**INTRODUCTION**

**PYTHON:**

**About Python**

* + - Python is a high-level, general-purpose, open source, strictly typed programming language. The language provides constructs intended to enable clear programs on both a small and large scale.
    - Python was created By Guido van Rossum.
    - The Python Software Foundation (PSF) is the organization behind Python.

**Python versions:**

* + - First released in 1991.
    - Python 2.0 was released on 16 October 2000.
    - Python 3.0 was released on 3 December 2008.
    - Python 3.1 - 27 Jun 2009
    - Python 3.2 - 20 Feb 2011
    - Python 3.3 - 29 Sep 2012
    - Python 3.4 - 16 Mar 2014
    - Python 3.5 - 13 Sep 2015
    - Python 3.6 - 23 Dec 2016
    - Python 3.7 - 27 Jun 2018

**Current Versions:**

* + - 3.6.3
    - 2.7.14

**Python features:**

Some of the features of python include :-

* + - Easy to understand
    - Dynamic
    - Object oriented
    - Multipurpose
    - Strongly typed
    - Open Sourced

# **Python is mainly used in many domains:**

* + - Web Development
    - Data Analysis
    - Machine Learning
    - Internet Of Things
    - GUI Development
    - Image processing
    - Data visualization
    - Game Development

**IDLE:**

IDLE is an integrated development environment for Python, which has been bundled with the default implementation of the language.

# **Anaconda**

Anaconda is a open source Distribution for data science and machine learning using python. It includes hundreds of popular data science packages and the conda package and virtual environment manager for Windows, Linux, and MacOS. Conda makes it quick and easy to install, run, and upgrade complex data science and machine learning environments like scikit-learn, TensorFlow, and SciPy. Anaconda Distribution is the foundation of millions of data science projects as well as Amazon Web Service Machine Learning AMIs and Anaconda for Microsoft on Azure and Windows.

**Packages:**

**NumPy**

NumPy is the fundamental package for scientific computing with Python. It contains among other things:

* + - * a powerful N-dimensional array object
      * sophisticated (broadcasting) functions
      * tools for integrating C/C++ and Fortran code
      * useful linear algebra, Fourier transform, and random number capabilities.

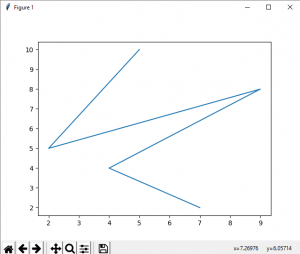
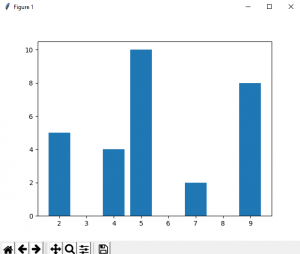
Besides its obvious scientific uses, NumPy can also be used as an efficient multi-dimensional container of generic data. Arbitrary data-types can be defined. This allows NumPy to seamlessly and speedily integrate with a wide variety of databases.

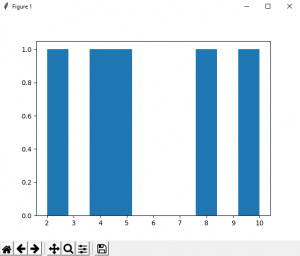
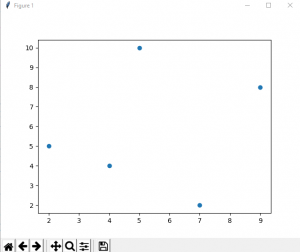
**Matplotlib**

Matplotlib is a Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms. Matplotlib can be used in Python scripts, the Python and IPython shell, the jupyter notebook, web application servers, and four graphical user interface toolkits.

Matplotlib tries to make easy things easy and hard things possible. You can generate plots, histograms, power spectra, bar charts, error charts, scatterplots, etc., with just a few lines of code.

For simple plotting the pyplot module provides a MATLAB-like interface, particularly when combined with IPython. For the power user, you have full control of line styles, font properties, axes properties, etc, via an object oriented interface or via a set of functions familiar to MATLAB users.





**Scikit-learn**

Scikit-learn provides machine learning libraries for python.Some of the features of Scikit-learn includes:

* + - * Simple and efficient tools for data mining and data analysis
      * Accessible to everybody, and reusable in various contexts
      * Built on NumPy, SciPy, and matplotlib
      * Open source, commercially usable - BSD license

**Pandas**

Pandas is an open source, BSD-licensed library providing high- performance, easy-to-use data structures and data analysis tools for the Python programming language.

Pandas library is well suited for data manipulation and analysis using python. In particular, it offers data structures and operations for manipulating numerical tables and time series.

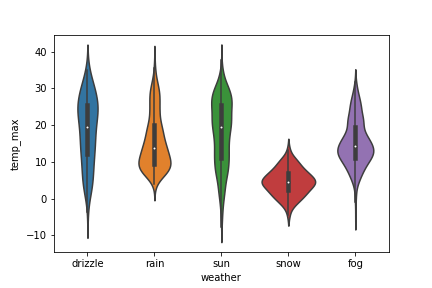
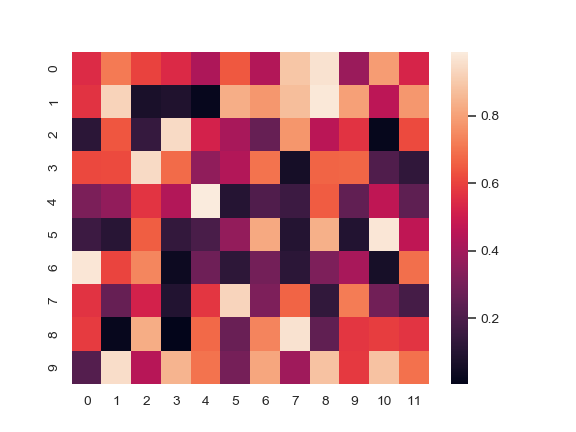
## Advantages of Using Pandas

The following are some of the advantages of the [Pandas library](https://pandas.pydata.org/):

1. It can present data in a way that is suitable for data analysis via its Series and DataFrame data structures.
2. The package contains multiple methods for convenient data filtering.
3. Pandas has a variety of utilities to perform Input/Output operations in a seamless manner. It can read data from a variety of formats such as CSV, TSV, MS Excel, etc.

# **Seaborn**

Seaborn is a Python visualization library based on matplotlib. It provides a high-level interface for drawing attractive statistical graphics. **E.g:-**

 **violinplot heatmap**

# 

**TRAINING WORK UNDERTAKEN**

## COLLECTING DATA FROM KAGGLE

Kaggle is a platform for predictive modelling and analytics competitions in which statisticians and data miners compete to produce the best models for predicting and describing the datasets uploaded by companies and users. This crowd sourcing approach relies on the fact that there are countless strategies that can be applied to any predictive modelling task and it is impossible to know beforehand which technique or analyst will be most effective. On 8 March 2017, Google announced that they were acquiring Kaggle. They will join the Google Cloud team and continue to be a distinct brand. In January 2018, Booz Allen and Kaggle launched Data Science Bowl, a machine learning competition to analyze cell images and identify nuclei.

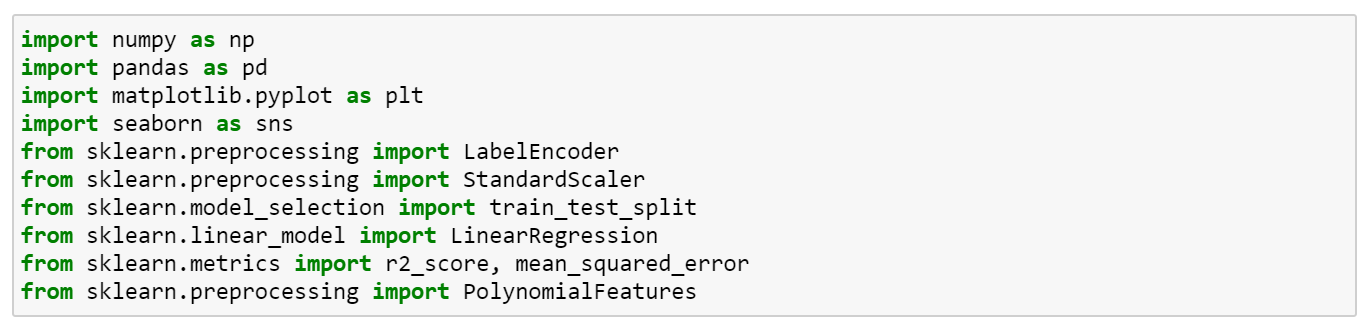
# **DATA SCIENCE**

Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from data in various forms, both structured and unstructured, similar to data mining. Data science is a "concept to unify statistics, data analysis, machine learning and their related methods" in order to "understand and analyze actual phenomena" with data. It employs techniques and theories drawn from many fields within the context of mathematics, statistics, information science, and computer science.

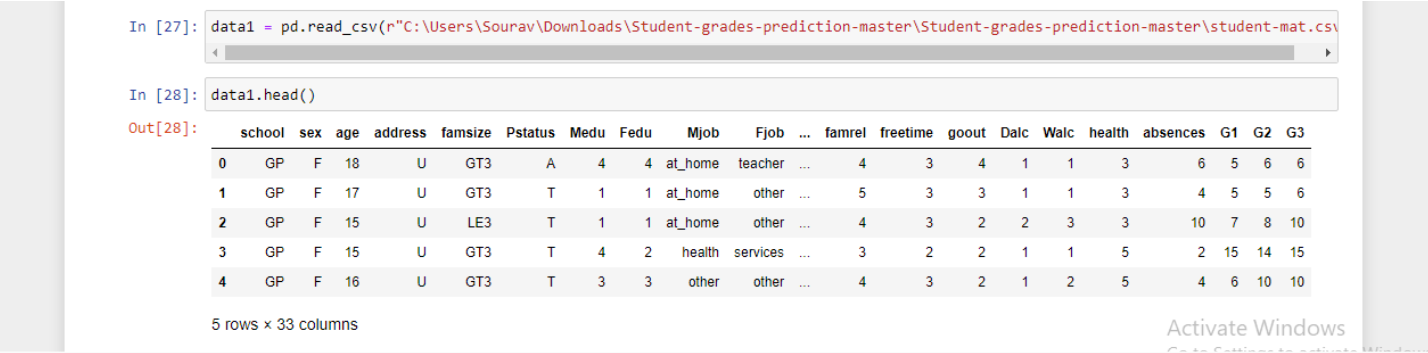
Turing award winner JiGray imagined data science as a "fourth paradigm" of science (empirical, theoretical, computational and now data-driven) and asserted that "everything about science is changing because of the impact of information technology" and the data deluge. When Harvard Business Review called it "The Sexiest Job of the 21st Century" the term became a buzzword, and is now often applied to business analytics, business intelligence, predictive modeling, or any arbitrary use of data, or used as a glamorized term for statistics. In many cases, earlier approaches and solutions are now simply rebranded as "data science" to be more attractive, which can cause the term to become "dilute[d] beyond usefulness." While many university programs now offer a data science degree, there exists no consensus on a definition or suitable curriculum contents. Because of the current popularity of this term, there are many "advocacy efforts" surrounding the field. To its discredit, however, many data science and big data projects fail to deliver useful results, often as a result of poor management and utilization of resources.

**SOURCE CODE & OUTPUT**

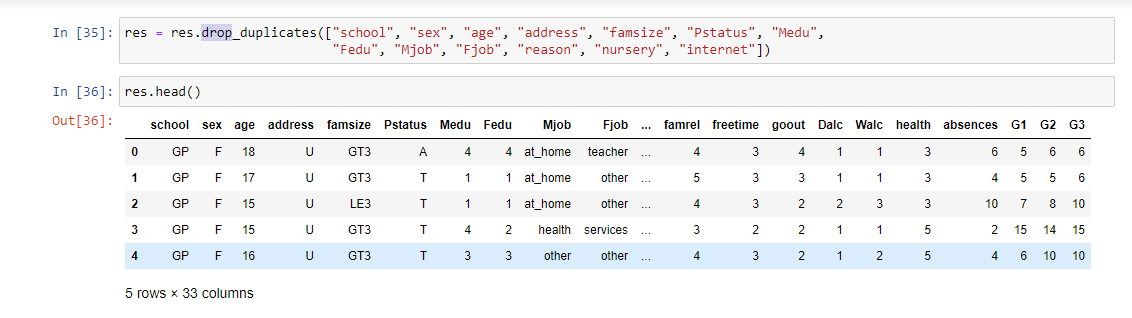
**IMPORT PACKAGES:**



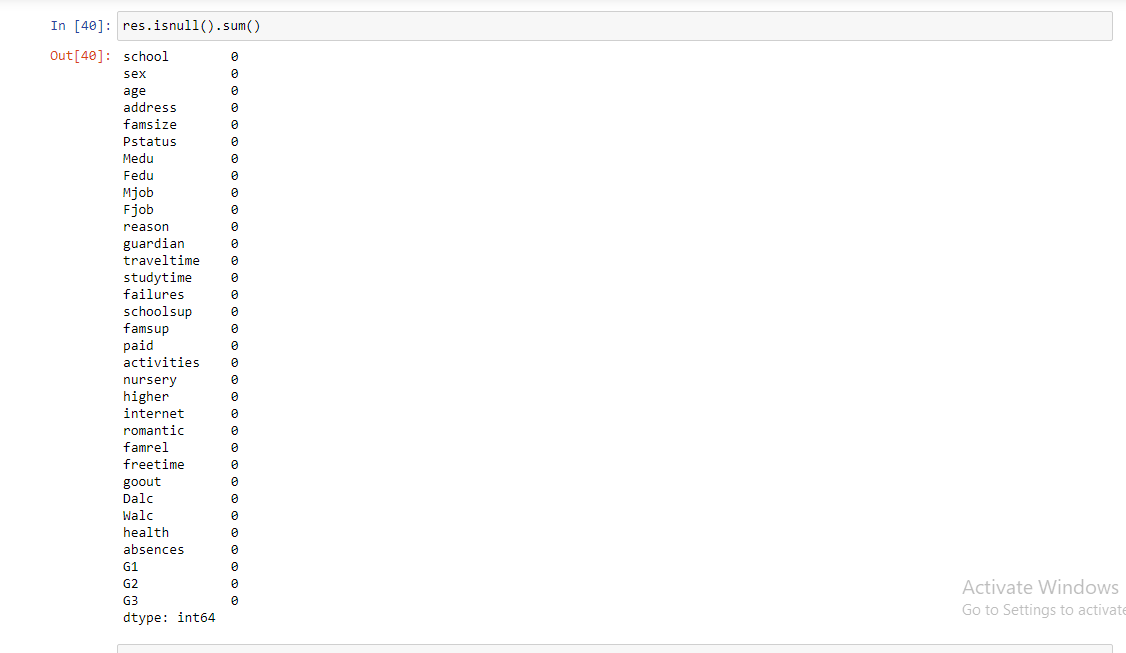
**READING THE DATASET:**

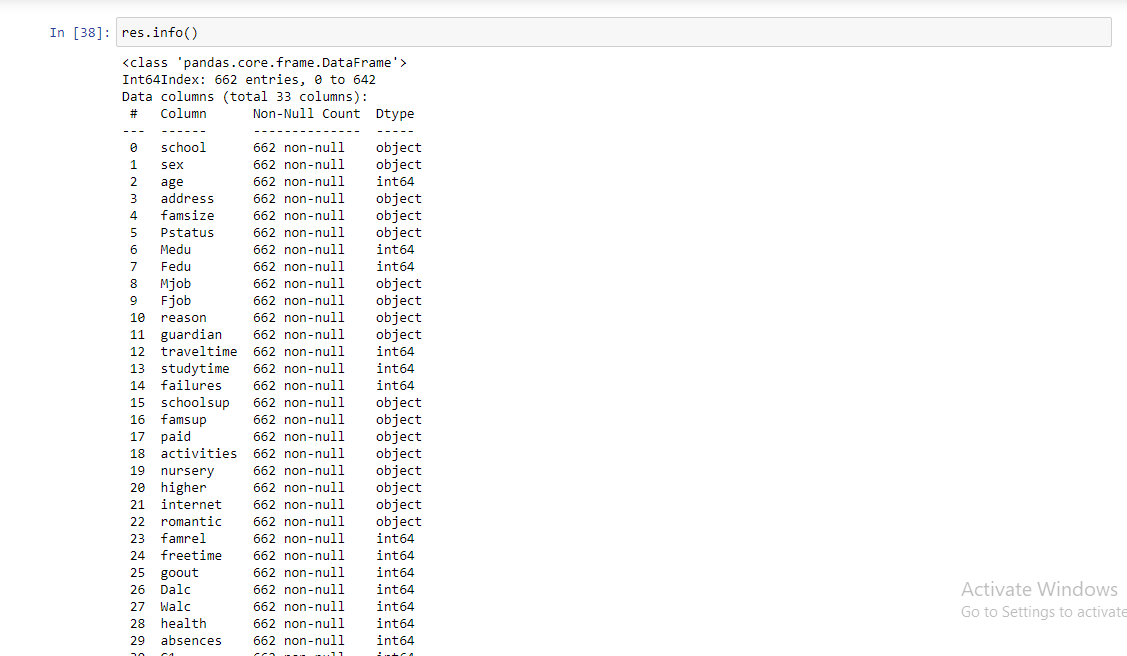


**DROPPING THE NON REQUIRED COLUMNS:**

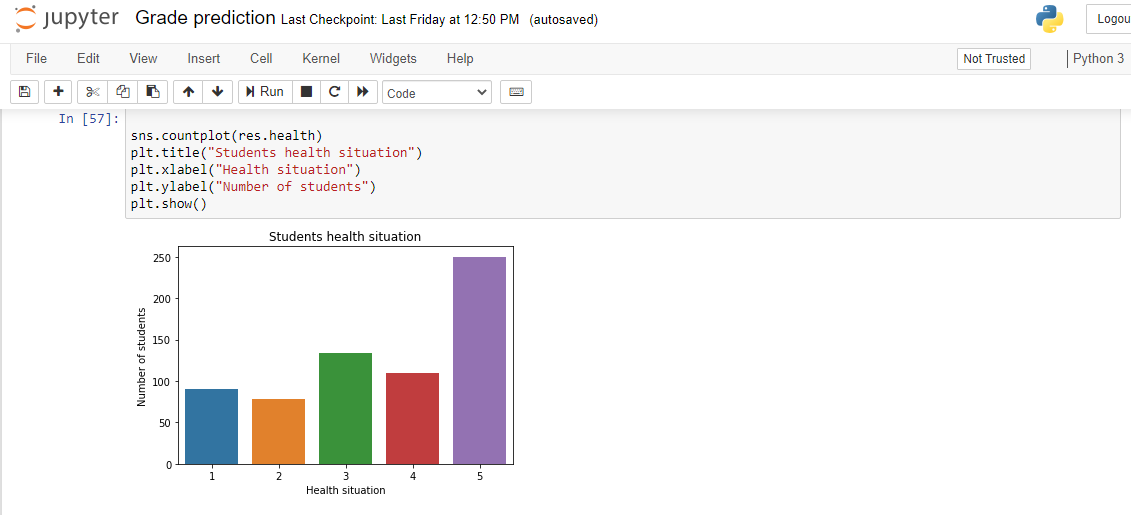
**DATA PREPROCESSING:-**

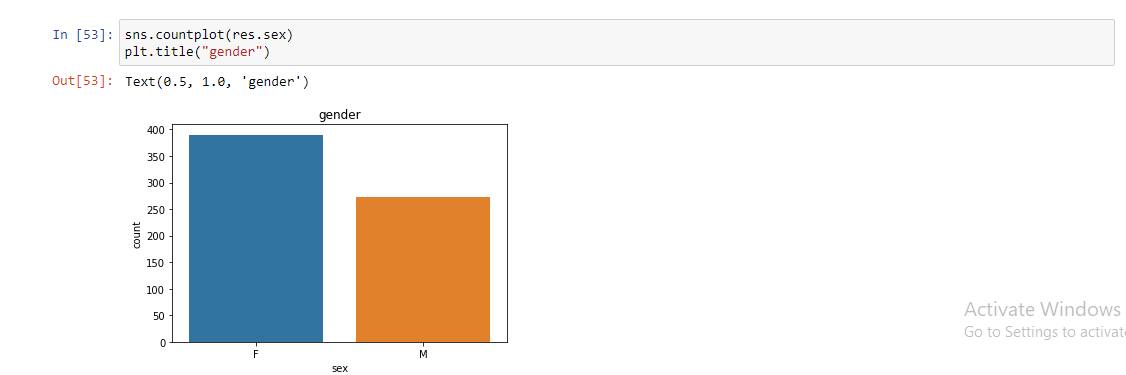
**LOOKING FOR NULL VALUES IF ANY:**

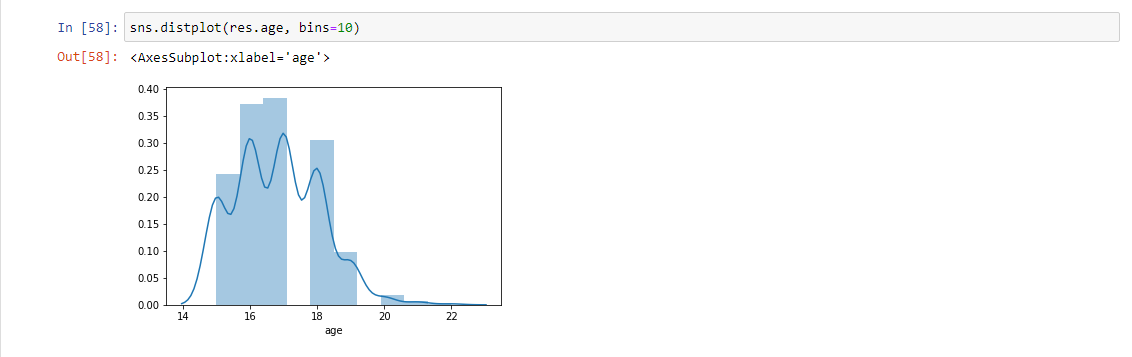
****

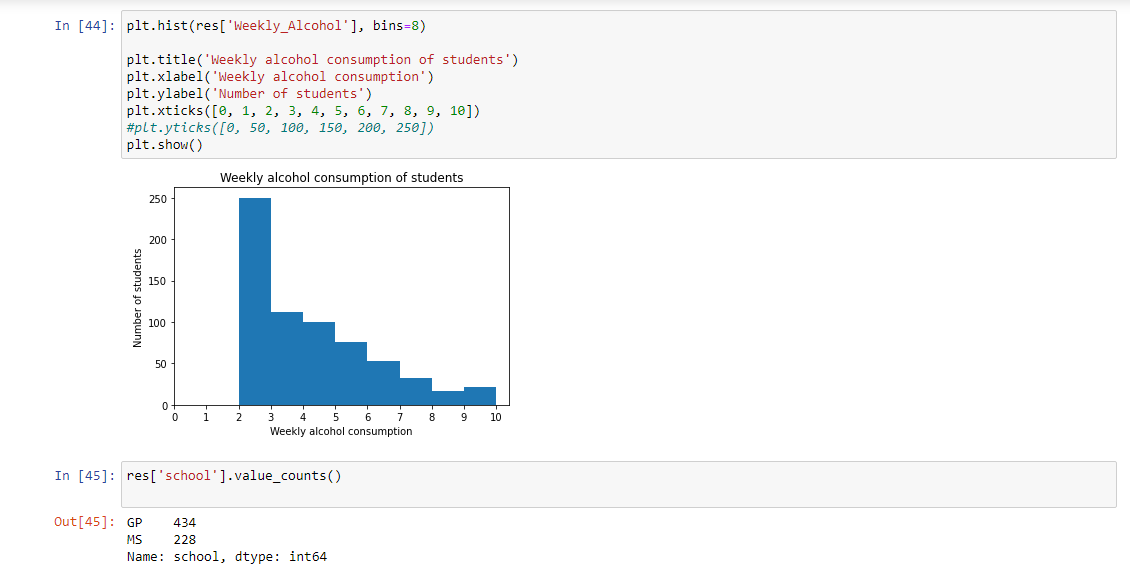
****

**DATA VISUALIZATION:**

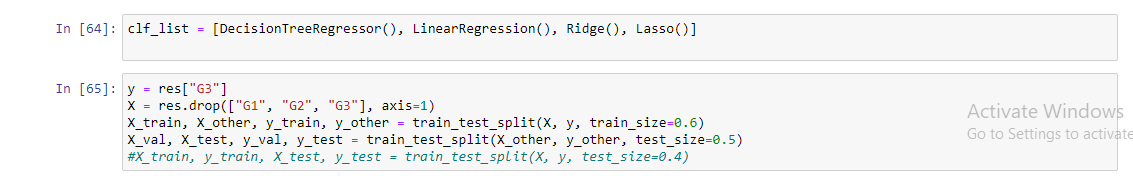
****

****

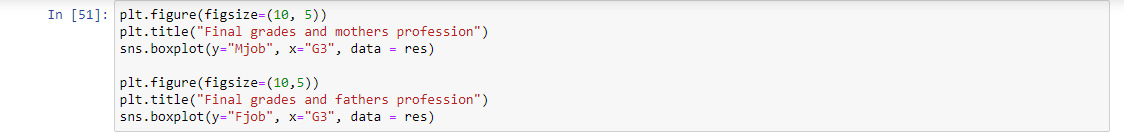
****

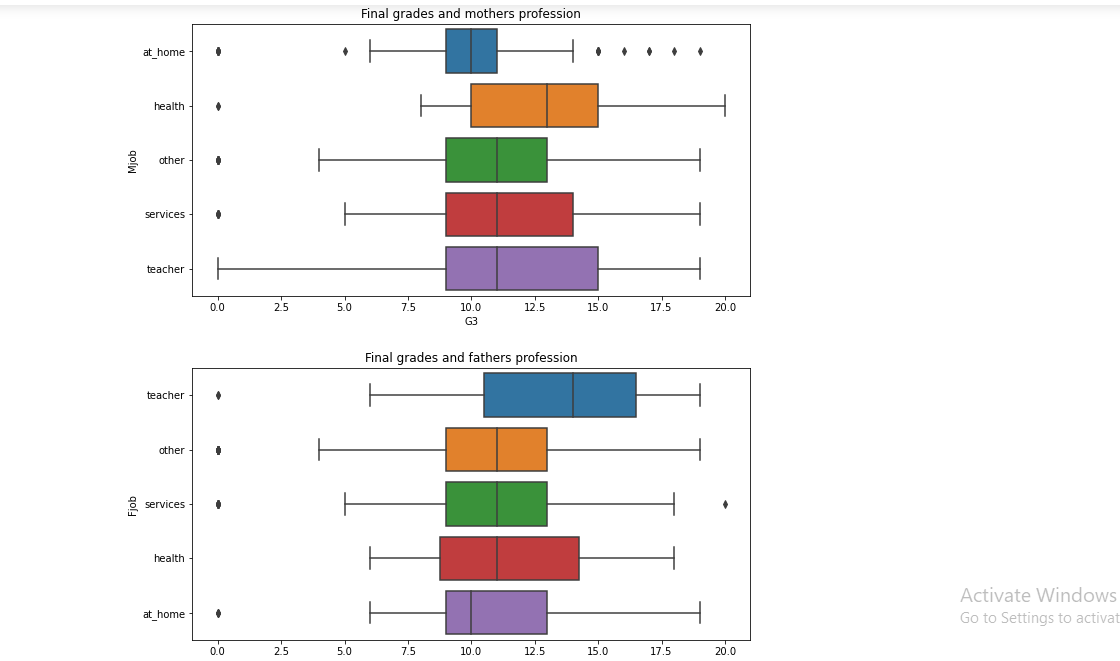
****

**PREDICTION USING LINEAR REGRESSION:**

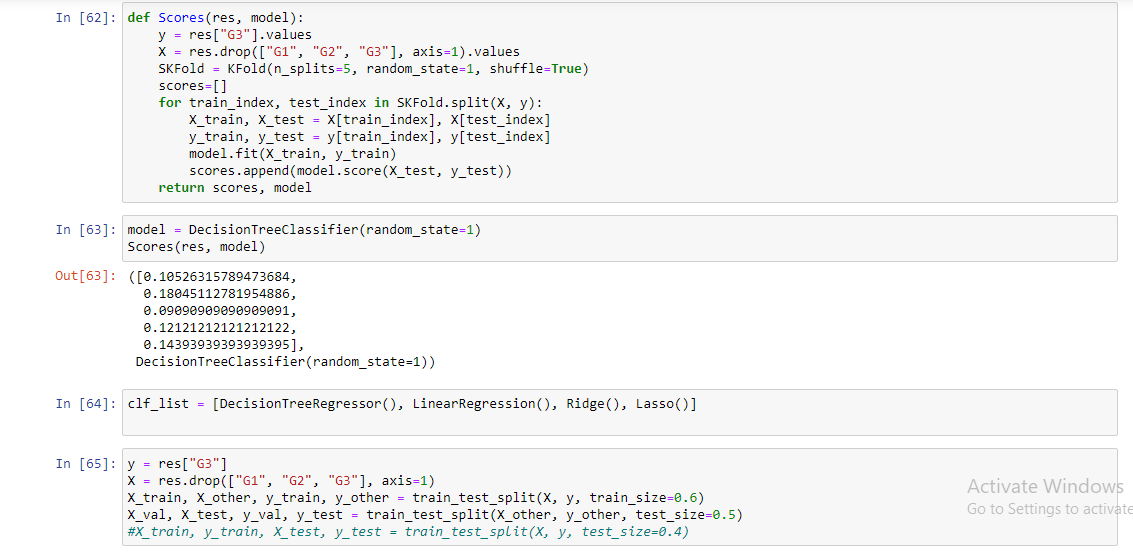
****

**FINAL GRADE PREDICTION:**

****

****

**Building Model:**

****

**CONCLUSION:**

In this paper, the classification, rule-based learning, ensemble methods, and neural network based algorithms are employed in student information to predict the students' division on the premise of previous information. The accuracy was calculated using a confusion matrix. The confusion matrix proved to be a good metric for unbalanced data. Ensemble methods work best with such structured data as compared to its counterpart i.e neural networks. The neural networks require a large amount of data for training.Since the dataset used is less complex and is in a structured format, ensemble methods work best as compared to Convolutional neural networks and Multilayer perceptron for both 3 categories and 5 categories prediction.This study may be quite helpful for both students and teachers for improving the performance and improving the future results of the weak students.

**REFERENCES**

### [**https://www.kaggle.com/**](https://www.kaggle.com/)

### **https://www.python.org/**

### <https://anaconda.org/anaconda/python/>

### <https://www.numpy.org/>

### https://matplotlib.org/

### <http://scikit-learn.org/>

### <https://pandas.pydata.org/>