Lab 2: Iris dataset classification using a Decision Tree

Yusif Ibrahimov[1]

[1] *French-Azerbaijani University, Computer Science department. Email: yusif.ibrahimov@ufaz.az*

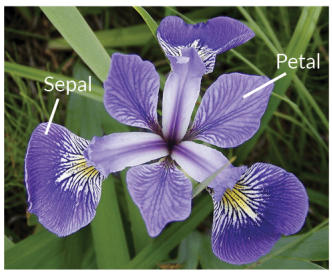
Abstract

The objective of this lab is to implement a Decision Tree in order to classify Irises (yes, the flowersa). **Specific objectives:**

- Observe the data, understand their nature and how to adapt them (if needed) so you can use them in a Decision Tree model.
- Understand how Decision Trees work so as to implement this model in a computer program.
- Evaluate the results and put them into perspective with what we know about the data.

1 Iris flower data set

The Iris flower data set or Fisher's Iris data set is a multivariate data set introduced by the British statistician, eugenicist, and biologist Ronald Fisher in his 1936 paper The use of multiple measurements in taxonomic problems as an example of linear discriminant analysis.It is sometimes called Anderson's Iris data set because Edgar Anderson collected the data to quantify the morphologic variation of Iris flowers of three related species.Two of the three species were collected in the Gaspé Peninsula "all from the same pasture, and picked on the same day and measured at the same time by the same person with the same apparatus".Fisher's paper was published in the journal, the Annals of Eugenics, creating controversy about the continued use of the Iris dataset for teaching statistical techniques today.



Iris Versicolor

Iris Setosa

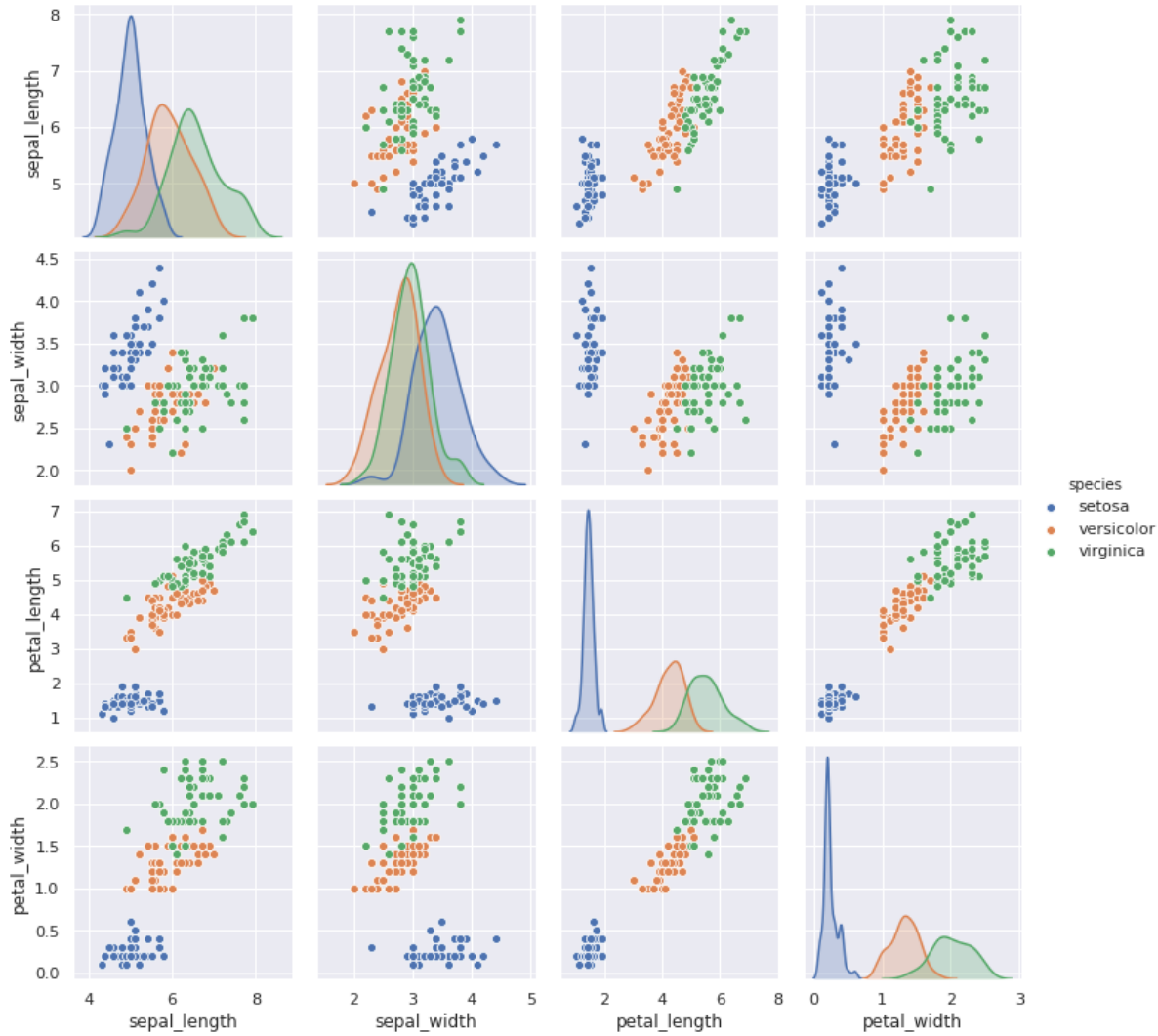
Iris Virginica

The data set consists of 50 samples from each of three species of Iris (Iris setosa, Iris virginica and Iris versicolor). Four features were measured from each sample: the length and the width of the sepals and petals, in centimeters. Based on the combination of these four features, Fisher developed a linear discriminant model to distinguish the species from each other.

Question: Is the Decision Tree model used for supervised or unsupervised classification? Explain your answer

Answer: Decision Tree model used for **supervised classification**. Referring to the definition of the supervised learning, we know that the supervised learning algorithm learns learns on a labeled data set. Supervised learning algorithms can evaluate the performance and accuracy based on the accuracy. Here' in this example, we can say that, we have 3 labels, and based on that labels, we learn the conditions and provide the tree to classify the new unseen data.

Te detailed visualization of the Iris dataset:



2 Buidling a Decision Tree

Each instance of the dataset has 4 attributes. Building a Decision Tree is basically determining which of these attributes has the highest discriminative power. Once you have determined this attribute, you must determine which attribute has the second highest discriminative power, and so on and so forth. Once this process is set you will be able, upon the examination of a new instance, to predict the species (the class) of this “unknown” instance.

As we said that , Discriminative power helps us to distinguish two categories, and aids to determine the which division is successfull.It is our information gain. To identify them let’s have a look formulas:

$$H(g) = -\sum_{s=0}^{S=0} p(s)\log_2 p(s) \quad and \quad P(S, g_1, g_2) = H(S) - [\frac{\#g_1}{\#S}H(g_1) + \frac{\#g_2}{\#S}H(g_2)]$$

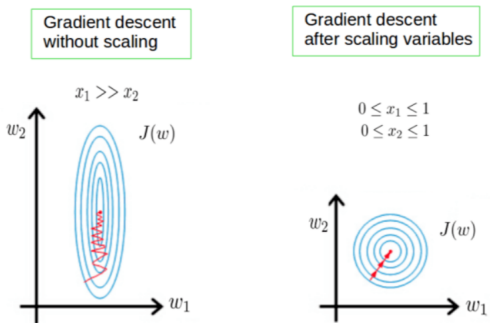
Question: What is the nature of the attributes of the dataset?

Answer: There are 4 attributesin the dataset, includeing , ’sepal_length’,’sepal_width’,’petal_length’, ’petal_width’.These variables are assisting on the determinin the group of the flower. These variables are **continuous** floating point variables.

Question: What is the nature of the attributes of the dataset?

Answer: Feature scaling is one of criticial data pre-processing step in Machine Learning.It can make difference between weak and strong ML model. There are two types of the scalin: Normalization and Standardization.We use Normalization when we need to bound our set with [0,1] or [-1,1] with the formula of $x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}}$, and the usage of the Standardization in need of the zero mean and 1 variance.Formula : $x_{new} = \frac{x - \mu}{\sigma}$

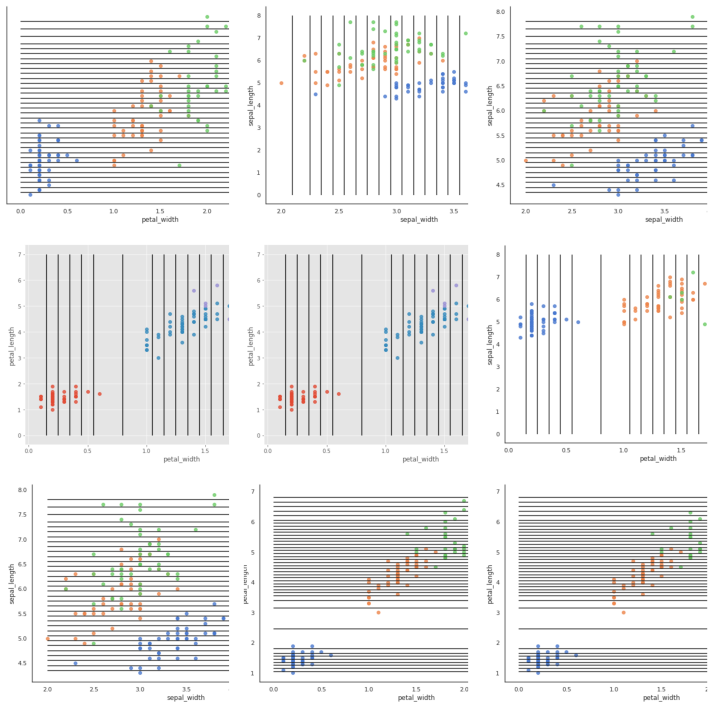
Our algorithms works with features (numbers) and there can be huge differences between the numbers and if we apply the feature scaling, it make easier the gradient descent process. Like:



But, we are talking about the tree models and we don't use the gradient descent. Although that the feature scaling is very sensitive for some models, in our case it's not mandatory, However, I don't prefer to use scaling for this model, even it may affect my accuracy, because, When I look my Tree, it changes the values of the tree and visually not pretty to demonstrate.

Question: How are you going to use real value attributes to build your Decision Tree?:

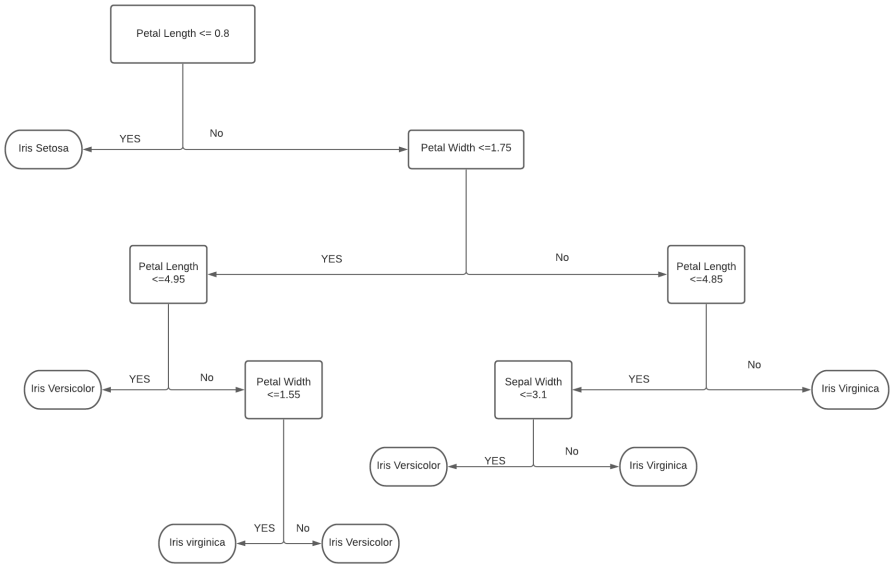
Answer: We will use the discretization in order to work with real continuous variables. We will calculate the medians (or the mean of two neighbors) of distince values for all combination of two distinct features. Source code provided inside of the notebook. Here are the potential discrete points for each case.



We gave all the potential discrete values and splits for 6 cases (not all). We will look for each single split and calculate the information gains, because each line divides the data into two parts, and we will choose the best one, then continue.This is our technique of the choosing the column and value.We don't use the discriminativa power for three different groups(setosa, virginica and versicolor), we use it with two groups, left and right parts.

3 Results and analysis

After generating the tree with the help of the information gain and and constructing the tree with the help of the dictionary, we will have a look of the tree itself. First, we will do it with the train data. The result is (not the dictionary itself) :



Classification Report and Confusion Matrix

Training Accuracy is: 100.0

Wonderful for training accuracy.Very Low Bias

...

...

...

Test Accuracy is: 98.0

Wonderful for test accuracy.

It seems No Overfitting

	precision	recall	f1-score	support
setosa	1.00	1.00	1.00	14
versicolor	0.95	1.00	0.98	21
virginica	1.00	0.93	0.97	15
accuracy			0.98	50
macro avg	0.98	0.98	0.98	50
weighted avg	0.98	0.98	0.98	50

Visualization of confusion matrix: