WEKA PRESENTATION

Thang “Summer” Tran

Agenda

- Diabetes Dataset
- Using WEKA to do diabetes Analysis

What is Diabetes?

- Diabetes is a chronic medical condition characterized by elevated blood sugar (glucose) levels.
- It occurs when the body cannot produce enough insulin or effectively use the insulin it produces.

Types of Diabetes

- Type 1 Diabetes:
  - Usually diagnosed in childhood or adolescence.
  - The immune system attacks and destroys insulin-producing cells in the pancreas.
  - Requires lifelong insulin injections or an insulin pump.
- Type 2 Diabetes:
  - Often diagnosed in adulthood, but increasingly found in children and adolescents.
  - The body becomes resistant to insulin or doesn't produce enough.
  - Managed through lifestyle changes, oral medications, and, in some cases, insulin.
- Gestational Diabetes:
  - Occurs during pregnancy and typically resolves after childbirth.
  - Women with gestational diabetes have a higher risk of developing Type 2 diabetes later in life.

Diabetes Dataset

- Original owners: **National Institute of Diabetes and Digestive and Kidney Diseases**
- Donor of the database: Vincent Sigillito, Applied Physics Laboratory, The Johns Hopkins University
- Date received: 9 May 1990

Using WEKA to do diabetes Analysis

SimpleKMeans

```
Clustered Instances

0       515 ( 67%)
1       253 ( 33%)


Class attribute: class
Classes to Clusters:

   0   1  <-- assigned to cluster
 380 120 | tested_negative
 135 133 | tested_positive

Cluster 0 <-- tested_negative
Cluster 1 <-- tested_positive

Incorrectly clustered instances :       255.0    33.2031 %
```

Using WEKA to do diabetes Analysis

MakeDensityBasedClusterer

```
Clustered Instances

0       498 ( 65%)
1       270 ( 35%)


Log likelihood: -29.34739


Class attribute: class
Classes to Clusters:

   0   1  <-- assigned to cluster
 372 128 | tested_negative
 126 142 | tested_positive

Cluster 0 <-- tested_negative
Cluster 1 <-- tested_positive

Incorrectly clustered instances :       254.0     33.0729 %
```

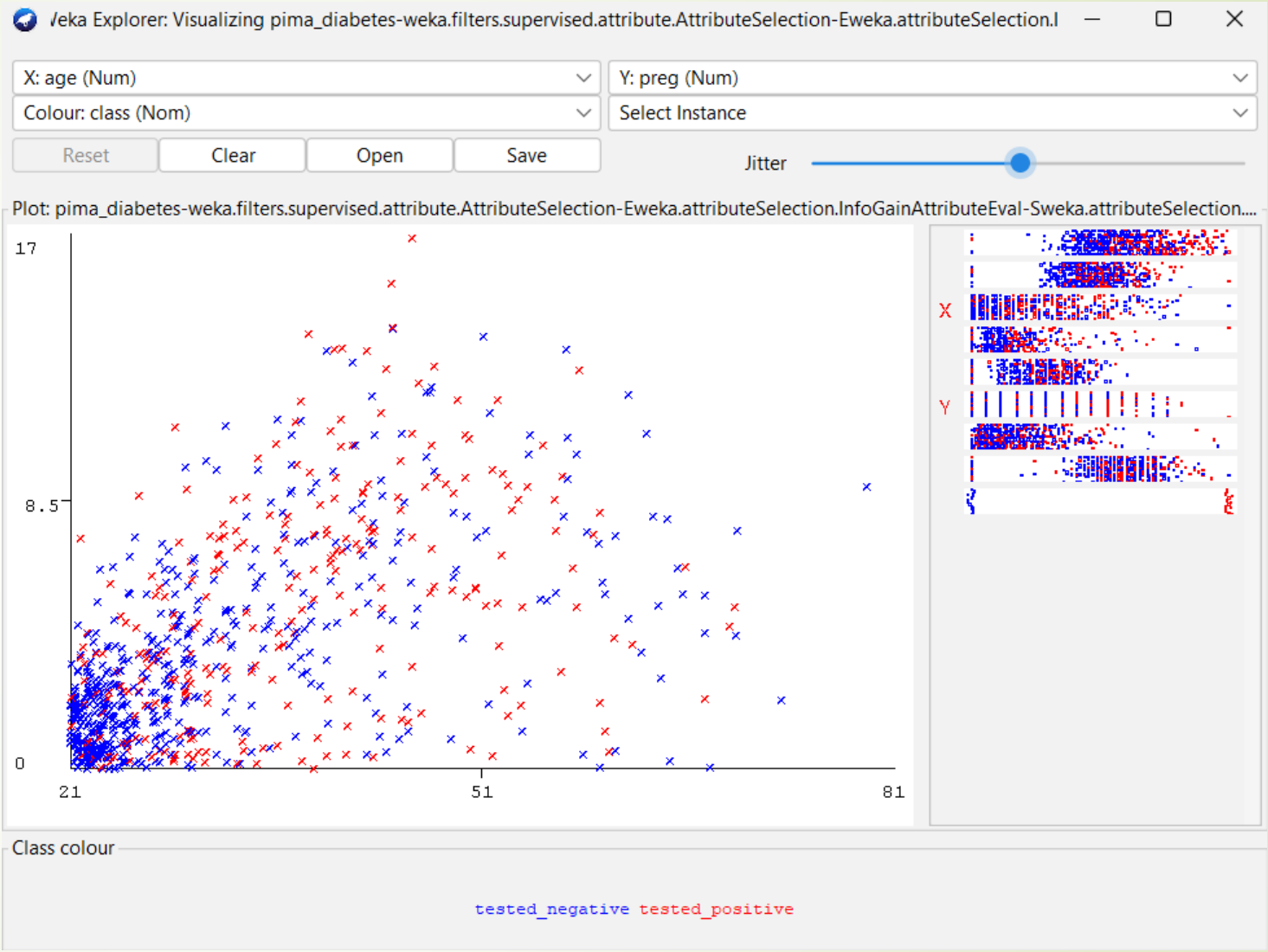
Training Data

- Most Accurate: LMT and SimpleLogistic
  - 79.4788 % Correctly
- ROC Area is 0.853 which is good AUC

```
=== Summary ===

Correctly Classified Instances         244               79.4788 %
Incorrectly Classified Instances        63               20.5212 %
Kappa statistic                          0.5093
Mean absolute error                      0.3192
Root mean squared error                  0.3821
Relative absolute error                 70.3178 %
Root relative squared error             80.5296 %
Total Number of Instances              307

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.921    0.448    0.798      0.921    0.855      0.525    0.853     0.899     tested_negative
                 0.552    0.079    0.784      0.552    0.648      0.525    0.853     0.778     tested_positive
Weighted Avg.    0.795    0.322    0.793      0.795    0.784      0.525    0.853     0.858

=== Confusion Matrix ===

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

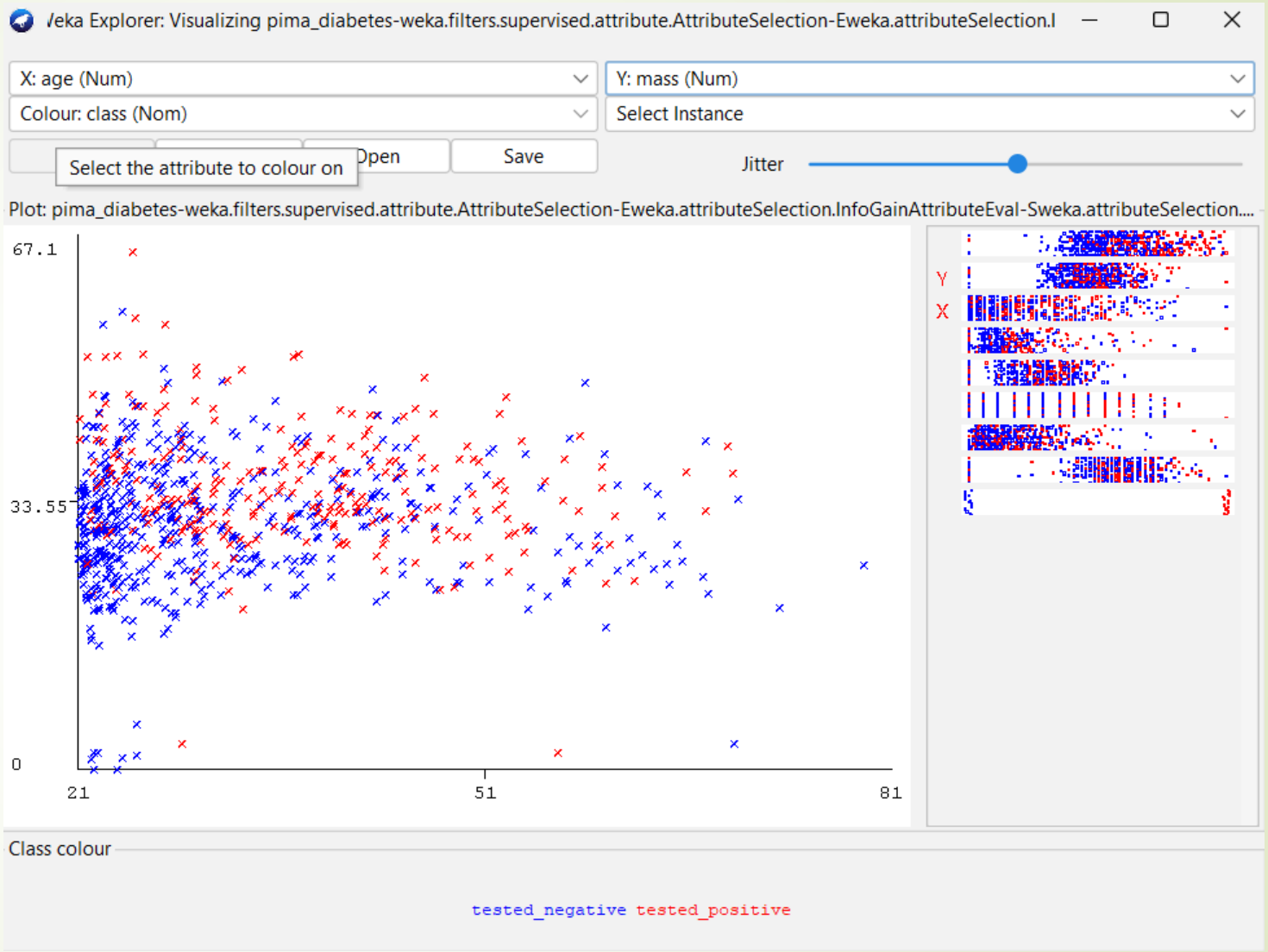
Relation between preg and Age

- The percentage tested positive increased when the older got pregnant
- Mostly higher chance for tested positive when pregnant after 30 years old



Relation between mass and age

- Average weight is 33.55
- People most likely tested positive with diabetes when they get older (30-50) or overweight



Thank you