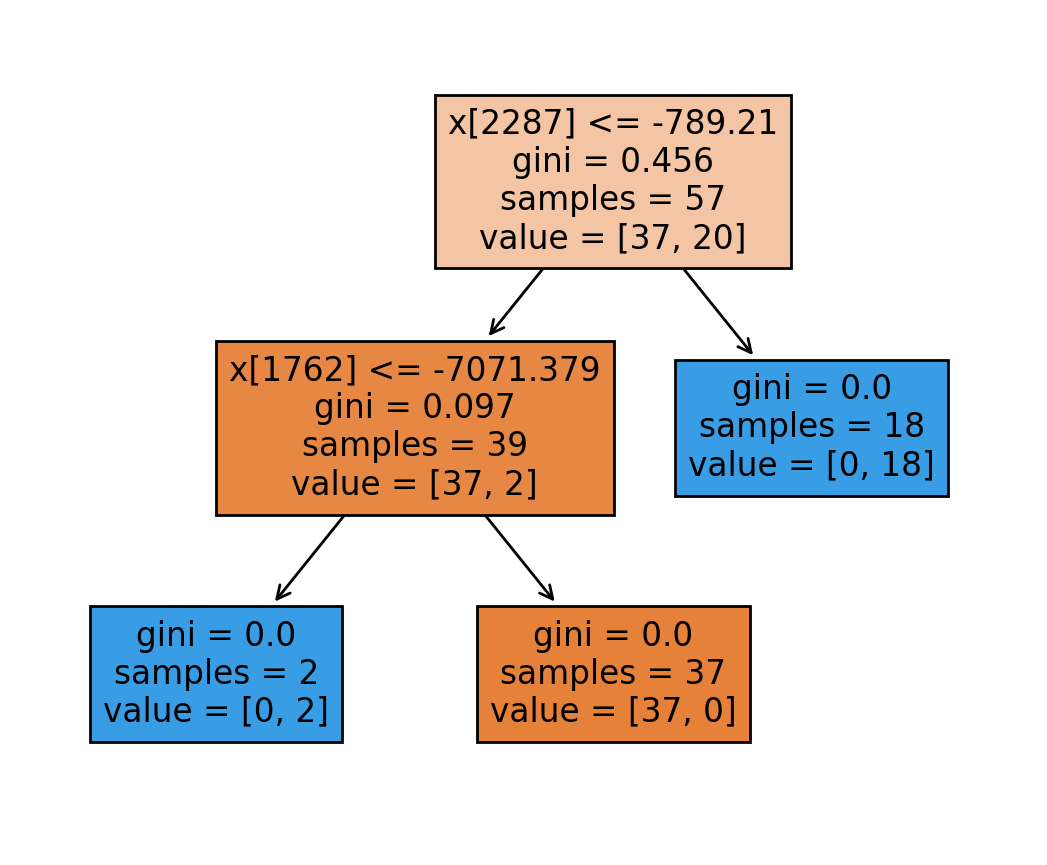
# Gene expression data analysis

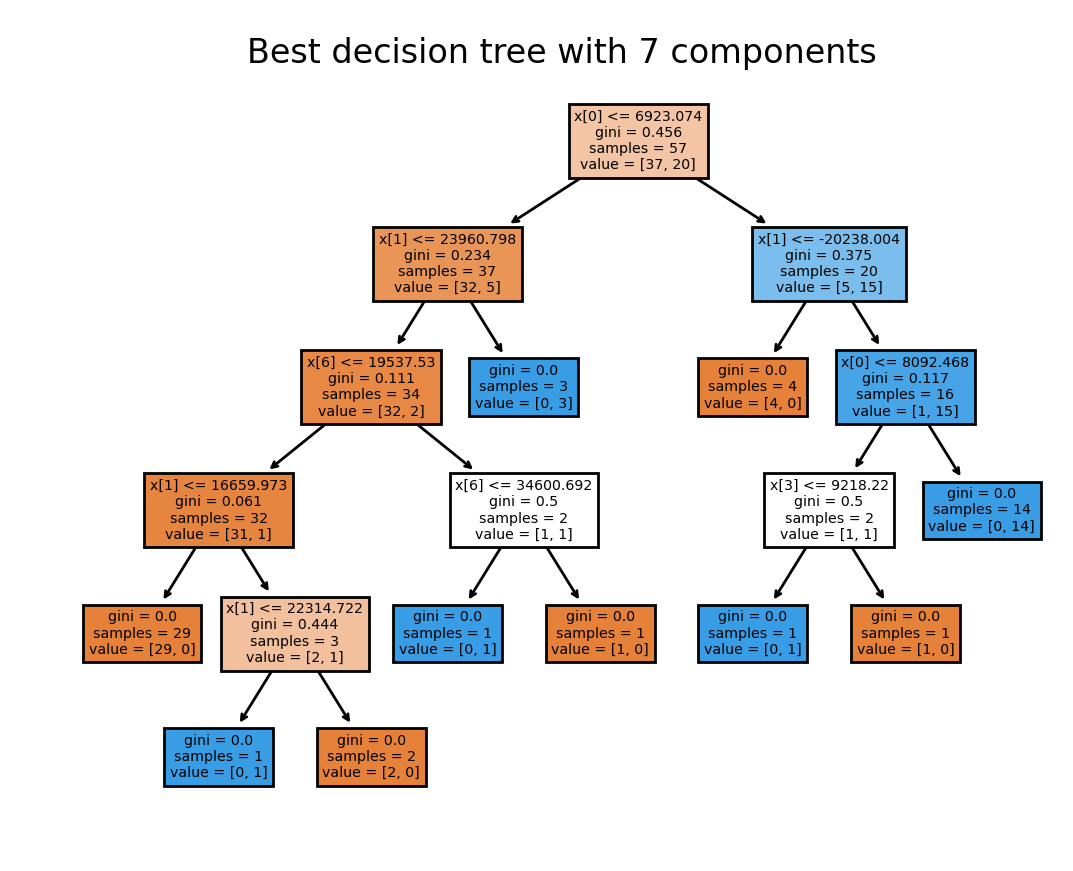
Tereza Kubová

## Part 1

1. Train a decision tree on subjected data. Use Python scikit-learn (this is up to you) classification tree fit it in to the data.
2. Show the tree and enumerate its training accuracy.  
   Training accuracy: 100 %  
   
3. How would you interpret this model? Which gene is crucial for the decision?  
   Pro rozhodování je důležitý gen pod číslem 2288 – kvůli číslování pythonu od nuly (gen PDIA2), dále také gen pod číslem 1763 (gen TMSB4X).
4. Is this gene really the one causing the cancer? Look up in the article Golub et al.,   
   1999.  
   Podle článku je to gen HOXA9.
5. Estimate real accuracy of the tree. Use e.g., crossvalidation (alternatively, you can split the data).  
   Accuracy: 86.67 %
6. Compare it with the training accuracy.  
   Přesnost na testovacích datech je menší, než na trénovacích, což je logické, protože testovací data klasifikátor nikdy neviděl. Proto se i testovací data oddělují, aby se zjistila reálná přesnost klasifikátoru.

## Part 2

1. Learn a basis-matrix V of the data. Use PCA method.
2. For a range of component numbers K:
   1. project the original data X to the top K components of V. The result are data Z with reduced dimensionality.
   2. Create a tree out of these reduced data. Show it and enumerate its training accuracy.
3. Compare all the trees resulting from the reduced data and pick the “best” according to its accuracy and structure. Follow the Occam razor.



1. Estimate the real accuracy of the “best” chosen tree. Again, by e.g. crossvalidation.

Accuracy: 93.34 %