Using WEKA and ML to learn more about Diabetes diagnosis

CS 4961

Ashley Muñoz & Jake Schultz

Data

Patients

- Females at least 21 years old
- Pima Indian Heritage

Attributes of interest

- Plas
  - Plasma glucose concentration 2 hours in an oral glucose tolerance test
- Mass
  - Body mass index (bmi)
- Age

```
Title: Pima Indians Diabetes Database

Sources:
(a) Original owners: National Institute of Diabetes and Digestive and
                     Kidney Diseases
(b) Donor of database: Vincent Sigillito (vgs@aplcen.apl.jhu.edu)
                       Research Center, RMI Group Leader
                       Applied Physics Laboratory
                       The Johns Hopkins University
                       Johns Hopkins Road
                       Laurel, MD 20707
                       (301) 953-6231
(c) Date received: 9 May 1990
```

Ashley

Information Gain

InfoGainAttributeEval:

- Ranks the Attributes by their worth
- Measures Information Gain
- Attributes with high Information Gain are ranked higher and have more impact

What is Information Gain?

- Measures how likely an event is to occur based on the value of a variable
- High Information Gain means an event is likely to occur

Jake

Attributes and Rankings

1. Plas
2. Mass
3. Age
4. Insu
5. Skin
6. Preg
7. Pedi
8. Pres
9. Class

- Insu
  - 2 hour serum insulin
- Skin
  - Triceps skin fold thickness
- Preg
  - Num of times pregnant
- Pedi
  - Diabetes pedigree function
- Pres
  - Diastolic blood pressure

21:30:33 - Ranker + InfoGainAttributeEval

```
Ranked attributes:
 0.1901    1 plas
 0.0749    2 mass
 0.0725    3 age
 0.0595    4 insu
 0.0443    5 skin
 0.0392    6 preg
 0.0208    7 pedi
 0.014     8 pres
```

Ashley

Logistic Regression

- Binary classification algorithm
- Learns coefficient for each input value
- Input values are linearly combined into a regression function
- Regression function is then transformed using a Logistic Function

We use the SimpleLogistic function

- Dependent Variable is nominal
  - Tested Positive/Tested Negative
- Tests whether getting a particular value of the dependent variable is associated to the measured variables.

Jake

Confusion Matrix

Takes the number of accurate results of a threshold and compares it to the number of inaccurate results.

True Negative	False Negative
False Positive	True Positive

20:53:20 - functions.SimpleLogistic

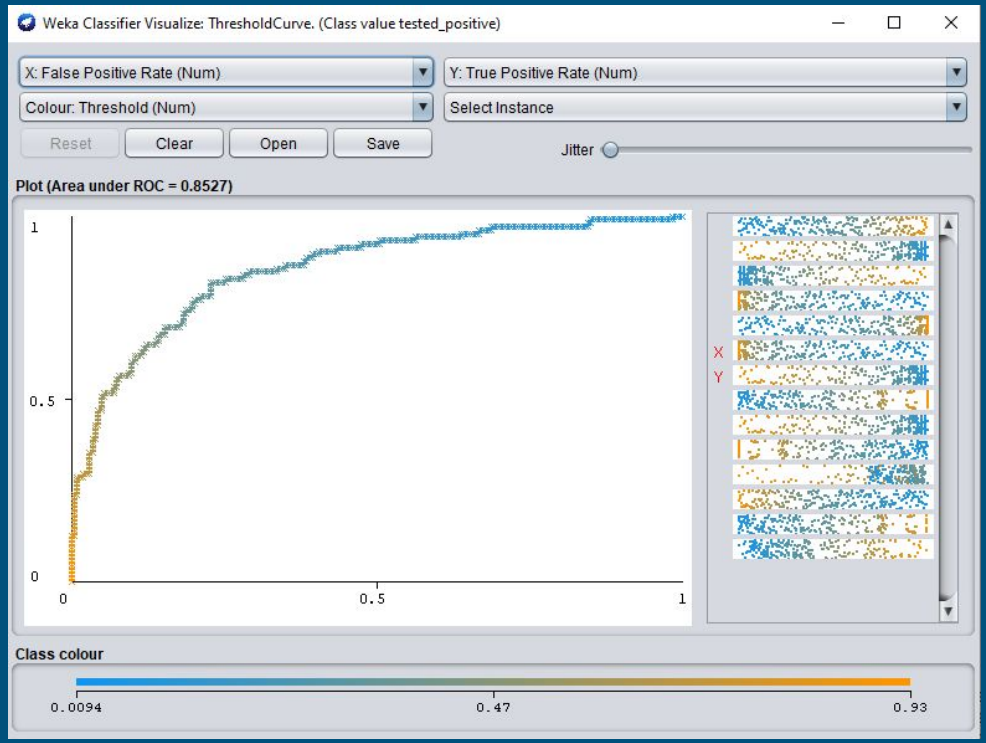
```
=== Confusion Matrix ===

   a   b   <-- classified as
 186  16 |   a = tested_negative
  47  58 |   b = tested_positive
```

Jake

ROC curve

- Takes the confusion matrix for each threshold value from 0-1
- 0 is everything is classified as False
- 1 is everything is classified as True
- The goal is to get the curve very close to 1 for as many values as possible
- This limits the number of false positives while maximizing True positives

Simple Logistic

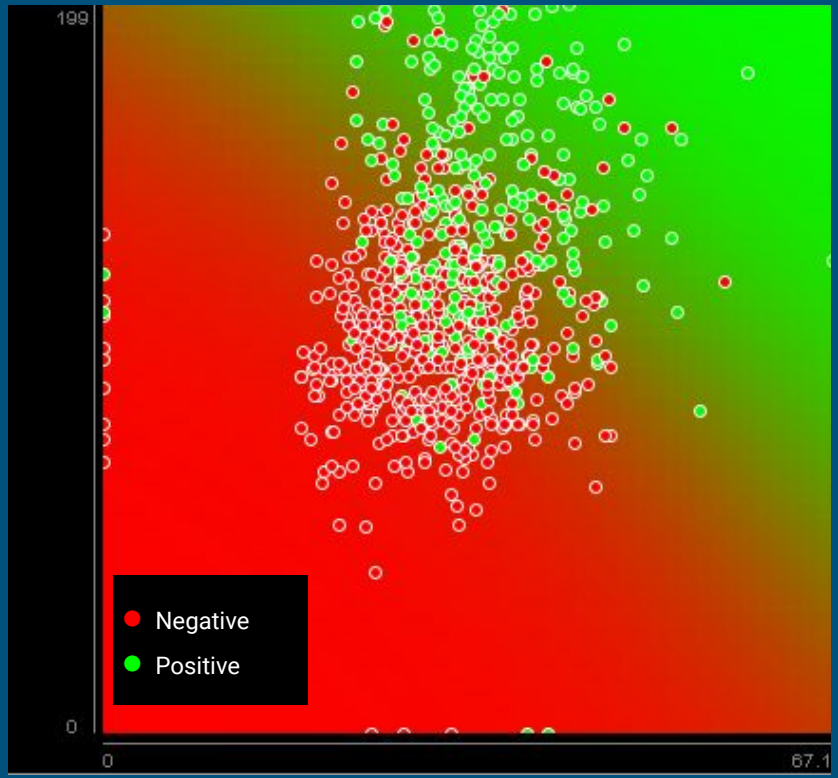


Jake

Classification Boundary

- Helps visualize where a particular person will be categorized
- Prominent diagonal line near upper middle of dataset

Mass vs Plasma
Simple Logistic

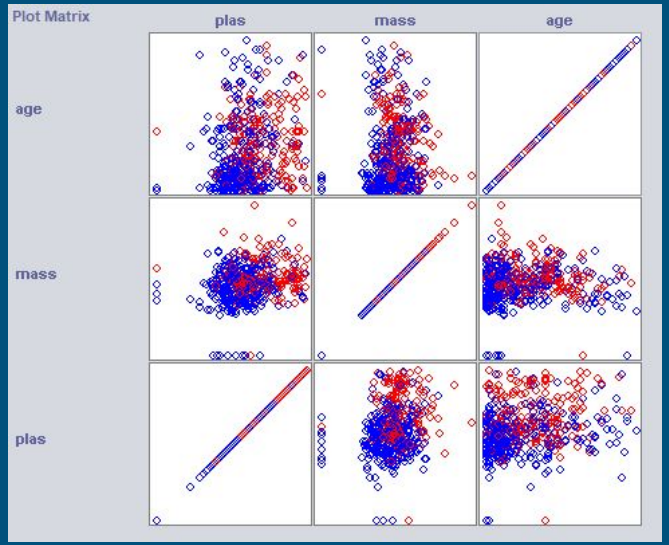


Jake

Observations?

Plot Matrix

- Focus on
  - Plasma concentration
  - Mass
  - Age
- No clear correlation between the values

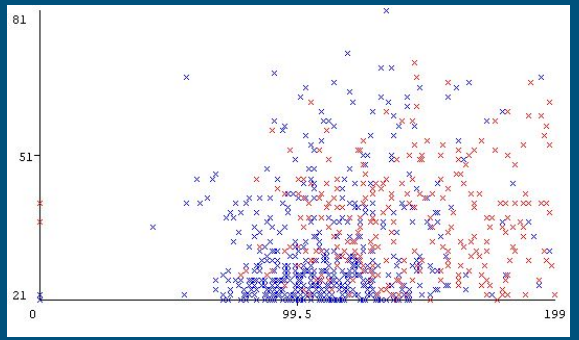


Ashley

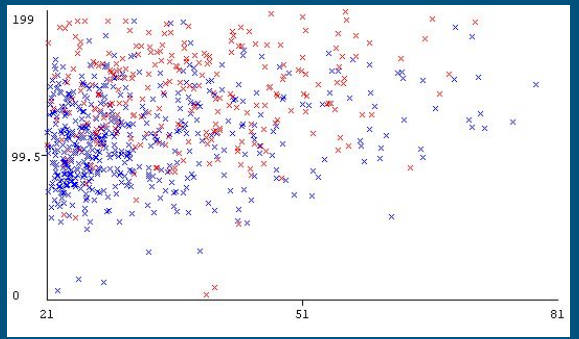
Observations?(cont.)

- Plas v. Age
  - Higher the patient's plasma the more likely the diagnosis
- Age v. Plas
  - The younger the patient and the higher plasma the likely the diagnosis
- Values cluster
  - Around 21-51 age range
  - Around 99.5 - 199 plasma range
    - Plasma average?

Plas v. Age



Age v. Plas

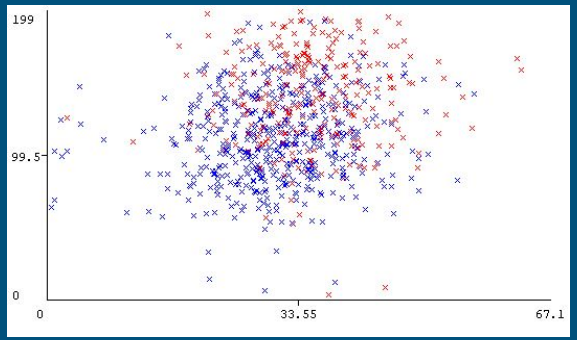


Ashley

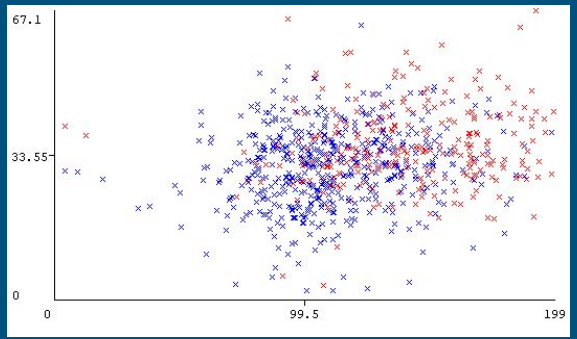
Observations?(cont.)

- Mass v. Plas
  - A range around a BMI 33.2 and Plasma 99.5 where no diagnosis
    - beyond Plasma 99.5
- Plas v. Mass
  - BMI alone is no clear indicator
  - Though, a higher BMI and Plasma the more likely a diagnosis
- Values cluster
  - Around 33.55 BMI range
    - BMI Average?

Mass v. Plas



Plas v. Mass



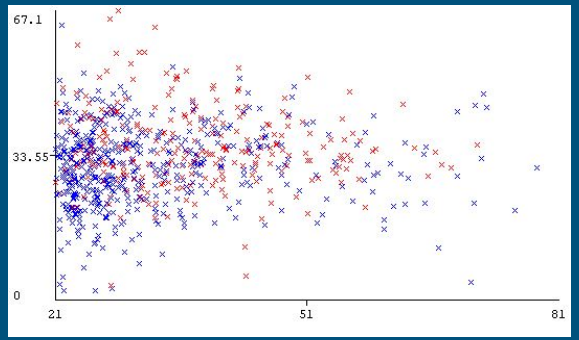
Ashley

Observations?(cont.)

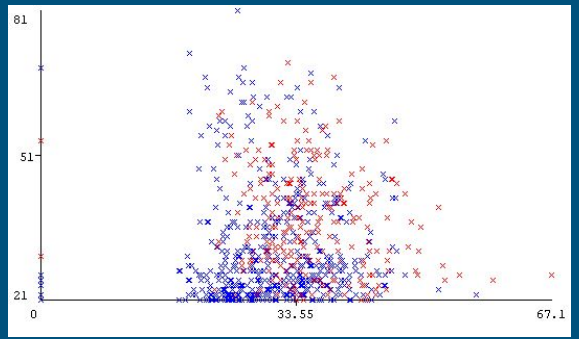
- Age v. Mass
  - The older someone is and the lower their BMI the less likely diagnosis

- Mass v. Age
  - No definite conclusions

Age v. Mass



Mass v. Age



Ashley

Summary

- Plasma concentration seems to be very significant to diabetes diagnosis
- Age and Mass seem to help supplement these conclusions significantly especially when combined with Plasma Concentration
- Though
  - The data clustered in some spots
    - Seems that there was more data on a specific groups
    - i.e. there was more data for the ages 21 - 51
  - It is important to look at wider and diverse data sets
  - It is important to look at ALL attributes that may assist with diagnosis

Conclusion!

THANK YOU FOR LISTENING