Fundamental of Data Science - L3BC

Arvin Yuwono	(2502009721)
Christopher Owen	(2502019180)
Kenneth Samuel Djasmin	(2502009620)

Conduct a mini research on:

1. What is Information Gain?

   Information gain is the measurement of reduction according to entropy. It is usable to conclude the best features or attributes for the highest information gain in a class. A larger information gain suggests a lower entropy group or groups of samples. The more homogenous the samples, the higher the information gain will be.

2. What is Entropy?

   Entropy is the measurement of how similar the values of the data are. A higher entropy value means the data's value is more differ with each other. It basically shows how chaotic your data is.

3. Why Decision Tree (ID3) use these metrics?

   Because, technically, ID3 means to divide the data repeatedly becoming more groups using top-down greedy algorithm. With computing the entropy and information gain, the decision tree would be able detect the homogeneity and highest information gain to split the data to select the best feature.

4. Advantages & Disadvantages of Decision Trees?

   Advantages:

   - Less data preparation
   - No need for data normalization
   - No need for data scaling
   - Intuitive and easily explainable

   Disadvantages:

- Change in data means change in structure
- Complex calculation is possible
- Higher model training time
- Inadequate for regression and prediction

5. When do you think Decision Tree will be better than any other method?

   Decision Tree is suitable for categorizing data. Data analytics and machine learning are complex data and need to be first processed for analysis and visualization. It is of no wonder that decision trees are commonly found in usage of prediction analysis, data classification, and regression. A good example is in operations research, where analysis for better strategy decisions are required.

6. Find out what is CART decision tree?

   Classification and Regression Trees is an algorithm that uses Gini Index(Classification) as a metric to split the dataset to create the decision tree. The main process revolves around calculating Gini index, comparing Gini index, and choosing lowest Gini index. This algorithm type can be used to solve classification and regression problems. It is also by far a popular choice among data scientists, especially considering that scikit-learn, a Python module, uses an optimized version of it.

Work with a toy dataset

1. Find a small toy dataset

Toy_dataset

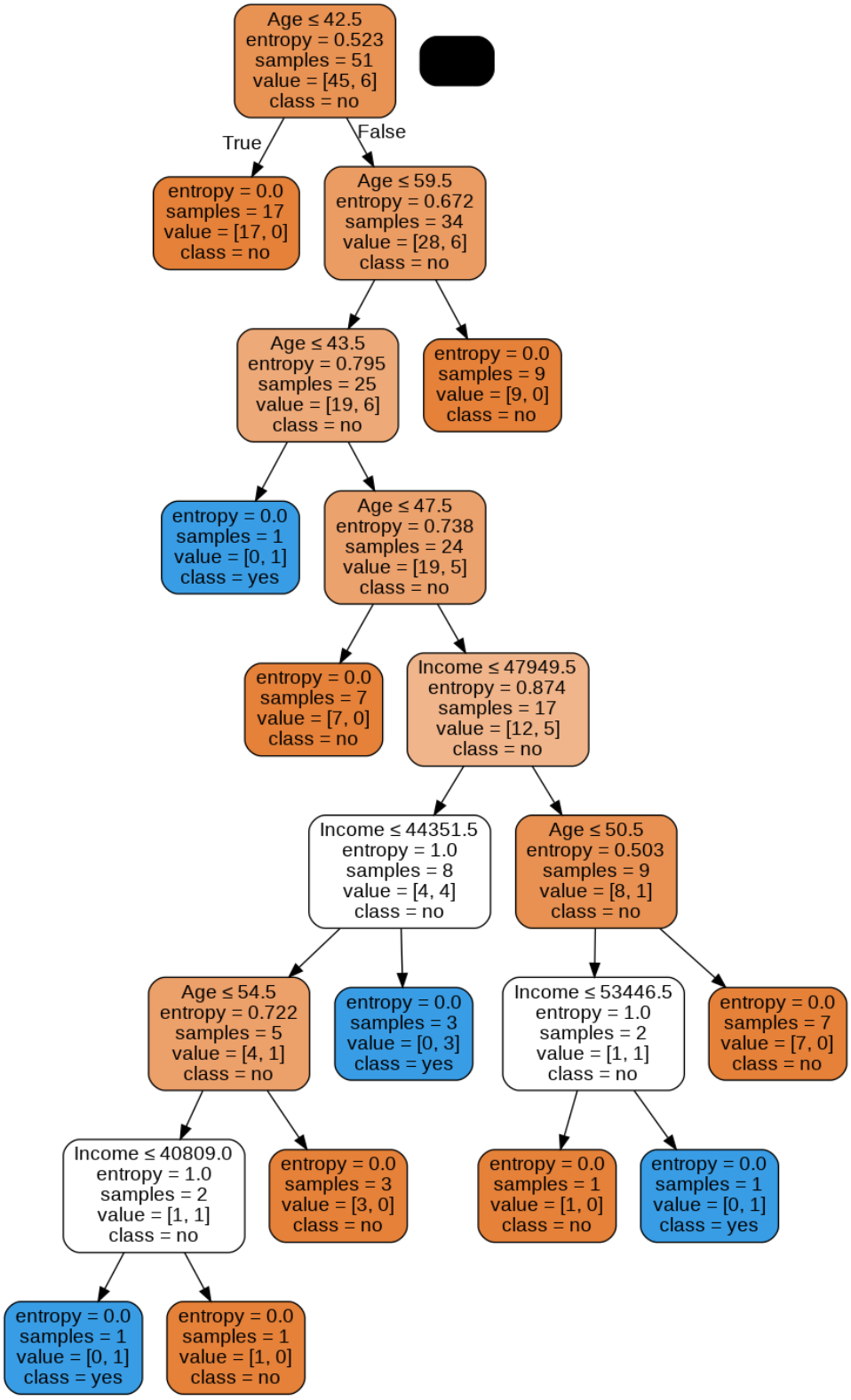
https://docs.google.com/spreadsheets/d/19IHJhuMfjSKRSmNrDODbiWuUgGLCo-lWbSsxh-nJw0k/edit?usp=sharing

2. Pick a sample of new data and show what is the predicted class.

Google collab

https://colab.research.google.com/drive/1FB8oNZAm3DbgHqB_NuMyaTZrvixQVrT4?usp=sharing

3. How to interpret your tree?



Age <= 42.5: The decision tree checks whether the age is less than 42.5. Afterwards, the data will go either true or false.

Entropy = 0.523: The data sample is neither completely homogeneous nor equally divided, but is somewhere between in the middle.

Sample = 51: The data set contains 51 samples in total

Value = [45,6]: The first element of the list displays the number of samples in illness class with the value “no”, while the second element of the list displays the number of samples in illness class with the value “yes”.

Class = no: Illness class with the value of “no” dominates the dataset, therefore “no” is selected as the class’ value.

The way to read the tree is as stated below:

- If the age is smaller than 42.5, then class=No (Age<=42.5 True)
- If the age is smaller than 42.5, and the age is smaller than 59.5, and the age is not smaller than 43.5, and income is not smaller than 47949.5, and income is larger than 44351.5, then class = Yes (Age<=42.5 True, Age<=59.5 True, Age<= 43.5 True, Age<=47.5 False, Income <= 47949.5 True, Income<= 44351.5 False, the leaf = Yes[0,3])