**Predicting the quality of wine**

**Introduction To Python**

**Python:**

Python is a high-level, interpreted, interactive and object-oriented scripting language. Python is designed to be highly readable. It uses English keywords frequently where as other languages use punctuation, and it has fewer syntactical constructions than other languages.

Python was developed by Guido van Rossum in the late eighties and early nineties at the National Research Institute for Mathematics and Computer Science in the Netherlands.

Python is derived from many other languages, including ABC, Modula-3, C, C++, Algol-68, SmallTalk and Unix shell and other scripting languages.

Libraries used in python:

* Pandas
* Numpy
* Matplotlib
* Scikitlearn
* Seaborn

**Softwares used:**

* Anaconda Navigator
* Python 3.7
* IBM cloud

***SOFTWARES:***

Anaconda Navigator is a desktop graphical user interface (GUI) included in Anaconda distribution that allows you to launch applications and easily manage conda packages, environments and channels without using command-line commands. Navigator can search for packages on Anaconda Cloud or in a local Anaconda Repository. It is available for Windows, macOS, and Linux.

In order to run, many scientific packages depend on specific versions of other packages. Data scientists often use multiple versions of many packages and use multiple environments to separate these different versions.

The command line program conda is both a package manager and an environment manager, to help data scientists ensure that each version of each package has all the dependencies it requires and works correctly.

**Installation:**

We can get anaconda navigator from the link given below

<https://www.anaconda.com/distribution/#windows>

**Navigator** is an easy, point-and-click way to work with packages and environments without needing to type conda commands in a terminal window. You can use it to find the packages you want, install them in an environment, run the packages and update them, all inside Navigator.

The following applications are available by default in Navigator:

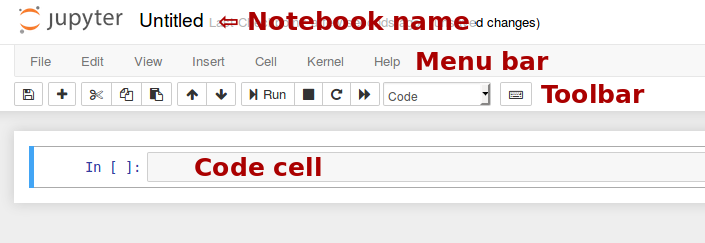
* JupyterLab
* Jupyter Notebook
* QTConsole
* Spyder
* VSCode
* Glueviz
* Orange 3 App
* Rodeo
* RStudio

**JUPYTER NOTEBOOK** extends the console-based approach to interactive computing in a qualitatively new direction, providing a web-based application suitable for capturing the whole computation process: developing, documenting, and executing code, as well as communicating the results. The Jupyter notebook combines two components:

* **A web application**: a browser-based tool for interactive authoring of documents which combine explanatory text, mathematics, computations and their rich media output.
* **Notebook documents**: a representation of all content visible in the web application, including inputs and outputs of the computations, explanatory text, mathematics, images, and rich media representations of objects.

**STRUCTURE OF NOTEBOOK**:

When you create a new notebook document, you will be presented with the **notebook name**, a **menu bar**, a **toolbar** and an empty **code cell**.



**Notebook name**: The name displayed at the top of the page, next to the Jupyter logo, reflects the name of the .ipynb file. Clicking on the notebook name brings up a dialog which allows you to rename it. Thus, renaming a notebook from “Untitled0” to “My first notebook” in the browser, renames the Untitled0.ipynb file to My first notebook.ipynb.

**Menu bar**: The menu bar presents different options that may be used to manipulate the way the notebook functions.

**Toolbar**: The tool bar gives a quick way of performing the most-used operations within the notebook, by clicking on an icon.

**Code cell**: the default type of cell; read on for an explanation of cells.

### **Keyboard shortcuts:**

All actions in the notebook can be performed with the mouse, but keyboard shortcuts are also available for the most common ones. The essential shortcuts to remember are the following:

* **Shift+Enter: run cell**

Execute the current cell, show any output, and jump to the next cell below. If Shift-Enteris invoked on the last cell, it makes a new cell below. This is equivalent to clicking the **Cell**, **Run** menu item, or the Play button in the toolbar.

* **Esc: Command mode**

In command mode, you can navigate around the notebook using keyboard shortcuts.

* **Enter: Edit mode**

In edit mode, you can edit text in cells.

**Machine learning**

**Machine Learning** is the field of study that gives computers the capability to learn without being explicitly programmed. ML is one of the most exciting technologies that one would have ever come across. As it is evident from the name, it gives the computer that which makes it more similar to humans: **The ability to learn**. Machine learning is actively being used today, perhaps in many more places than one would expect.

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Abstract**

Nowadays industries are using product quality certifications to promote their products. This is a time taking process and requires the assessment given by human experts which makes this process very expensive. This paper explores the usage of machine learning technique like random forest classifier for product quality. This random forest classifier takes independent variables and predicts the values of dependent variable.

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Introduction To Our Project**

Today, all type of industries are improving by adapting new technologies and applying these in all areas. These technologies are also helpful to enhance the production and making the whole process smooth. But, still there are different areas, which demands human expertise such as product quality assurance. Nowadays, it became an expensive process as the demand of the product is growing over the time. Therefore this project explores different machine learning technique like random forest classifier for product quality assurance. This technique perfoms quality assurance process with the help of available characteristics of product and automate the process by minimizing human interfere. The work also identifies the important features to predict the values of dependent variable.

**Literature review**

In literature, some researchers have used machine learningtechniques to assess wine quality, but still a huge scope is available for improvement. Sun et al predicted six geographic wine origins based on neural networks fed with 15 input variables. They used 170 samples of data from Germany for their experiments. They got 100% predictive rate.

In this paper random forest classifier is implemented to determine dependency of wine quality on 11 different characteistics.

**DATA COLLECTION**

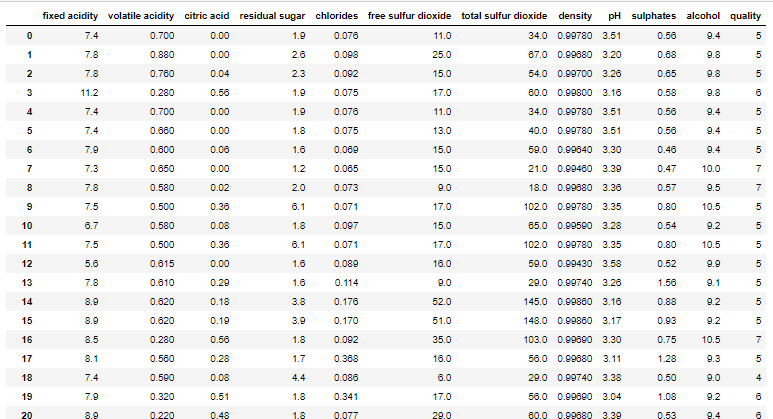
We have searched different websites for the required dataset and we finally found it on kaggle website.

**DATASET**

We have searched various websites for dataset, and finally we found our required dataset in kaggle website.

We can access the dataset using the link given below

<https://www.kaggle.com/c/ml210-wine-quality/data>

****

Independent attributes

* Fixed acidity
* Volatile acidity
* Citric acid
* Residual sugar
* Chlorides
* Free sulphur dioxide
* Total sulphur dioxide
* Density
* pH
* sulphates
* alcohol

Dependent attribute

* quality

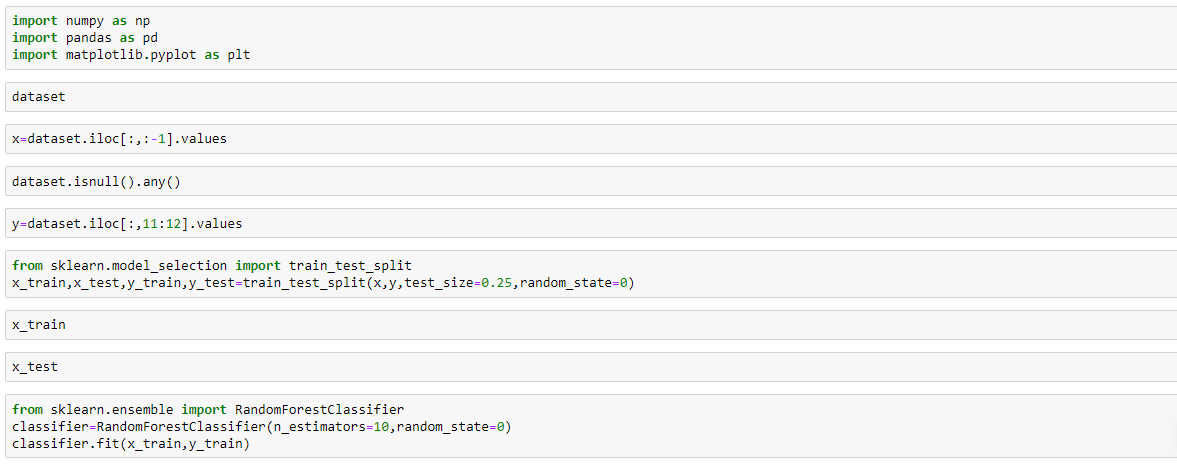
Methodology

**Classification is** a data mining function that assigns items in a collection to target categories or classes. The goal of **classification is** to accurately predict the target class for each case in the data.

Hence we used **random forest classifier** for predicting the dependent variable of our project. We also tried using remaining **classification** modelsbut the output was most accurately given by **random forest classifier** model.

Our Model:

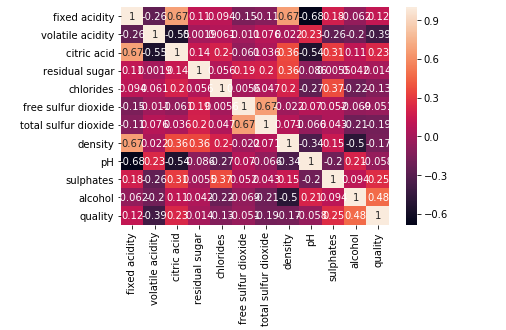
We have taken our required dataset from kaggle website and then imported it into our model using **pandas** library and sliced it into **independent** and **dependent variables.**





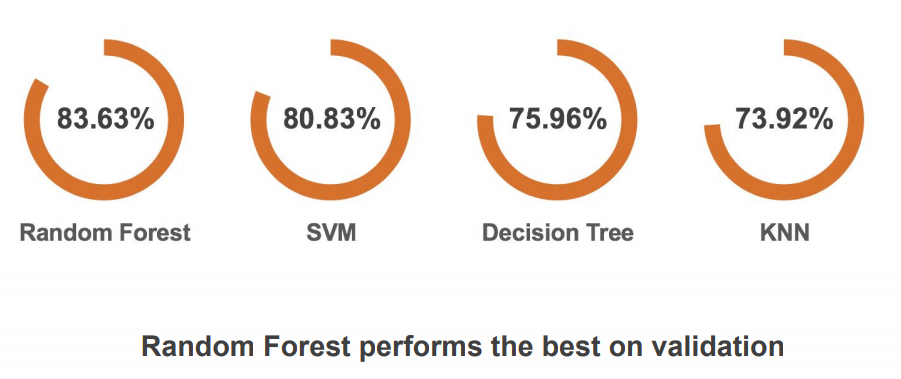








We used different models to predict the dependent variable and the accuracy results are as follows



**Findings & Suggestions:**

We have used a small dataset for our **random forest classifier** model and got the accurate output. But we found that when we have a large dataset the output is not accurate as it were for small datasets hence at these kind of situations it is better to use **Artificial Neural Networks** to get the output as accurate as possible.

**Conclusion:**

The interest has been increased in wine industries in recent years which demands growth in this industry. Therefore companies are investing in new technologies to improve wine production and selling. In this direction, wine quality certification plays a very important role for both processes and requires wine testing by human experts. This paper explores the usage of machine learning technique such as random forest classifier for predicting the values. In future, large dataset can be taken fir experiments and other machine learning techniques may be explored for wine quality prediction.