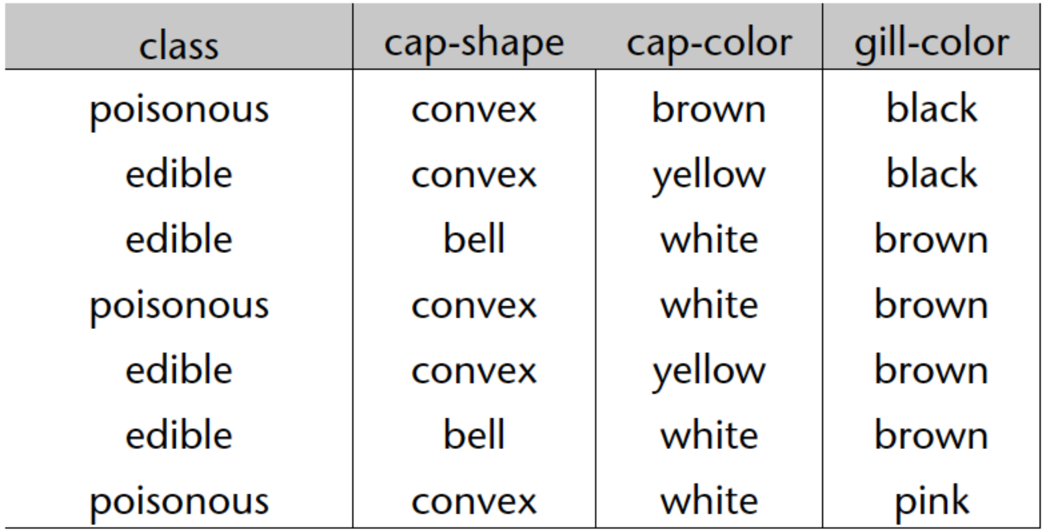
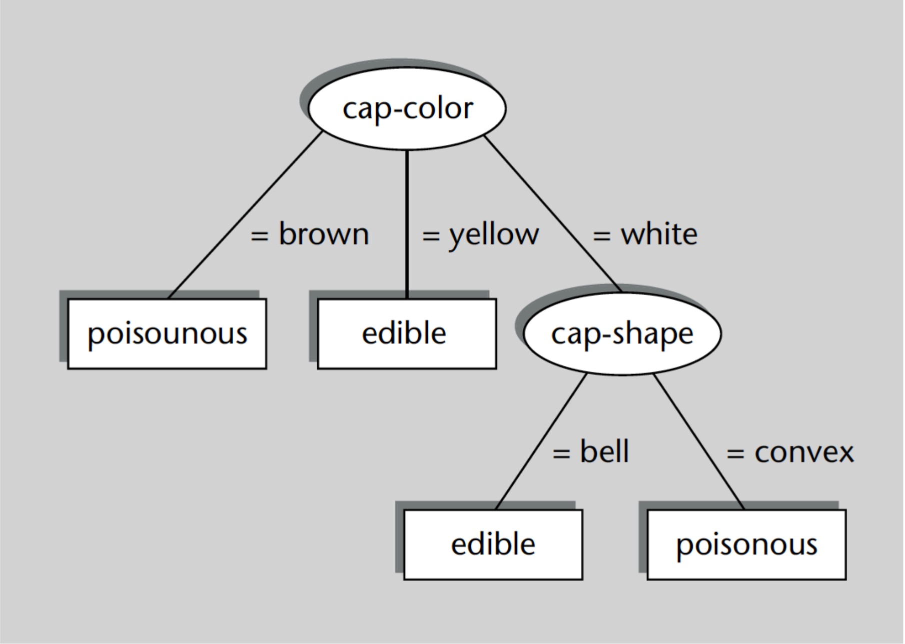
Tabla 19. Conjunto de entrenamiento



class	cap-shape	cap-color	gill-color
poisonous	convex	brown	black
edible	convex	yellow	black
edible	bell	white	brown
poisonous	convex	white	brown
edible	convex	yellow	brown
edible	bell	white	brown
poisonous	convex	white	pink

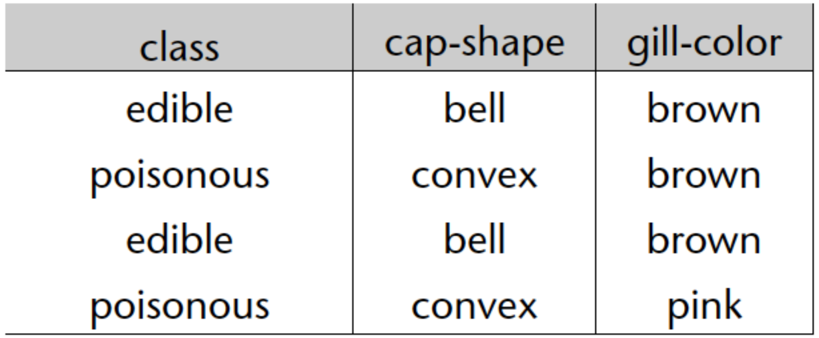
Fuente: problema «mushroom» del repositorio UCI (Frank y Asunción, 2010)



Purity of each attribute:

cap-shape: cap-shape = convex: 3 right, 2 wrong ; cap-shape = bell: 2 right, 0 wrong -> goodness(cap-shape) = (3+2)/7 = 0.71
cap-color: brown: 1 right, 0 wrong ; yellow: 2 right, 0 wrong ; white: 2 right, 0 wrong ; -> goodness(cap-color) = (1+2+2)/7=0.71
gill-color: black: 1 right, 1 wrong ; brown: 3 right, 1 wrong ; pink: 1 right, 0 wrong -> goodness(gill-color) = (1+3+1)/7=0.71

randomly select one -> cap-color:



class	cap-shape	gill-color
edible	bell	brown
poisonous	convex	brown
edible	bell	brown
poisonous	convex	pink

cap-shape: bell: 2 right; convex: 2 right-> goodness(cap-shape) = (2+2)/4 = 1
gill-color: brown:2 right; pink: 1 right right-> goodness(gill-color) = (2+1)/4 = 0.75

Example: Constructing a decision tree:

For each feature, we should evaluate how well it splits the data in classes in such a way that each value of the feature contains only one of the classes.

One of the evaluation measures is to count how many observations are classified according to the majority of classes in each feature value (purity of the split).

For instance, the feature cap-shape has two features: convex and bell. If we choose convex, most of the observations are poisonous (3 over a total of 5). A value bell classifies the two observations as edible, so the total goodness of the feature will be (3+2)/7 = 0.71.

We proceed in a similar way with the other features and the result is that all are equally good. So we can choose one of them randomly. Image we select the feature cap-color to split the data. Then we will have three groups according to the three possible values of the feature cap-color. For the value cap-color = brown, the tree ends classifying the observation as poisonous. Something similar occurs for the value yellow, for which the two observations are classified as edible. In the case cap-color = white we should further develop the tree using the remaining features.

For this reduced set ob 4 observations, we should evaluate the goodness of each of the features. Selecting the feature cap-shape will have a goodness of (2+2)/4 = 1 since the valuers bell correctly classify all observations as edible and the value convex as poisonous. The feature gill-color can not do it better: The performance is (2+1)/4=0.75. Then we choose cap-shape and proceed to complete the decision tree as in the figure of the previous slide.