Diabetes

MATH 3220

Research Report

By: Jennifer Lamb

Outline

- **Fact**
- **Problem**
- **Decision Tree**
- **C4.5**
- **Addressing Messy Data**
- **Results**
- **Medical Research**
- **Real Life Examples**
- **Works Cited**

Fact

- “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)

Problem: Studying what?

Pima Indians Diabetes DataBase

- Who? Women at least 21 years old
- What? 8 categorizes, one class
- When? 1990
- Where? Living Near Phoenix
- Why? Classify as Healthy/Sick
- How? Decision Trees

Problem: Studying what?

Continued

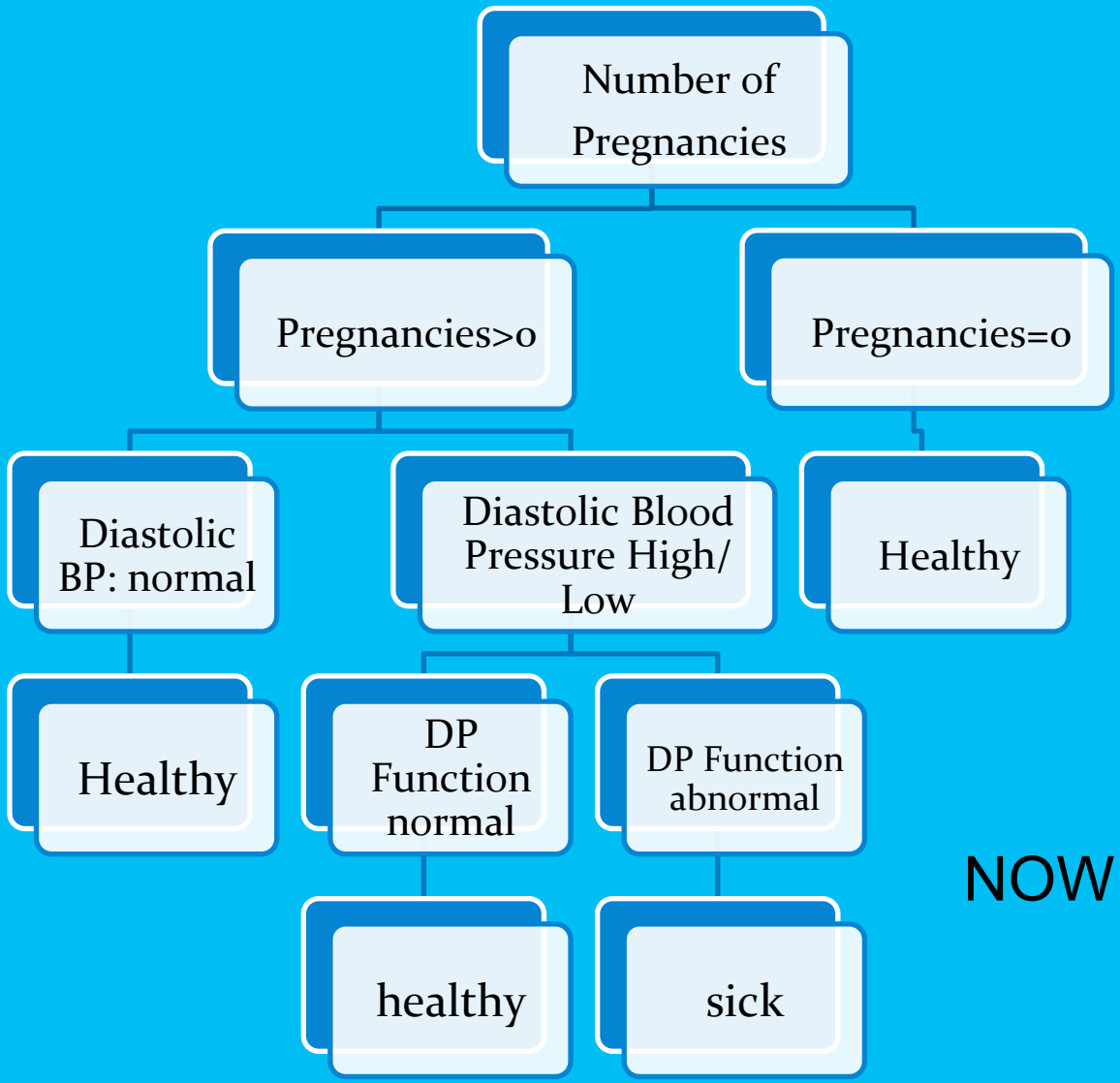
1. Pregnancies
2. PG Concentration
3. Diastolic BP (mm Hg)
4. Tri Fold Thickness (mm)
5. Serum Insulin (μU/ml)
6. BMI (weight in kg/ (height in m)^2)
7. DP Function
8. Age

Problem: Studying what?

- Type 1 Diabetes: Little-no insulin, build up of sugar in bloodstream instead of cells.
- Type 2 Diabetes: Cells become resistant to insulin and pancreas unable to make enough to fight resistance
- Gestational Diabetes: placenta produces hormones produced during pregnancy make cells resistant to insulin

Decision Tree Example

WHY?



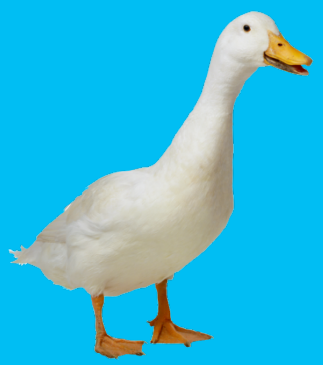
NOW WHAT?

C4.5

- C4.5 reads the input data
- C4.5 builds decision trees: information entropy

  data is the set: $S = s_1, s_2, \ldots$

  sample (s_i) equals the vector ($x_1, x_2, \ldots$) [important notes]



The building data is reproduced with best data :vector $C = c_1, c_2, \ldots$ where $c_1, c_2, \ldots$ to represent a building block of where each sampling information should be placed.

- 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

Addressing Messy Data

- Option: 1:Data must be discarded and tests must be defined as not being able to be put into classes.
- Option 2:Otherwise, the algorithm must be altered to deal with the missing values (Quinlan 7)
- 1. Adjusting gain ratio
- 2.weighting factor
- 3. probability of classification under multiple leaf nodes must be performed.
- Pruning is done after completed the creation of the tree.

Tests:

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%

Results Continued

- 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 Simplified

- PG Concentration > 166 : Sick (79.0/15.7)
- PG Concentration <= 166 :
- | BMI <= 25.4 : Healthy (122.0/7.7)
- | BMI > 25.4 :
- | | PG Concentration <= 99 : Healthy (141.0/19.4)
- | | PG Concentration > 99 :
- | | | Age <= 24 : Healthy (98.0/24.6)
- | | | Age > 24 :
- | | | | BMI > 45.4 : Sick (21.0/7.1)
- | | | | BMI <= 45.4 :
- | | | | | DP Function > 0.73 : Sick (55.0/19.7)
-

Compared to Medical Data

1. Pregnancy: Gestational Diabetes- age 25+
2. Age: 45+
3. Triceps Skin Fold Thickness: Normal 23mm
4. BMI: Ideal Range between 18.5-24.9
5. 2-hour Serum Insulin: Greater than 150 mu U/ml relates to insulin therapy
6. Diastolic BP: 60-80 mm normal
7. Plasma Glucose: normal when less than/equal to 110 mg/dL
8. Diabetes Pedigree Function:
   a. =0.5 for parent, full sibling
   b. =0.25 half sibling, grandparent, aunt, or uncle
   c. =0.125 half aunt, half uncle, or first cousin

Real Life -- Examples

- medical algorithm: finding a healthcare treatment for the patient.
- Big Bang Theory – Friends Decision Tree
- Use All the time without knowing

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:2fxQuarYoHUJ:www.ncbi.nlm.nih.gov/pmc/articles/PMC2245318/pdf/procascamc00018-0276.pdf Diabetes pedigree function&hl=en&gl=us&pid=bl&srcid=ADGEESgzx2ij8HXVr-APWnOTboWgdQQVHVSoVKoYDj7Zr-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.