A study of **pattern recognition of Iris flower** based on Artificial Intelligence

A Project Report

Submitted in fulfillment of the requirements for the award of

The Certificate for Internship Program 2019

By

**SMARTBRIDGE**

Team members

1.P.Aditya kumar

2.V.Swapna Sri

3.D.V.Vyshnavi

**1.Classification of iris flower based on Artificial neural networks**

* 1. **Introduction**

In this project we are going to make a classification of iris flower based on the attributes using artificial neural networks.

**Technical operations:**

1.Python

2.Artificial Intelligence

**Python**

**Python** is an [interpreted](https://en.wikipedia.org/wiki/Interpreted_language), [high-level](https://en.wikipedia.org/wiki/High-level_programming_language), [general-purpose](https://en.wikipedia.org/wiki/General-purpose_programming_language) [programming language](https://en.wikipedia.org/wiki/Programming_language). Created by [Guido van Rossum](https://en.wikipedia.org/wiki/Guido_van_Rossum) and first released in 1991.  Its language constructs and [object-oriented](https://en.wikipedia.org/wiki/Object-oriented_programming) approach aims to help programmers write clear, logical code for small and large-scale projects.

Python is [dynamically typed](https://en.wikipedia.org/wiki/Dynamic_programming_language). It supports multiple [programming paradigms](https://en.wikipedia.org/wiki/Programming_paradigm), including [procedural](https://en.wikipedia.org/wiki/Procedural_programming), object-oriented, and [functional programming](https://en.wikipedia.org/wiki/Functional_programming). Python is often described as a "batteries included" language due to its comprehensive [standard library](https://en.wikipedia.org/wiki/Standard_library).

**Features:**

Python is a [multi-paradigm programming language](https://en.wikipedia.org/wiki/Multi-paradigm_programming_language). [Object-oriented programming](https://en.wikipedia.org/wiki/Object-oriented_programming) and [structured programming](https://en.wikipedia.org/wiki/Structured_programming) are fully supported, and many of its features support [functional programming](https://en.wikipedia.org/wiki/Functional_programming) and [aspect-oriented programming](https://en.wikipedia.org/wiki/Aspect-oriented_programming) (including by [metaprogramming](https://en.wikipedia.org/wiki/Metaprogramming) and [metaobjects](https://en.wikipedia.org/wiki/Metaobject) (magic methods)). Many other paradigms are supported via extensions, including [design by contract](https://en.wikipedia.org/wiki/Design_by_contract) and [logic programming](https://en.wikipedia.org/wiki/Logic_programming).

**Uses:**

Large organizations that use Python include [Wikipedia](https://en.wikipedia.org/wiki/Wikipedia), [Google](https://en.wikipedia.org/wiki/Google), [Yahoo!](https://en.wikipedia.org/wiki/Yahoo!),[CERN](https://en.wikipedia.org/wiki/CERN),[NASA](https://en.wikipedia.org/wiki/NASA), [Facebook](https://en.wikipedia.org/wiki/Facebook), [Amazon](https://en.wikipedia.org/wiki/Amazon_(company)), [Instagram](https://en.wikipedia.org/wiki/Instagram), [Spotify](https://en.wikipedia.org/wiki/Spotify) and some smaller entities like [ILM](https://en.wikipedia.org/wiki/Industrial_Light_%26_Magic) and [ITA](https://en.wikipedia.org/wiki/ITA_Software). The social news networking site [Reddit](https://en.wikipedia.org/wiki/Reddit) is written entirely in Python.

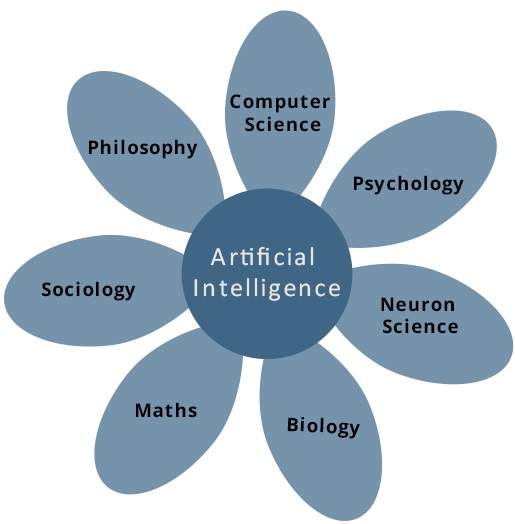
Python can serve as a [scripting language](https://en.wikipedia.org/wiki/Scripting_language) for [web applications](https://en.wikipedia.org/wiki/Web_application), e.g., via [mod\_wsgi](https://en.wikipedia.org/wiki/Mod_wsgi) for the [Apache web server](https://en.wikipedia.org/wiki/Apache_web_server). With [Web Server Gateway Interface](https://en.wikipedia.org/wiki/Web_Server_Gateway_Interface), a standard API has evolved to facilitate these applications.

**Artificial Intelligence**

## What is Artificial Intelligence?*.*

Artificial Intelligence is a way of **making a computer, a computer-controlled robot, or a software think intelligently**, in the similar manner the intelligent humans think.

## Goals of AI

* To Create Expert Systems
* To Implement Human Intelligence in Machines
* 

**Applications of AI:**

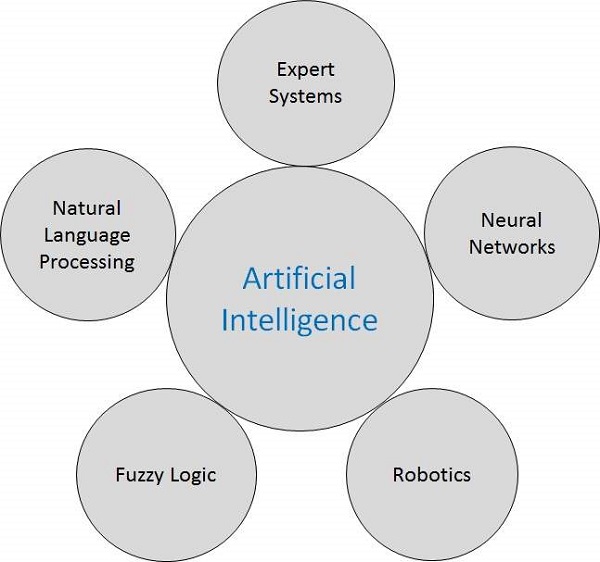
AI has been dominant in various fields such as −

* Gaming
* Natural Language Processing
* Expert Systems
* Vision Systems
* Speech Recognition
* Handwriting Recognition
* Intelligent Robots

## Intelligence:

The ability of a system to calculate, reason, perceive relationships and analogies, learn from experience, store and retrieve information from memory, solve problems, comprehend complex ideas, use natural language fluently, classify, generalize, and adapt new situations.

The domain of artificial intelligence is huge in breadth and width. While proceeding, we consider the broadly common and prospering research areas in the domain of AI −



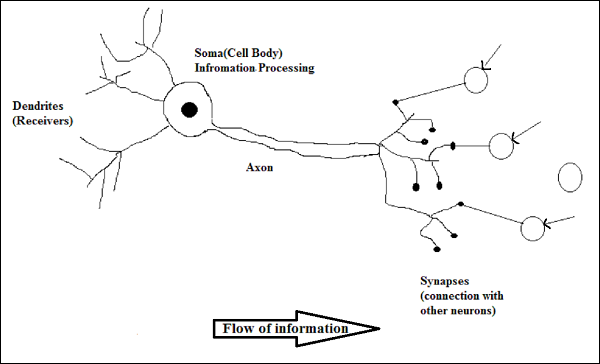
## Expert Systems

The expert systems are the computer applications developed to solve complex problems in a particular domain, at the level of extra-ordinary human intelligence and expertise.

**Artificial Neural Network**

Neural networks are parallel computing devices, which are basically an attempt to make a computer model of the brain. The main objective is to develop a system to perform various computational tasks faster than the traditional systems. This tutorial covers the basic concept and terminologies involved in Artificial Neural Network. Sections of this tutorial also explain the architecture as well as the training algorithm of various networks used in ANN

**Biological Neuron:**A nerve cell (neuron) is a special biological cell that processes information. According to an estimation, there are huge number of neurons, approximately 1011 with numerous interconnections, approximately 1015.

****

## Applications of Neural Networks

* **Aerospace**
* **Automotives**
* **Military**
* **Electronics**
* **medical**
* **Control**
* **Anomaly Detection**

# Data Preprocessing

# Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues.

# Data preprocessing steps:

#### Step 1: Import Libraries

First step is usually importing the libraries that will be needed in the program. A library is essentially a collection of modules that can be called and used. A lot of the things in the programming world do not need to be written explicitly ever time they are required. There are functions for them, which can simply be invoked. This is a [list](https://medium.com/activewizards-machine-learning-company/top-15-python-libraries-for-data-science-in-in-2017-ab61b4f9b4a7)for most popular Python libraries for Data Science. Here’s a snippet of me importing the pandas library and assigning a shortcut “pd”.

import pandas as pd

#### Step 2: Import the Dataset

A lot of datasets come in CSV formats. We will need to locate the directory of the CSV file at first (it’s more efficient to keep the dataset in the same directory as your program) and read it using a method called read\_csv which can be found in the library called pandas.

import pandas as pd

dataset = pd.read\_csv('Medium.csv')

After inspecting our dataset carefully, we are going to create a matrix of features in our dataset (X) and create a dependent vector (Y) with their respective observations. To read the columns, we will use iloc of pandas (used to fix the indexes for selection) which takes two parameters — [row selection, column selection].

X = dataset.iloc[:, :-1].values

: as a parameter selects all. So the above piece of code selects all the rows. For columns we have :-1, which means all the columns except the last one. You can read more about the usage of iloc [here](https://www.shanelynn.ie/select-pandas-dataframe-rows-and-columns-using-iloc-loc-and-ix/).

#### Step 3: Taking care of Missing Data in Dataset

Sometimes you may find some data are missing in the dataset. We need to be equipped to handle the problem when we come across them. Obviously you could remove the entire line of data but what if you are unknowingly removing crucial information? Of course we would not want to do that. One of the most common idea to handle the problem is to take a mean of all the values of the same column and have it to replace the missing data.

The library that we are going to use for the task is called [Scikit Learn](http://scikit-learn.org/stable/index.html)preprocessing. It contains a class called Imputer which will help us take care of the missing data.

from sklearn.preprocessing import Imputer

#### Step 4: Encoding categorical data

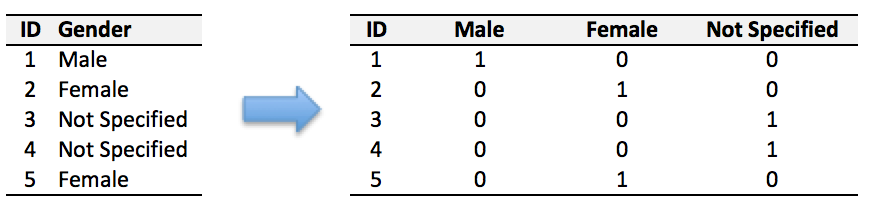
Sometimes our data is in qualitative form, that is we have texts as our data. We can find categories in text form. Now it gets complicated for machines to understand texts and process them, rather than numbers, since the models are based on mathematical equations and calculations. Therefore, we have to encode the categorical data.

So the way we do it, we will import the scikit library that we previously used. There’s a class in the library called LabelEncoder which we will use for the task.

from sklearn.preprocessing import LabelEncoder

As I have mentioned before, the next step is usually to create an object of that class. We will call our object labelencoder\_X.

So instead of having one column with n number of categories, we will use n number of columns with only 1s and 0s to represent whether the category occurs or not.



Example of a Dummy encoding

To accomplish the task, we will import yet another library called OneHotEncoder.

from sklearn.preprocessing import LabelEncoder, OneHotEncoder

#### Step 5: Splitting the Dataset into Training set and Test Set

Now we need to split our dataset into two sets — a Training set and a Test set. We will train our machine learning models on our training set, i.e our machine learning models will try to understand any correlations in our training set and then we will test the models on our test set to check how accurately it can predict. A general rule of the thumb is to allocate 80% of the dataset to training set and the remaining 20% to test set. For this task, we will import test\_train\_split from model\_selection library of scikit.

from sklearn.model\_selection import train\_test\_split

# 1.2 Objectives of Reasearch

# After the project has been settled, the computer should have the ability to aggregate three different classifications of Iris flower to three categories. The whole workflow of machine learning should work smoothly. The users do not need to tell the computer which class the Iris belongs to, the computer can recognize them all by itself. Moreover, the case study of Iris recognition will show how to implement machine learning by using Scikit-learn software.

# 1.3 Problem Statement

# Fisher’s Iris data base (Fisher, 1936) is perhaps the best known database to be found in the pattern recognition literature. The data set contains 3 classes of 50 instances each.

# Attributes are:

# Training data:

# 1). sepal length in cm

# 2). sepal width in cm

# 3). petal length in cm

# 4). petal width in cm

# Testing data:

# 5). class: - Iris Setosa - Iris Versicolour - Iris Virginica

# The goal of the seminar is to demonstrate the process of building a neural network based classifier that solves the classification problem.

# 2.Review of literature

# 2.1 Basic introduction:

# Learning is a very important feature of Artificial Intelligence. Many scientists tried to give the definition for learning.

# The structure of machine learning system consists of four main parts:

# 1.Environment

# 2. Learning

# 3.Knowledge base

# 4. Execute.

# The environment represents a combination of information from external information source. That would include any information from persons or references materials and so on.

# The knowledge base can be treated as the brain of the whole machine learning system. Different kinds of form and content of knowledge can have different influence on the designing of a machine learning system. Knowledge representation modes are eigenvector, First-order logic statements, production rule, and semantic system. Every mode has its own advantages and disadvantages. Therefore, when users want to design a machine leaning system, a good knowledge representation mode is very important for the whole system.

# Moreover, a machine learning system cannot create new knowledge from nothing. It always needs original knowledge to understand the information from environment. Then the computer can use this information to learn new knowledge by step.

# 3.Data Collection

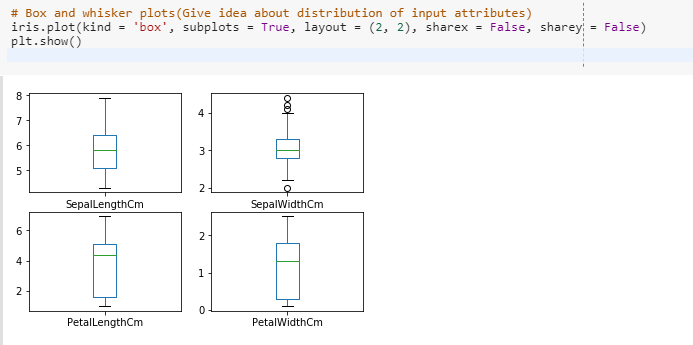
# The data set contains three classes of 50 instances each, where each class refers to a type of iris plant. Each class is linearly separable from the other two classes. The attribute information will include sepal length, sepal width, and petal length and petal width. All of them have the same unit, cm.

**4.Methodologies**

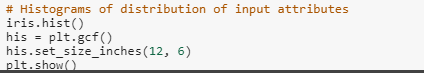
**4.1 Exploratory data analysis**

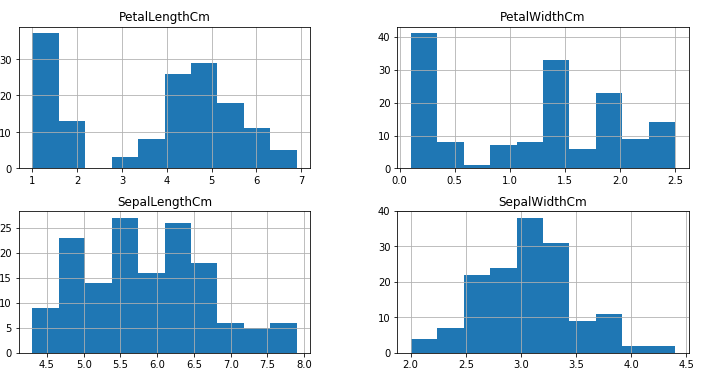
**4.1.1 Figures and Tables**

**Iris dataset plotting using box plots and whisker plots:**



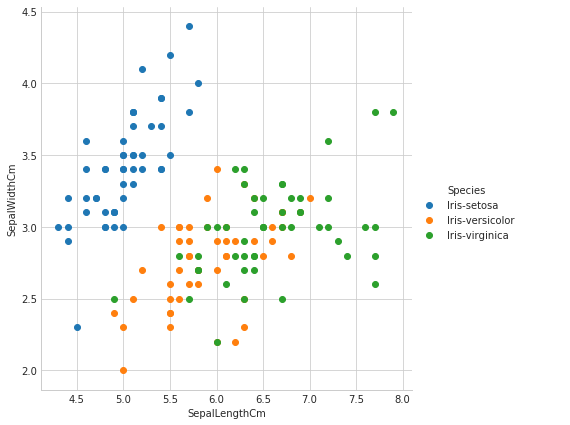
**Iris dataset plotting using Histograms:**



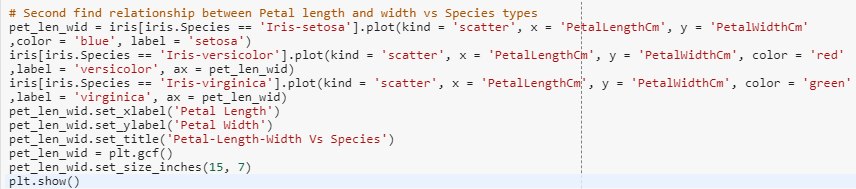


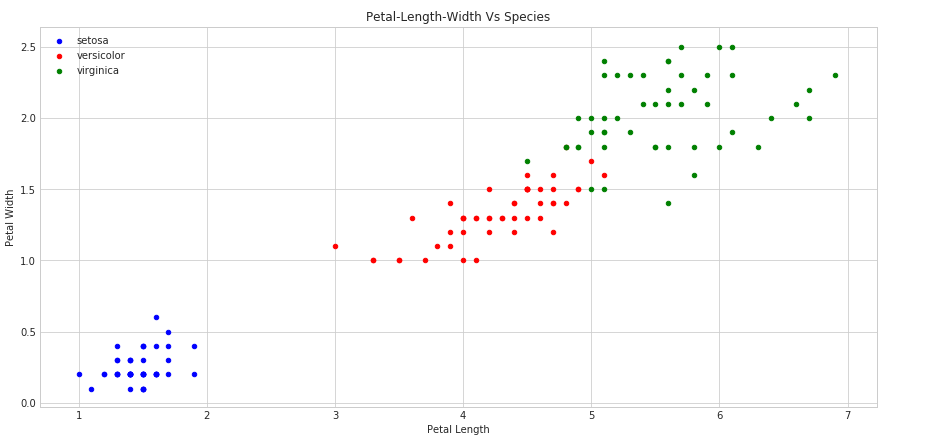
**Iris dataset plotting using White grids on attributes sepal length and sepal width:**

## 



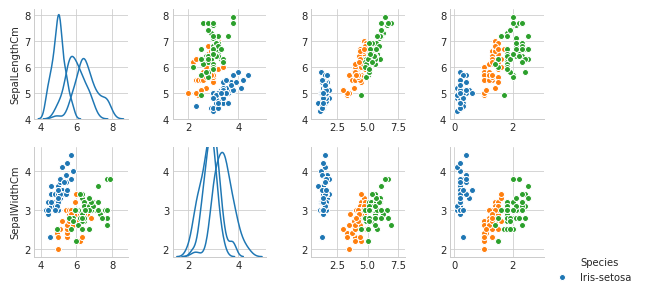
**Iris dataset plotting using White grids on attributes petal length and petal width:**

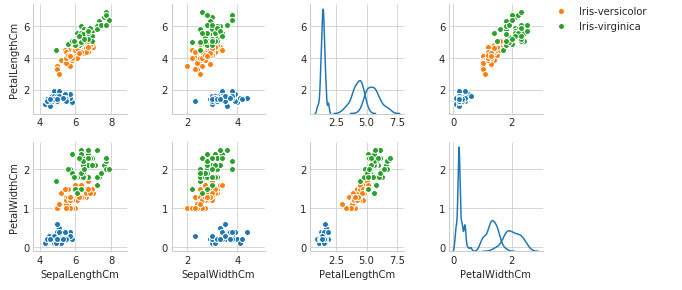




**Scatter plots:**







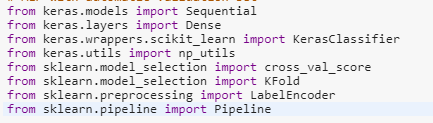
**Data Modelling:**

In Python, there is no data type called array. In order to implement the data type of array with python, numpy and scipy are the essential libraries for analyzing and calculating data. Numpy is mainly used for the matrix calculation. By using them in Python programming, they can be used with two simple commands:

>>> import numpy

>>> import scipy

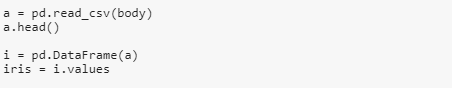
Preparing the Iris flower data set:

In the species of this table, 0 represents setosa, 1 represents versicolor, 2 represents virginica.

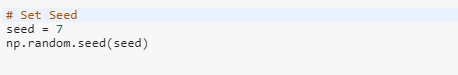
Importing the datasets:

Here’s a snippet of me importing assigning a shortcut “pd”.

A lot of datasets come in CSV formats. We will need to locate the directory of the CSV file at first and read it using a method called read\_csv which can be found in the library called pandas.



Setting seed:

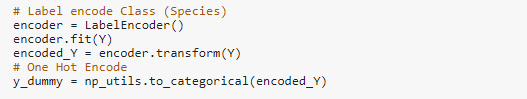


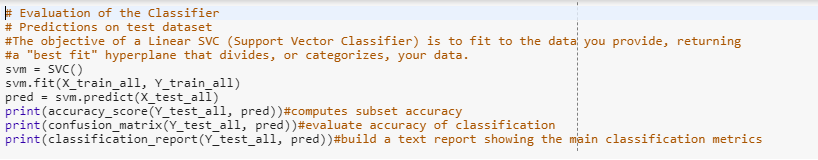
So the way we do it, we will import the scikit library that we previously used. There’s a class in the library called LabelEncoder which we will use for the task.

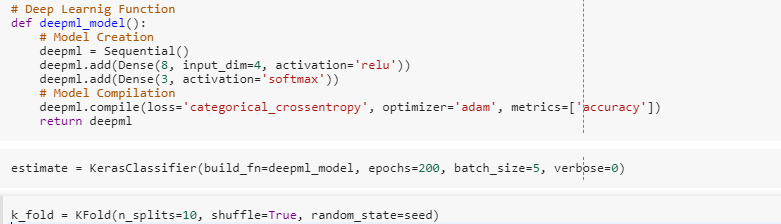
from sklearn.preprocessing import LabelEncoder

As I have mentioned before, the next step is usually to create an object of that class. We will call our object labelencoder\_X.

labelencoder\_X = LabelEncoder()

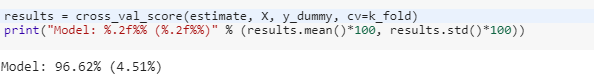


Now we need to split our dataset into two sets — a Training set and a Test set. A general rule of the thumb is to allocate 80% of the dataset to training set and the remaining 20% to test set. For this task, we will import test\_train\_split from model\_selection library of scikit.from sklearn.model\_selection import train\_test\_split



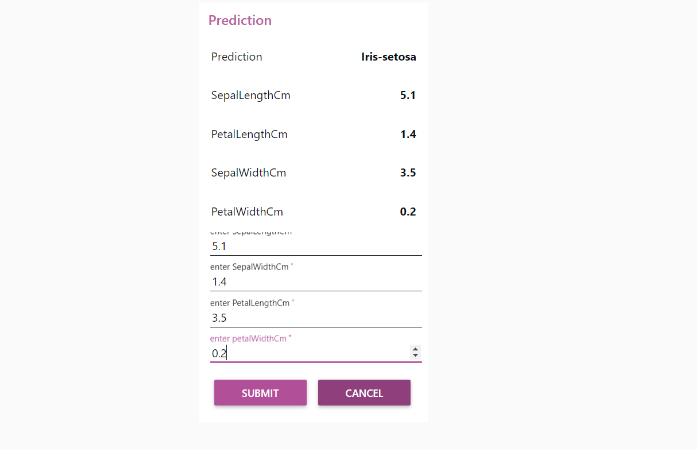
### Evaluating the Model:

The results are pretty good for 200 epochs and batch size 5 for such simple model.



The model gave us 96.62% accuracy.

**Prediction:**

****

Based on the given input values the flower is classified.

**Conclusion:**

Conclusion With the rapid development of technology, AI has been applied in many fields. Machine learning is the most fundamental approach to achieve AI. This thesis describes the work principle of machine learning, two different learning forms of machine learning and an application of machine learning. In addition, a case study of Iris flower recognition to introduce the workflow of machine learning in pattern recognition is shown. In this case, the meaning of pattern recognition and how the machine learning works in pattern recognition has been described. The K-means algorithm, which is a very simple machine learning algorithm from the unsupervised learning method is used. The work also shows how to use SciKit-learn software to learn machine learning.

## References:

* <https://pythonspot.com/matplotlib-scatterplot/> # scatter plot ideas and python script
* <https://docs.python.org/3/library/csv.html> #handling csv data
* <https://stackoverflow.com/questions/15389290/how-to-find-out-the-average-on-a-csv-file>
* <http://www.pythonforbeginners.com/systems-programming/using-the-csv-module-in-python/># how to access and split up the data
* <http://www.olsondata.com/?p=58> #from google search iris data set in python,really good document breaking down the analysis and some code to possibly access it and make a scatter plot
* <https://simply-python.com/2017/07/22/analyzing-iris-data-set-with-scikit-learn/> #bit more complex
* <http://www.learn4master.com/algorithms/visualize-iris-dataset-using-python> #nice document downloding panda?
* <https://medium.com/codebagng/basic-analysis-of-the-iris-data-set-using-python-2995618a6342> #really good document using panda and explains how to access the data.
* <https://en.wikipedia.org/wiki/Iris_flower_data_set>
* <http://dataconomy.com/2015/01/whats-the-difference-between-supervised-and-unsupervised-learning/>
* <https://issuu.com/denhams/docs/statistical_analysis_report>
* <https://www.techopedia.com/definition/32880/iris-flower-data-set>