Mushroom Dataset:

The dataset contains 8,124 instances in all. There are 2 classes and 22 attributes. The first column of the dataset contains the class label ‘p’ for poisonous or ‘e’ for edible type of mushroom. The attributes are all nominal and described in the PDF file enclosed with the assignment. A typical record of the dataset consisting of comma separated string of characters looks like as follows:

p,x,s,n,t,p,f,c,n,k,e,e,s,s,w,w,p,w,o,p,k,s,u

Class	Cap-shape	Cap-Surface	Cap-Color	Bruises	Odor	Gill-Attachment	Gill-Spacing	Gill-Size	Gill-Color	Stalk-Shape	Stalk-Root	Stalk-Surface-Above-Ring	Stalk-Surface-Below-Ring	Stalk-Color-Above-Ring	Stalk-Color-Below-Ring	Veil-Type	Veil-Color	Ring-Number	Ring-Type	Spore-Print-Color	Population	Habitat
p	x	s	n	t	p	f	c	n	k	e	e	s	s	w	w	p	w	o	p	k	s	u

A note about missing values: The attribute number 11 contains missing values represented as ‘?’. You can ignore these. This means that you won’t consider it in counting when, e.g., computing the probabilities of class given the value of attribute 11.

Dividing/Splitting the dataset into Training and Test sets:

For the given problems, you would first need to divide/split the dataset into training and test sets. To do this, you would randomly pick 80% of the records (about 6,499 of them) for training and set aside the remaining 1,625 for the test set. You can use random number generator to figure which records to include in the training and test sets. If that sounds too difficult, simply take the first 6,499 for training and the rest for testing.

How to use test set for computing accuracy?

For testing your model, you would input each record containing attribute values into your model (Decision Tree or Naïve Bayes Classifier) and get the ‘predicted’ class label. Now, compare that predicted label with the ‘true’ label that you already knew (since it was part of the record!). If the predicted label is the same as the true label, you have got the ‘hit’. If it isn’t the case, you have got the ‘miss’. Accuracy, then, is the number of hits divided by the total number of records in the test set.