Aprendizagem 2024/25

Homework I – Group 98

(João Pinho - ist1102765, Tiago Sousa - ist1110691)

Part I: Pen and paper

1. Decision tree

   In order to complete the decision tree, we must select the variable that has the most information gain in the subset of D where $y_1 >= 0.3$.

y_1	y_2	y_3	y_4	y_{out}
0.3	0	1	0	B
0.76	0	1	1	A
0.86	1	0	0	A
0.93	0	1	1	C
0.47	0	1	1	C
0.73	1	0	0	A
0.89	1	2	0	B

   Considering X as the dataset and Z the target variable, information gain is calculated as follows:

$$IG(y) = H(Z) - H(Z|y)$$
$$H(Z|y) = \sum_i p(X_i)H(Z|X_i)$$

   Applying these formulas to our dataset, we get:

$$H(y_{out}) = \frac{3}{7}\log_2\left(\frac{7}{3}\right) + \frac{4}{7}\log_2\left(\frac{7}{2}\right) = 1.56$$
$$IG(y_2) = H(y_{out}) - H(y_{out}|y_2) = 1.56 - \left(\frac{2}{7} - \frac{2}{7}\log_2\left(\frac{1}{4}\right) - \frac{2}{7}\log_2\left(\frac{2}{3}\right) - \frac{1}{7}\log_2\left(\frac{1}{3}\right)\right) = 0.31$$
$$IG(y_3) = H(y_{out}) - H(y_{out}|y_3) = 1.56 - \left(0 + \frac{2}{7} - \frac{2}{7}\log_2\left(\frac{1}{4}\right) + 0\right) = 0.70$$
$$IG(y_4) = H(y_{out}) - H(y_{out}|y_4) = 1.56 - \left(\frac{4}{7} - \frac{1}{7}\log_2\left(\frac{1}{3}\right) - \frac{2}{7}\log_2\left(\frac{2}{3}\right)\right) = 0.59$$

For y_1, we have to consider every possible split:

$$split(0.4) : IG(y_1) = 1.56 - \left(-\frac{3}{7}\log_2\left(\frac{1}{2}\right) - \frac{2}{7}\log_2\left(\frac{1}{3}\right) - \frac{1}{7}\log_2\left(\frac{1}{6}\right)\right) = 0.31$$

$$split(0.5) : IG(y_1) = 1.56 - \left(\frac{2}{7} - \frac{3}{7}\log_2\left(\frac{3}{5}\right) - \frac{2}{7}\log_2\left(\frac{1}{5}\right)\right) = 0.29$$

$$split(0.74) : IG(y_1) = 1.56 - \left(-\frac{3}{7}\log_2\left(\frac{1}{3}\right) - \frac{2}{7}\log_2\left(\frac{1}{2}\right) - \frac{2}{7}\log_2\left(\frac{1}{4}\right)\right) = 0.02$$

Note that the pairs of splits (0.4 e 0.9, 0.5 e 0.87, 0.74 e 0.8) are equivalent because they represent the same class distributions.

The variable that maximizes IG is y_3, and so it is the next node of the decision tree. Note that in our subset of D, when $y_3 = 0$, y_{out} is always A, and $y_3 = 2$ only happens once, where y_{out} is B. These are leaf nodes in our tree. There are 4 instances remaining when $y_3 = 1$, which we further analyse:

y_1	y_2	y_3	y_4	y_{out}
0.3	0	1	0	B
0.76	0	1	1	A
0.93	0	1	1	C
0.47	0	1	1	C

$$H(y_{out}) = -\frac{1}{2}\log_2\left(\frac{1}{4}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right) = 1.5$$

$$IG(y_2) = 0$$

$$IG(y_3) = 0$$

$$IG(y_4) = 1.5 - \left(-\frac{1}{4}\log_2\left(\frac{1}{3}\right) - \frac{1}{2}\log_2\left(\frac{2}{3}\right)\right) = 0.81$$

For y_1, possible splits are:

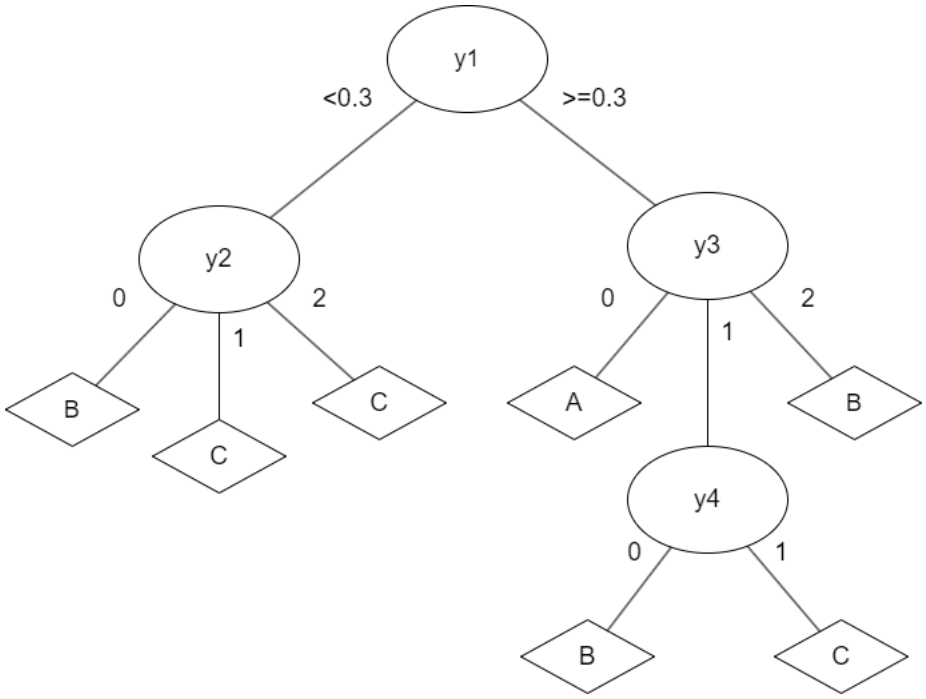
$$split(0.4) : IG(y_1) = 1.5 - \left(-\frac{1}{4}\log_2\left(\frac{1}{3}\right) - \frac{1}{2}\log_2\left(\frac{2}{3}\right)\right) = 0.81$$

$$split(0.5) : IG(y_1) = 1.5 - \left(-\frac{1}{2}\log_2\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right)\right) = 0.5$$

$$split(0.8) : IG(y_1) = 1.5 - \left(-\frac{3}{4}\log_2\left(\frac{1}{3}\right)\right) = 0.31$$

Both y_1 split at 0.4 and y_4 maximize the IG, so we'll choose y_4 to become the next node of the tree. There's only one instance where $y_4 = 0$, which will be a leaf node of class B, and other 3 instances which are split in two C classes and one A, which will be a leaf node of class C.

The complete decision tree



2. Confusion Matrix

The following table shows the real labels of the observations of our dataset and the labels predicted by our model.

Observation	y_{out}	$\widehat{y_{out}}$
x_1	C	C
x_2	B	B
x_3	C	C
x_4	B	B
x_5	C	C
x_6	B	B
x_7	A	C
x_8	A	A
x_9	C	C
x_{10}	C	C
x_{11}	A	A
x_{12}	B	B

The confusion matrix allows us to compare our classifier model's predictions against the ground truth. The columns refer to the real labels of an observation and the lines to the prediction.

$$\text{Confusion Matrix} = \begin{array}{c} \\ A \\ B \\ C \end{array} \begin{array}{c} \begin{array}{ccc} A & B & C \end{array} \\ \begin{pmatrix} 2 & 0 & 0 \\ 0 & 4 & 0 \\ 1 & 0 & 5 \end{pmatrix} \end{array}$$

With this matrix, we can see our decision tree has correctly labeled all instances of class B, and only mislabelled one A as a C. Therefore, there were 11 True Positives, and one A False Negative/ C False Positive.

3. F1 measures

   The formulas for Precision, Recall and the F1-measure are as follows:

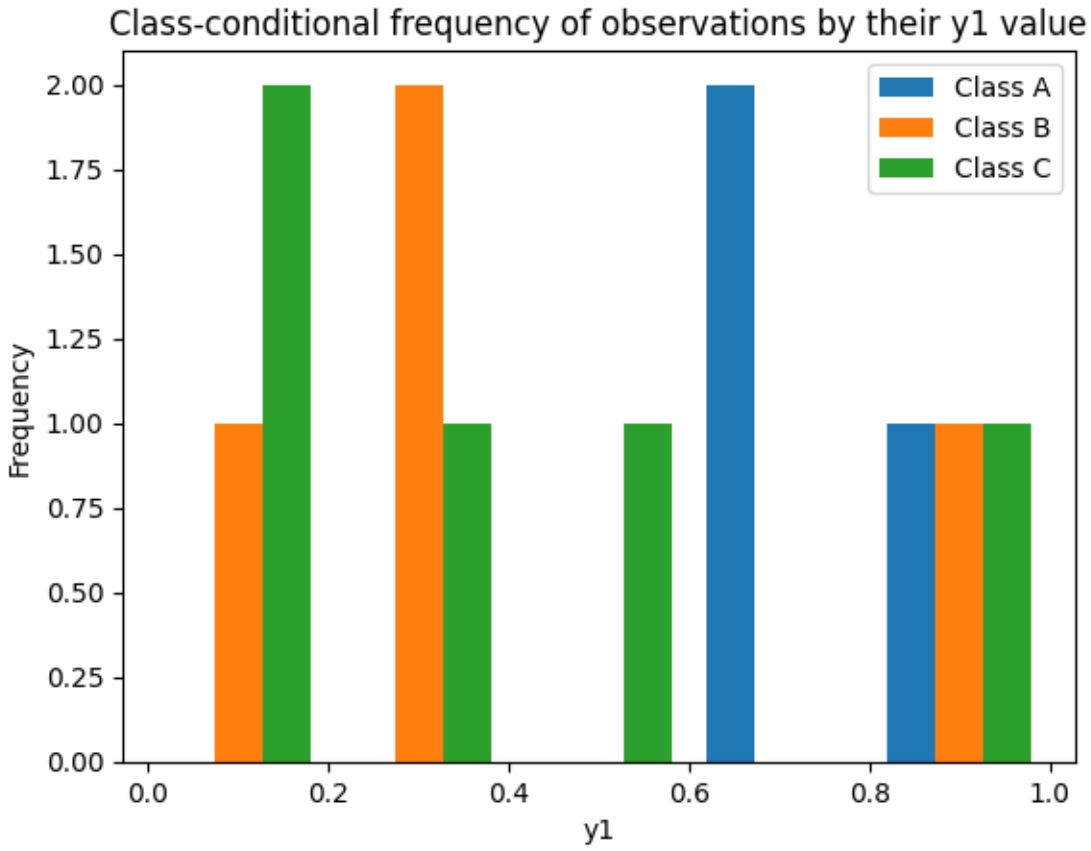
$$Precision = \frac{TP}{TP + FP}$$
$$Recall = \frac{TP}{TP + FN}$$
$$\frac{1}{F_1} = \frac{1}{2}\left(\frac{1}{P} + \frac{1}{R}\right)$$

   Applying these to our data, we get the following:

$$Precision(A) = 1 = Precision(B)$$
$$Precision(C) = \frac{5}{6}$$
$$Recall(A) = \frac{2}{3}$$
$$Recall(B) = 1 = Precision(C)$$
$$\frac{1}{F_{1A}} = \frac{1}{2}\left(1 + \frac{3}{2}\right) \Rightarrow F_{1A} = \frac{4}{5}$$
$$\frac{1}{F_{1B}} = \frac{1}{2}\left(1 + 1\right) \Rightarrow F_{1B} = 1$$
$$\frac{1}{F_{1C}} = \frac{1}{2}\left(\frac{6}{5} + 1\right) \Rightarrow F_{1C} = \frac{10}{11}$$

   Class A has the lowest F_1 score.

4. Class-conditional histograms



We found that in each of the four bins from 0 to 0.8, there is one class with a higher frequency. For these bins, we chose that class as the classification of a new observation in that bin. For the bin in the interval [0.8,1.0], as all the classes were tied, we chose the class that comes first by alphabetical order (A). The following function classifies an observation according the discriminant rules derived from these histograms:

$$\widehat{y_{out}}(y_1) = \begin{cases} \text{A if } 0.6 \leq y_1 \leq 1.0 \\ \text{B if } 0.2 \leq y_1 < 0.4 \\ \text{C if } 0.0 \leq y_1 < 0.2 \vee 0.4 \leq y_1 < 0.6 \end{cases}$$

Part 2: Programming

1. Discriminative Power and Class conditional-probability distributions

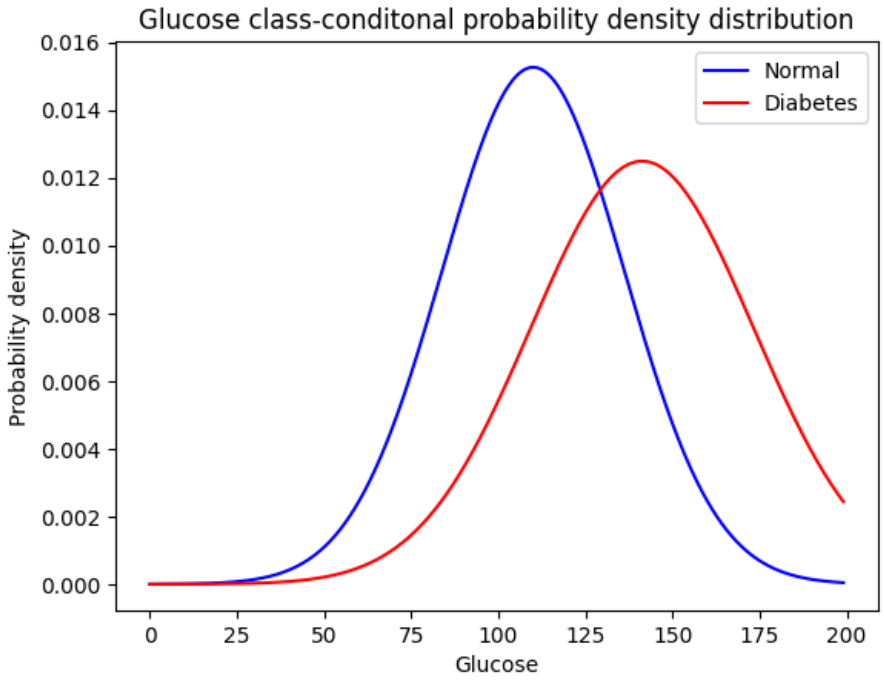
   We obtained the following results on the Analysis of Variance statistical test (we divided the results in two tables to ensure they are readable):

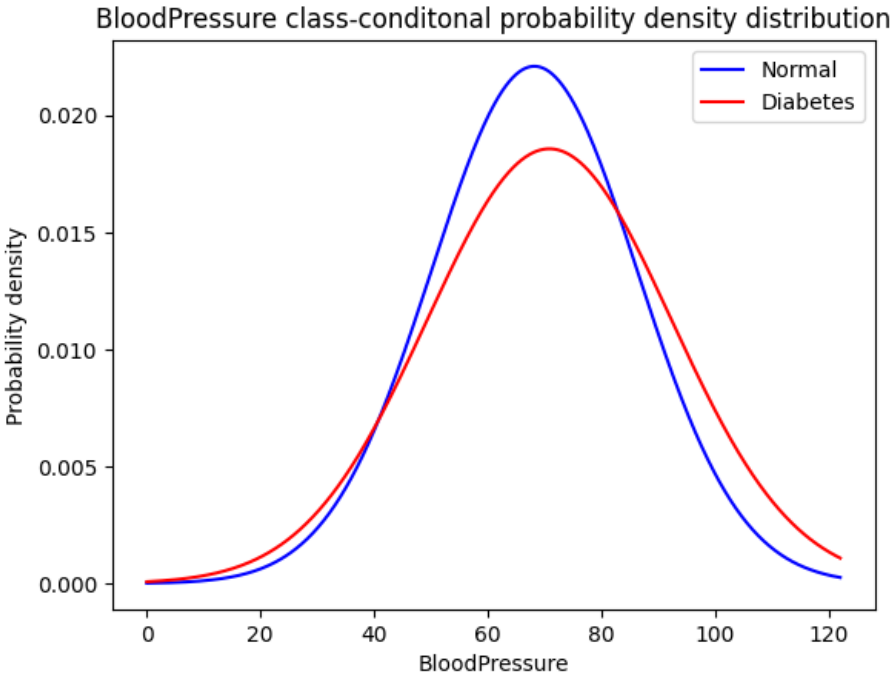
Features	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI
f-values	39.670	213.162	3.257	4.304	13.281	71.772
p-values	5.065E-10	8.935E-43	7.151E-02	3.835E-02	2.862E-04	1.230E-16

Features	DiabetesPedigreeFunction	Age
f-values	23.871	46.141
p-values	1.255E-06	2.210E-11

   The Glucose variable is the one showing the highest f-value in the Analysis of Variance statistical test, and hence it has the best discriminative power. The low p-value supports the statistical significance of this result. The BloodPressure variable shows the smallest f-value, and hence has the worst discriminative power.

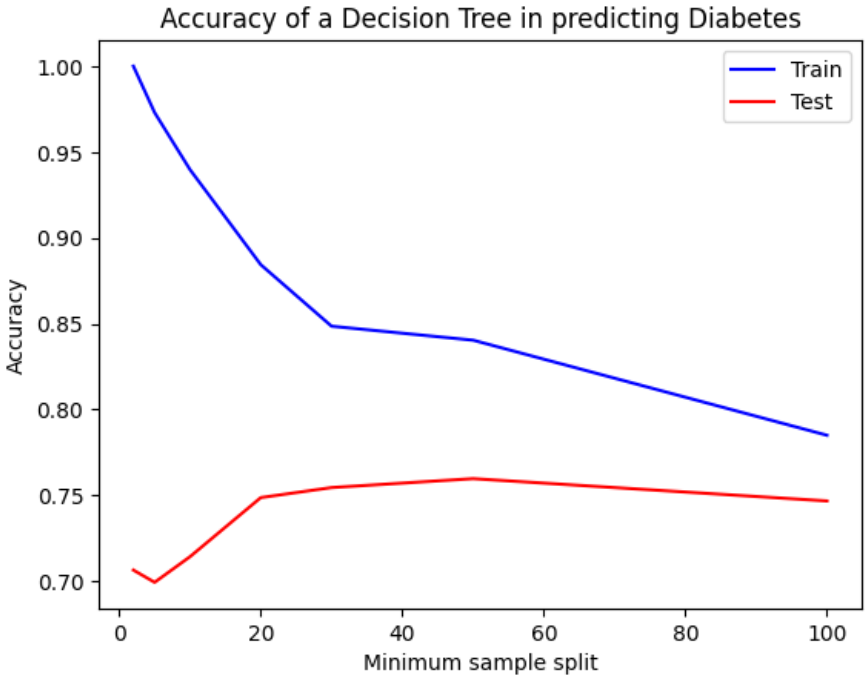
   We then divided the dataset by class (Normal vs Diabetes) and plotted the normal distributions of the Glucose and BloodPressure variables for each class.





2. Decision Tree accuracy

   We trained and tested a decision tree with a 80-20 train-test split for each minimum sample split value and averaged the accuracies for each value over 10 runs.

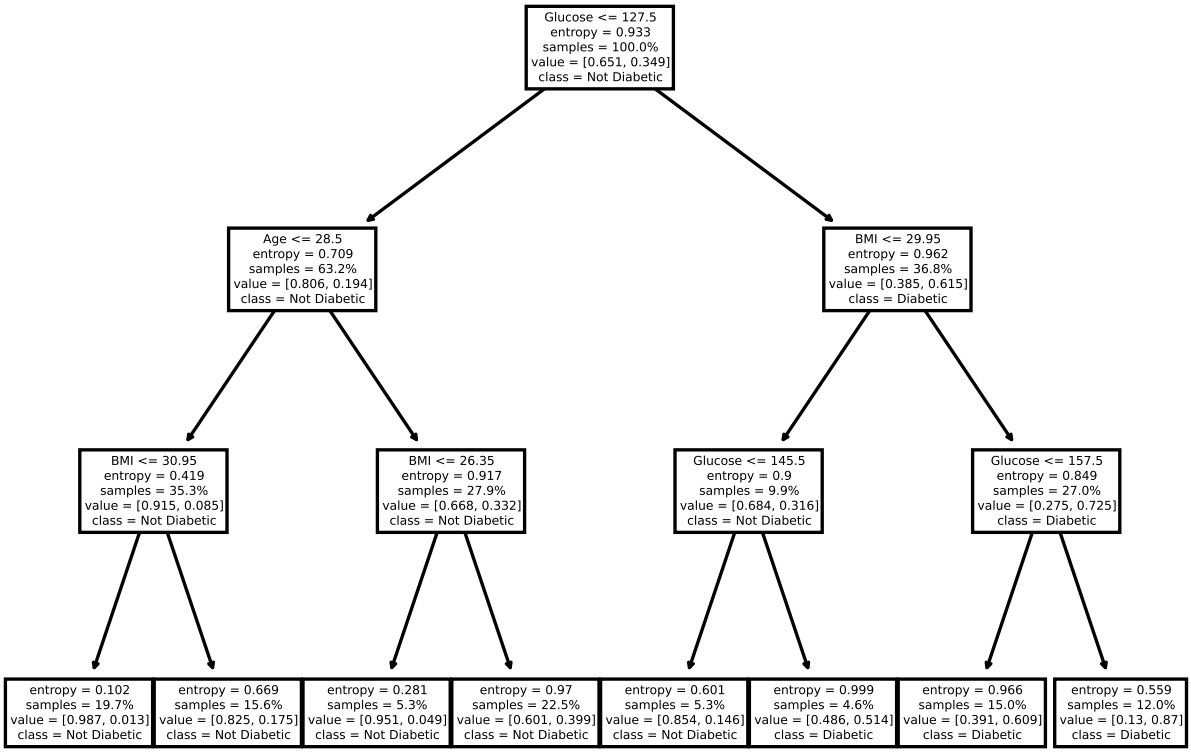


3. Critical analysis of results

We can observe that the testing accuracy increased as the minimum sample split number decreased from 100 to 50, suggesting the model was underfitted for minimum sample split numbers bigger than 50. A lower value of minimum sample split means the model will need a smaller amount of observations in a node to create a split, and hence the tree will have more nodes and leaves and will be more specialized / less general, therefore reducing the underfitting. However, the testing accuracy decreased and the training accuracy increased as the minimum sample splits decreased from 50 to 2, suggesting the model was overfitting. A small minimum sample split number leads to a big number of nodes and leaves and therefore a model too specialized to the training dataset.

Hence, for this Decision Tree Classifier and for this dataset, among the tested values, a minimum sample split of 50 is the best to balance the tree generalization capacity and the test accuracy.

4. **Plot of the decision tree**



ii) The initial probability of having diabetes in the dataset ($p(d)$) is 0.349. Considering glucose levels(g), age(a) and BMI(b), the plot shows us the following posterior probabilities of having diabetes given the symptoms:

$$
\begin{aligned}
p(d|g \leq 127.5) &= 0.194 \\
p(d|g \leq 127.5 \wedge a \leq 28.5) &= 0.085 \\
p(d|g \leq 127.5 \wedge a \leq 28.5 \wedge b \leq 30.95) &= 0.013 \\
p(d|g \leq 127.5 \wedge a \leq 28.5 \wedge b > 30.95) &= 0.175 \\
p(d|g \leq 127.5 \wedge a > 28.5) &= 0.332 \\
p(d|g \leq 127.5 \wedge a > 28.5 \wedge b \leq 26.35) &= 0.049 \\
p(d|g \leq 127.5 \wedge a > 28.5 \wedge b > 26.35) &= 0.399 \\
p(d|g > 127.5) &= 0.615 \\
p(d|g > 127.5 \wedge b \leq 29.95) &= 0.316 \\
p(d|145.5 \geq g > 127.5 \wedge b \leq 29.95) &= 0.146 \\
p(d|g > 145.5 \wedge b \leq 29.95) &= 0.514 \\
p(d|g > 127.5 \wedge b > 29.95) &= 0.725 \\
p(d|157.5 \geq g > 127.5 \wedge b > 29.95) &= 0.609 \\
p(d|g > 157.5 \wedge b > 29.95) &= 0.870
\end{aligned}
$$

Therefore, the most significant diabetes characteristic features are, in order from highest to lowest chance of having diabetes:

- Glucose levels above 157.5 and BMI above 29.95 $\Rightarrow$ 0.870
- Glucose levels between 127.5 and 157.5 and BMI below 29.95 $\Rightarrow$ 0.609
- Glucose levels above 145.5 and BMI below 29.95 $\Rightarrow$ 0.514
- Glucose levels below 127.5, age above 28.5 and BMI above 26.35 $\Rightarrow$ 0.399
- Glucose levels below 127.5, age below 28.5 and BMI above 30.95 $\Rightarrow$ 0.175
- Glucose levels between 127.5 and 145.5 and BMI below 29.95 $\Rightarrow$ 0.146
- Glucose levels below 127.5, age above 28.5 and BMI below 26.35 $\Rightarrow$ 0.049
- Glucose levels below 127.5, age below 28.5 and BMI below 30.95 $\Rightarrow$ 0.013