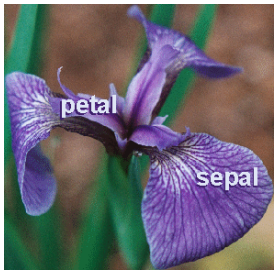
Lesson 15 (Iris Decision Trees)
R. A. Fisher's iris data set is one of the oldest and most used examples in data mining. The data set contains three class (50 records each) of the iris plant. Each class corresponds to one of three species.

Attributes (measured in centimeters)
- sepal length
- sepal width
- petal length
- petal width

Class (plant species)
- setosa
- versicolour
- virginica



iris plant


setosa


versicolor


virginica

(a) Preprocessing:
- Check data types using `df.dtypes`.
- Convert species to categorical data type using `.astype('category')`.
- Randomize the order of the records using the command below.

```
df = df.sample(frac=1.0,random_state=0)
```

(b) Fit a depth 2 decision tree for predicting iris species. Define classes as shown below.

```
class_names = df.species.cat.categories
```

(c) Compute species group averages for the attributes and use these attribute averages to check that your decision tree is reasonable.

(d) Use cross-validation and the hyper-parameters shown below to determine a tree that minimizes generalization error.

`max_depth` — maximum depth of the tree.
`min_samples_split` — mimimum size of a node that is allowed to be split. (Regulates the complexity of the tree.)