**Name:** Yakov Finkelshtein

# Report

Table of Contents

[Report 1](#_Toc33224080)

[Dataset 3](#_Toc33224081)

[Problem definition 3](#_Toc33224082)

[Technical method of the proposed project 3](#_Toc33224083)

[Programming Language 3](#_Toc33224084)

[Programming Tool 3](#_Toc33224085)

[Algorithms 3](#_Toc33224086)

[Methodology 3](#_Toc33224087)

[Training code 4](#_Toc33224088)

[Testing code 7](#_Toc33224089)

[Conclusion: 11](#_Toc33224090)

# Dataset

I downloaded the wine dataset on the given <https://scikitlearn.org/stable/modules/generated/sklearn.datasets.load_wine.html>

After downloading the dataset I changed format in CSV file as required.

Following steps applied to convert into CSV file:

* Open in Notepad++
* Save in Txt file
* Open in excel file
* Save in CSV file

# Problem definition

To devise a wine classifier for overcoming the challenges to identify wine classification by implementing machine learning algorithms.

# Technical method of the proposed project

There are some technical things.

## Programming Language

* Python 3.8

## Programming Tool

* Anaconda Jupyter Notebook

## Algorithms

Random forest, decision trees, KNN and logistic regression algorithms.

# Methodology

**Testing code**

**Training code**

**Dataset**

**Testing with 5folds on trained algorithms**

**Data pre-processing and Training with 5 folds**

**Got Accuracy, confusion matrix, precision recall, f1**

**Parameter tuning of algorithms**

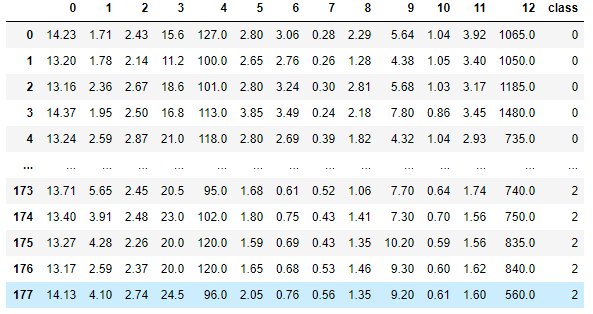
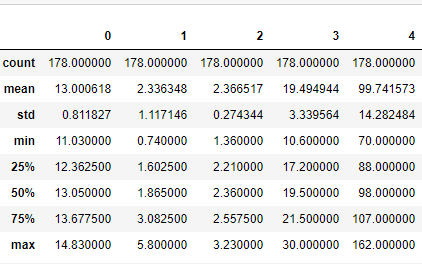
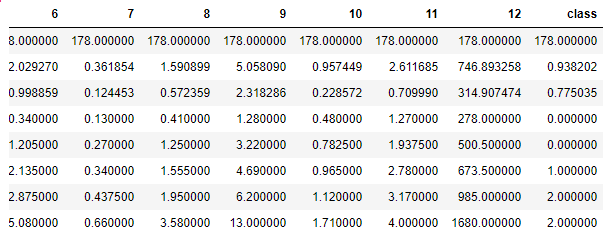
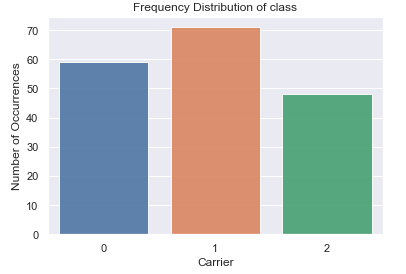
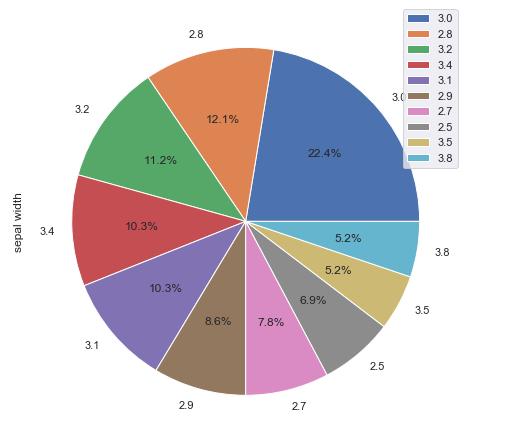
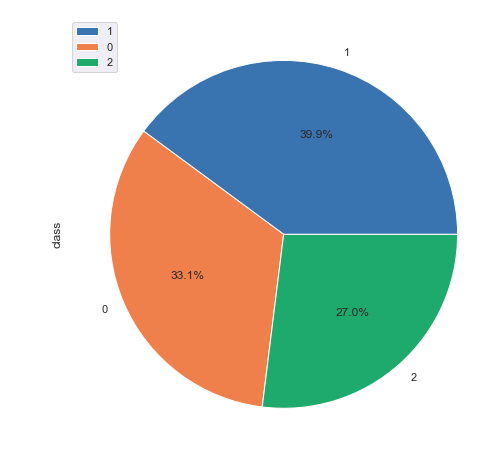
**Average accuracies, graphs of comparison of models**

**Save the models accordingly**

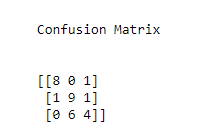
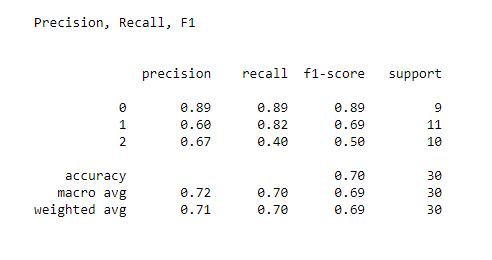
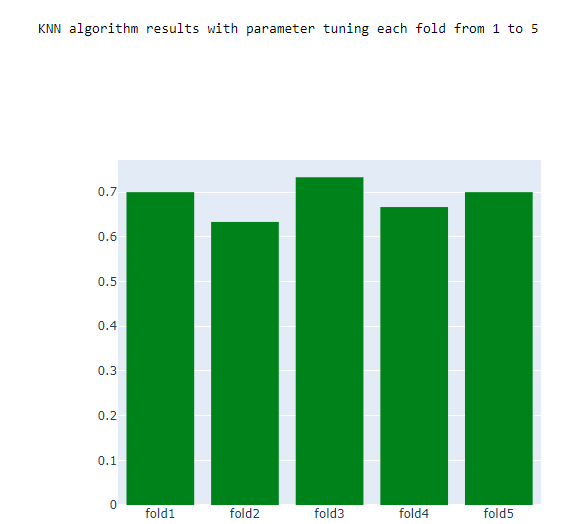
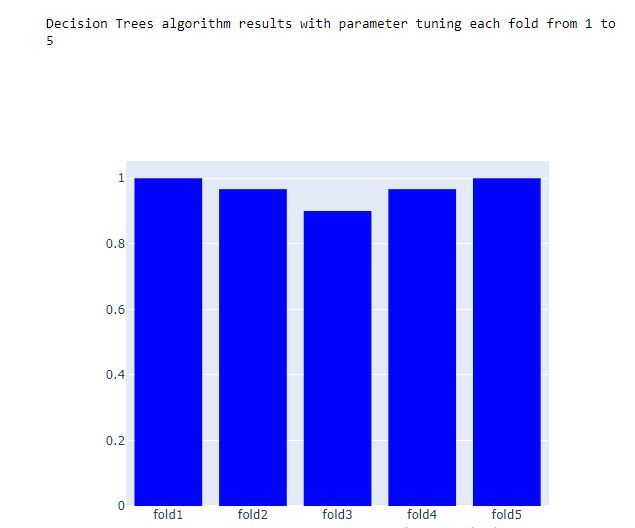
I created two python files and did the followings.

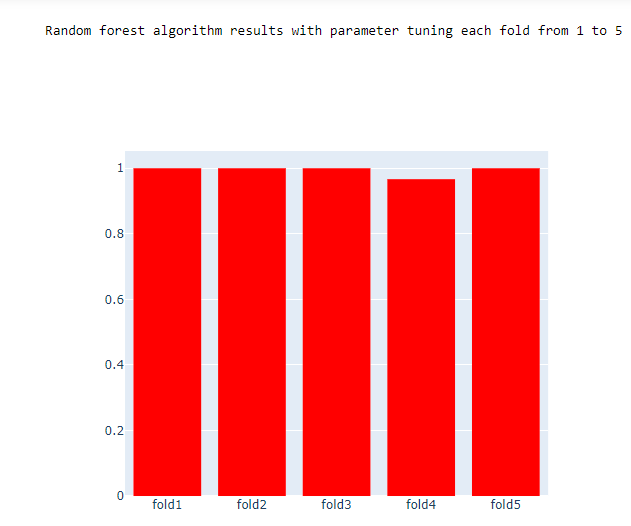
* Training file
* Testing file

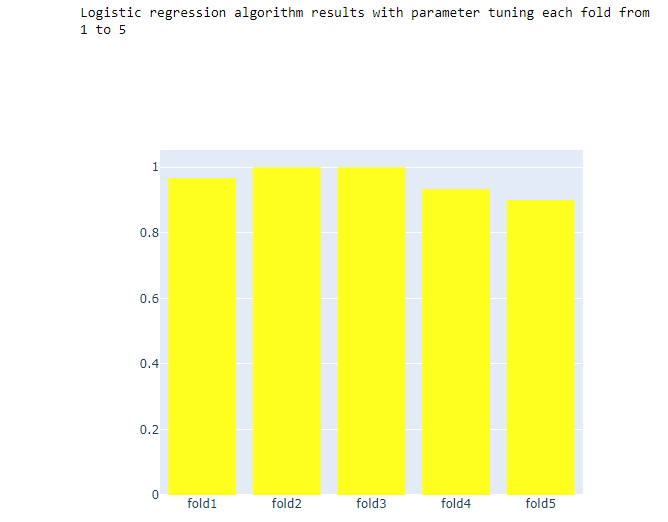
## Training code

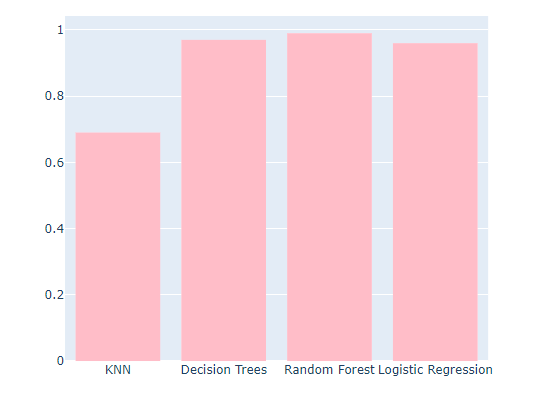
* I imported the required python files
* Loaded the CSV data file
* 
* Analyse the data
* 
* 
* Visualization of data
* 
* 
* Make the data in random form because we wanted to split the data into five parts
* Get the class and attributes separate
* Split the class into five parts
* Split the attributes into five parts same as class
* Train the KNN, decision trees, random forest and logistic regression algorithms with four parts each time with different sequence of 5 times as mentioned in the code
* Parameter tuning of all algorithms with each fold
* Save the trained files separately each time of each algorithm in txt files

## Testing code

* I imported the required python files
* Loaded the CSV data file
* Analyse the data
* Make the data in random form because we wanted to split the data into five parts
* Get the class and attributes separate
* Split the class into five parts
* Split the attributes into five parts same as class
* Test the KNN, decision trees, random forest and logistic regression algorithms by using the trained model on part1 to part5 separately
* Compute the accuracy of each algorithm with each split part
* 
* Compute the confusion matrix of each algorithm with each split part
* 
* Compute the precision of each algorithm with each split part
* Compute the recall of each algorithm with each split part
* Compute the F1 of each algorithm with each split part
* 
* Compute the average accuracy of each algorithm
* 
* Show the accuracies in graphs
* 
* 





* Show the comparison of all algorithms in graph
* 

# Conclusion:

I train and test KNN, decision trees, random forest and logistic regression algorithms with 5 folds of the dataset and tuned parameters of the algorithms and check if the algorithms are over fitting. Evaluation measures was following:

* Accuracy
* Confusion matrix
* Precision
* Recall
* F1

All algorithms performed well as we can see their results and in comparison, there is no big difference as well as with their average accuracies. But if we look deeply then Random forest and decision trees algorithms performed well with 98% accuracy of both. We can differentiate and compare the results and performance of the algorithms more and more efficiently if we will have large amount of data with lot of attributes. In machine learning, we always try to put maximum and unique features to get the good accuracy.