Task 3 – A data mining system for a Hospital (25 marks)

A hospital has been collecting a great deal of data on their patients and have heard that use of data mining could improve their service. They would like you to create a brief report that includes the following.

(i) **What data mining is and an appropriate application for the hospital.**

Data mining is the process of analysing data from different perspectives and summarizing it into valuable information that can determine/forecast the success rate of business such as their profit, loss, breakeven, and costs etc. through the analytics of big data such as facts, numbers or text that are gathered by different organisations. For organisations to data mine, they use data mining software. There are a lot of software which has their own features with pros & cons. These software s allows people within big organisations to analyse the big data they have gathered/stored. When these data are implemented, they are trained, tested and evaluated which will produce results and allows users to view data from many different dimensions/angle, categories and relationship between the data.

One of the data mining software which can do this perfectly is called Weka. Weka is an appropriate application for this hospital because it has a very graphical user interface which makes the system easy to access which means data will be easy to implement, however the best one feature of this application are the algorithms. The hospital needs to find out if a patient has diabetes or not and therefore have collected data which will show the result using intelligent algorithms which Weka has built in. This will produce correct results which are reliable and can be used to determine diabetes faster and quicker, faster than it would have taken it traditional way of checking for it physically. This saves the hospital time and money as the process helps determine results of diabetes faster and it could also help save lives as diabetes can be diagnosed quicker and effectively.

(ii) **How you would go about creating the system using the data mining lifecycle below.**

To create the system, I must follow the data mining Lifecyle. The life cycle will help ensure that the data is gathered and processed correctly to ensure correct results at the end.

Problem Definition: with the help of the data mining lifecycle, I can identify the problem and its definition which is to find out how many patients of the hospital are diabetic. To do this we need to create system and to create a system we need to gather data, big data of each patients of the hospital.

Data gathering and preparation: After we have identified the problem, the next stage is to collect enough amount of data which is then prepared to be broken down so that there are no anomalies in the data that we have gathered of the patients. This means that the data must be put into the right tables and the attributes has to be identified. After this process, hypotheses can be formed on how the system is going to be created with the data collected. Now we will begin the data preparation for the system, different activities are constructed to plan the system model by feeding the normalised data into the modelling tool.

Model building and evaluation – In this stage, we begin with building the model/models with various techniques with different activities. Values are set to be met with data set parameters. Different techniques will require different data type to be set and perhaps even change the model and therefore must be careful with the patient's data, to make sure no duplication was formed in the process. After this process, model/models are created. Now it is time evaluate, which means the model is trained to produce close to 100% percentage correct outputs.

Use Knowledge - The trained model can be now be used by doctors to diagnose the patients of the hospital who are diabetic with the right medicines. The model will use relationships to link and differentiate patients of the hospital who are diabetic and who are not. These are the stages of the life cycle which will ensure that by following it, the system is created faster and efficiently.

(iii) **If the small amount of data (diabetes.arff) collected so far by the hospital is appropriate for assessing if a person has diabetes**.

The data collected by the hospital will have been broken down into attributes of the patients which are their age, diabetes pedigree function, 2-hour serum insulin, triceps skin fold thickness, diastolic blood pressure, BMI, plasma glucose concentration and the number of time they were pregnant which included only female patients. I am going to research if the gathered by the hospital is appropriate if a person has diabetes by finding out how strong or if there is one, relationship between the attributes of the patients from the hospital and the attributes of the patients who are also diabetic.

[1] First attribute I researched was age, research showed me that a human being grows older they risk for type 2 diabetes. This applies to all gender age but not very common to young ages because people who are young eat healthy, stay active because they grow as they age and because they are active, they can maintain their weight, however for old people it is not the same, because as people grow old, the body becomes old and people start to get lazy and prefer comfort life style.

[2] Research showed me that diabetes pedigree function does have a relationship as if a person's family member such as mother, father, sister or brother diabetes then that person is in a higher risk of getting diabetes as well.

[3] Our bodies need some circulation insulin at all time, even when we don't eat. Otherwise, our livers keep making glucose dumping it into the blood. Livers do this to prevent blood glucose form going too low. A high insulin level is a sign of prediabetes and therefore strongly linked to finding out if a person is likely to be diabetic or not.

[1] http://www.diabetes.org/are-you-at-risk/lower-your-risk/nonmodifiables.html?referrer=https://www.google.co.uk/
[2] http://www.diabetes.org/are-you-at-risk/lower-your-risk/nonmodifiables.html?referrer=https://www.google.co.uk/
[3] http://www.diabetesselfmanagement.com/blog/do-you-know-your-insulin-level/

[4] Triceps skin fold thickness is an examination of the relation of circumference (waist and hip) and skinfold-thickness (subscapular and triceps) measurements to lipid and insulin concentration levels. A high level of insulin is a sign of prediabetes and has a strong relationship to find out results of diabetic.

[5] Diastolic blood pressure means the lowest pressure when your heart relaxes between beats. Having a high blood pressure means that high chance of being diabetic. If you have a high blood pressure, you will need to take medication to lower blood pressure. This attribute is therefore appropriate.

[6] Plasma glucose concentration is a glucose test that can help determine diabetic patients. If the patient has a glucose concentration of 7.0 mmol/l and above (126 mg/dl and above then the person can be diagnosed as diabetic, therefore this data is appropriate.

[7] Number of times women were pregnant does not have any link leading people women being diabetic, however, people have been diabetic and pregnant at the same time, it requires hard work and dedication if the baby is to be born healthy but the person can be diabetic and pregnant.

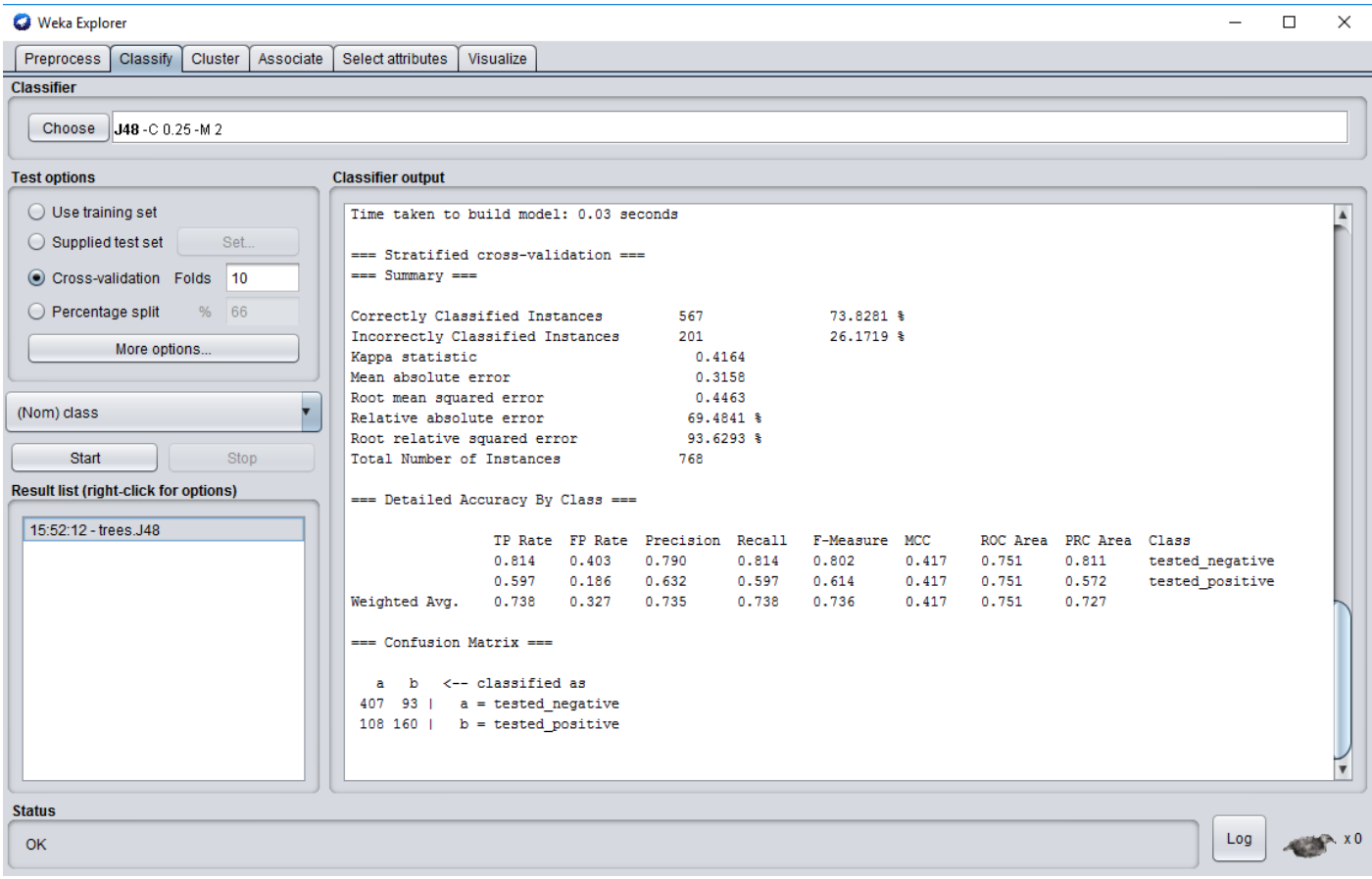
[8] BMI - An increase in body fat is generally associated with increased risk of metabolic diseases such as type 2 diabetes mellitus.

In conclusion, I think the small amount of data collected so far by the hospital is appropriate for assessing if a person has diabetes or not, because 7 out of 8 attributes helps to form strong links between each other to give an overall percentage or individual percentages of attributes which then will tell the hospital doctors the likely chance of a person is diabetic or not.

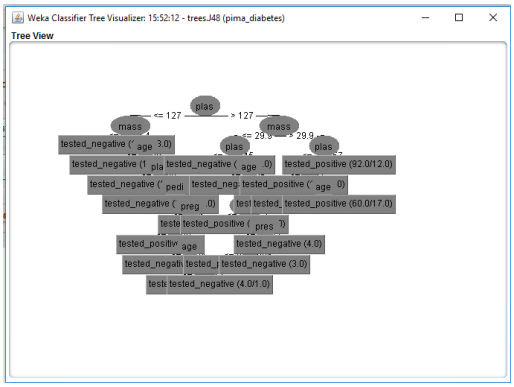
(iv) **The use of a data mining model such as a multilayer perceptron or decision tree to determine whether a person has diabetes. Note, you will need to use a data mining tool like WEKA to create your model and use the diabetes.arff data to train and test this model.**

I used the data provide by the hospital to import into a data mining tool called Weka. I then ran the Weka file with all its attributes to create a model, and I then began to train and test this model. I have screenshotted it below.

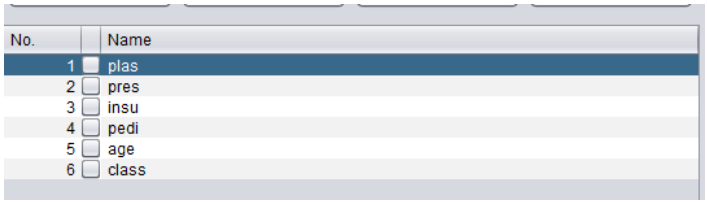
[4] http://ajcn.nutrition.org/content/69/2/308.full
[5] http://patient.info/health/diabetes-and-high-blood-pressure
[6] https://www.diabetes.co.uk/fasting-plasma-glucose-test.html
[7] https://www.diabetes.org.uk/Guide-to-diabetes/Life-with-diabetes/Pregnancy/
[8] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1890993/



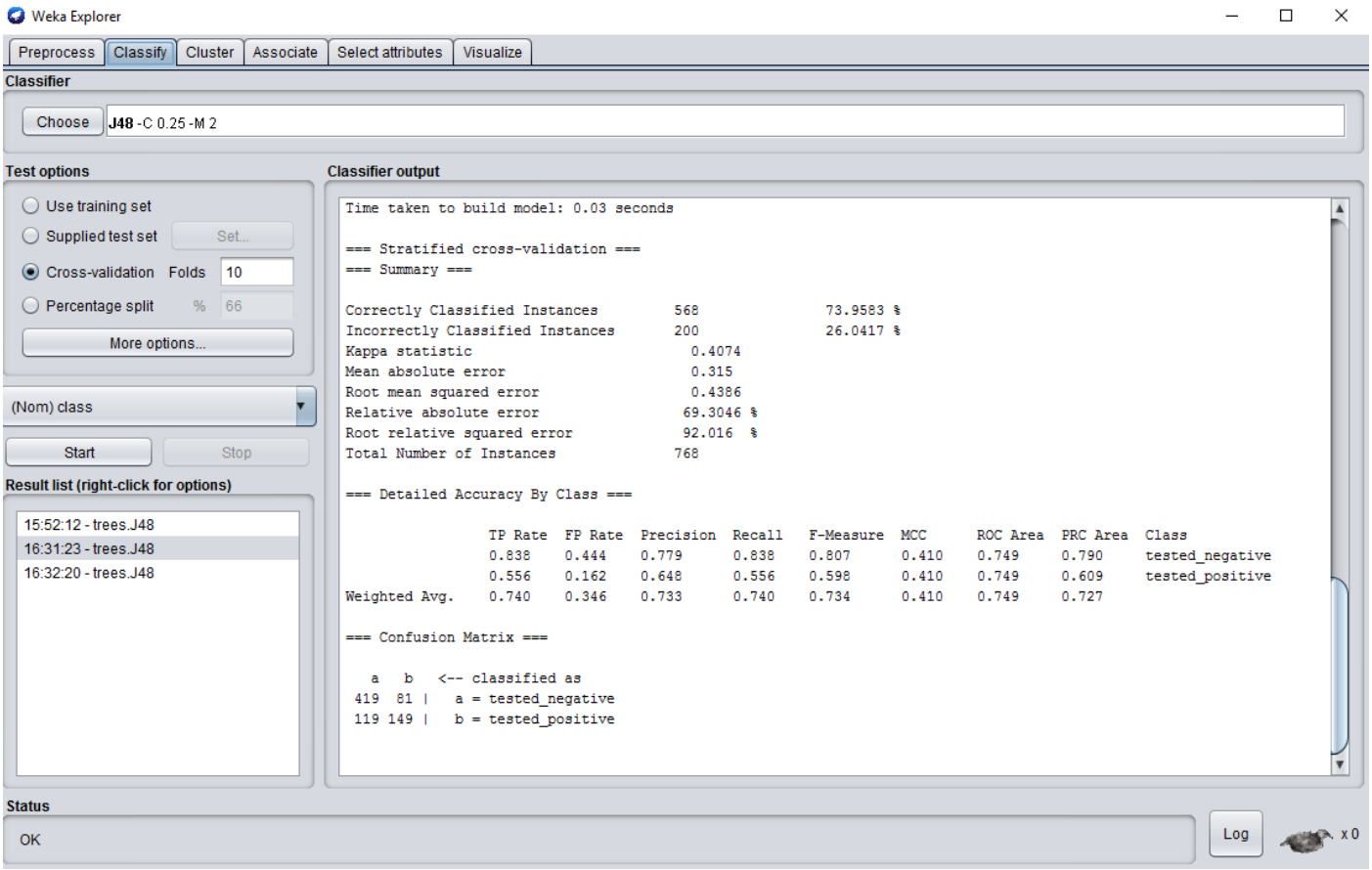
By analysing the result got from Weka J48 Tree, there were 567 instances which were correct and 201 instances which were incorrect. This meant that out of 768 instances, 73.8281% was correct instances of the patients and 26.1719% instances. Further into the tree I could see the Confusion Matrix I could see that out of 407 instances were correctly tested positive correctly and 93 of the were tested negative but they were positive, this meant that there was contradiction between data and it was giving patients wrong information which meant that 93 patients could not have been diagnosed with the right medicines of diabetes because they were falsely detected as not diabetic. Moreover, 160 were tested positive for diabetes correctly, however 108 of them were tested positive for diabetes incorrectly, which meant that they could be diagnosed as being diabetic and this means patients haven given wrong medicine for a condition which was not even there.



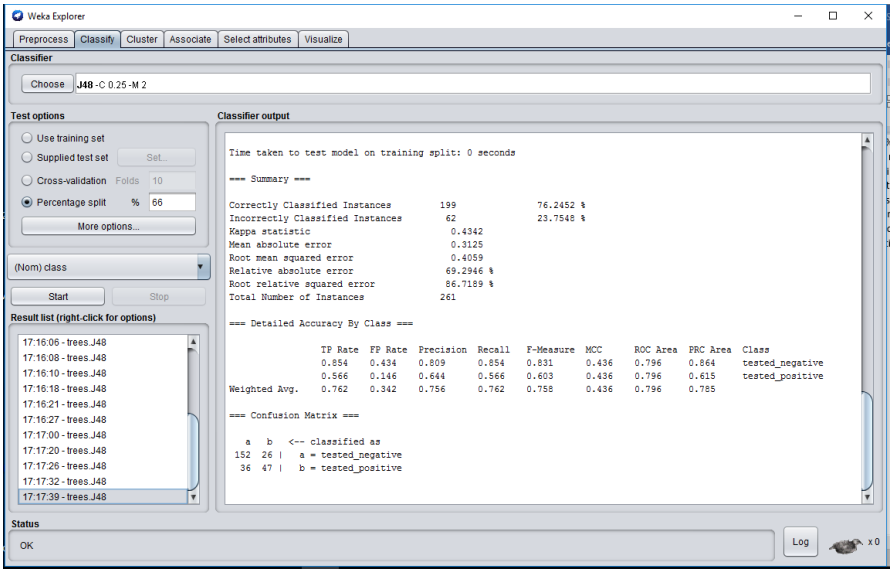
After the initial analysis, I looked at the attributes which Weka was using to produce J48 tree which in turn was generating the instances, percentages and confusion matrix. I noticed there were a lot of attributes which Weka was using and so I thought to remove some attributes and store only the important attributes. I removed the attributes which were not needed such as 'Pregency' and 'Skin' which did not have an effect on risk of having diabetes and decided only to keep attributes which could really indicate diabetes 'Plas', 'Pres', 'insul', 'pedi' and class.



'Plas' – Plasma
'Pres' – Diastolic blood pressure
'Insus' – 2 Hour serum insulin
'Pedi' – Diabetes Pedigree function
'Age' – Patient's Age
'Class' – Class variable



Removing the unrelated attributes gave me a bit better percentage 73.9583% of correctly classified instances with 568 correct instances and reduced the incorrect instances to 200 giving me the percentage of 26.0417%. The confusion matrix was also better after removing the attributes, 419 instances were negative and 81 instances were positive tested but were actually negative. Moreover, 119 instances which were negative were negative but classified as positive and 149 were positive but classified as positive. This model with compared when the model tree with all attributes, I can see that getting rid of the attributes made more sense as this gave me a clearer result than the other results. It gave me a better result of 0.1302 % which seem a small percentage but could prove to be crucial when finding out the correct diabetic patients.



I did another test technique where I took the model and splatted the percentage of the model into 66%, by doing this I want to find out how high or low the percentage would go if the model was splatted into 66%. The data I got this from was bit better as I managed to get the percentage up by 2.4669%. I also managed to the get lower defective results, only getting 26 incorrectly tested positive results from a possible 152. I only got it down to 16 false negatives and out of 47 tested negatives.

In conclusion, the use of a data mining tool is a great way for the hospital to find out if a person has diabetes. However I recommend the hospital to make more data to train and test the model to get the result close to perfection because at the moment data mining tool like Weka is showing that the data contains anomalies which are makes the predicts the wrong instances which in returns outputs the incorrect results. This will make the hospital doctors to diagnose wrong patients with diabetes and also not detect those people who actually have diabetes.

References

N.p., 2016. Web. 9 Dec. 2016.

N.p., 2016. Web. 9 Dec. 2016.

"Age, Race, Gender & Family History". *American Diabetes Association*. N.p., 2016. Web. 9 Dec. 2016.

"Age, Race, Gender & Family History". *American Diabetes Association*. N.p., 2016. Web. 9 Dec. 2016.

Bays, H. E., R. H. Chapman, and S. Grandy. "The Relationship Of Body Mass Index To Diabetes Mellitus, Hypertension And Dyslipidaemia: Comparison Of Data From Two National Surveys". N.p., 2016. Print.

"Diabetes And High Blood Pressure. Hypertension In Diabetes | Patient". *Patient*. N.p., 2016. Web. 9 Dec. 2016.

"Do You Know Your Insulin Level? - Diabetes Self-Management". *Diabetes Self-Management*. N.p., 2016. Web. 9 Dec. 2016.

Freedman, David et al. "Relation Of Circumferences And Skinfold Thicknesses To Lipid And Insulin Concentrations In Children And Adolescents: The Bogalusa Heart Study". *Ajcn.nutrition.org*. N.p., 2016. Web. 9 Dec. 2016.