International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2015): 6.391

Towards Effective Data Preprocessing for Classification Using WEKA

Kariuki Paul Wahome[1], Wekesa Bongo[2], Dr. Rimiru Richard Maina[3]

[1, 2, 3]Jomo Kenyatta University of Agriculture & Technology /Department of Computing, Nairobi, Kenya

Abstract: ***Trend statistics through countless studies depict that there is an exponential growth of data form terabytes to petabytes and beyond in the world. This reality brings into perspective the apparent need for data mining which is the process of discovering previously unknown facts and patterns. Increasingly, data mining is gaining popularity due to the need by organizations to acquire useful information and develop hypothesis from the massive data sets they have in their data centers. Preprocessing comes in handy in the KDD process since it serves as the first stage while classification is the most common data mining task. This paper uses WEKA data mining tool which facilitates various data mining tasks through different algorithms to put into a kaleidoscope the importance of data preprocessing and the task of classification. Special focus is given to the procedure and results obtained after carrying out the two processes on WEKA.***

Keywords: Data Preprocessing, Classification, Data Mining, WEKA

1. Introduction

Today, there is a lot of data being collected and warehoused ranging from web data, ERP reports, electronic commerce sales and purchases, remote sensors at different locations, credit card transactions, multimedia data, scientific simulations, bioinformatics and so much more. Indeed, "we are drowning in data but starving for knowledge yet organizations have made huge investestments in data centers and other technologies but their Return on Investment (ROI) is not as expected". This can be attributed to factors like the exponential decline in the cost of computers, laptops and other portable devices like tablets, iPad and smartphones which generate a considerably large amount of data. Provision of cheap, fast and readily available bandwidth; as well as the continuous act of striving to bridge the digital divide gap by provision of technology enabling factors like electricity and literacy to mention but a few.

Due to the high data dimensionality, enormity, heterogeneous and distributed states of current data, traditional data mining techniques and pure statistics alone cannot handle this chunks of data. There is need to use and embrace modern automated techniques like WEKA which are a convolution of statistical, mathematical, machine learning and modelling techniques.

Waikato Environment for Knowledge Analysis (WEKA) is an open source data mining tool developed by university of Waikato in New Zealand for data mining education and research. WEKA is developed in JAVA and it has many advantages over other data mining tools. Key among them is that it is open source and available under the GNU license, it is a light program with a straight forward GUI interface and it is highly portable. It supports tasks like preprocessing of data, selection of attributes, classification, clustering, visualization and many other knowledge discovery techniques.

Generally, there exist more than 100 machine learning algorithms for classification, 75 for data preprocessing, 25 for feature selection and 20 for clustering and association rule mining. In this paper, the Iris data set from UCI data sets will be used to demonstrate different activities on WEKA tool in the KDD process as show below.

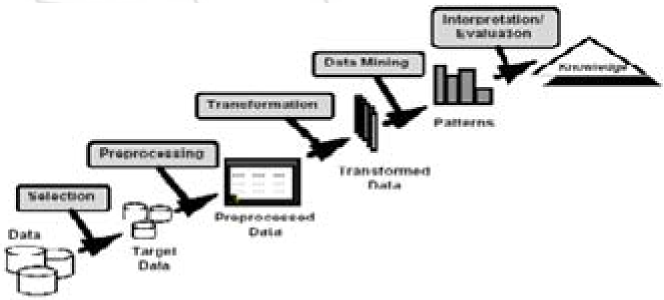


Figure 1: KDD Process

2. Data Preprocessing

Why is data preprocessing important data? Ideally, data in the real world is of low quality and hence referred to as "dirty". This is because it contains noise and outliers, inconsistencies or it is incomplete.

Presence of Noise and Outliers: Noise is a data quality problem which signifies presence of incorrect values or modification of a signal during or after transmission. Outliers on the other hand are data observation points which lie outside the overall distribution patterns. This means that they poses discriminant characteristics which differ from the other objects in the data set, (Moore and McCabe 1999).

Presence of Inconsistencies: Inconsistencies are discrepancies about a certain data object in the same data set. The most common types of inconsistencies range from redundancies, presence of duplicates and naming problems.

Incomplete Data: This is the most common data quality problem handled during preprocessing. Reasons for this problem are majorly due to some attributes not being applicable to all cases as well as values for some attributes not being collected. To solve these problems, the following data preprocessing tasks are undertaken; Data Cleaning, Integration, Reduction and Transformation.

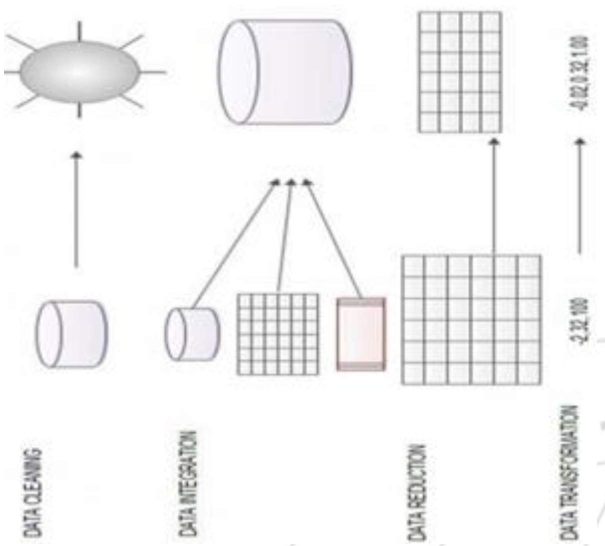


Figure 2: Data Preprocessing Tasks

Data Preprocessing eliminates unnecessary records and fills in missing gaps. Therefore, it is prudent to employ a multi-dimensional data quality assessment techniqueafter preprocessing. This can be done by measuring the accuracy, completeness, consistency, timeliness, believability, interpretability of the data (Swasti et al, 2013).

In a nutshell, data preprocessing seeks to solve data quality issues by answering the following questions

1) Does the data have any quality problems?
2) What methodologies can be used to detect problems associated with the data?
3) Which solutions can be applied on the present problems?

3. Preprocessing in WEKA

Among the four data preprocessing tasks i.e. data cleaning, data integration, data reduction, data transformation and data, the first can be comfortably handled in WEKA.

Data Cleaning: Cleaning is the process of filling in missing values, smoothing noisy data, identifying or removing outliers and resolving any inconsistencies present in the data.

Data Integration: this is the process of assimilating data from different sources like databases, files or data cubes in warehouse. The challenge of integration is probability of redundancy and integration of different schemas.

Data Transformation: Transformation in data preprocessing refers to converting data from one data format to another. Methods like smoothing, aggregation, normalization and generalization can be used to transform data.

WEKA has many inbuilt filters which are categorized into two; supervised and unsupervised filters. In both cases, WEKA provides different filters for attributes and instances.

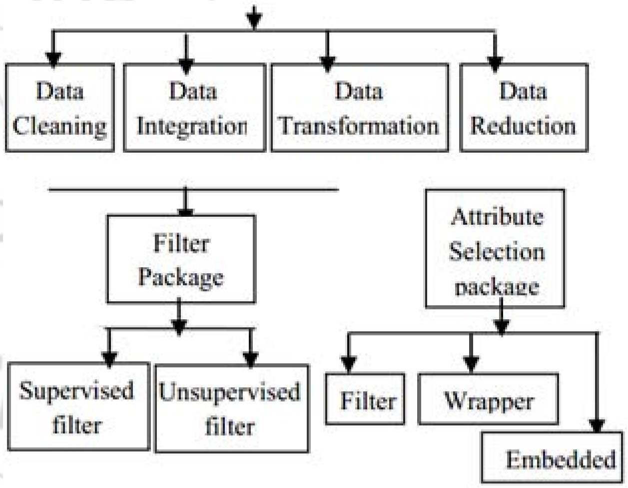


Figure 3: Data Preprocessing Filters (Shweta, 2014)

The Chronic Kidney Disease (CKD) data set from UCI data set will be used to carry out the experiments in WEKA.

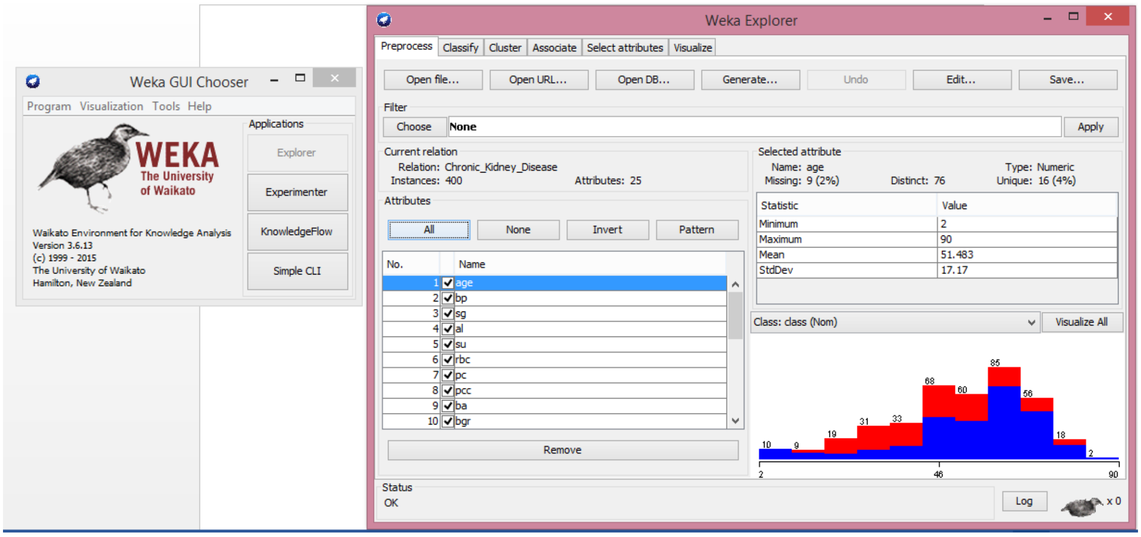


Figure 4: WEKA User Interface

Procedure for data preprocessing in WEKA
1) Choose Explorer on the WEKA GUI Chooser
2) Click open file to load the data set
3) In the explorer, choose a filter (supervised or unsupervised)
4) Select all the attributes and click apply
5) Analyze the effect of preprocessing on the data.

The Chronic Kidney Disease (CKD) data set is collection of attributes which can be used to determine whether somebody has CKD or not. Preprocessing the data show that there are some attributes which have missing values as shown below.

Viewer

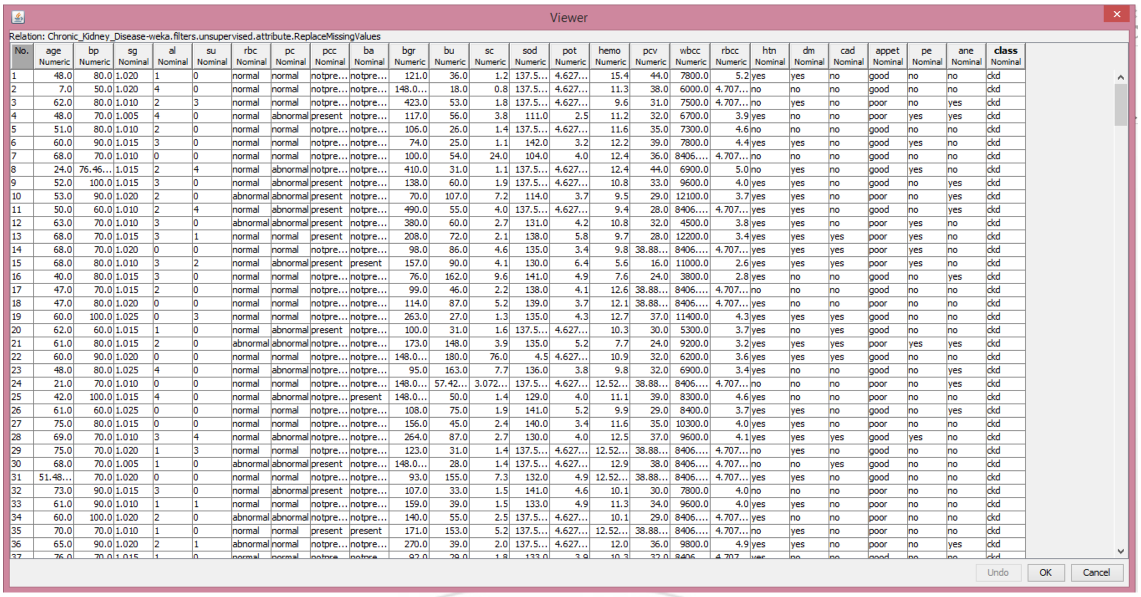
Relation: Chronic_Kidney_Disease

No.	age Numeric	bp Numeric	sg Nominal	al Nominal	su Nominal	rbc Nominal	pc Nominal	pcc Nominal	ba Nominal	bgr Numeric	bu Numeric	sc Numeric	sod Numeric	pot Numeric	hemo Numeric	pcv Numeric	wbcc Numeric	rbcc Numeric	htn Nominal	dm Nominal	cad Nominal	appet Nominal	pe Nominal	ane Nominal	class Nominal
1	48.0	80.0	1.020	1	0		normal	notpre...	notpre...	121.0	36.0	1.2			15.4	44.0	7800.0	5.2	yes	yes	no	good	no	no	ckd
2	7.0	50.0	1.020	4	0		normal	notpre...	notpre...		18.0	0.8			11.3	38.0	6000.0		no	no	no	good	no	no	ckd
3	62.0	80.0	1.010	2	3	normal	normal	notpre...	notpre...	423.0	53.0	1.8			9.6	31.0	7500.0		no	yes	no	poor	no	yes	ckd
4	48.0	70.0	1.005	4	0	normal	abnormal	present	notpre...	117.0	56.0	3.8	111.0	2.5	11.2	32.0	6700.0	3.9	yes	no	no	poor	yes	yes	ckd
5	51.0	80.0	1.010	2	0	normal	normal	notpre...	notpre...	106.0	26.0	1.4			11.6	35.0	7300.0	4.6	no	no	no	good	no	no	ckd
6	60.0	90.0	1.015	3	0			notpre...	notpre...	74.0	25.0	1.1	142.0	3.2	12.2	39.0	7800.0	4.4	yes	yes	no	good	yes	no	ckd
7	68.0	70.0	1.010	0	0		normal	notpre...	notpre...	100.0	54.0	24.0	104.0	4.0	12.4	36.0			no	no	no	good	no	no	ckd
8	24.0		1.015	2	4	normal	abnormal	notpre...	notpre...	410.0	31.0	1.1			12.4	44.0	6900.0	5.0	no	yes	no	good	yes	no	ckd
9	52.0	100.0	1.015	3	0	normal	abnormal	present	notpre...	138.0	60.0	1.9			10.8	33.0	9600.0	4.0	yes	yes	no	good	no	yes	ckd
10	53.0	90.0	1.020	2	0	abnormal	abnormal	present	notpre...	70.0	107.0	7.2	114.0	3.7	9.5	29.0	12100.0	3.7	yes	yes	no	poor	no	yes	ckd
11	50.0	60.0	1.010	2	4		abnormal	present	notpre...	490.0	55.0	4.0			9.4	28.0			yes	yes	no	good	no	yes	ckd
12	63.0	70.0	1.010	3	0	abnormal	abnormal	present	notpre...	380.0	60.0	2.7	131.0	4.2	10.8	32.0	4500.0	3.8	yes	yes	no	poor	yes	no	ckd
13	68.0	70.0	1.015	3	1		normal	present	notpre...	208.0	72.0	2.1	138.0	5.8	9.7	28.0	12200.0	3.4	yes	yes	yes	poor	yes	no	ckd
14	68.0	70.0						notpre...	notpre...	98.0	86.0	4.6	135.0	3.4	9.8				yes	yes	yes	poor	yes	no	ckd
15	68.0	80.0	1.010	3	2	normal	abnormal	present	present	157.0	90.0	4.1	130.0	6.4	5.6	16.0	11000.0	2.6	yes	yes	yes	poor	yes	no	ckd
16	40.0	80.0	1.015	3	0		normal	notpre...	notpre...	76.0	162.0	9.6	141.0	4.9	7.6	24.0	3800.0	2.8	yes	no	no	good	no	yes	ckd
17	47.0	70.0	1.015	2	0		normal	notpre...	notpre...	99.0	46.0	2.2	138.0	4.1	12.6				no	no	no	good	no	no	ckd
18	47.0	80.0						notpre...	notpre...	114.0	87.0	5.2	139.0	3.7	12.1				yes	no	no	poor	no	no	ckd
19	60.0	100.0	1.025	0	3		normal	notpre...	notpre...	263.0	27.0	1.3	135.0	4.3	12.7	37.0	11400.0	4.3	yes	yes	yes	good	no	no	ckd
20	62.0	60.0	1.015	1	0		abnormal	present	notpre...	100.0	31.0	1.6			10.3	30.0	5300.0	3.7	yes	no	yes	good	no	no	ckd
21	61.0	80.0	1.015	2	0	abnormal	abnormal	notpre...	notpre...	173.0	148.0	3.9	135.0	5.2	7.7	24.0	9200.0	3.2	yes	yes	yes	poor	yes	yes	ckd
22	60.0	90.0						notpre...	notpre...		180.0	76.0	4.5		10.9	32.0	6200.0	3.6	yes	yes	yes	good	no	no	ckd
23	48.0	80.0	1.025	4	0	normal	abnormal	notpre...	notpre...	95.0	163.0	7.7	136.0	3.8	9.8	32.0	6900.0	3.4	yes	no	no	good	no	yes	ckd
24	21.0	70.0	1.010	0	0		normal	notpre...	notpre...										no	no	no	poor	no	yes	ckd
25	42.0	100.0	1.015	4	0	normal	abnormal	notpre...	present		50.0	1.4	129.0	4.0	11.1	39.0	8300.0	4.6	yes	no	no	poor	no	no	ckd
26	61.0	60.0	1.025	0	0		normal	notpre...	notpre...	108.0	75.0	1.9	141.0	5.2	9.9	29.0	8400.0	3.7	yes	yes	no	good	no	yes	ckd
27	75.0	80.0	1.015	0	0		normal	notpre...	notpre...	156.0	45.0	2.4	140.0	3.4	11.6	35.0	10300.0	4.0	yes	yes	no	poor	no	no	ckd
28	69.0	70.0	1.010	3	4	normal	abnormal	notpre...	notpre...	264.0	87.0	2.7	130.0	4.0	12.5	37.0	9600.0	4.1	yes	yes	yes	good	yes	no	ckd
29	75.0	70.0		1	3			notpre...	notpre...	123.0	31.0	1.4							no	yes	no	good	no	no	ckd
30	68.0	70.0	1.005	1	0	abnormal	abnormal	present	notpre...		28.0	1.4			12.9	38.0			no	no	yes	good	no	no	ckd
31		70.0						notpre...	notpre...	93.0	155.0	7.3	132.0	4.9					yes	yes	no	good	no	no	ckd
32	73.0	90.0	1.015	3	0		abnormal	present	notpre...	107.0	33.0	1.5	141.0	4.6	10.1	30.0	7800.0	4.0	no	no	no	poor	no	no	ckd
33	61.0	90.0	1.010	1	1		normal	notpre...	notpre...	159.0	39.0	1.5	133.0	4.9	11.3	34.0	9600.0	4.0	yes	yes	no	poor	no	no	ckd
34	60.0	100.0	1.020	2	0	abnormal	abnormal	notpre...	notpre...	140.0	55.0	2.5			10.1	29.0			yes	no	no	poor	no	no	ckd
35	70.0	70.0	1.010	1	0	normal		present	present	171.0	153.0	5.2							no	yes	no	poor	no	no	ckd
36	65.0	90.0	1.020	2	1	abnormal	normal	notpre...	notpre...	270.0	39.0	2.0			12.0	36.0	9800.0	4.9	yes	yes	no	poor	no	yes	ckd

Undo OK Cancel

The highlighted fields depict that the data set is dirty because it contains missing values. Therefore, it is important to clean it before doing any other data mining task. This can be done by applying a filter like Replace-Missing-With-User-Constant which is an example of an unsupervised filter. However, the choice of the filter may vary depending on the data and the needs of the expert. Below is a figure showing the preprocessed data after successfully applying the filter. All the missing values were replaced and now the data can be used for classification or any other data mining task.

International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2015): 6.391



Viewer

Relation: Chronic_Kidney_Disease-weka.filters.unsupervised.attribute.ReplaceMissingValues

No.	age Numeric	bp Numeric	sg Nominal	al Nominal	su Nominal	rbc Nominal	pc Nominal	pcc Nominal	ba Nominal	bgr Numeric	bu Numeric	sc Numeric	sod Numeric	pot Numeric	hemo Numeric	pcv Numeric	wbcc Numeric	rbcc Numeric	htn Nominal	dm Nominal	cad Nominal	appet Nominal	pe Nominal	ane Nominal	class Nominal
1	48.0	80.0	1.020	1	0	normal	normal	notpre...	notpre...	121.0	36.0	1.2	137.5...	4.627...	15.4	44.0	7800.0	5.2	yes	yes	no	good	no	no	ckd
2	7.0	50.0	1.020	4	0	normal	normal	notpre...	notpre...	148.0...	18.0	0.8	137.5...	4.627...	11.3	38.0	6000.0	4.707...	no	no	no	good	no	no	ckd
3	62.0	80.0	1.010	2	3	normal	normal	notpre...	notpre...	423.0	53.0	1.8	137.5...	4.627...	9.6	31.0	7500.0	4.707...	no	yes	no	poor	no	yes	ckd
4	48.0	70.0	1.005	4	0	normal	abnormal	present	notpre...	117.0	56.0	3.8	111.0	2.5	11.2	32.0	6700.0	3.9	yes	no	no	poor	yes	yes	ckd
5	51.0	80.0	1.010	2	0	normal	normal	notpre...	notpre...	106.0	26.0	1.4	137.5...	4.627...	11.6	35.0	7300.0	4.6	no	no	no	good	no	no	ckd
6	60.0	90.0	1.015	3	0	normal	normal	notpre...	notpre...	74.0	25.0	1.1	142.0	3.2	12.2	39.0	7800.0	4.4	yes	yes	no	good	yes	no	ckd
7	68.0	70.0	1.010	0	0	normal	normal	notpre...	notpre...	100.0	54.0	24.0	104.0	4.0	12.4	36.0	8406....	4.707...	no	no	no	good	no	no	ckd
8	24.0	76.46...	1.015	2	4	normal	abnormal	notpre...	notpre...	410.0	31.0	1.1	137.5...	4.627...	12.4	44.0	6900.0	5.0	no	yes	no	good	yes	no	ckd
9	52.0	100.0	1.015	3	0	normal	abnormal	present	notpre...	138.0	60.0	1.9	137.5...	4.627...	10.8	33.0	9600.0	4.0	yes	yes	no	good	no	yes	ckd
10	53.0	90.0	1.020	2	0	abnormal	abnormal	present	notpre...	70.0	107.0	7.2	114.0	3.7	9.5	29.0	12100.0	3.7	yes	yes	no	poor	no	yes	ckd
11	50.0	60.0	1.010	2	4	normal	abnormal	present	notpre...	490.0	55.0	4.0	137.5...	4.627...	9.4	28.0	8406....	4.707...	yes	yes	no	good	no	yes	ckd
12	63.0	70.0	1.010	3	0	abnormal	abnormal	present	notpre...	380.0	60.0	2.7	131.0	4.2	10.8	32.0	4500.0	3.8	yes	yes	no	poor	yes	no	ckd
13	68.0	70.0	1.015	3	1	normal	normal	present	notpre...	208.0	72.0	2.1	138.0	5.8	9.7	28.0	12200.0	3.4	yes	yes	yes	poor	yes	no	ckd
14	68.0	70.0	1.020	0	0	normal	normal	notpre...	notpre...	98.0	86.0	4.6	135.0	3.4	9.8	38.88...	8406....	4.707...	yes	yes	yes	poor	yes	no	ckd
15	68.0	80.0	1.010	3	2	normal	abnormal	present	present	157.0	90.0	4.1	130.0	6.4	5.6	16.0	11000.0	2.6	yes	yes	yes	poor	yes	no	ckd
16	40.0	80.0	1.015	3	0	normal	normal	notpre...	notpre...	76.0	162.0	9.6	141.0	4.9	7.6	24.0	3800.0	2.8	yes	no	no	good	no	yes	ckd
17	47.0	70.0	1.015	2	0	normal	normal	notpre...	notpre...	99.0	46.0	2.2	138.0	4.1	12.6	38.88...	8406....	4.707...	no	no	no	good	no	no	ckd
18	47.0	80.0	1.020	0	0	normal	normal	notpre...	notpre...	114.0	87.0	5.2	139.0	3.7	12.1	38.88...	8406....	4.707...	yes	no	no	poor	no	no	ckd
19	60.0	100.0	1.025	0	3	normal	normal	notpre...	notpre...	263.0	27.0	1.3	135.0	4.3	12.7	37.0	11400.0	4.3	yes	yes	yes	good	no	no	ckd
20	62.0	60.0	1.015	1	0	normal	abnormal	present	notpre...	100.0	31.0	1.6	137.5...	4.627...	10.3	30.0	5300.0	3.7	yes	no	yes	good	no	no	ckd
21	61.0	80.0	1.015	2	0	abnormal	abnormal	notpre...	notpre...	173.0	148.0	3.9	135.0	5.2	7.7	24.0	9200.0	3.2	yes	yes	yes	poor	yes	yes	ckd
22	60.0	90.0	1.020	0	0	normal	normal	notpre...	notpre...	148.0...	180.0	76.0	4.5	4.627...	10.9	32.0	6200.0	3.6	yes	yes	yes	good	no	no	ckd
23	48.0	80.0	1.025	4	0	normal	abnormal	notpre...	notpre...	95.0	163.0	7.7	136.0	3.8	9.8	32.0	6900.0	3.4	yes	no	no	good	no	yes	ckd
24	21.0	70.0	1.010	0	0	normal	normal	notpre...	notpre...	148.0...	57.42...	3.072...	137.5...	4.627...	12.52...	38.88...	8406....	4.707...	no	no	no	poor	no	yes	ckd
25	42.0	100.0	1.015	4	0	normal	abnormal	notpre...	present	148.0...	50.0	1.4	129.0	4.0	11.1	39.0	8300.0	4.6	yes	no	no	poor	no	no	ckd
26	61.0	60.0	1.025	0	0	normal	normal	notpre...	notpre...	108.0	75.0	1.9	141.0	5.2	9.9	29.0	8400.0	3.7	yes	yes	no	good	no	yes	ckd
27	75.0	80.0	1.015	0	0	normal	normal	notpre...	notpre...	156.0	45.0	2.4	140.0	3.4	11.6	35.0	10300.0	4.0	yes	yes	no	poor	no	no	ckd
28	69.0	70.0	1.010	3	4	normal	abnormal	notpre...	notpre...	264.0	87.0	2.7	130.0	4.0	12.5	37.0	9600.0	4.1	yes	yes	yes	good	yes	no	ckd
29	75.0	70.0	1.020	1	3	normal	normal	notpre...	notpre...	123.0	31.0	1.4	137.5...	4.627...	12.52...	38.88...	8406....	4.707...	no	yes	no	good	no	no	ckd
30	68.0	70.0	1.005	1	0	abnormal	abnormal	present	notpre...	148.0...	28.0	1.4	137.5...	4.627...	12.9	38.0	8406....	4.707...	no	no	yes	good	no	no	ckd
31	51.48...	70.0	1.020	0	0	normal	normal	notpre...	notpre...	93.0	155.0	7.3	132.0	4.9	12.52...	38.88...	8406....	4.707...	yes	yes	no	good	no	no	ckd
32	73.0	90.0	1.015	3	0	normal	abnormal	present	notpre...	107.0	33.0	1.5	141.0	4.6	10.1	30.0	7800.0	4.0	no	no	no	poor	no	no	ckd
33	61.0	90.0	1.010	1	1	normal	normal	notpre...	notpre...	159.0	39.0	1.5	133.0	4.9	11.3	34.0	9600.0	4.0	yes	yes	no	poor	no	no	ckd
34	60.0	100.0	1.020	2	0	abnormal	abnormal	notpre...	notpre...	140.0	55.0	2.5	137.5...	4.627...	10.1	29.0	8406....	4.707...	yes	no	no	poor	no	no	ckd
35	70.0	70.0	1.010	1	0	normal	normal	present	present	171.0	153.0	5.2	137.5...	4.627...	12.52...	38.88...	8406....	4.707...	no	yes	no	poor	no	no	ckd
36	65.0	90.0	1.020	2	1	abnormal	normal	notpre...	notpre...	270.0	39.0	2.0	137.5...	4.627...	12.0	36.0	9800.0	4.9	yes	yes	no	poor	no	yes	ckd

Undo OK Cancel

4. Classification

In data mining, classification is the process of determining a label or a membership for a particular instance based on a training model. It seeks to predict the class attribute of an instance whose label was previously unknown. In WEKA, classification is categorized into supervised and unsupervised although for the two, the procedure is similar; Building the model (classifier) by determining the class label for every object then training the model with requisite data which is represent as a decision tree, association rules or mathematical formulas.

Once the model is developed and trained, it is then presented with a previously unknown and unclassified instance to predict its class label. WEKA provides statistics about the accuracy of the model in percentages.

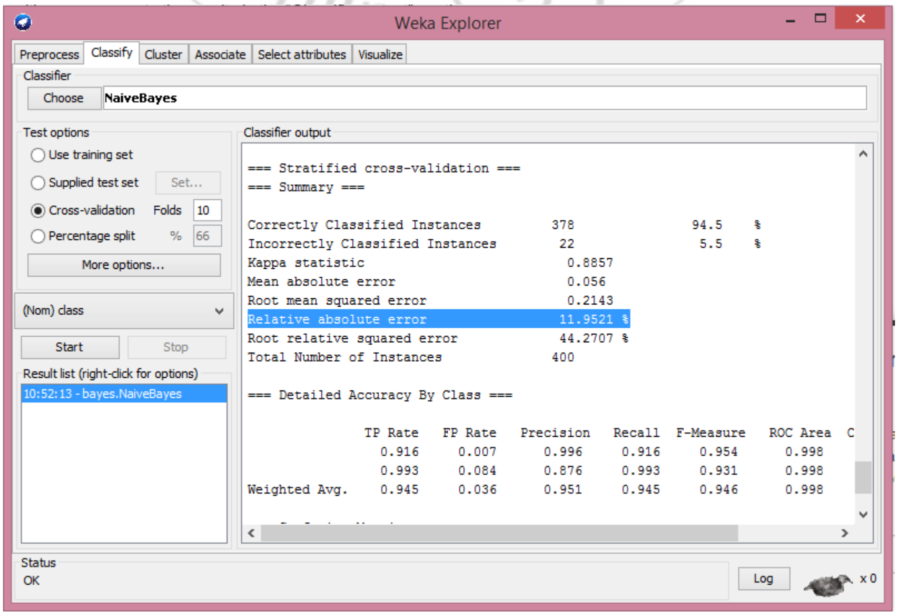
The assumption is that after data has successfully been preprocessed, it produces a set of attributes X_i X_{ii}…..X_n and Y such that the objective is to learn a function f:(X_i….X_n)→Y so that this function can be used to predict y (which is a discrete attribute or class label) for a given record (x_i…xn).

5. Classification in WEKA

The following are the supported classifiers in WEKA; Bayes, Functions, Lazy, Meta, Mi, Misc, Rules, and trees.

1) Load the data in WEKA through the GUI or command line interface
2) Choose the classifier
3) Determine the classification algorithm
4) Visualize the classification by generating a tree

This experiment used Naïve Bayes with a cross validation test options set to 10 folds meaning that the data was split into 10 distinct parts where the first 9 instances are used for training and the remaining 1 instance is used to assess how the algorithm performs. This process is iterated such that each of the 10 split parts is given a chance to be trained and tested.



As seen from the implementation, Naïve Bayes correctly, identified two classes, CKD and NOTCKD. It is clear that preprocessing improved the accuracy and efficiency of the model to correctly classify 378 instances translating to 94.5%. The incorrectly classified instances were 22 translating to a 5.5%.

```
=== Confusion Matrix ===

   a   b   <-- classified as
 229  21 |   a = ckd
   1 149 |   b = notckd
```

The confusion matrix presents a table with a comparison between the correctly classified and incorrectly classified instances. It is clear that 21 instances were incorrectly classified as CKD while 1 instance was incorrectly classified as NOTCKD. The confusion matrix can be used to justify the accuracy achieved by the classifier.

6. Conclusion and Future Work

In this paper, a preamble to data preprocessing is presented. Special focus is given to literature on data preprocessing for classification.A description of data preprocessing and classification experiments are run in WEKA. It is clear that to achieve high accuracy with a classifier, preprocessing is a critical task. As well, the choice of the classification algorithm also determines the accuracy attained after performing any data mining tasks.

References

[1] Holmes, A. Donkin, I. H. Witten, WEKA: A Machine Learning Workbench, In Proceedings of the Second Australian and New Zealand Conference on Intelligent Information Systems, 357-361, 1994.
[2] C. Velayutham and K. Thangavel, "Unsupervised Quick Reduct Algorithm Using Rough Set Theory", JOURNAL OF ELECTRONIC SCIENCE AND TECHNOLOGY, VOL. 9, NO. 3, SEPTEMBER 2011
[3] Thair Nu Phyu, " Survey of classification techniques in data mining"; Proceedings of the International MultiConference of Engineers and Computer Scientists 2009 Vol I
[4] B.G.ObulaReddy, Dr. MaligelaUssenaiah,"Literature Survey On Clustering Techniques"; IOSR Journal of Computer Engineering (IOSRJCE) ISSN: 2278-0661 Volume 3, Issue 1 (July-Aug. 2012), PP 01-12
[5] Yugalkumar , G. Sahoo; "Study of Parametric Performance Evaluation of Machine Learning and Statistical Classifiers", I.J. Information Technology and Computer Science, 2013, 06, 57-64
[6] Weka: http://www.cs.waikato.ac.nz/~ml/weka/
[7] Asuncion A, Newman D (2007) UCI machine learning repository
[8] Famili, Data Pre-Processing and Intelligent Data Analysis, IJSR,1997.
[9] Sven F., Crone,The impact of preprocessing on data mining: An evaluation of classifier sensitivity in direct marketing, IEEE, 2005.
[10] S. B. Kotsiantis, D. Kanellopoulos and P. E. PintelasData Preprocessing for Supervised Leaning, IEEE, 2006.
[11] Kamiran,ToonCalders, Faisal,Data preprocessing techniques for classification without discrimination IJCSE, 2011.
[12] F. Mary Harin Fernandez and R. Ponnusamy, Data Preprocessing and Cleansing in Web Log on Ontology for Enhanced Decision Making,IJSR, 2016.
[13] http://weka.wikispaces.com/Use+WEKA+in+your+Java+code#Classification-Building a Classifier.