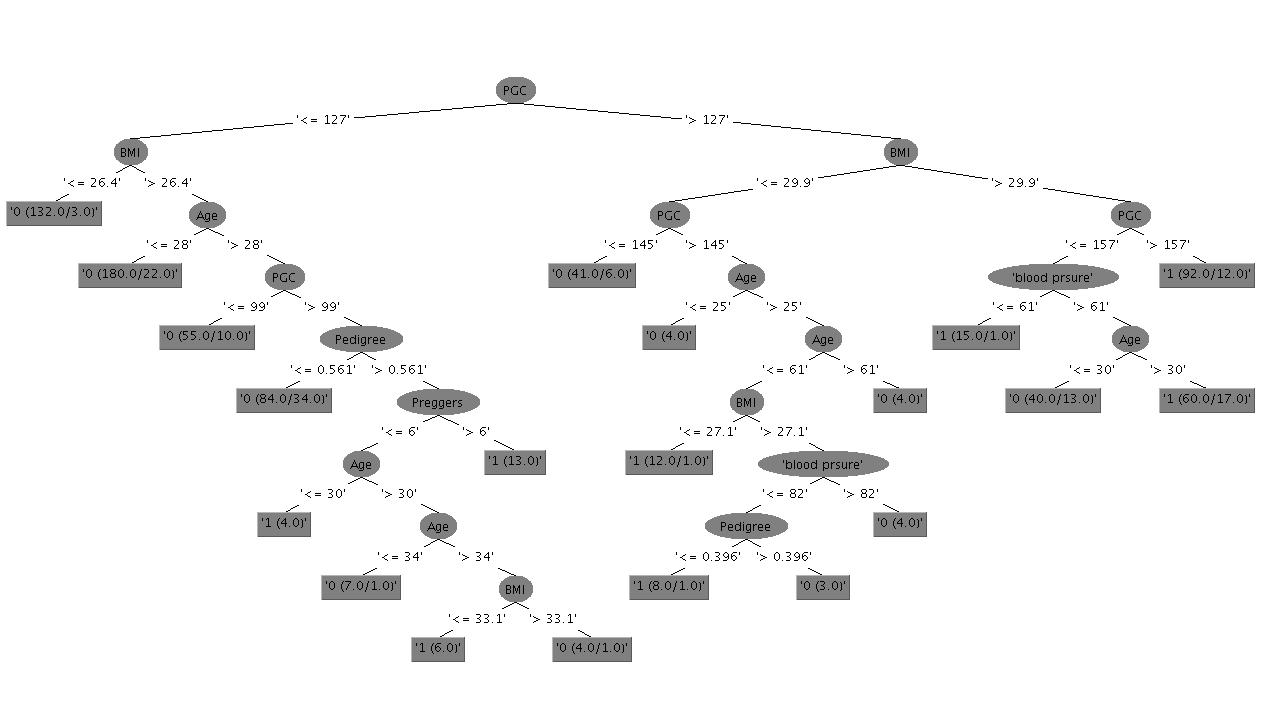
1. Draw part of the tree (enough so that you understand the representation used). Write down at least 2 different IF THEN rules that could be derived from this tree.



IF PCG <= 127

&& BMI <=26.4

THEN !Diabetes

IF PGC >127

&& BMI <= 29.9

&& PGC >145

&& AGE > 25

&& BMI <= 27.1

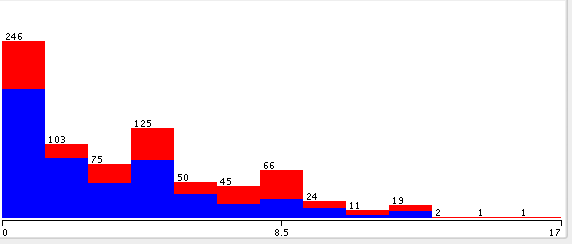
THEN Diabetes

2. From your earlier attribute analysis, record the number of examples of each class and a brief description of each attribute’s value spread (from the histogram).

Preggers

The results appear in a falling curve

As pregnancies increases the probability of diabetes increases

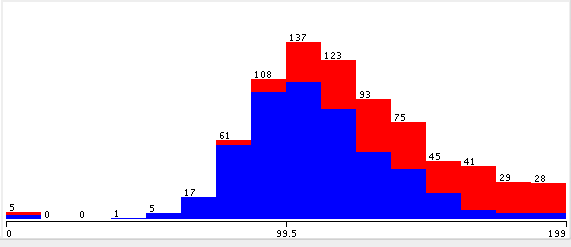


Plasma Glucose

The results appear in a bell curve in the middle of the graph

As the plasma glucose increases, the probability of having diabetes also increases

The lower the plasma glucose, the less chance of having diabetes

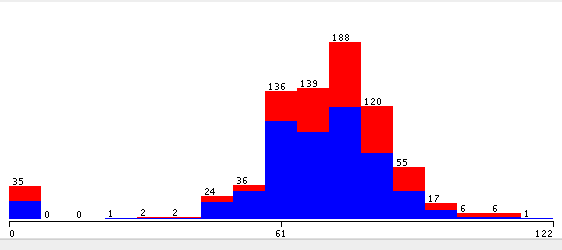


Blood Pressure

Results appear in a bell curve in the middle of the result set.

35 Results with 0 blood pressure, look to be invalid data

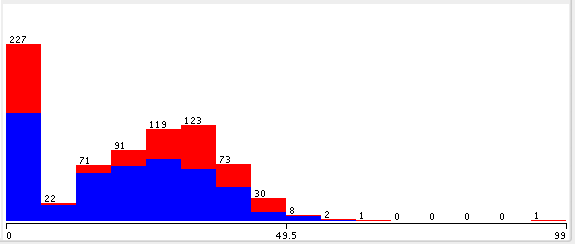
The majority of each range is spread about 1/3 towards not having diabetes



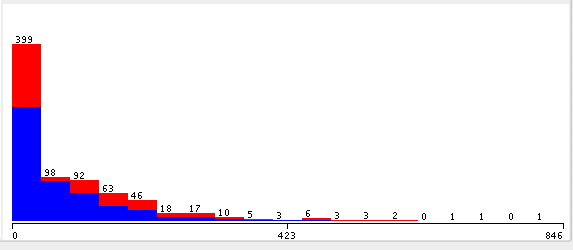
Skin Thickness

There is one piece of invalid data

The spread of !diabetes : diabetes is about 60:40 at the lower end of the graph



Insulin

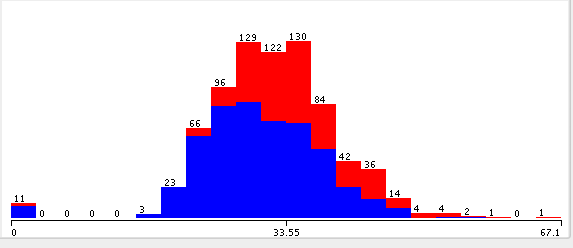


BMI

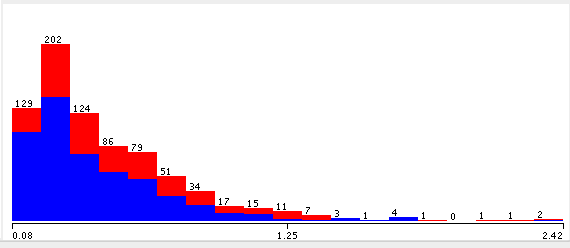
Results appear in a bell curve in the middle of the result set

11 results with a 0 BMI which is most likely invalid data

The higher the BMI the higher probability of having diabetes



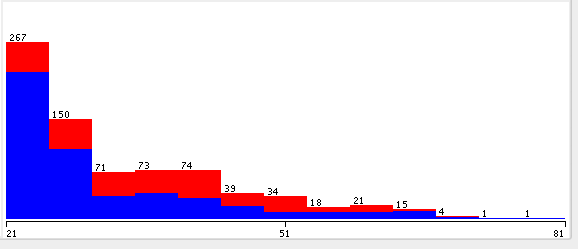
Pedigree



Age

The amount of subjects with diabetes decreases as the age increases

At the lower end of the results there is a spread of about 1/6 to not having diabetes



3. How many leaves are there? 20

4. Does the algorithm make use of the same attribute to split on at different levels?

Yes – It has split on the same attributes in multiple places (Such as BMI)

5. What was the percentage accuracy on the test data set? 76.2452%

1. How long did the algorithm take to build the tree? 0.13s