ASSIGNMENT A2

Title: To find decision based on given scenario using Decision Tree

Problem Statement:
A dataset collected in a cosmetics shop showing details of customers and whether or not they respond to an offer special to buy a new lipstick is shown. Use this dataset to build a Decision Tree, with Buys as the target variable to help lip-stick buying in the future. Find the root node of the decision tree. According to the decision tree you have made from the previous training dataset, what is the decision for the following test data: [Age < 21, Income = Low, Gender = Female, Marital Status = Married]?

Objective: To understand how the decision tree classifier algorithm works and use it on the given dataset

Outcome: To find the decision based on a given scenario of people with income, gender and marital status information from dataset using Decision Tree Classifier

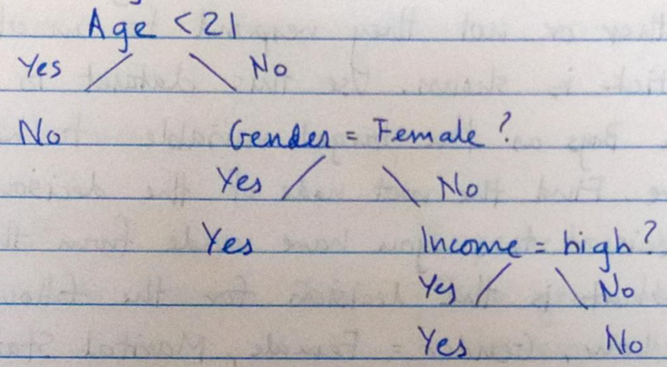
Requirements: python3, pandas, numpy, jupyter, Unix/Linux machine, text editor

Theory:
Decision tree is a simple representation for classifying examples - It is a supervised machine learning algorithm where data is continuously split according to a certain parameter.

Decision tree consists of:
- Nodes: Test for value of a certain attribute

- Edges / branch: Correspond to the outcome of a test and connect to the next node / leaf
- Leaf nodes: Terminal nodes that predict the outcome (represent class labels or class distribution)

Example: Will buy lipstick?



Algorithm: The core algorithm for building decision trees, ID3, ~~was~~ developed by J R Quinlan employs a top-down greedy search through the space of possible branches with no backtracking

- Entropy: Decision trees are built top-down from a root node and involves partitioning of data into subsets that contain instances with similar values (homogenous). ID3 uses entropy to calculate homogenity - completely homogenous (entropy 0) and if same is equally divided, it has an entropy of one.

To build a decision thee, two kinds of entropies need to be calculated:

(a) Entropy using the frequency table of one attribute

$$E(s) = \sum_{i=1}^{c} -p_i \log_2 p_i$$

(b) Entropy using the frequency table of two attributes

$$E(T,X) = \sum_{c \in X} P(c)\, E(c)$$

· Information Gain: Based on the decrease in entropy after a dataset is split on an attribute. Constructing a decision tree is all about finding the attribute that returns the highest information gain

$$\text{Gain}(T,X) = \text{Entropy}(T) - \text{Entropy}(T,X)$$

Test Cases

Test	Expected output	Actual output
Generate Decision Tree	Tree generated as expected	Tree generated as expected
Predict Buy for Age <21, Income = high, Gender= Male, Single	Yes	Yes

Conclusion: Successfully built decision tree for given dataset, and predicted class for test data given