Homework 02

Gabriele Lorenzo
s314913

December 2022

Answer #1:

Using the whole dataset to train the Decision Tree, setting the default configuration and setting the minimal gain to 0.01, we can analyze that:

(a)

As we can see in Figure 1:, the most discriminative attribute for class prediction is the *Node-caps* attribute. Said attribute has been divided in its three values: **"yes"**, **"no"** and **"?"**.

(b)

The height of the generated Decision Tree is 6. We recall that: the height of a tree is defined as the maximum depth of any leaf node from the root node. That is, the length of the longest path (links traversed) from the root node *(Node-Caps)* to any leaf node.

(c)

There are multiple examples of partial pure partitions, as we can see in Figure 2. In general there is a pure partition when all the data in the partition is correctly classified. In other words: if all the elements are accurately divided in different classes, that's a pure split.

Answer #2:

In the default configuration, the Decision Tree uses the *gain ratio splitting criterion*. Using this criterion, the algorithm computes the gain value before each partition. If the calculated gain value is higher than the *minimal gain* value (configurated in the parameters of the Decision Tree), the node is partitioned.

Another important value is the *maximal depth* value. This value represents the maximal depth that the tree can grow to. In pratical terms, it represents the maximum number of ”vertical” nodes of the tree.

Some examples of different parameters configurations: Figure 3, Figure 4, Figure 5, Figure 6, Figure 7.

Answer #3:

Here are some examples of the confusion matrices with different configurations of the Decision Tree model: Figure 8, Figure 9, Figure 10, Figure 11, Figure 12, Figure 13.

As we can see, the model with the highest accuracy (Figure 13) is the one with balanced values of *minimal gain* and *maximal depth*.

Answer #4:

Here are some examples of the confusion matrices with different configurations of the K-NN model: Figure 14, Figure 15, Figure 16, Figure 17, Figure 18, Figure 19.

As we can see, the model has an average accuracy above 70%, with the highest value for k=9.

Here is the confusion matrix for the Naive Bayes model: Figure 20.

If we compute the average accuracy value for the K-NN model with the different configurations, we get an average accuracy value of: 72.32, which is sligtly lower than the accuracy value of the Naive Bayes model (72.45). In conclusion, the Naive Bayes model perform better on average with respect to the K-NN model.

Answer #5:

Here is the correlation matrix of the attributes: Figure 21.

(a)

The Naïve independence assumption does not hold for the Breast dataset. In fact we can see that the highest absolute correlation value is 0.465, between the *inv-nodes* attribute and the *node-caps* attribute (quite high compared to the maximum theoretical value of 1).

(b)

The pair of most correlated attributes are *inv-nodes* and *node-caps* with a correlation value of -0.465.

Appendix:

Figure 1: Decision Tree

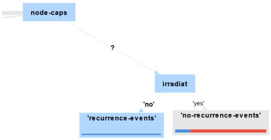


Figure 2: Example of left pure partition

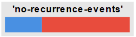


Figure 3: Decision Tree with minimal gain=0.01 and maximal depth=1

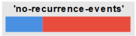


Figure 4: Decision Tree with minimal gain=0.1 and maximal depth=10

Figure 5: Decision Tree with minimal gain=0.03 and maximal depth=4

Figure 6: Decision Tree with minimal gain=0.05 and maximal depth=10

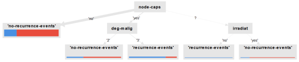
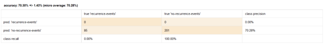


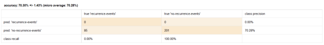
Figure 7: Decision Tree with minimal gain=0.06 and maximal depth=3



accuracy: 70.30% +/- 1.43% (micro average: 70.28%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	0	0	0.00%
pred. 'no-recurrence-events'	85	201	70.28%
class recall	0.00%	100.00%	

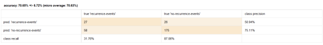
Figure 8: Confusion Matrix with minimal gain=0.01 and maximal depth=1



accuracy: 70.30% +/- 1.43% (micro average: 70.28%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	0	0	0.00%
pred. 'no-recurrence-events'	85	201	70.28%
class recall	0.00%	100.00%	

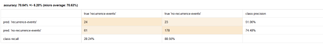
Figure 9: Confusion Matrix with minimal gain=0.1 and maximal depth=10



accuracy: 70.65% +/- 6.72% (micro average: 70.63%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	27	26	50.94%
pred. 'no-recurrence-events'	58	175	75.11%
class recall	31.76%	87.06%	

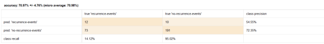
Figure 10: Confusion Matrix with minimal gain=0.03 and maximal depth=4



accuracy: 70.64% +/- 6.20% (micro average: 70.63%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	24	23	51.06%
pred. 'no-recurrence-events'	61	178	74.48%
class recall	28.24%	88.56%	

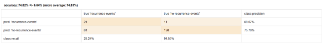
Figure 11: Confusion Matrix with minimal gain=0.05 and maximal depth=10



accuracy: 70.97% +/- 4.76% (micro average: 70.98%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	12	10	54.55%
pred. 'no-recurrence-events'	73	191	72.35%
class recall	14.12%	95.02%	

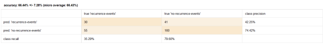
Figure 12: Confusion Matrix with minimal gain=0.06 and maximal depth=3



accuracy: 74.82% +/- 6.64% (micro average: 74.83%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	24	11	68.57%
pred. 'no-recurrence-events'	61	190	75.70%
class recall	28.24%	94.53%	

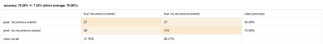
Figure 13: Confusion Matrix with minimal gain=0.04 and maximal depth=3



accuracy: 66.44% +/- 7.28% (micro average: 66.43%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	30	41	42.25%
pred. 'no-recurrence-events'	55	160	74.42%
class recall	35.29%	79.60%	

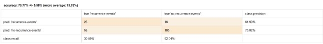
Figure 14: Confusion Matrix with k=1



accuracy: 70.26% +/- 7.23% (micro average: 70.28%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	27	27	50.00%
pred. 'no-recurrence-events'	58	174	75.00%
class recall	31.76%	86.57%	

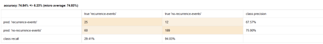
Figure 15: Confusion Matrix with k=3



accuracy: 73.77% +/- 5.98% (micro average: 73.78%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	26	16	61.90%
pred. 'no-recurrence-events'	59	185	75.82%
class recall	30.59%	92.04%	

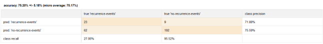
Figure 16: Confusion Matrix with k=5



accuracy: 74.84% +/- 6.23% (micro average: 74.83%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	25	12	67.57%
pred. 'no-recurrence-events'	60	189	75.90%
class recall	29.41%	94.03%	

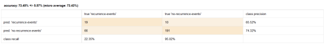
Figure 17: Confusion Matrix with k=7



accuracy: 75.20% +/- 5.18% (micro average: 75.17%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	23	9	71.88%
pred. 'no-recurrence-events'	62	192	75.59%
class recall	27.06%	95.52%	

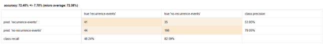
Figure 18: Confusion Matrix with k=9



accuracy: 73.45% +/- 5.57% (micro average: 73.43%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	19	10	65.52%
pred. 'no-recurrence-events'	66	191	74.32%
class recall	22.35%	95.02%	

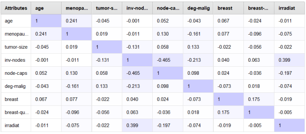
Figure 19: Confusion Matrix with k=11



accuracy: 72.45% +/- 7.70% (micro average: 72.38%)

	true 'recurrence-events'	true 'no-recurrence-events'	class precision
pred. 'recurrence-events'	41	35	53.95%
pred. 'no-recurrence-events'	44	166	79.05%
class recall	48.24%	82.59%	

Figure 20: Confusion Matrix with Naive Bayes model



Attributes	age	menopa...	tumor-s...	inv-nod...	node-ca...	deg-malig	breast	breast-...	irradiat
age	1	0.241	-0.045	-0.001	0.052	-0.043	0.067	-0.024	-0.011
menopau...	0.241	1	0.019	-0.011	0.130	-0.161	0.077	-0.096	-0.075
tumor-size	-0.045	0.019	1	-0.131	0.058	0.133	-0.022	-0.056	-0.022
inv-nodes	-0.001	-0.011	-0.131	1	-0.465	-0.213	0.040	0.063	0.399
node-caps	0.052	0.130	0.058	-0.465	1	0.098	0.024	-0.036	-0.197
deg-malig	-0.043	-0.161	0.133	-0.213	0.098	1	-0.073	0.018	-0.074
breast	0.067	0.077	-0.022	0.040	0.024	-0.073	1	0.175	-0.019
breast-qu...	-0.024	-0.096	-0.056	0.063	-0.036	0.018	0.175	1	-0.005
irradiat	-0.011	-0.075	-0.022	0.399	-0.197	-0.074	-0.019	-0.005	1

Figure 21: Correlation Matrix