Predict Adult Income with Decision Tree

QUOC DUNG NGUYEN

Abstract

The adult data set from UCI Machine Learning Repository is used to predict the income of an adult by using a supervised machine learning method – Decision Tree – with the accuracy of 83%.

Motivation

The problem is if we can predict an adult income based on their characterises like age, sex, education, occupation, race…

Dataset(s)

For this project I used the following dataset:

- Adult Data Set from UCI Machine Learning Repository
- Data contains 15 columns and 32561 rows

- https://archive.ics.uci.edu/ml/datasets/Adult

Data Preparation and Cleaning

Data contains lots of missing values, so they need to be removed.

In order to use Decision Tree method, we also need to convert categorial values to numeric values.

Redundant columns have to be removed before building and training models.

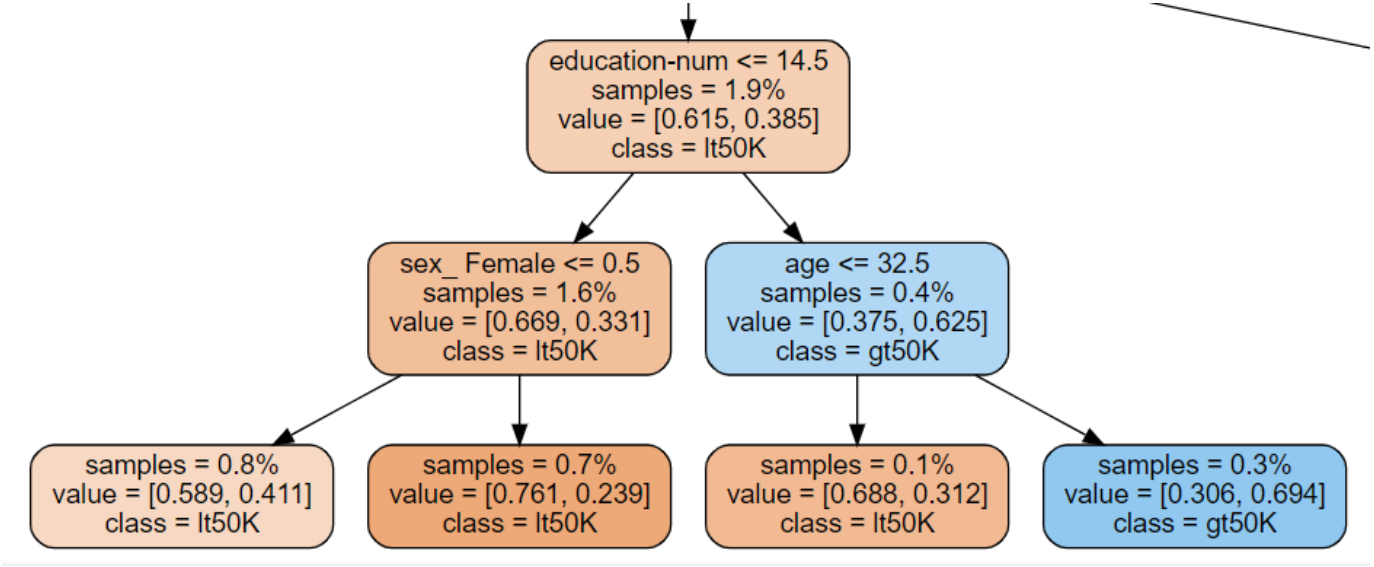
Research Question(s)

Can we use supervised machine learning to predict whether income of an adult is greater than 50K / year?

Methods

Supervised Decision Tree method will be used to analyze the data because we already had the label of income from the current dataset.

Findings



Decision Tree model is built and trained to predict an adult income with the accuracy of 83%. (The tree is too big that should view from the notebook file)

Limitations

The data is too old. In order to have a updated result, we need recent adult census dataset.

Conclusions

We can predict an income of an adult based on their demographic such as age, sex, occupation, race… with the accuracy of 83%.

References

https://archive.ics.uci.edu/ml/datasets/Adult