The Effect of Vertical Reduction on Classification Algorithms

Report Prepared by [redacted]

For CP3300 - Knowledge Discovery and Data Mining at James Cook University, 2014

1. Thesis

1.1: Project Overview

We have been approached by a health care clinic that is concerned about the increasing number of heart related disease that have been evident within their catchment area. We aim to improve their current detection of illnesses related to the cardiovascular system. This will be achieved by analysis of attributes collect and creating results found by conducting knowledge mining algorithm testing.

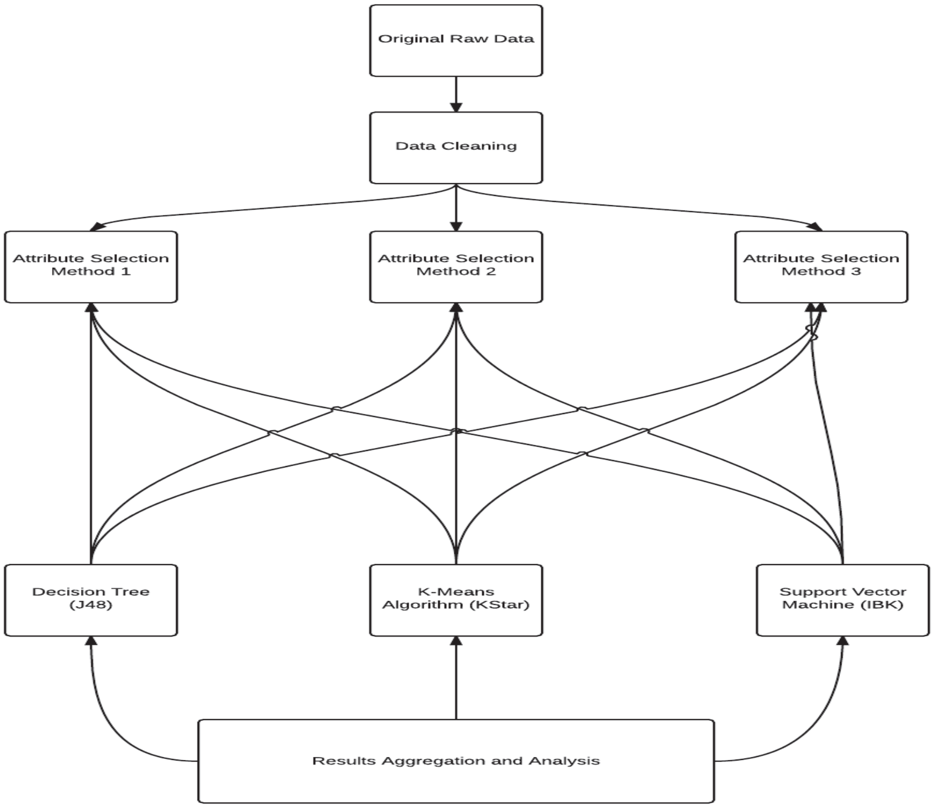

Original Raw Data
Data Cleaning
Attribute Selection Method 1
Attribute Selection Method 2
Attribute Selection Method 3
Decision Tree (J48)
K-Means Algorithm (KStar)
Support Vector Machine (IBK)
Results Aggregation and Analysis

Figure 1.0: Experiment Layout

The experiments are comprised of 6 variables where 3 variables consist of different attributes and 3 variables consist of different classification algorithms. This is known as vertical reduction and is used to help cut down on excessive processing time yet can also negatively impact the accuracy of the algorithms. To determine the optimal attribute and algorithm selections, a total of 9 outcomes will be produced by the experiments and conclusions can then be drawn from this data.

1.2: Project Objectives

Analyze the effect that attribute selection and algorithm selection has on the results with respect to efficiency and effectivity of the overall accuracy of the prediction of a class value.

1.3: Project Environmental Configuration

In order to construct an unbiased platform for the algorithms to perform, a uniform split between training and testing data must be established. In this case the training data is used to populate the algorithms with classification routes, and the testing data is tuples that are used to run through the algorithm to test the efficiency and accuracy metric.

In this case we have deemed it appropriate to select a 68% subset for training and 32% for testing. This was chosen due to the extremely stable results that were generated during initial testing stages. This was to solve the issue of overfitting, which is when a algorithm is over trained and as a result has a detrimental impact on its performance.

1.4: Acknowledgements

This data set was obtain from the **U**niversity of **C**alifornia (**UCI**) data repository. It is an amalgamation of data obtained from 4 different localities: Long Beach, Cleveland, Switzerland and Hungary. We would like to thank the respective health professionals for their efforts and diligence.

2. Data Set Description

2.1: Dataset Overview

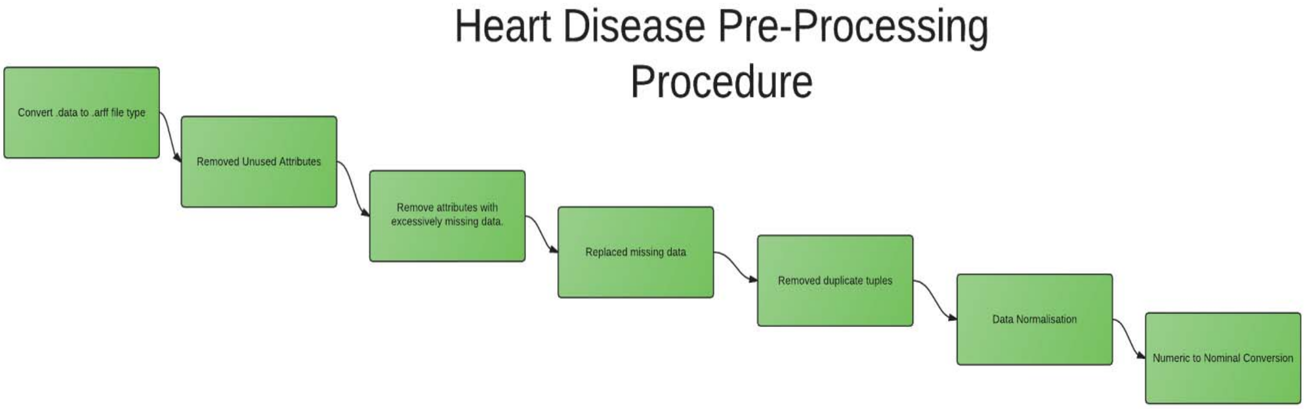
The Heart Disease Dataset is a collection of biometric parameters collected from a patient by a qualified health practitioner. For the purpose of this experiment, the data set used does not affect the overall results in any way as this experiment is focused solely on the algorithms and attributes used and thus what these actual attributes represent does not have any impact on the results that are produced from these experiments.

2.2: Gross Dataset Attributes

See Appendix 1

3. Pre-Experiment Preparation

3.1: Required Pre-Processing for Classification Algorithms


Heart Disease Pre-Processing Procedure
Convert .data to .arff file type
Removed Unused Attributes
Remove attributes with excessively missing data.
Replaced missing data
Removed duplicate tuples
Data Normalisation
Numeric to Nominal Conversion

3.1.1: File Conversion

As WEKA is only capable to deal with **A**ttribute **R**elation **F**ile-**F**ormat (**.arff**) and **C**omma **S**eparated **V**alues (.**csv**) files, the original .data file had to be converted to suit these file formats. A simple online Regular Expression (RegEx) matching system was used to quickly filter the .data file into the appropriate format to match a .arff file and was then imported into WEKA for data pre-processing.

3.1.2: Preprocessing

The data preprocessing included the approximation of missing data, data normalization, and discretization through the use of the algorithms provided by WEKA. As the raw data contains 76 attributes, these attributes then had to be split into different groups to allow the attribute selection methods to have maximum effectivity in determining the rate at which attribute selection impacts the results of certain algorithms.

Firstly, the original data set contained a large number of missing and incomplete data. This would have had a large detrimental factor on the outcomes of the experiment. In order to remedy this two approaches were applied to the dataset. Firstly, due to some external error there was some attributes that were missing a large amount of entries. This was compounded by the complete absence of any entry for attributes in a couple of cases. To resolve this issue attributes that had over 85% missing were deemed to be of little use and were removed from the data set. The secondary approach to resolving this issue was to approximate and replace with the “Replace Missing” algorithm within Weka.

The next issue that was discovered was the presence of outliers within the dataset. Outliers are defined as data records that skew the the overall representation of the data away from a normalised bell-curve. Data mining algorithms are impacted by outliers, some more than others. Therefore, in order to produce an unbiased and consistent baseline it was important to minimise their impact. In order to remedy this problem the Normalisation algorithm was employed within the dataset. This allowed for the “Smoothing” of data, therefore leading to reduction in negative outcomes produced by outliers.

Finally, in order to maximise the results that were produced by the data mining algorithms the data was discretised. Discretization is a process in which continuous data is partitioned into groups based on the density of values around it. This effectively turns the data into discrete. This was done due to the fact that many machine learning algorithms, particularly classification algorithms are capable of producing better models when presented with discrete data, then with continuous.[1]

1 - Kotsiantis, S.; Kanellopoulos, D (2006). "Discretization Techniques: A recent survey". *GESTS International Transactions on Computer Science and Engineering* **32** (1): 47–58.

3.1.3: Attribute Selection Methods

To determine the effect of vertical selection on classification algorithms, the dataset attributes had to be split into certain monitored groups and each group then had to be tested with multiple algorithms to determine what effect, if any, vertical selection has on classification algorithms. The attribute selections were divided into three (3) different groups through educated guesses which were also influenced by the data set description provided by the UCI repository. The attribute groupings were as follows:

- All attributes (76 attributes)
- "*The ideal 14*" which were used by the UCI repository and were seen to have the most impact in determining if a patient had heart disease.
- The necessary attributes as chosen by the WEKA algorithms (the "*idea 14*" dataset is run through WEKA's "CfsSubsetEval" algorithm and the remaining attributes are then used for the grouping)

3.1.5: Algorithm Selection

As the dataset being dealt with is a medical dataset, a classification algorithm must be used to classify entries based on a class value. The algorithms selected are to be used to benchmark and compare the effectiveness and efficiency of vertical reduction on classification algorithms. The algorithms selected are as follows:

- K-nearest neighbour
- **S**upport **V**ector **M**achine (**SVM**)
- J48 Decision Tree

3.5: K-Nearest Neighbour Algorithm

3.5.1: Algorithm Description

The **K**-**N**earest **N**eighbour (**KNN**) algorithm is a non-parametric method in which the input consists of the k closest training examples in the future space; and when the K-Nearest Neighbour algorithm is used for classification purposes, the output is a class membership. A new object is classified by a majority vote of it's neighbours, with the object being assigned to the class most common among its k nearest neighbours.

3.6: Support Vector Machine Algorithm

3.6.1: Algorithm Description

Support **V**ector **M**achines (**SVM**) are supervised learning models with associated learning algorithms that are used to analyze data and recognize patterns. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on.

3.7: Decision Tree Algorithm

3.6.1: Algorithm Description

Decision Tree is a predictive strategy that applies weight to each attributes value in a dataset. This allows for these weights to be utilized to construct a "leaf", a representation of a attribute value. This leaf is then connect to other leaf nodes by a learnt pattern that is constructed by the analysis of the relationship that is held with other attributes. The end goal of this analysis is to determine what values from children nodes equate to the probability of a connection with another node.

4. Attribute Selection Method Groupings

4.1: Gross Data Set

See appendix 1.

4.2: Recommended Data Set Attributes

See appendix 2.

4.3: Weka Optimized Data Set

See appendix 3.

5. Discussion Preface

Classification algorithms aim to predict the outcome of a class values, based on collected of historical observations. This allows the algorithm to conduct machine learning to construct a logical method of assessing newly collected data. This allows for data collected in the future to be compared against the rules developed from learned patterns.

The aim of this report is to determine the effect that vertical selection plays in the efficiency and accuracy of classification algorithms. This experiment was conducted on a dataset consisting of an original 78 attributes. This number could be considered to be an over representation of the attributes required to correctly determine the presence or absence of heart disease in a patient.

Herein lies the premise of the experiment: Does attribute count affect the results of classification algorithms? What problems can be encountered and what countermeasures can be employed by a data miner to mitigate these risk?

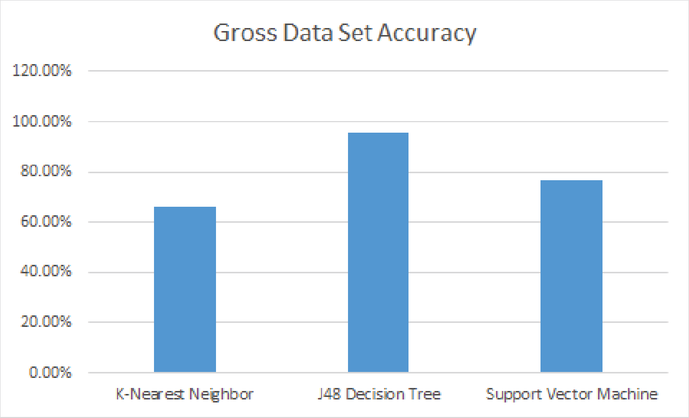
There are two metrics that will be be analysed to form the basis of the critical analysis of the algorithms: Accuracy and Efficiency. Accuracy will be determined by the percentage of correctly identified tuples after processing by the algorithm. Efficiency on the other hand can be determined by the actual amount of computational time required to retrieve the result.

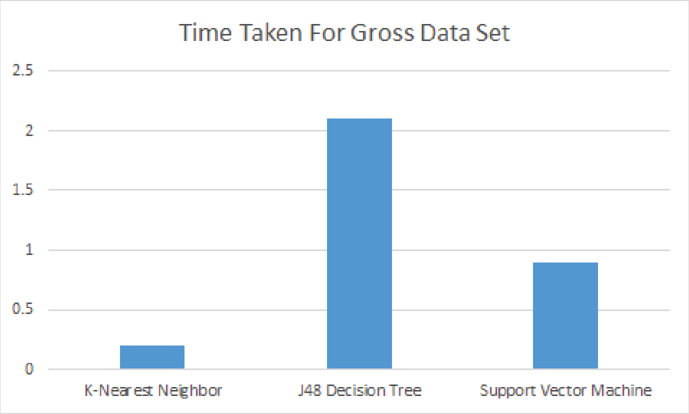
For this test it has been deemed appropriate to elect three different varieties of classification algorithms. These included: **S**upport **V**ector **M**achine (**SVM**), Decision Tree and K-Nearest Neighbor Lazy Learning algorithm. This allows for an inclusion of three of the most prominent and well used algorithms. The reasoning for this was that, although many novice data miners are aware of the impact that vertical weight induce, it is an interesting premise to test practically.

5.1 Gross Data Set Results

The gross dataset provides a relatively decisive result based on the calculated output from the various machine learning algorithms.

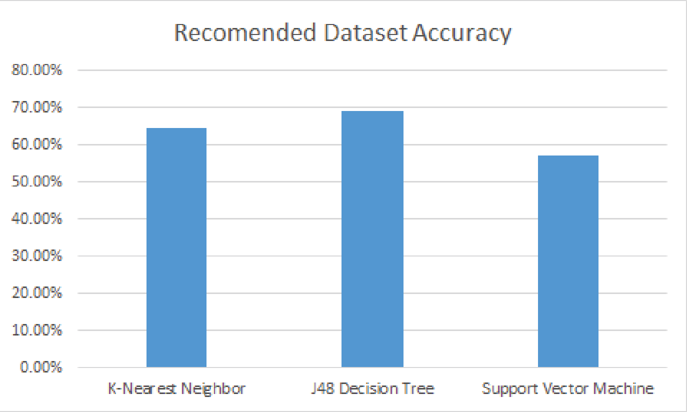
Firstly, it becomes clear that on this data set the J48 decision tree algorithm provides the most accurate result, yielding an approximate 95% accuracy. In comparison, both the K-Nearest Neighbor and the Support Vector Machine algorithms produced similar results, with a yield of 66% and 76% respectively.



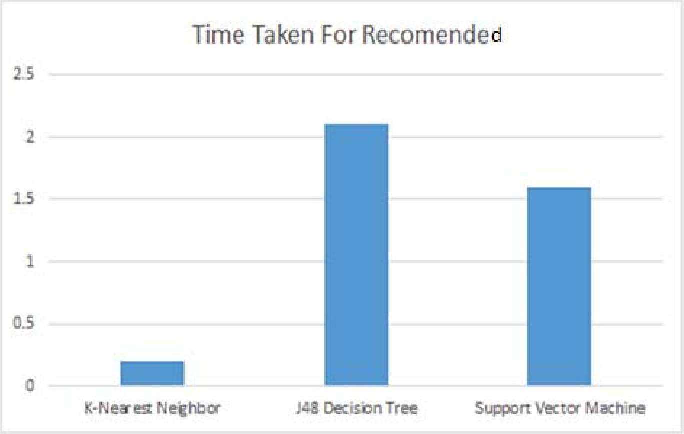


However, the high accuracy of the J48 algorithm comes at the cost of computational time. In this case the J48 took approximately three fold the time take by the SVM algorithm and more than twenty times longer than the K-Nearest Neighbor Algorithm.

5.2 Recommended Data Set Results

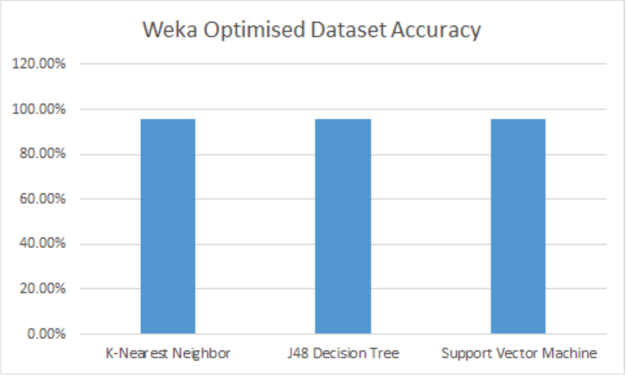


The accuracy of the selected algorithms for this data set support the results gained from the previous experiment. The J48 algorithm produced the most accurate results again, however the actual percentage of correctly classified observations was at a much lower percentage that that was produced from the first experiment.



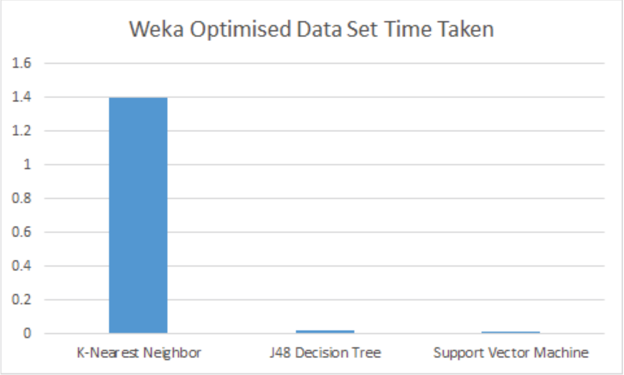
In this experiment the time taken for both the SVM and K-Nearest Neighbor has been inverted in comparison with the results from the previous test. In this test the SVM algorithm produced its result within 0.2 seconds, where the K-Nearest Neighbor took more than 1.5 seconds. In addition it should be noted that the J48 Algorithm still took the longest to produce results at approximately 2 seconds. This is inline with the results that were gained in the previous testing.

5.3 Weka Optimised Data Set Results



This experiment produced the most interesting results of all conducted experiments. Firstly, the optimised dataset produced precisely the same accuracy, regardless of selected algorithm.

Another notable anomaly that was discovered during this experimentation process was that both the J48 Decision Tree and Support Vector machine algorithms produced results that were approaching zero seconds (0.02 seconds), however the K-Nearest Neighbor produced results in an extremely contrasting elapse of time.



5.4 Overall Analysis of Experiment

It has become apparent from the experiments conducted that vertical selection has a strong impact on not only the accuracy of the results, but also the time taken to produce a machine learned on some algorithms. The extent to which this impact varies between each algorithm.

Firstly, in tests that were not conducted on machine reduced attributes the relationship between the attributes strongly affected the outcome of the experimentation. This conclusion was drawn based on the output from the comparison between the first and second experiments.

In the results for the first experiment, both the J48 Decision Tree and the Support Vector Machine produced accurate results (95% and 79% respectively) . In this experiment the dataset consisted of all 76 attributes that were provided. However, when the same algorithms process the dataset that consisted of less attributes the accuracy fell. In the second experiment, the accuracy of both the J48 Decision Tree and Support Vector Machine dropped by approximately 20%.

This drop is considered to be an extremely interesting finding. At this point it is important to reiterate the exact circumstances that the attributes for this data set were selected from reduction. The original dataset provided an outline for what attributes were considered by medical professionals to be the main indicators in diagnosis of heart disease within a real-life setting. However, in this dataset the information provided by these attributes alone affected the accuracy of the classification algorithm negatively. This is an indication that the recorded values for the tuples is too loosely related in order for the classification algorithm to form associations.

On analysis, the finding indicates to the researchers that the actual correlation of the data impacts strongly on how well the classification algorithm performs. This extends beyond the practical semantics of the data and is conducive on how strongly the data is represented within the dataset.

In regards to the J48 decision tree it has become apparent that it is overall one of the most accurate machine learning algorithms when applied to this dataset. Consistently, it produced the most accurate results. However, in all tests apart from the optimised set it also was the most inefficient algorithm in regards to computational time. This could be a result of an actual weakness within the algorithm, or due to other external factors. These factors could include programmatic errors within the algorithms programming code such as redundancies or memory leaks.

The experiment on attributes reduced via the CFS Subset Evaluation Algorithm produced extremely noteworthy results. This consisted of not only the results from the accuracy prospective but also the efficiency. This experiment was designed to see what impact that intelligent application of vertical reduction algorithms have, as opposed to the use of manual remove.

Firstly, in regards to the accuracy; there was a consistent 95.5882% accuracy for all algorithms. This is an extremely interesting observation, as although all algorithms aim to produce a classification rule between the tuples, the computational approach varies between each specific algorithm. However, from the researchers rudimentary knowledge it is difficult to draw a concrete reason for this anomaly, it can be suggested that this is a result of the pruning procedure that the vertical reduction algorithm employs.

The efficiency of the final experiment also produced an interesting anomaly within the testing results. For both the J48 Decision Tree and Support Vector Machine the results were produced from the algorithm in time elapses that approached zero, recording a response time of 0.02 seconds for each algorithm. In contrast, the K-Nearest Neighbor Algorithm produced computational output in a comparatively lengthy time, at approximately 1.4 seconds. From the external research conducted, no further concrete conclusion can be drawn as to the root cause of this anomaly. However, it can be inferred from the experience of the research group that this was the result of a bias placed on the data be the CFS Subset Evaluation Algorithm.

It has also been noted that the K-Nearest Neighbor, although has not produced the most accurate results, performed consistently within all tests. This anomaly was detected during the testing phase and was isolated as an interesting finding. Upon further investigation, it was proved to be inline with the general expectations of the algorithm. Historically[1], K-Nearest Neighbor has proven to be an extremely consistently performing algorithm in regards to accuracy. However, generally has a lower overall accuracy in comparison to other algorithms. From this, a conclusion has been drawn that indicates that K-Nearest Neighbor is not as affected by vertical reduction to the extent that was experienced by both SVM and J48.

This experiment also produced a valuable introduction into the impact that noisy or loosely related data has on the overall viability of any data mining experiment. The concept of "Garbage-in, Garbage-out" is introduced early in any Data Mining textbook, however in order for this to be fully quantified and appreciated it is extremely useful to experience a real-life example.

From this conclusion, it can be drawn from this experiment that careful consideration must be made to attribute selection for future data mining projects in which classification algorithms are employed. Not only should attributes be carefully tested in different configuration, testing should be conducted on multiple vertical reduction algorithms to determine what attributes provide the best result.

1 - Mills, Peter. "Efficient statistical classification of satellite measurements". *International Journal of Remote Sensing*.

6. Appendix

Appendix 1: Gross Data Set Attributes

No#	Key	Description
1	id	patient identification number
2	ccf	social security number (I replaced this with a dummy value of 0)
3	age	age in years
4	sex	sex (1 = male; 0 = female)
5	painloc	chest pain location (1 = substernal; 0 = otherwise)
6	painexer	(1 = provoked by exertion; 0 = otherwise)
7	relrest	(1 = relieved after rest; 0 = otherwise)
8	pncaden	(sum of 5, 6, and 7)
9	cp	chest pain type
10	trestbps	resting blood pressure (in mm Hg on admission to the hospital)
11	htn	Hypertension
12	chol	serum cholesterol in mg/dl
13	smoke	I believe this is 1 = yes; 0 = no (is or is not a smoker)
14	cigs	cigarettes per day
15	years	(number of years as a smoker)
16	fbs	(fasting blood sugar > 120 mg/dl) (1 = true; 0 = false)
17	dm	(1 = history of diabetes; 0 = no such history)

18	famhist	family history of coronary artery disease (1 = yes; 0 = no)
19	restecg	resting electrocardiographic results
20	ekgmo	(month of exercise ECG reading)
21	ekgday	(day of exercise ECG reading)
22	ekgyr	(year of exercise ECG reading)
23	dig	(digitalis used during exercise ECG: 1 = yes; 0 = no)
24	prop	(Beta blocker used during exercise ECG: 1 = yes; 0 = no)
25	nitr	(nitrates used during exercise ECG: 1 = yes; 0 = no)
26	pro	(calcium channel blocker used during exercise ECG: 1 = yes; 0 = no)
27	diuretic	(diuretic used used during exercise ECG: 1 = yes; 0 = no)
28	proto	exercise protocol
29	thaldur	duration of exercise test in minutes
30	thaltime	time when ST measure depression was noted
31	met	mets achieved
32	thalach	maximum heart rate achieved
33	thalrest	resting heart rate
34	tpeakbps	peak exercise blood pressure (first of 2 parts)
35	tpeakbpd	peak exercise blood pressure (second of 2 parts)
36	dummy	Not Supplied
37	trestbpd	resting blood pressure

38	exang	exercise induced angina (1 = yes; 0 = no)
39	xhypo	(1 = yes; 0 = no)
40	oldpeak	ST depression induced by exercise relative to rest
41	slope	the slope of the peak exercise ST segment
42	rldv5	height at rest
43	rldv5e	height at peak exercise
44	ca	number of major vessels (0-3) colored by flourosopy
45	restckm	Not Supplied
46	exerckm	irrelevant
47	restef	rest raidonuclid (sp?) ejection fraction
48	restwm	rest wall (sp?) motion abnormality
49	exeref	exercise radinalid (sp?) ejection fraction
50	exerwm	exercise wall (sp?) motion
51	thal	3 = normal; 6 = fixed defect; 7 = reversible defect
52	thalsev	not used
53	thalpul	not used
54	earlobe	not used
55	cmo	month of cardiac cath (sp?) (perhaps "call")
56	cday	day of cardiac cath (sp?)
57	cyr	year of cardiac cath (sp?)
58	num	diagnosis of heart disease (angiographic disease status)

59	lmt	Not Supplied
60	ladprox	Not Supplied
61	laddist	Not Supplied
62	diag	Not Supplied
63	cxmain	Not Supplied
64	ramus	Not Supplied
65	om1	Not Supplied
66	om2	Not Supplied
67	rcaprox	Not Supplied
68	rcadist	Not Supplied
69	lvx1	not used
70	lvx2	not used
71	lvx3	not used
72	lvx4	not used
73	lvf	not used
74	cathef	not used
75	junk	not used
76	name	last name of patient

Appendix 2: Recommended Data Set Attributes

Key	Description
Age	*Age in years*
Sex	*Sex (1 = male; 0 = female)*
CP	*Chest Pain Type* *-- Value 1: typical angina* *-- Value 2: atypical angina* *-- Value 3: non-anginal pain* *-- Value 4: asymptomatic*
trestbps	*Resting Blood Pressure (in mm Hg on admission to the hospital)*
chol	*Serum cholesterol in mg/dl*
fbs	*Fasting blood sugar > 120 mg/dl (1 = true; 0 = false)*
restecg	*Resting Electrocardiographic Results* *-- Value 0: normal* *-- Value 1: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV)* *-- Value 2: showing probable or definite left ventricular hypertrophy by Estes' criteria*
thalach	*maximum heart rate achieved*
exang	*Exercise Induced Angina (1 = yes; 0 = no)*
oldpeak	*ST depression induced by exercise relative to rest*
slope	*the slope of the peak exercise ST segment* *-- Value 1: upsloping* *-- Value 2: flat* *-- Value 3: downsloping*
thal	*3 = normal; 6 = fixed defect; 7 = reversible defect*
num	*diagnosis of heart disease (angiographic disease status)* *-- Value 0: < 50% diameter narrowing* *-- Value 1: > 50% diameter narrowing*

Appendix 3: Weka Optimized Data Set

No#	Key
1	painexer
2	lmt
3	ladprox
4	laddist
5	cxmain
6	om1
7	rcaprox
8	rcadist