< Diabetes>
(MATH 3220 Final Project)
Jennifer Lamb

Executive Summary

Diabetes is a rapidly growing disease that many suffer from across the states. Determining if one suffers from this illness is very distinct. However, determining the cause for having diabetes is not so clear. For this reason, it is important to create an algorithm to be able to see if diabetes can be diagnosed by having certain characteristics.

The diabetes.xls dataset was used and put into Quinlan's C4.5 program. This dataset was obtained from a study of female Pima Indians (768 cases) who were checked for diabetes. The attributes obtained from these females included: number of pregnancies, plasma glucose at 2 hours in an oral glucose tolerance test, diastolic blood pressure, triceps skin fold thickness, body mass index, diabetes pedigree function, age, and whether or not the woman had diabetes. Looking at these eight attributes and the class, whether or not the female had diabetes, the data was placed into the C4.5 program. By doing so, a decision tree was created. This tree creates certain criteria a person must fit under in order to be diagnosed as diabetic (healthy or sick).

Looking at the data, there was a lot of missing data. Multiple tests were run on the given dataset by altering the data. After performing multiple runs on the data, it was concluded that the most accurate data was the data that had averaged all of the missing data out. The decision tree came back with an error of 21.7%. Comparing the decision tree to medicinal data showed that there were errors. However, there is always room for error in a dataset. Additionally, it also important to remember that even though people who appear to have the same characteristics as those who are diagnosed with diabetes does not mean that the person who is being compared will have diabetes, as well.

Problem Description

Diabetes is a serious illness that many Americans face in today's society. This illness occurs when the pancreas does not produce enough or none insulin at all to control the amount of glucose in the blood. "Diabetes is the seventh leading cause of death in the U.S. […] The disease can cause serious health problems which may include heart disease, blindness, kidney failure, and lower-extremity amputations" (Dyess). Lower-extremity amputations are due to circulation problems and nerve damage. Due to the seriousness of this illness, it is important to understand the causes for one to become diabetic.

In order to determine such causes, the Diabetes.xls dataset was used. This dataset was taken from a medical study that dates back to 1994. It contains eight categorizes used for classifying the women of Pima Indian Heritage living near Phoenix as healthy or sick (having or not having diabetes). These attributes include number of pregnancies, PG concentration (plasma glucose at 2 hours in an oral glucose tolerance test), Blood Pressure (mm Hg), triceps skin folds thickness (mm), 2-hour serum insulin (μ U/ml),

body mass index, diabetes pedigree function, and age was gathered from each patient. These eight attributes are closely examined and are tested to see if by examining such characteristic, one can be classified as healthy or sick (a.k.a. diabetic).

Analysis Technique

The methodology of this lab is very important to follow, in order to configure solutions for the Diabetes Problem. Ultimately by following the method, patients with distinct characteristics can be determined as diabetic. Using a decision tree was vital to producing the results. The decision tree runs through a series of labels, where one is categorized under the description and continues down the tree until the patient is or is not diagnosed with diabetes. This is process is described in J. Ross Quinlan's book, C4.5: PROGRAMS FOR MACHINE LEARNING. "A decision tree can be used to classify a case by starting at the root of the tree and moving through it until a leaf is encountered. At each non-leaf decision node, the case's outcome for the test at the node is determined" (Quinlan 6). Once the node is determined, the root that is the outcome from this node and the process continues down until a leaf is encountered; which in this case would be healthy or sick. This tree is created through C4.5, a decision tree algorithm.

C4.5 reads the input data given and then creates a decision tree based off this information. The tree data is then converted into a set of concrete rules for running the data. The data from the tree then triggers questions to accurately ask the user for the inputted information. The consult additionally works through the rules data to compute the user's responses into questions.

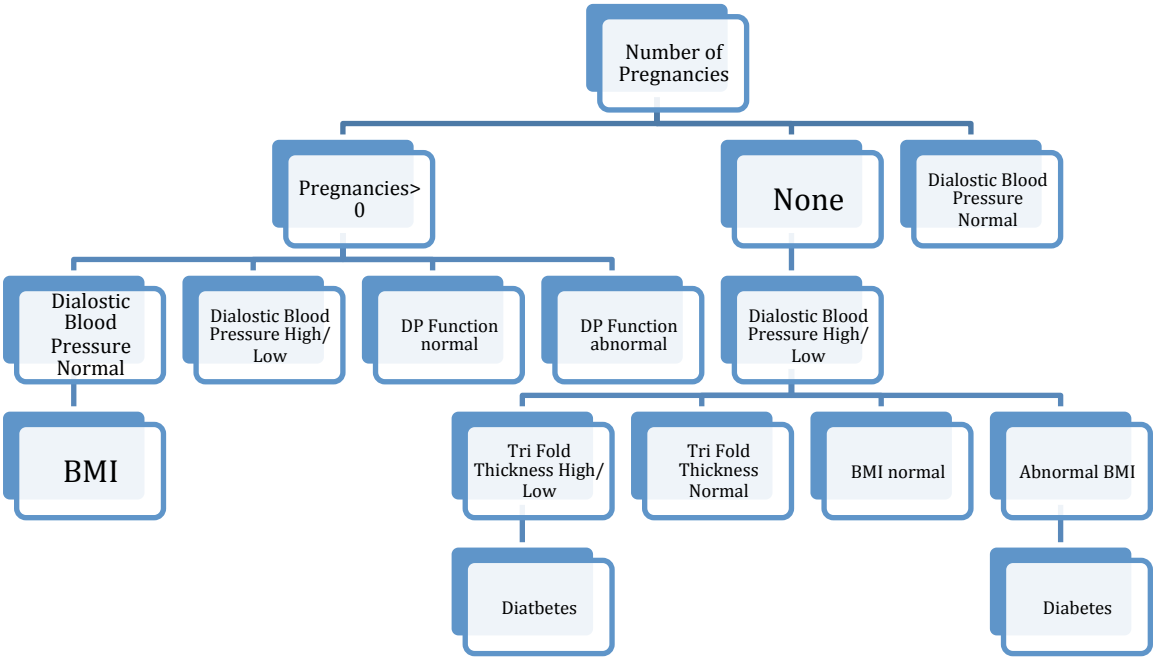
C4.5 builds decision trees using the concept of information entropy from data known to find designated outcomes (data is the set: $S = s_1,s_2,...$). Each sample (s_i) equals the vector ($x_1,x_2,...$) that represent the important notes of the data, which thus classifies the information. The building data is reproduced with best data, called vector $C = c_1,c_2,...$ where $c_1,c_2,...$ to represent a building block of where each sampling information should be placed (C4.5 algorithm).

Information entropy:

$$H(X) = \sum_{i=1}^{n} p(x_i)\, I(x_i) = -\sum_{i=1}^{n} p(x_i) \log_b p(x_i),$$

X must be a random variable with p denoting the probability mass function of X

Using C4.5, each node of the tree the algorithm chooses the next leaf of the tree by comparing the data to the training data. The result is found by computing the difference in entropy.



This is a rough idea of the decision tree; not all areas are covered on the decision tree. The areas that are being looked at are the following: number of pregnancies, plasmas glucose (PG Concentration), Diastolic BP, Tri Fold Thickness, Serums Ins, BMI, DP function, and Age. From this decision tree, it will be determined if the person has diabetes. Once the results are found, the person is then compared with their real diagnostic. From this, percent accuracy will be found to determine if C4.5 is an accurate algorithm for testing whether or not a person is diagnosed with diabetes.

The medical algorithm is similar to the procedure described above. It represents this procedure by finding a healthcare treatment for the patient. This algorithm works by using, "any computation, formula […] useful in healthcare" (Medical Algorithm). They use decision trees to find healthcare treatment (Medical Algorithm). By using a decision tree, one steps through the important information needed for medical treatment to find the treatment needed. Additionally, using this tree is not the only step that will reach this goal.

Quinlan addresses missing values by posing two possible options for the user to choose. One, data must be discarded and tests must be defined as not being able to be put into classes. Otherwise, the algorithm must be altered to deal with the missing values (Quinlan 7). As eliminating data is not looked upon, the second option is addressed. Adjusting gain ratio, weighting factor, and probability of classification under multiple

leaf nodes must be performed. Gain ratio takes normal gain measure and changes it to the number of cases that will be produced from the split of the child node. So by multiplying it by a the ratio of number of instances with a given attribute that has known data value over the total instances that has the attribute, missing data will not hurt the algorithm (Quinlan 28-9). The weighting takes in effect that the data could be split into multiple nodes; therefore by increasing weight, one decreases the error.

Pruning is done after completed the creation of the tree. It aims to reduce classification errors and makes tree more generalized so it is easier to read.

In order to create an accurate decision tree, the dataset must be tested several times with modifications. This is important in creating a decision tree, as 48% of the data is either missing or was not calculated. Therefore, the data will be altered to try and receive the most accurate decision tree. Initially the dataset will be tested with the original data. However, on the second run the data will be changed by addressing the missing data. This will be done is by marking all missing values with a "?" instead of a zero; all of the zeros in the dataset will be changed accordingly. However, the number of pregnancies will be left alone as it is the only possible attribute that one can have a zero in. Additionally, the dataset will then be tested in c4.5 by taking out all of the incomplete data. Another test on the data will be by excluding pregnancy from the testing data as it may be irrelevant in determining if one is diabetic or not. C4.5 will also be tested with all the missing data values averaged from the known attributes, and lastly it will be tested against no pregnancy values with the unknown values changed to the averages of the known.

Assumptions

It is very important to note that there is always room for error in the decision tree. Diabetes can be diagnosed even though they may not have certain characteristics that define a patient with diabetes. Additionally, given that there are missing data sets there will be certain precautions with the given tree and solution found. The data with the missing data set may have been purposely taken out for sake of error or may have been accidentally lost. Given this fact, certain precautions will be taken. These include, performing C4.5 with different modifications; seen above.

It may also be assumed that diagnostics are clear-cut.

Also it is assumed that the population of women at least 21 years old of Pima Indian Heritage living near Phoenix in 1990 (the population for this data set) was not affected by other contributing factors that cause diabetes.

Results

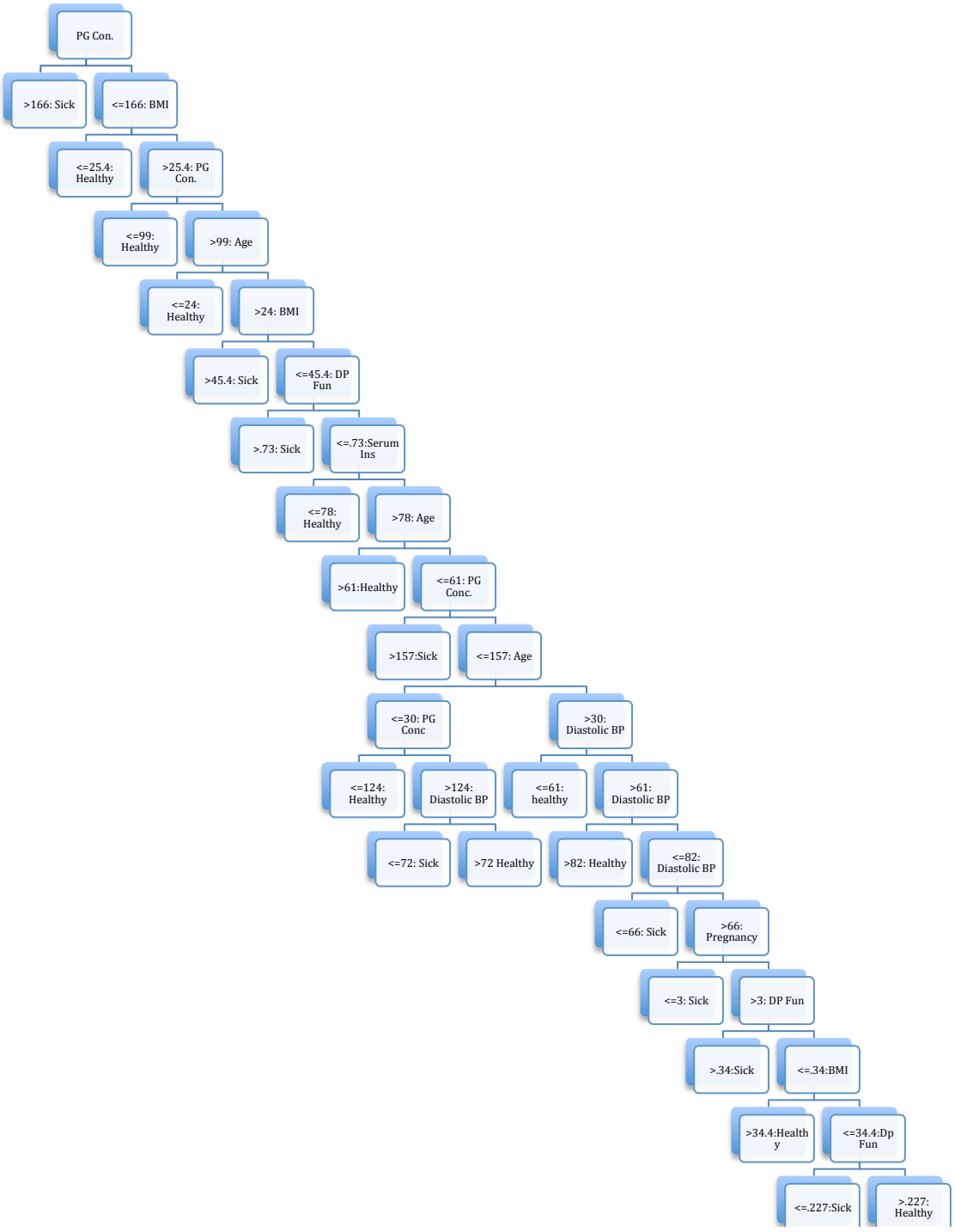
Run	Training Data Errors
Original Data	30.6%

Replaced “0”s with “?”s	25.9%
No incomplete Data	33.2%
No Pregnancy	22.7%
Averaged All incomplete data out	21.7%
No Pregnancy, Averaged all incomplete data out	32.4%

Weight: 8
Pruning confidence level 12%
---------- ----------
Size Errors Size Errors Estimate
67 109(14.2%) 39 121(15.8%) (24.4%) <<
---- ---- ----- ---- ----- ---------
Tested 768, errors 167 (21.7%) <<
(a) (b) <-classified as
---- ----
132 136 (a): class Sick
32 468 (b): class Healthy

*Decision tree on following page

Note: Decision Tree is simplified.



Issues

Grouping the attributes into one clean tree was an issue as there were a lot of data. Not only was this due to a large dataset, but also the dataset had a lot of noise and a 48% of dataset was missing.

Appendices

Medical diagnostic: Looking into eight attributes and comparing it to normal levels.

1. Pregnancies: One may be diagnosed with Gestational Diabetes during pregnancy. As the hormones produced during pregnancy can make one's cells more resistant to insulin. Those who are older than 25 years old are at risk. Additionally, if one has had Gestational Diabetes during one pregnancy, they are at greater risk at next pregnancy (Diabetes-Bing Health).
2. Age: Increase risk in diabetes as one gets older. This is especially true after one passes the age of forty-five (Diabetes-Bing Health).
3. Triceps skin-fold thickness: Normal value 23mm for women. Anything higher is, thus, above normal and may result in showing that one is overweight. It is important to maintain a healthy weight, as being overweight is a direct link to being diabetic (Triceps Skin-fold Thickness).
4. BMI: The ideal range is between 18.5-24.9. Where 25-29.9 is when a person is considered overweight, 30-39.9 indicates obesity, and 40+ indicate morbid obesity ("Am I overweight or obese"). Again, there is a direct link between being overweight and having diabetes.
5. 2-hour Serum Insulin: Numbers greater than 150µU/ml is related to insulin therapy, meaning that they are pre-diabetic or diabetic (Kronmal and others).
6. Diastolic Blood Pressure: 60-80 mm-Hg normal. Pre-hypertension is blood pressure of 80-89 and 90+ represents high blood pressure ("High Blood Pressure (HBP), Blood Pressure Readings" 1).
7. Plasma Glucose at 2 hours: "A person is said to have a normal response when the 2-hour glucose level is less than or equal to 110 mg/dL" (Norman 1).
8. Diabetes Pedigree Function: "equals 0.500 when the relative. is a parent or full sibling, equals 0.250 when the relative. is a half sibling, grandparent, aunt or uncle, and equals 0.125 when the relative, is a half aunt, half uncle or first cousin" (Smith and others).

Works Cited

"Am I Overweight or Obese?" *WebMD - Better Information. Better Health.* Ed. Judi Goldstone/MD. Web. 6 Dec. 2010. <http://www.webmd.com/diet/diagnosing-obesity>.

“C4.5 algorithm.” *Wikipedia, the Free Encyclopedia*. 15 Oct. 2010. <http://en.wikipedia.org/wiki/C4.5_algorithm>.

"Diabetes - Bing Health." *Bing*. Mayo Foundation for Medical Education and Research. Web. 5 Dec. 2010. <http://www.bing.com/health/article/mayo-126781/Diabetes?q=diabetes&qpvt=Diabetes>.

Dyess, Drucilla. "Percentage of Americans with Diabetes Is on the Rise." *Health News*. 26 June 2008. Web. 5 Dec. 2010. <http://www.healthnews.com/disease-illness/percentage-americans-suffering-diabetes-is-rise-1282.html>.

"High Blood Pressure (HBP), Blood Pressure Readings." *National Heart, Lung and Blood Institute*. Nov. 2008. Web. 12 Dec. 2010. <http://www.nhlbi.nih.gov/health/dci/Diseases/Hbp/HBP_WhatIs.html>.

Kronmal, Richard A., Joshua I. Barzilay, Russell P. Tracy, Peter J. Savage, Trevor J. Orchard, and Gregory L. Burke. "The Relationship of Fasting Serum Radioimmune Insulin Levels to Incident Coronary Heart Disease in an Insulin-Treated Diabetic Cohort." *The Journal of Clinical Endocrinology & Metabolism* 89 (2004): 1-10. *Journal of Clinical Endocrinology & Metabolism*. The Endocrine Society, 2004. Web. 7 Dec. 2010. <http://jcem.endojournals.org/cgi/content/full/89/6/2852>.

"Medical Algorithm." *Wikipedia, the Free Encyclopedia*. 10 Oct. 2010. <http://en.wikipedia.org/wiki/Medical_algorithm>.

Norman/MD, James. "Diagnosing Diabetes: Glucose Tolerance Test and Blood Glucose Levels. - The Two Primary Tests and Their Results, Which Combine to Make the Diagnosis of Diabetes." *Endocrine Diseases: Thyroid, Parathyroid Adrenal and Diabetes - EndocrineWeb*. 29 Mar. 2009. Web. 11 Dec. 2010. <http://www.endocrineweb.com/conditions/diabetes/diagnosing-diabetes>.

Smith, Jack W., JE Everhart, WC Dicksont, WC Knowler, and RS Johannes. "Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus." *Google Docs - Online Documents, Spreadsheets, Presentations, Surveys, File Storage and More*. Web. 7 Dec. 2010. <http://docs.google.com/viewer?a=v&q=cache:2fxQuarY0HUJ:www.ncbi.nlm.nih.gov/pmc/articles/PMC2245318/pdf/procascamc00018-0276.pdf Diabetes pedigree function&hl=en&gl=us&pid=bl&srcid=ADGEESgzx2ij8HXVr-APWnOTboWgdQQVHVS0VKoYDj7Zr-7_M4XUS8_ZlmyGinYK65LH-RW6TO-q2NRaXpmFc4Kh5X_V1ZqMs7pl-BD7Zmb3-rC96-YP9nJla2Sw>.

"Triceps Skin-fold Thickness - Definition of Triceps Skin-fold Thickness in the Medical Dictionary - by the Free Online Medical Dictionary, Thesaurus and Encyclopedia." *Medical Dictionary*. Web. 4 Dec. 2010. <http://medical-dictionary.thefreedictionary.com/triceps skin-fold thickness>.

Quinlan, J. R. C4.5: PROGRAMS FOR MACHINE LEARNING. San Mateo, CA: Morgan Kaufmann, 1993.