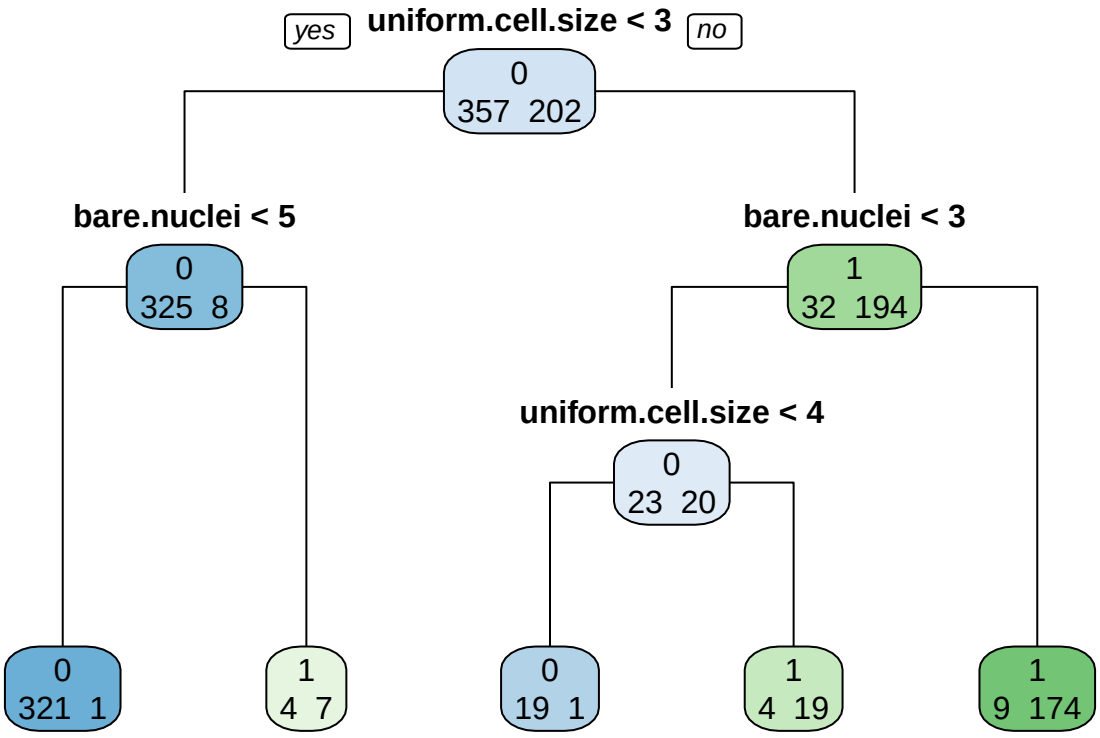
Decision Trees - Breast Cancer Biopsy Data

Decision trees are a Machine Learning method that is underrated, because of its (usually) low accuracy.

Although, its explanatory power can be very good to explore problems and gain new insights. Let's see this example, with data from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg. He assessed biopsies of breast tumours for 699 patients up to 15 July 1992. (R Documentation, 2019)



Interpreting results:

All branches for the left side are the "yes" answer for the group dividing question. "no" for the right. The numbers inside each cell are the total of benign (left side) and malignant (right side)

1. It is well known that non-uniformity of cell sizes and shapes are correlated with malignancy
2. The presence of a dual population of epithelial and myoepithelial cells and of numerous bare nuclei within a breast aspirate is generally indicative of a benign lesion. (McCluggage et al, 1997)
3. Chromatin alterations are also suggestive of malignancy. The bigger bland.chromatin, higher the probability of malignancy

One interesting conclusion from the model: Tumor samples with non-uniform cell sizes, non-uniform shapes and low number of bare nuclei are ~ 20x more likely to be malignant.

Testing Model Parameters:

```
# TESTING & EVALUATING MODEL:

# testing predictions on the model:
t_pred <- predict(mytree,test,type="class")
t <- test['malignant']

#building a confusion matrix
confMat <- table(test$malignant,t_pred)

print('Confusion Matrix:')
```

```
## [1] "Confusion Matrix:"
```

```
print(confMat)
```

```
##    t_pred
##      0  1
##   0 92  9
##   1  2 37
```

```
# getting model accuracy:
accuracy <- sum(diag(confMat))/sum(confMat)
print('model accuracy (in %):')
```

```
## [1] "model accuracy (in %):"
```

```
print(accuracy) # 0.9285714 -> 92.8%  Not bad!!!
```

```
## [1] 0.9214286
```

We have a confusion matrix which shows that our model got 5 false positives and 5 false negative results, for a total of 140 testing samples.

In terms of success metrics, it means that the model has an accuracy of 92.8%.

Find the source code for this report at https://git.io/fh7qR