Autor: Antonio Ortiz

# Intent of the application

The purpose of this application is to explore the Iris dataset and use the features in it to perform predictions using different classification to predict the label.

This is with the purpose of exploring different classification techniques, the results they provide, and compare them with other techniques.

# Functional description

This application will use the Iris dataset, a very well-known dataset, to perform basic exploration of the data and observe the distribution of the data.

Afterwards, several classification techniques will be used on the data to be able to classify the observations into one of the three available categories:

* Setosa
* Versicolor
* Virginica

# Dataset to be used

The Iris dataset is a very well-known dataset with information about three different types of flowers.

This dataset includes 150 observations, each with 4 features, and one label.

These observations are evenly distributed in 3 values for the label.

The dataset to be used, the Iris dataset, can be found here: https://www.ritchieng.com/machine-learning-iris-dataset/

# Mathematical background

## KNN (K-Nearest Neighbors)

It is a non-parametric supervised learning method, used for both classification and regression.

In the case of this application, it will be used for classification.

The only argument that the algorithm requires is the number of clusters, and it is optional.

In the case of this application, two attempts were made: one where only four classes were considered (those that actually had members) and another attempt in which the full categories (8) were taken into consideration. The results are the same, as the later 4 categories did not have a single element in them.

## Decision Tree

Non-parametric, supervised learning algorithm. It can be used for both regression and classification, but for the purposes of this application, it will be used for classification.

It uses a tree-like model of decisions in order to get to the result.

It has a hierarchical, tree structure, which consists of a root node, branches, internal nodes and leaf nodes.

## Multiclass Logistic Regression

Also known as multinomial logistic regression, it is a classification technique that generalizes the logistic regression algorithm into multiclass problems.

It works by predicts the probabilities that a given element belongs to a class. The class with the highest probability is the class to which the given element belongs.

One of the conditions for this technique is that the dependent variable (label) is nominal, which means that the categories cannot be ordered in any meaningful, other than arbitrarily decided.

It also requires that each observation belongs to only one category.

## Random Forest

Ensemble learning method (it uses multiple instances of other learning algorithms) used for classification, regression, and others. In the case of this application, it will be used for classification.

It uses multiple instances of decision trees, and the output of the random forest is the solution to which most decision trees arrived to. By using the output of multiple decision trees, it maximizes the changes that the prediction is correct.

# Use case

This application can be used by anyone looking for a method to classify large numbers of flowers of one of these 3 classes: setosa, virginica, and versicolor.

The only information available from these flowers is their measurements, such as petal length and width, and sepal length and width.

# Variables

* Sepal.Length: The length of the sepal, which is the outer part of the flower that encloses a developing bud.
* Sepal.Width: The width of the sepal.
* Petal.Length: The length of the petal of the flower, which are leaves that surround the reproductive parts of a flower.
* Petal.Width: The width of the petal of the flower.

# Labels

* Species: Describes the species of the flower associated to the previous measurements.

# Data import

In this application, there is no input needed from the user.

## Proposed Libraries

Seaborn: Data visualization library based on matplotlib. Source: https://seaborn.pydata.org/. Version: 0.11.2

Sklearn: Machine learning tool for predictive data analysis. Supports both supervised and unsupervised learning. Version: 1.0.2

Source -> https://scikit-learn.org/stable/getting\_started.html

Numpy: Used for vectorization and indexing for scientific computing. Version: 1.25.1 Source -> https://github.com/numpy/numpy

Pandas: Data analysis and data manipulation library. Version: 2.0.3 Source -> https://pandas.pydata.org/getting\_started.html

Matplotlib: Library for creating static, animated, and interactive visualizations in Python. Source:3.7.2 https://matplotlib.org/stable/users/release\_notes.html.

# Plots

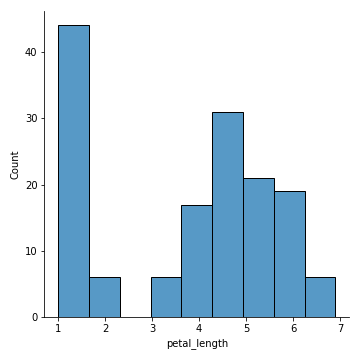


Figure 1 Distribution of the petal length feature

The Petal length variable is split in two parts, as there is a gap of values between the two sides. The right side is normally distributed, while the left side is composed of only two values. The range of the left side goes from 1 to 2, while the right side goes from 3 to 7, with no observations between 2 and 3.

Gráfico, Histograma

Descripción generada automáticamente

Figure 2 Distribution of the sepal width feature

The sepal width variable is normally distributed, with a range that goes from 2.9 to 4.5, and a spike at the value 3.0. The mean is approximately 3.0.

Gráfico, Histograma

Descripción generada automáticamente

Figure 3 Distribution of the sepal length feature

The sepal length variable is normally distributed, with a range that goes from 4.5 to 8.0. The mean is approximately 6.

Gráfico, Histograma

Descripción generada automáticamente

Figure 4 Distribution of the petal width feature

The Petal width variable is split in two parts, as there is a gap of values between the two sides. The right side is normally distributed with a skew to the left, while the left side is composed of only two values. The range of the left side goes from 0 to 0.5, while the right side goes from 1 to 2.5.

Imagen que contiene Gráfico

Descripción generada automáticamente

Figure 5 Correlation Matrix

With the heatmap, it can be seen that:

* Petal width and length are heavily related.
* Petal width and length are heavily related to the species.
* Sepal length is decently related to every variable and label, except for sepal width.
* Sepal width is inversely related to every variable and label, except for sepal length.

# Observations

In order to use the classification algorithms, the data had to be transformed, as the labels were categorical.

For some algorithms, the data had to be hot encoded. However, for some algorithms, the data could be used directly, or encoded into a numeric type.

With this transformed data, the algorithms were trained.

## Naïve Bayes

This algorithm required the data, the label, to be encoded into a numeric type, in the following manner:

* Setosa: 0
* Versicolor: 1
* Virginica: 2

The features were not encoded, as they are numerical.

The results were satisfactory, as out of 150 observations, 144 were correctly predicted.

There were 3 errors for Virginica, and 3 for Versicolor, but this is due to the way the data is presented, in which Setosa can be easily divided from the other two species with a straight line, but Virginica and Versicolor have some mixed elements, which may add to the noise.

## KNN

After the data was hot encoded, the model could be trained.

It was trained with the assumption that there were only 3 possible clusters, which was known knowledge due to the available data.

After the training was done, the results were basically the same as those from Naïve Bayes, with 144 correct predictions, and 6 errors, evenly split between virginica and versicolor.

## Decision Tree

The criterion for the training of the decision tree was “Entropy”.

It was trained with no maximum depth, which produced great results, as every observation was correctly predicted.

This, however, may be cause for concern, as it may be overfitted.

## Logistic Regression

This algorithm does not accept multiple labels as the output, but it accepts multiple classes for a single label, so the process for one hot encoding for the labels were reversed.

With this data, the algorithm was trained with a maximum iterations parameter of 1000.

The results were satisfactory were similar to those produced by Naïve Bayes and KNN, with 146 correct results, and only 4 errors.

However, this model was trained again, with no limits to the iterations.

The model reached the maximum number of iterations allowed by the package without reaching a convergence, producing the exact same errors.

## Random Forest

The same procedure was followed as the one in Decision Tree, and the results were the same.

With this algorithm, it was possible to see the most relevant categories and values for the predictions.

The most relevant categories were:

* Petal Length
* Petal Width

The other two features, the sepal length and width, were considered irrelevant.

This produced excellent results, as every observation was correctly predicted.

This model was trained again, without the features considered to be irrelevant.

Out of 150 observations, 149 were correctly predicted.

Therefore, by reducing the model in half, the loss of accuracy was negligible.

# Conclusion

The models with the best results were Decision Tree and Random Forest.

However, Random Forest is composed of several Decision Trees, which suggests that the best algorithm for the classification of flowers using their measurements is the Decision Tree.

This may be due to the nature of the algorithm, in which it creates multiple “rules” and decisions in order to get the results it needs. This grants the algorithm certain flexibility that other algorithms do not have.

An example of this is the KNN algorithm. It clusters nearby points or observations into a single cluster. This does a great job at correctly clustering observations, as long as they are all within close proximity. However, in the case of this dataset, some observations were mixed, which caused the algorithm to incorrectly cluster them into the wrong group.

A similar situation occurred with the Multiclass logistic regression algorithm, and the Naïve Bayes algorithm.

Finally, after comparing the results of all the techniques applied to the Iris dataset, it can be seen that most supervised clustering techniques worked as expected, such as Decision Tree, or KNN, while only some unsupervised clustering techniques worked as expected, such as BIRCH clustering, or mean shift clustering.

However, these comparisons may not be valid, as the unsupervised clustering techniques were trained using a dataset after going through PCA (Principal Component Analysis).

PCA reduces the dimensionality of the dataset, but this may result in loss of information, which may impact the accuracy of the clustering techniques.

For the comparisons to be valid, either both supervised and unsupervised techniques would have to be performed without applying PCA or applying PCA.

But as far as can be seen, the best techniques for clustering the Iris dataset, are supervised, and out of the supervised techniques, the best ones are Decision Tree, and Random Forest, which is just multiple Decision Trees.