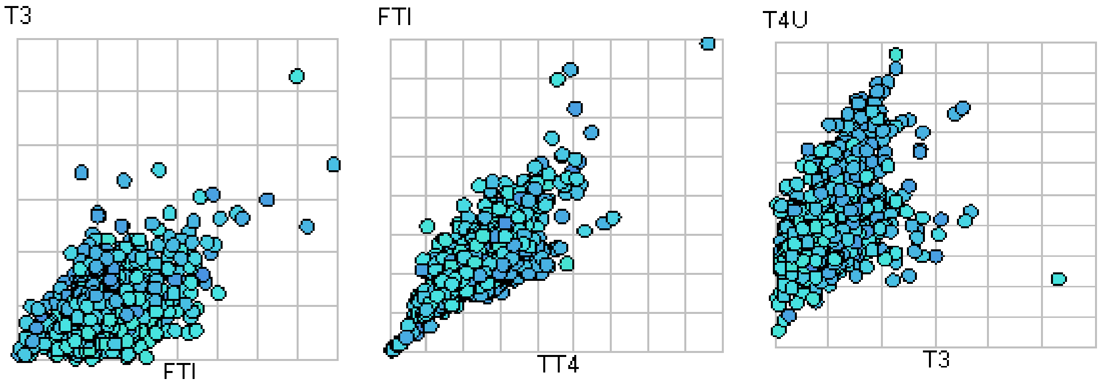


Figure 3: Identifying correlated attributes

Where the scatterplots differ, it highlights an attribute that will be good at distinguishing between the two classes. This boxplot illustrates a predictive attribute pregnant and TT4, other predictive values include **lithium and age, Pregnant and age** within the dataset**.**

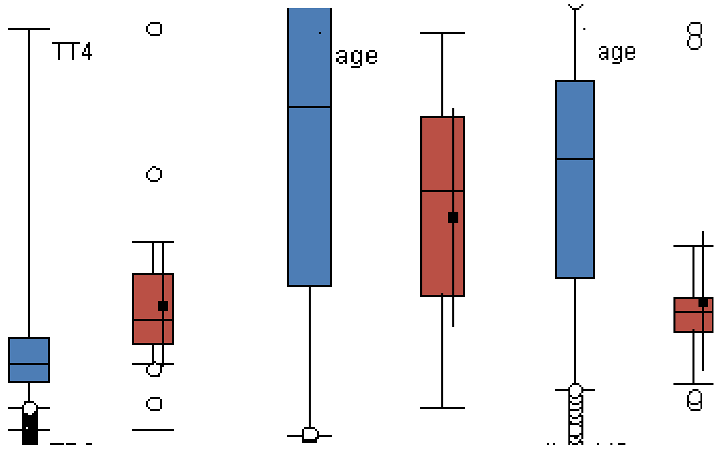


Figure 4: Using box plots to identify predictive attributes

The following shows attributes TT4 histogram. This is not a normal distribution and is skewed to the left. The said can be said for many attributes in this dataset such as age, gender Thyoxine, antitthyroid, sick, pregnant, thyroidSurgery, i131treatment hypothyroid, hyperthyroid, lithium, goitre, tumor, hypopituitary, psych TSHmeasured, TSH, T3, TT4 and FTI to name a few. It remains to be seen whether more data is needed to give a true representation of the domain.

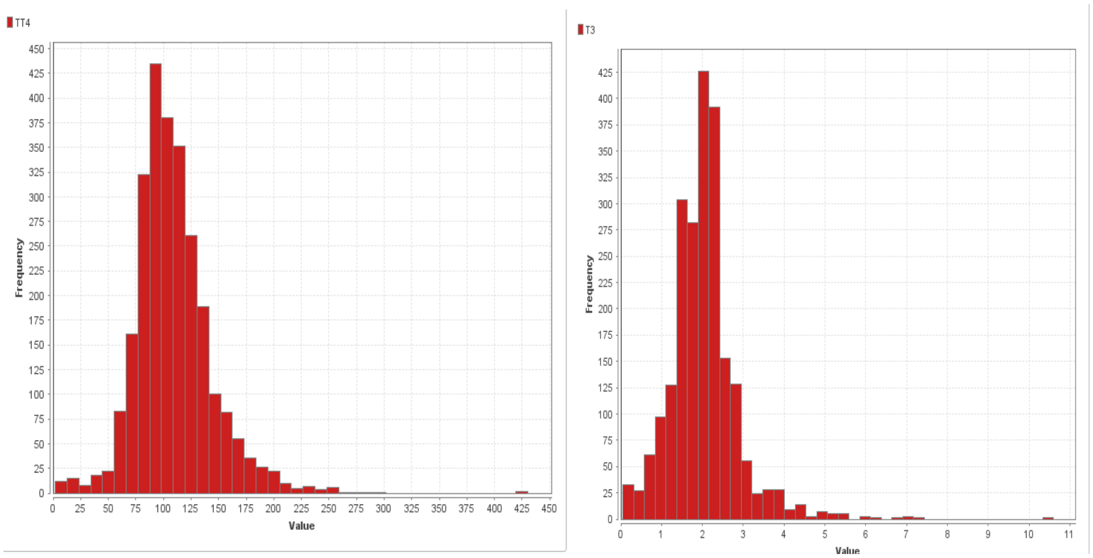


Figure 5: Using histogram to identify data distribution

2.3 Verify data quality

- There seems to be an error among the age attribute as there's a value of 450 present.
- 4556 missing values are present in the thyroid dataset.
- The above histogram diagram (figure 5) an outlier can be detected on TT4 and T3.
- Interesting that there are more than twice as many females in this dataset than there are men indicating bias.

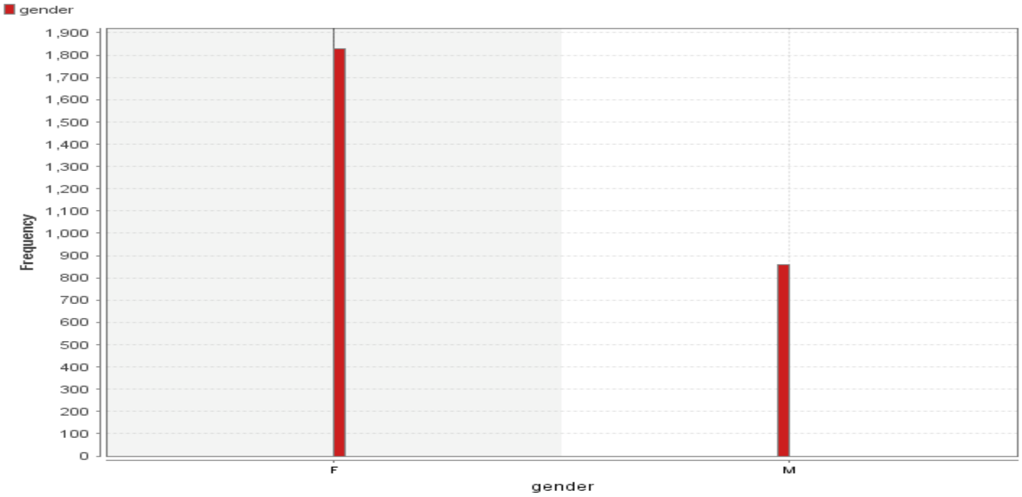


Figure 6: Using histogram to indicate biased

- No pollution detected.
- No missing labels
- No noise

There's more than enough attributes in predicating the class label in fact it remains to see if all of them are necessary. The presence of outliers will cause problems as it can indicate bad data (human error) also the algorithms that will run during data mining will perform better without the presents of outliers

3. Data Preparation.

By issuing a simple Cross Validation within our dataset, a baseline accuracy of 96.36% was discovered.

3.1 Select Data

3.1.1 Stratified Sampling.

When initially exploring the data, I identified that 30 columns (1 special attribute/29 regular) where present, because the number of columns didn't exceed over 50 there was little value in identifying columns that would not be useful at this stage.

However, the dataset contained too many rows of data (2800). It was decided that stratified sampling could be performed to maintain the same patterns in the data while reducing the number of rows in the data. By making use of the loop parameter operator we can determine the accuracy as the sample size increases. To record each iteration of this loop a log operator was used.

Before running the process, a number of different parameter settings where issued first was the Grid/Range for the loop operator. I set the min and max to 10 and 2800 because this was the range of the original data set and by setting the steps parameter to 20 it specified just the right number values within that range. The results of these parameters are illustrated in the diagram below. The optimal number of rows that could be deleted while maintaining the original patterns in the dataset was **2250**. It was important to consider not to overfitting/underfitting the data in any way. The accuracy of the overall model did not change as the nature of the sampling was to maintain the accuracy by reducing the dimensionality of the data. The objective of this data preparation technique was to improve the precision of the sample by reducing sampling error and to improve processing time and memory.

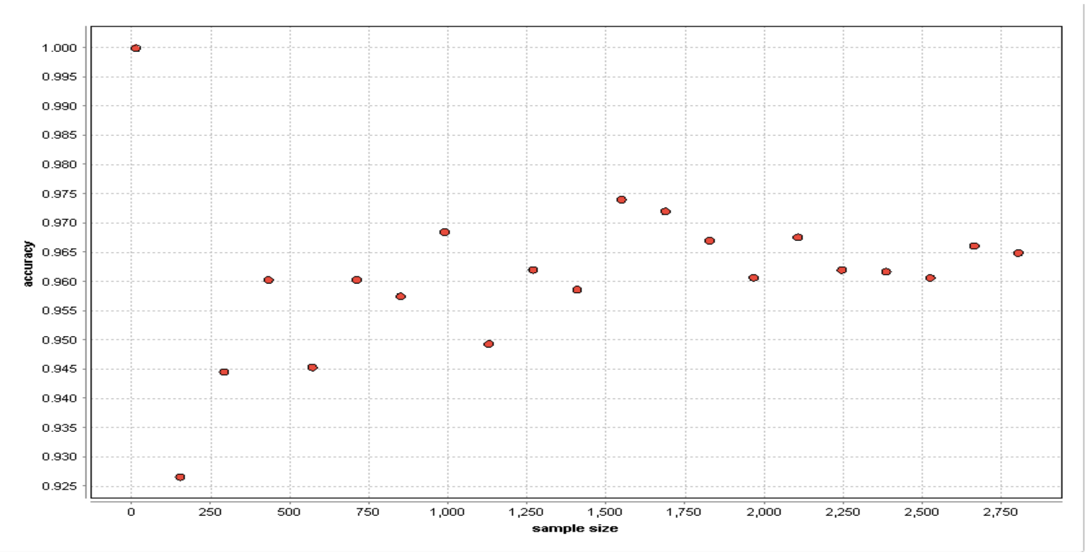


Figure 7: Scatterplot results of Stratified sampling

3.2 Clean Data

3.2.1 Missing values

After exploration of the data it was discovered that several attributes had no influence in predicting the class labels (Accuracy of the model did not change when attribute was deleted) so it was decided to remove it from the dataset by using the select attribute operator.

Attribute that did predict the class label
FTI
T3
TSH
TT4
Thyroxine
Age
Pregnant

referralSource
Tumor
Class
T4U

Before the mining starts it's important to handle the missing values in the dataset which was identified in data understanding as 4556.

As commonly accepted rule is that if a column of data has a high proportion of missing values (> 40%), it is removed from the data set. If a column has a low proportion of missing values (<5%), then just those rows that contain missing values can be removed, and any missing data between these two measures can be replaced. The following is a table identifies the attribute that need to be deleted, rows of data removed or fill in the missing values.

Replace missing values in attributes that did influence the model by using the average

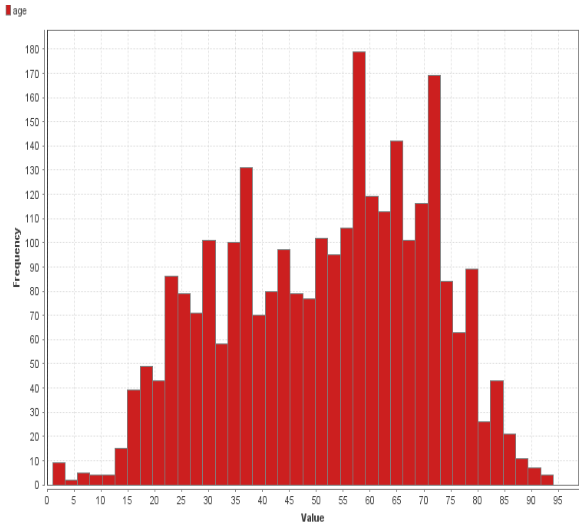
Attribute	Missing values
TSH	284
T3	585
TT4	184
T4U	297
FTI	295

When rerunning the mode again the accuracy increased to 96.74 an increase of 0.38 from the baseline.

3.3 Construct Data

3.3.1 Outlier Detection

During data understanding numerous outliers where detected, for example age had a single value in the range of 455 this was obviously an error, and heavily skewed the data. One outlier that was tried was detect outlier distance combined with filter examples was used to remove them. The same can be said for other attributes that showed outliers like TT4 and T3. By setting the parameter strings to outlier=true on the filter examples operator and by specify the number of outliers to find (10) we were able to filter out the outliers that where the furthest from their neighbours. The outlier operator that was implemented was local outlier factors (LOF). This is based on a concept of local density, where locality is given by the k nearest neighbour. In this case we set the parameter string as outlier>2.0 which eliminated outliers that where greater than this value. This method was ultimately better as it allowed us to choose which outliers we want to get rid of.



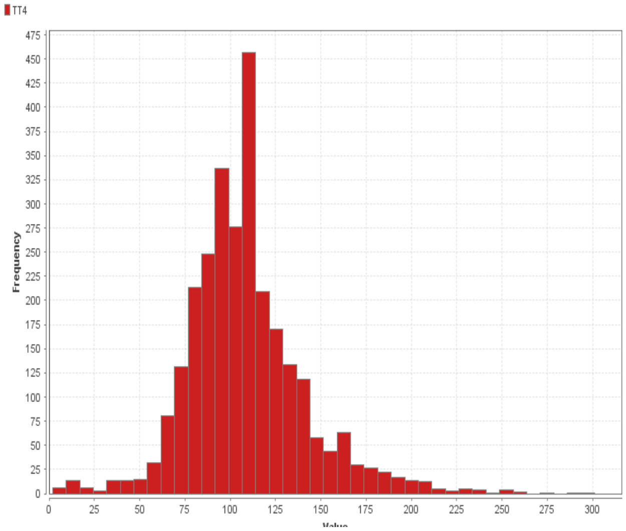


Figure 8: Results of Distance outlier detection

The results give us a much more even distribution of the data. It was important not to delete all outliers as in reality there is always outliers in a dataset, so it was important to only delete significant outliers. When rerunning the model an accuracy of 96.96 is gathered indicating an increase.

4. Modelling

4.1 Select modelling technique

Decision tree seems to be the most appropriate for this dataset as most of our data is categorical and the results are easily understood and visualized using a tree structure. The computation of the algorithm is inexpensive and because we deleted the irrelevant attributes training the model is much faster. Nearest Neighbour can be used as some of are attributes are numeric as well as categorical (by using distance measure) Good for datasets with a lot of variability, accuracy depends on the appropriate value for k.

4.2 Generate Test Design

Nominal Cross validation was the method used to spilt the dataset in to a training and test datasets. In the training phase, a model is built on the current training data set. The number of subsets and iterations the example set was divided in to were set to 15. When evaluating the results, a confusion matrix was used, it was important to get certain classes predicated correct for example it was crucial to predict decreased binding protein as this is the case where the patients are getting worst, so a balance between precision and recall was ultimately decided to be more important than the accuracy.

4.3 Build and Assess the model

4.3.1 Decision tree

By default, the model had pre-post pruning enabled, disabling this lead to the decision tree having way more sections added to it, to the point it was no longer human readable, and the accuracy of the model was less (96.76). It was later learned that pruning reduces the size of decision tree by removing sections that provide little power to classify instances therefore increasing predictive accuracy by the reduction of overfitting. Interesting how T4U is the root of the tree it remains to a domain expert to see if this is would be the case.

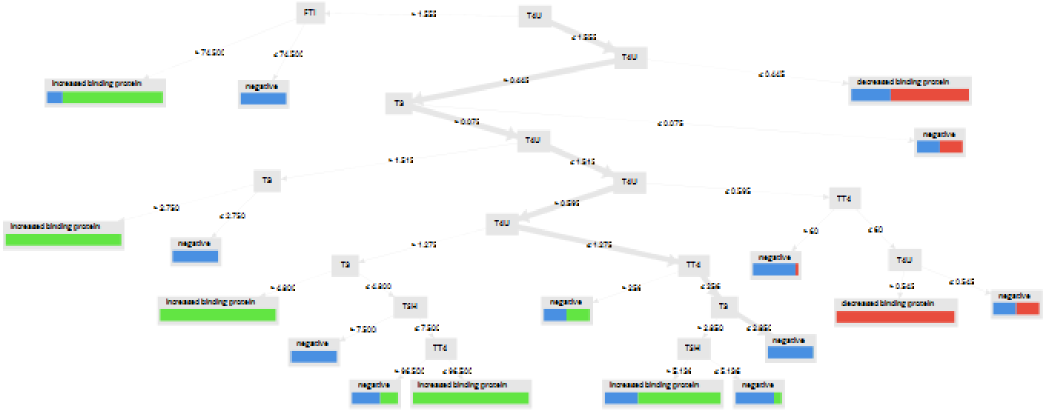
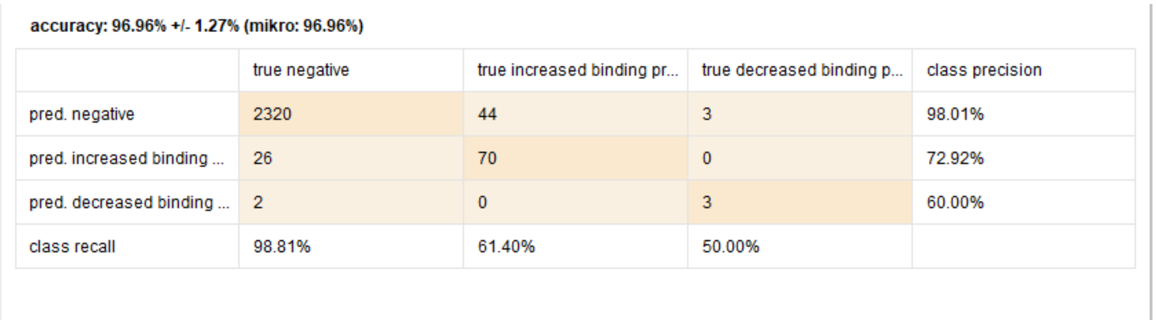


Figure 9: Decision tree after pruning

The second parameter that was configured was the information gain thresholds by changing this from 0.1 to 0.4 doing this the accuracy of the overall model went down to 95.14 and the decision tree only had a single root of negative. The best information gain value was 0.2 which had an accuracy of 97.20 a good balance between the recall and precision and a pretty good size decision tree. It was unsure if this was a good approach to take as according to rapid miner documentation information gain can be slightly bias in selecting attributes with a large number of values.



accuracy: 96.96% +/- 1.27% (mikro: 96.96%)

	true negative	true increased binding pr...	true decreased binding p...	class precision
pred. negative	2320	44	3	98.01%
pred. increased binding ...	26	70	0	72.92%
pred. decreased binding ...	2	0	3	60.00%
class recall	98.81%	61.40%	50.00%	

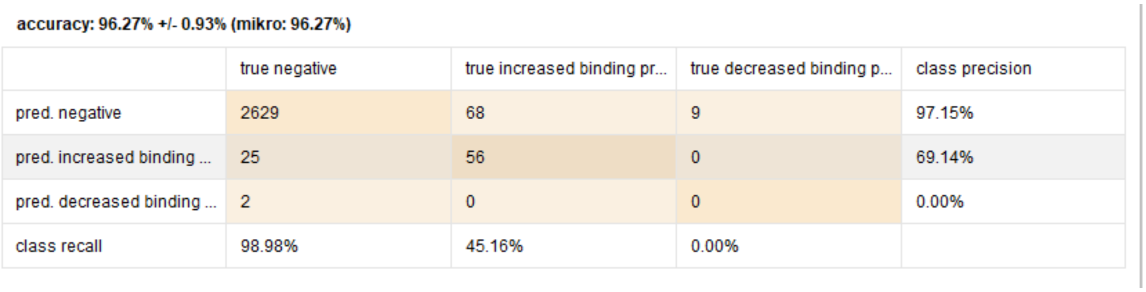
Figure 10: Optimized results of running the decision tree model

4.3.2 K nearest neighbour

When configuring the k operator, the only parameter that i modified was k which was original set to one with an accuracy of 95.34%. The best accuracy for k was 5 (96.27) but the best value for k that had a balance between recall and precision was 3

Value for K	**Accuracy**
K1	95.34
K2	96.02
K3	96.09

K4	96.20
K5	96.27
K6	96.20
K7	96.13
K8	96.20
K9	96.09
K10	96.06



accuracy: 96.27% +/- 0.93% (mikro: 96.27%)

	true negative	true increased binding pr...	true decreased binding p...	class precision
pred. negative	2629	68	9	97.15%
pred. increased binding ...	25	56	0	69.14%
pred. decreased binding ...	2	0	0	0.00%
class recall	98.98%	45.16%	0.00%	

Figure 11: Optimized results of running the K-NN model

A neural net operator was ran on this set but it required that all attribute be converted to numeric which resulted in an error. If the example set was all numeric a neural network would have been used.

5.Evaluation

In conclusion the better of the two algorithms that produced a better model was the decision tree s it gave an accuracy of 96.92% and had a way better balance when it came to recall and precision. In contrast KNN accuracy was only 96.27 and had little balance between class and precision.

In context of the business objective it's important to produce a model that can predict all three class with an acceptable degree of accuracy and this is true when using a decision tree. The only class that may need to be improved is decreased biding protein in which the model can only make a correct prediction 50% of the time but other then that the model works well with unseen data. The reason this algorithm worked better was it was more suited to nominal attributes which this dataset had plenty of

The learning objectives from this dataset has been immense from learning what each of the attribute meant to implementing a CRISP DM model from start to finish