Netaji Subhash Engineering College

Project Report of Industrial Training

On

Data Science & Machine Learning with Python

Name	Md Imran Khan
Stream	Computer Science & Engineering
Section	B
Class roll	60
University roll	10900117075
Semester	6th
Year	3rd
Session	2020-2021

CANDIDATE'S DECLARATION

I hereby declare that I have undertaken the industrial training at **"WEBTEK LABS"** in partial fulfillment of requirements for the degree of B.TECH **(COMPUTER SCEIENCE & ENGINEERING) at NETAJI SUBHASH ENGINEERING COLLEGE, KOLKATA**. The work which is being presented in the training report submitted to department of **COMPUTER SCEIENCE & ENGINEERING** at **NETAJI SUBHASH ENGINEERING**
COLLEGE, KOLKATA is an authentic record of training work.

Student Name: **Md. Imran Khan** Sign. of the Student

The 60 hours industrial training Viva-Voice examination on Data Science and Machine Learning with Python was held on 28th Feb. 2020 and the project (allotted) submitted was accepted.

Signature of Examiner

ACKNOWLEDGEMENT

The success and final outcome of this project required a lot of guidance and assistance from many people and I am extremely privileged to have got this all along the completion of my project. All that I have done is only due to such supervision and assistance and I would not forget to thank them. I owe my deep gratitude to our project guide **Mr. Abhishek Day and Mrs. Mousita Dhar**, who took keen interest on my project work and guided me all along, till the completion of my project work by providing all the necessary information for developing a good project. I heartily thank our internal project guide, **Mr. Malay Mitra**, **Professor** of **CSE (NSEC)** for his guidance and suggestions during this project work. I am thankful too and fortunate enough to get constant encouragement, support and guidance from all teaching staffs of **WebTek Labs** which helped me in successfully completing my project work. Also, I would like to extend my sincere esteems to all staff in laboratory for their timely support.

Md. Imran Khan

Sem: 6th Year: 3th

Computer Science & Technology

CERIFICATE OF APPROVAL

The project **"Data Visualisation of Titanic(.csv) Dataset"** prepared by **Md. Imran Khan** along with the other group members, is hereby approved as a creditable study for the Bachelor of Technology in **Computer Science & Technology** and presented in a manner of satisfactory to warrant its acceptance as a prerequisite to the degree for which it has been submitted. It is understood that by this approval the undersigned this project only for the purpose for which it is submitted.

Mr. Abhishek Day

(Project In Charge)

CONTENTS

Introduction

Python: Python is a clear and powerful object-oriented programming language, comparable to Pearl, Ruby, Scheme or Java.

Python Features:

- Uses an elegant syntax, making the programs we write easier to read.
- Is an easy-to-use language that makes it simple to get our program working. This makes python ideal for prototype development and other ad-hoc programming tasks, without compromising maintainability.
- Comes with a large standard library that support many common programming task such as connecting to web servers, searching text with regular expressions, reading and modifying files.
- Python's interactive mode makes it easy to test short snippets of code. There's also a bundled development environment called **IDLE.**
- Is easily extended by adding new modules implemented in a complied language such as C or C++.
- Can also be embedded into an application to provide a programmable interface.

- Runs anywhere, including **Mac OS X**, **Windows**, **Linux** and **Unix** with unofficial builds also available for **Android** and **iOS**.
- Is free software in two senses. It doesn't cost anything to download or use python, or to include it in your application. Python can also be freely modified and re-distributed, because while the language is copyrighted it's available under an open source license.

Python's Programming Language Features:

- A variety of basic data types are available: Numbers (floating point, complex, and unlimited-length long integers), strings (both ASCII and Unicode), lists, and dictionaries.
- Python support object-oriented programming with classes and multiple inheritance.
- Code can be grouped into modules and packages.
- The language supports raising and catching exception, resulting in cleaner error handling.
- Data types are strongly and dynamically typed. Mixed incompatible type (e.g. Attempting to add a string and number) causes an exception to be raised, so errors are caught sooner

Python Versions:

- First released in 1991.
- Python 2.0 was released on 16th October 2000.

- Python 3.0 was released on 3 December 2008.
- 2.7.14 was released on 2017.
- 3.8(current stable version)

Application of Python:

- Web Development
- Data Analysis
- Machine Learning
- Internet of Things
- GUI Development
- Image processing
- Data visualization

Anaconda: The open-source Anaconda Distribution is the easiest way to perform Python/R data science and machine learning on Linux, Windows, and Mac OS X with over 15 million user worldwide.

Anaconda's Features: It is the industry standard for developing, testing, and training on a single machine, enabling individual data scientists to:

- Quickly download 1,500+ Python/R data science packages.
- Manage libraries, dependencies, and environment with conda.
- Develop and train machine learning and deep learning models with scikit-learn Tensor Flow, and Theano.

- Analyze data with scalability and performance with NumPy, pandas and Numba.
- Visualize results with Matplot.lib, Bokeh, Datashader, and holoviews.

IPython: Python (Interactive Python) is a command shell for interactive computing in multiple programming language, originally developed for the Python programming language, that offers introspection, rich media, shell syntax, tab completion, and history.

IPython Features:

- Interactive shells (terminals and Qt-based).
- A browser-based notebook interface with support for code, text, mathematical.
- Expressions, inline plots and other media.
- Support for interactive data visualization and use of GUI toolkits.
- Flexible, embeddable interpreters to load into one's own projects.
- Tools for parallel computing.

Introduction To Machine Learning

Machine learning is an application of artificial intelligence (AI) that provides system the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves. The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for pattern in data and make better decision in the future based on the examples that we provide. The primary aim is to allow the

computers learn automatically without human intervention or assistance and adjust actions accordingly.

How Machines Learn: Although a machine learning model may apply a mix of different techniques, the method for learning can typically can be categorized as three general types:

- **Supervised learning:** The learning algorithm is given labelled data and the desired output. For example, pictures of dogs labelled “dog” will help the algorithm identify the rules to classify pictures of dogs.
- **Unsupervised learning:** The data given to the learning algorithm is unlabelled, and the algorithm is asked to identify patterns in the input data. For example, the recommendation system of an e-commerce website where the learning algorithm discovers similar items often bought together.
- **Reinforcement learning:** The algorithm interacts with a dynamic environment that provides feedback in term of rewards and punishment. For example, self -driving cars being rewarded to stay on the road.

Applications:

- Handwriting Recognition
- Medical Diagnosis
- Email Spam Filtering
- Face Detection

Tkinter

Python offers multiple options for developing GUI (Graphical User Interface). Out of all the GUI methods, tkinter is the most commonly used method. It is a standard Python interface to the Tk GUI toolkit shipped with Python. Python with tkinter is the fastest and easiest way to create the GUI applications. Creating a GUI using tkinter is an easy task.

Tkinter is the standard GUI library for Python. Python when combined with Tkinter provides a fast and easy way to create GUI applications. Tkinter provides a powerful object-oriented interface to the Tk GUI toolkit.

Creating a GUI application using Tkinter is an easy task. All you need to do is perform the following steps −

- Import the *Tkinter* module.
- Create the GUI application main window.
- Add one or more of the above-mentioned widgets to the GUI application.
- Enter the main event loop to take action against each event triggered by the user.

Tkinter also offers access to the geometric configuration of the widgets which can organize the widgets in the parent windows. There are mainly three geometry manager classes class.

- **pack() method:** It organizes the widgets in blocks before placing in the parent widget.
- **grid() method:** It organizes the widgets in grid (table-like structure) before placing in the parent widget.
- **place() method:** It organizes the widgets by placing them on specific positions directed by the programmer.

NumPy & Pandas

NumPy:

- NumPy (Numeric Python) is a linear algebra library for python.
- NumPy enriches the programming language Python with powerful data structure for efficient computation of multi-dimensional arrays and matrices.
- A NumPy array is a grid of values, all of the same types, and is indexed by a tuple of nonnegative integers.
- The number of dimensions is the rank of the array; the shape of an array is a tuple of integers giving the size of the array along each dimension.

Pandas:

- Pandas is the most popular python library that is used for data analysis.
- It provides highly optimized performance with back-end source code is purely written in C or Python.
- We analysis data in pandas with:
  - ✓ Series (1-d array)
  - ✓ Data Frame (2-d array)

- Using Pandas, we can accomplish five typical steps in the processing and analysis of data, regardless of the origin of data – load, prepare, manipulate, model and analyse.

Stepsof Machine Learning

To apply the learning process to real world tasks, we'll be a live step process. Regardless of the task at hand, any machine learning algorithm can be deployed by following these steps:

- **Data Collection:** The data collection step Involves gathering the learning material an algorithm will use to generate actionable knowledge. In most cases, the data will need to be combined into a single source like a text file, spreadsheet, or database.
- **Data exploration and preparation:** The quality of any machine learning project is based largely on the quality of its input data. Thus, it is important to learn more about the data and its nuances during a practice called data exploration. Additional work is required to prepare the data for the learning process. This involves fixing or cleaning so-called "messy" data, eliminating unnecessary data, and recoding the data to confirm to the learner's expected inputs.
- **Model training:** By the time the data has been prepared for analysis, you are likely to have a sense of what you are capable of learning from the data. The specific machine learning task chosen will inform

the selection of an appropriate algorithm, and the algorithm will represent the data in the form of a model.

- **Model evaluation:** Because each machine learning model results in a biased solution to the learning problem, it is important to evaluate how well the algorithm learns from its experience. Depending on the type of model used, you might be able to evaluate the accuracy of the model using a test dataset or you may need to develop measures of performance specific to the intended application.
- **Model improvement:** If better performance is needed, it becomes necessary to utilize more advanced strategies to augment the performance of the model. Sometimes, it may be necessary to switch to a different type of model altogether. You may need to supplement your data with additional data or perform additional preparatory work as in step two of this process.

Supervised Learning

Supervised learning as the name indicates the presence of a supervisor as a teacher. Basically supervised learning is a learning in which we teach or train the machine using data which is well labelled that means some data is already tagged with the correct answer. After that, the machine is provided with a new set of examples (data) so that supervised learning algorithm analyses the training data set of training examples) and produces a correct outcome from labelled data.

For instance, suppose we are given a basket filled with different kinds of fruits. Now the first step is to train the machine with all different fruits one by one like this:

- If shape of object is rounded and depression at top having colour Red then it will be labelled as - Apple.
- If shape of object is long curving cylinder having colour Green-Yellow then it will be labelled as - Banana.

Now suppose after training the data, we have given a new separate fruit say Banana from basket and asked to identify it.

Since the machine has already learned the things from previous data and this time have to use it wisely. It will first classify the fruit with its shape and colour and would confirm the fruit name as BANANA and put it in Banana category. Thus the machine learns the things from training data (basket containing fruits) and then apply the knowledge to test data (new fruit). Supervised Learning is where we have input variables(x) and an output variable(y) and we use an algorithm to learn the mapping function from the input to the output.

Process Flow: Supervised Learning

Supervised learning classified into two categories of algorithms:

Classification: A classification problem is when the output variable is a category, such as "Red" or "blue" or "disease" and "no disease".

- A classification model attempts to draw some conclusion from observed values. Given one or more inputs a classification model will try to predict the value of one or more outcomes.
- For example, when filtering emails "spam" or "not spam", when looking at transaction data, "fraudulent", or "authorized".

Regression: A regression problem is when the output variable is a real or continuous value, such as "dollars" or "weight" or "salary".

- Many different models can be used, the simplest is the linear regression.
- It tries to fit data with the best hyper-plane which goes through the points.

Data Visualisation of Titanic(.csv) Dataset

Aim:

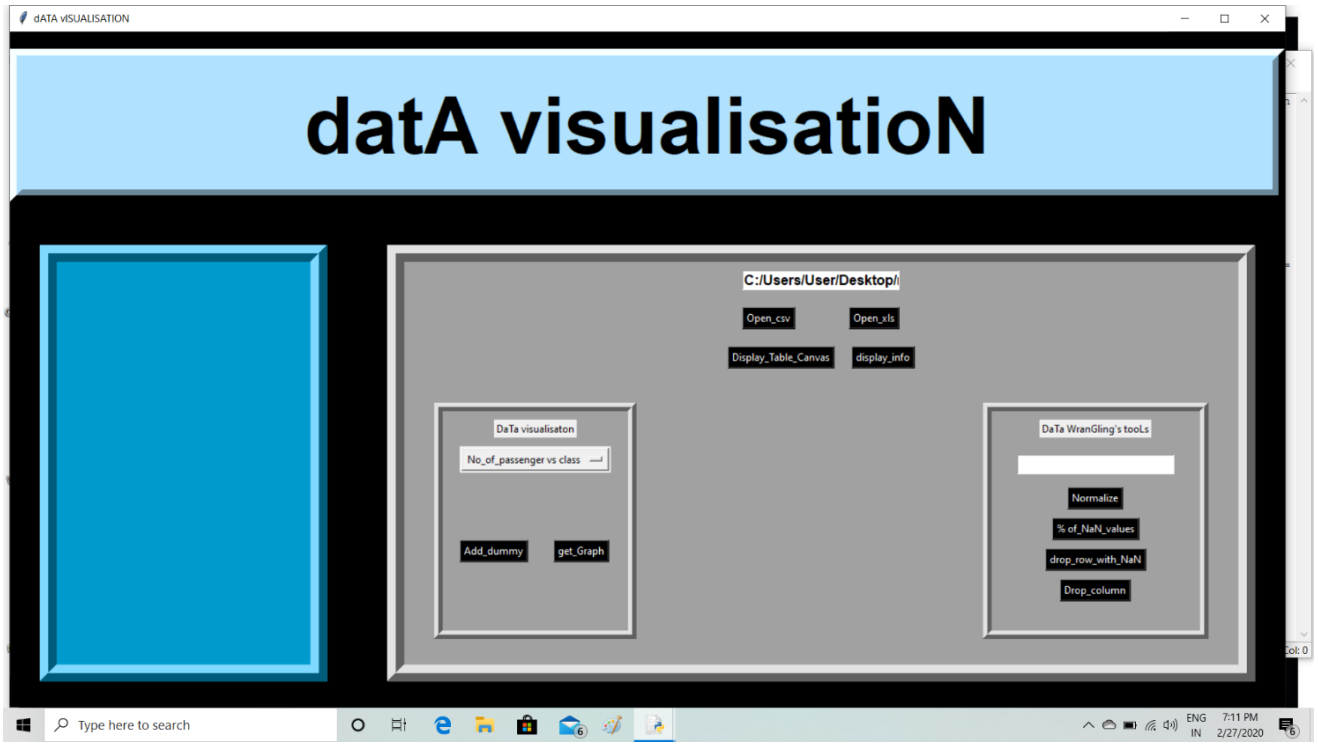
To perform data wrangling and data visualisation.

Objective:

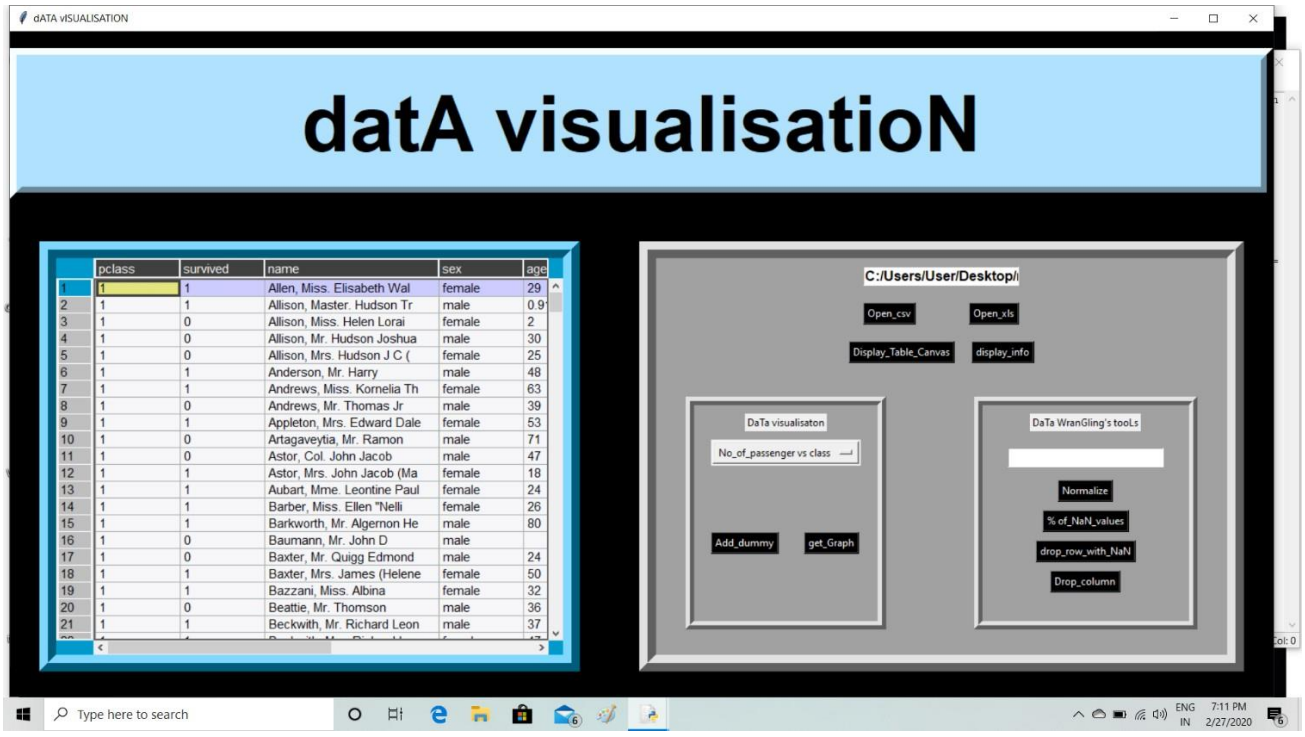
- To import desired data-set(titanic.csv)
- To display data-set in tabular form.
- To display information which includes column_names, no_of_NaN_values and data_type of respective columns.
- Perform data wrangling i.e drop_column, drop_row, add_dummy_column etc
- To perform data normalization
- To visualize data on different graphs plot and pie-charts.

tkinter	to create window and import built-in widgets i.e Button, optionmenu
matplotlib	to visualize data on to graphs and pie charts i.e. bar graph
re	to match pattern
pandas	for data wrangling i.e drop_column, drop_rows , add_dummy
numpy	to normalize data
tkintertable	to put data-set on canvas embedded on tkinter table

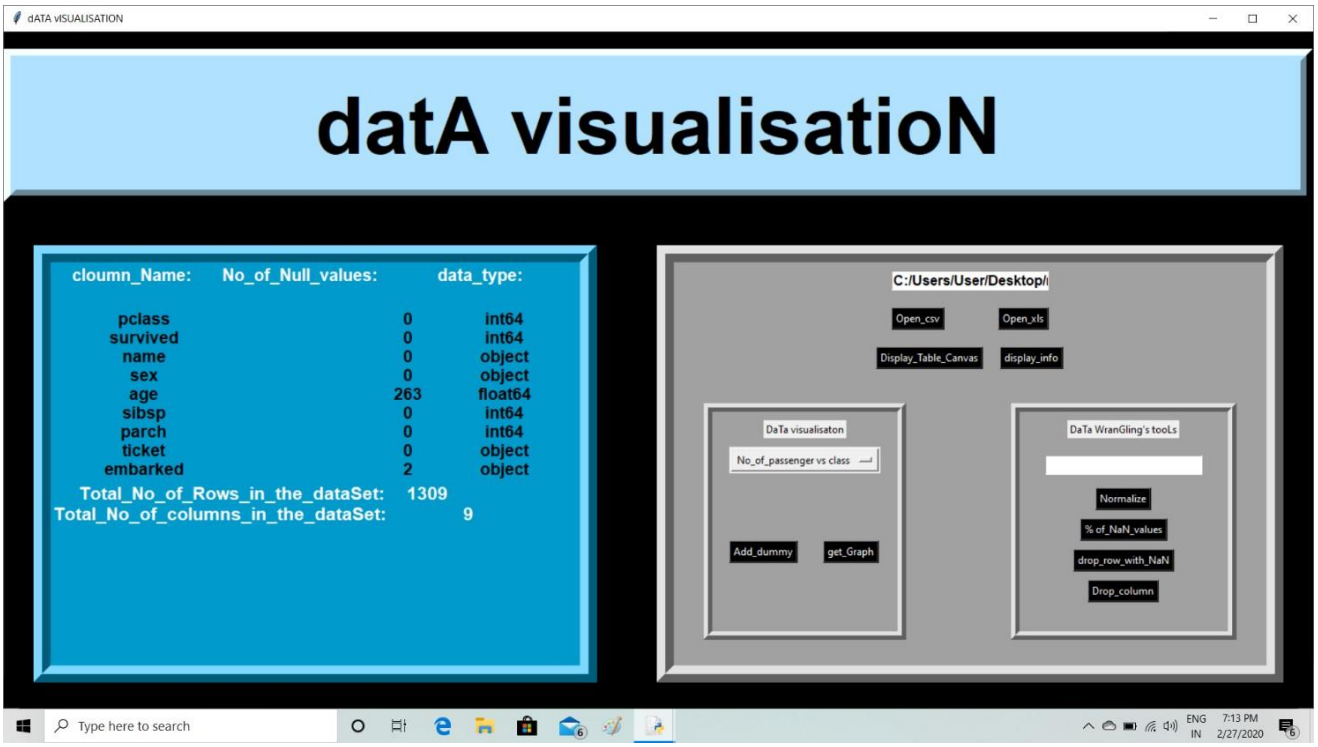
---------- overview of the main window ----------



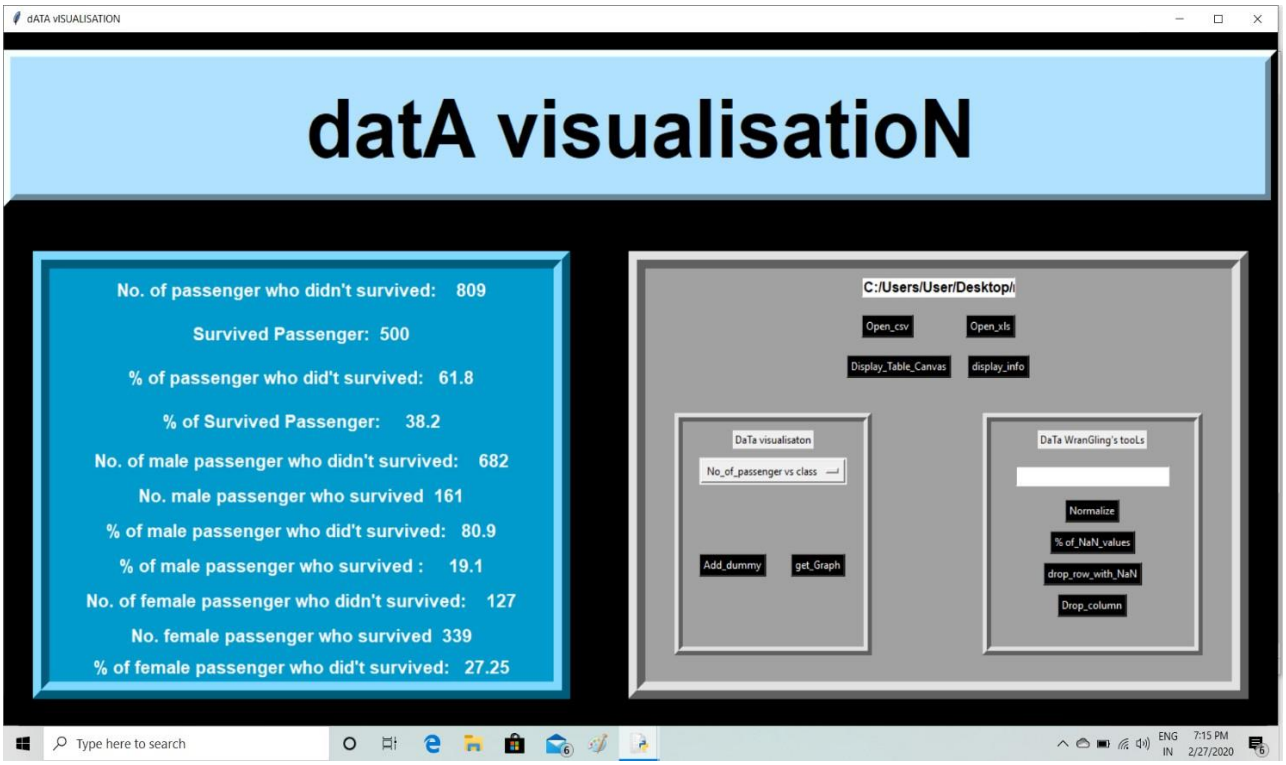
----------information displayed on Canvas ----------



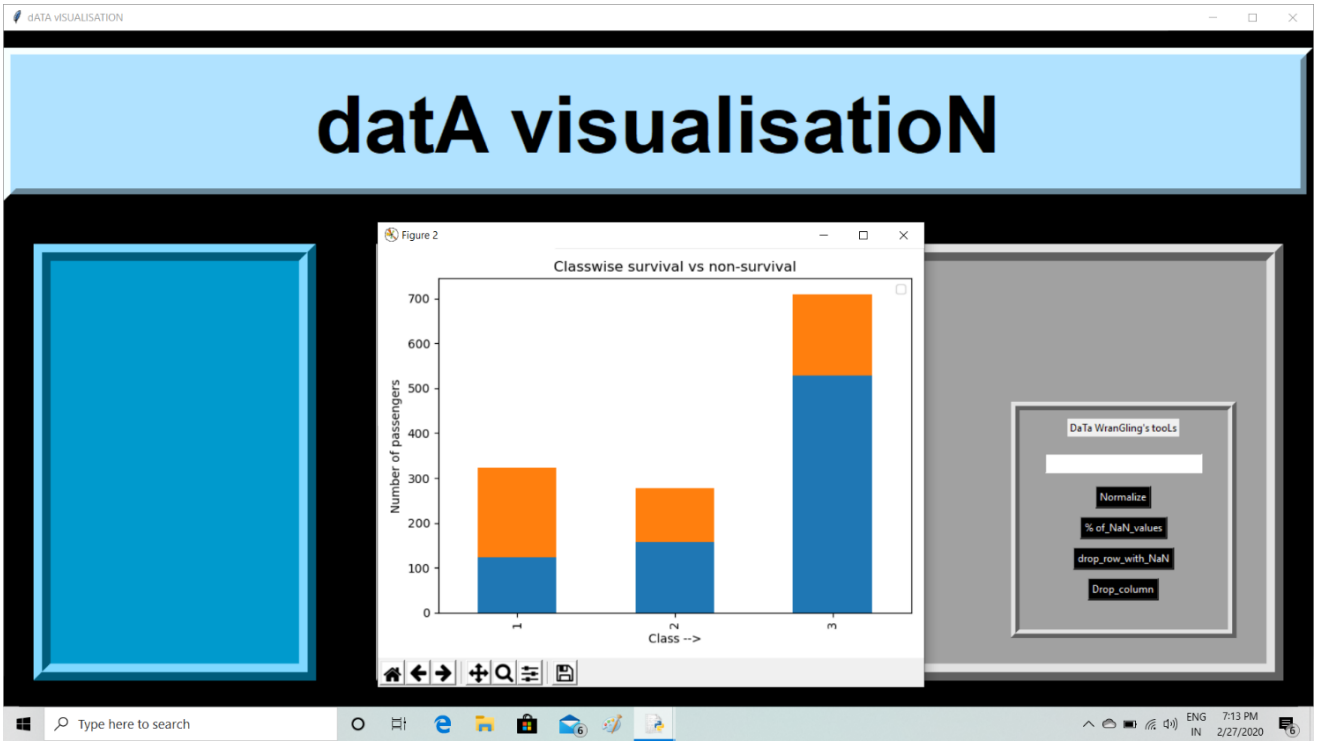
----------information of dataset----------



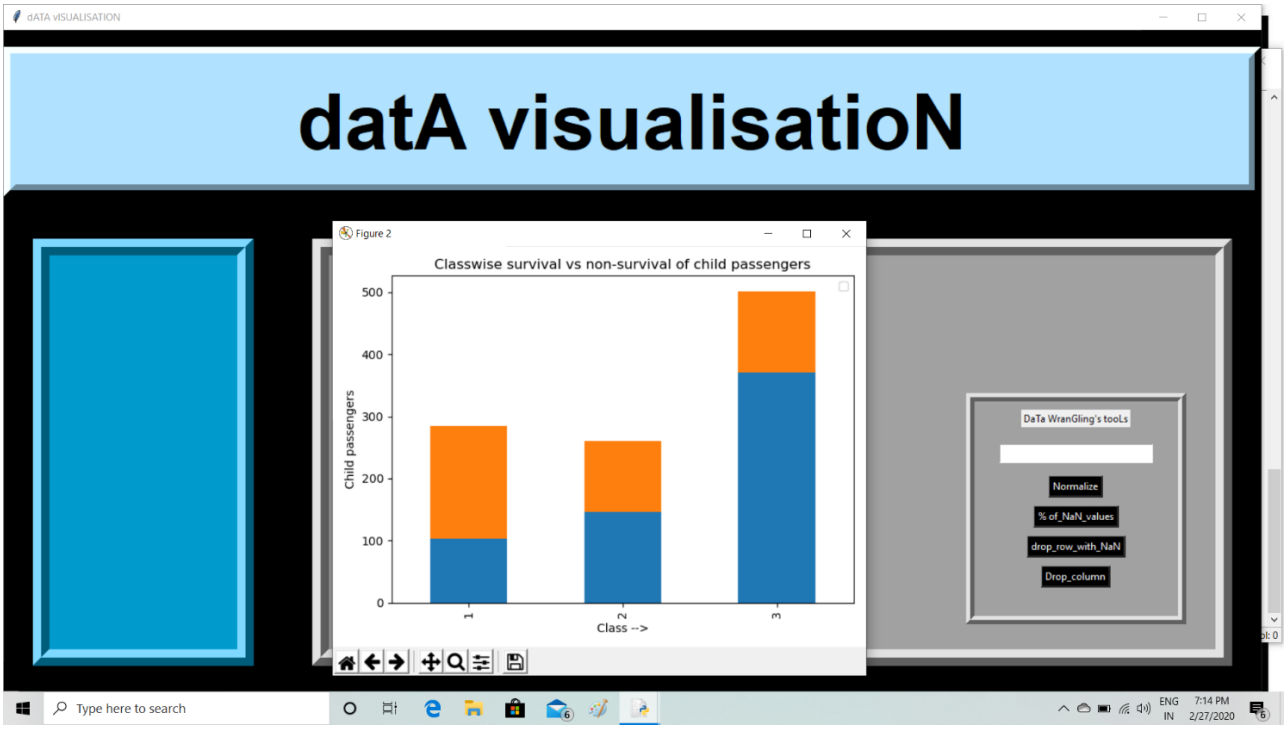
----------normalisation----------



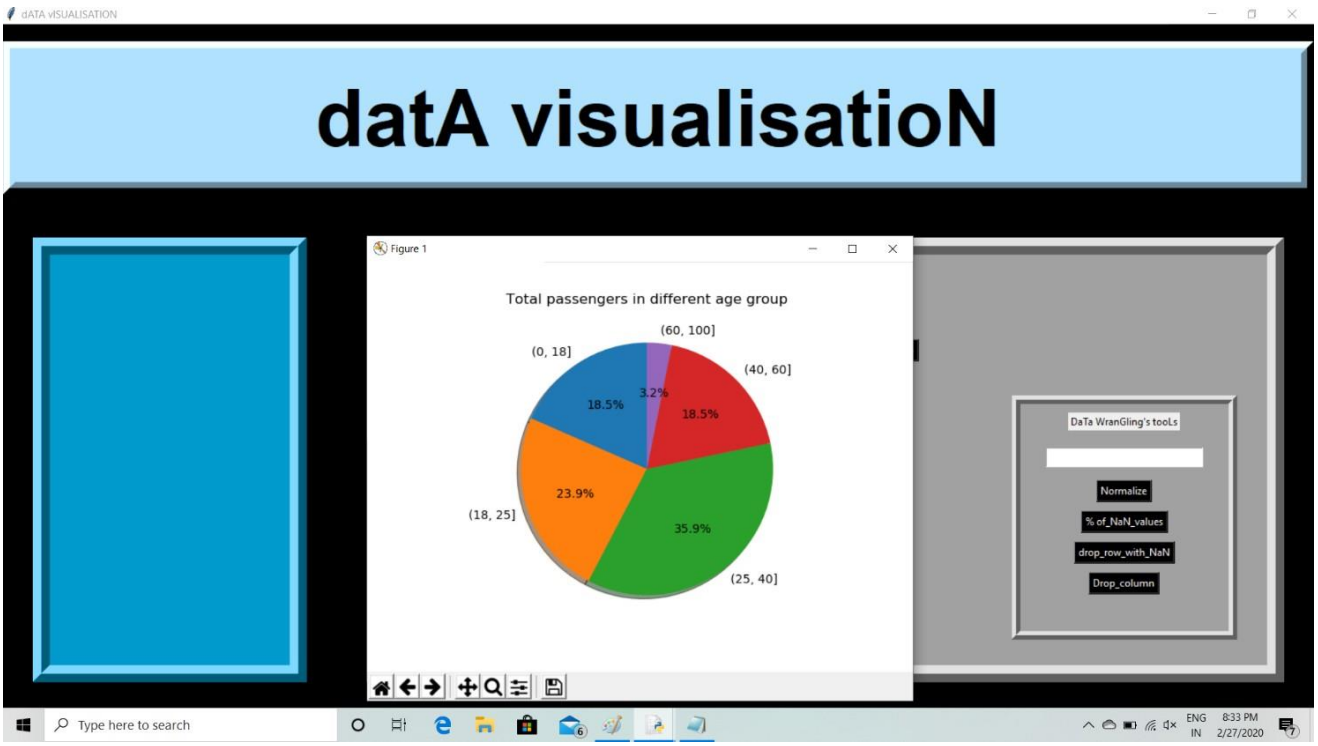
----------classwise survival vs non-survival----------



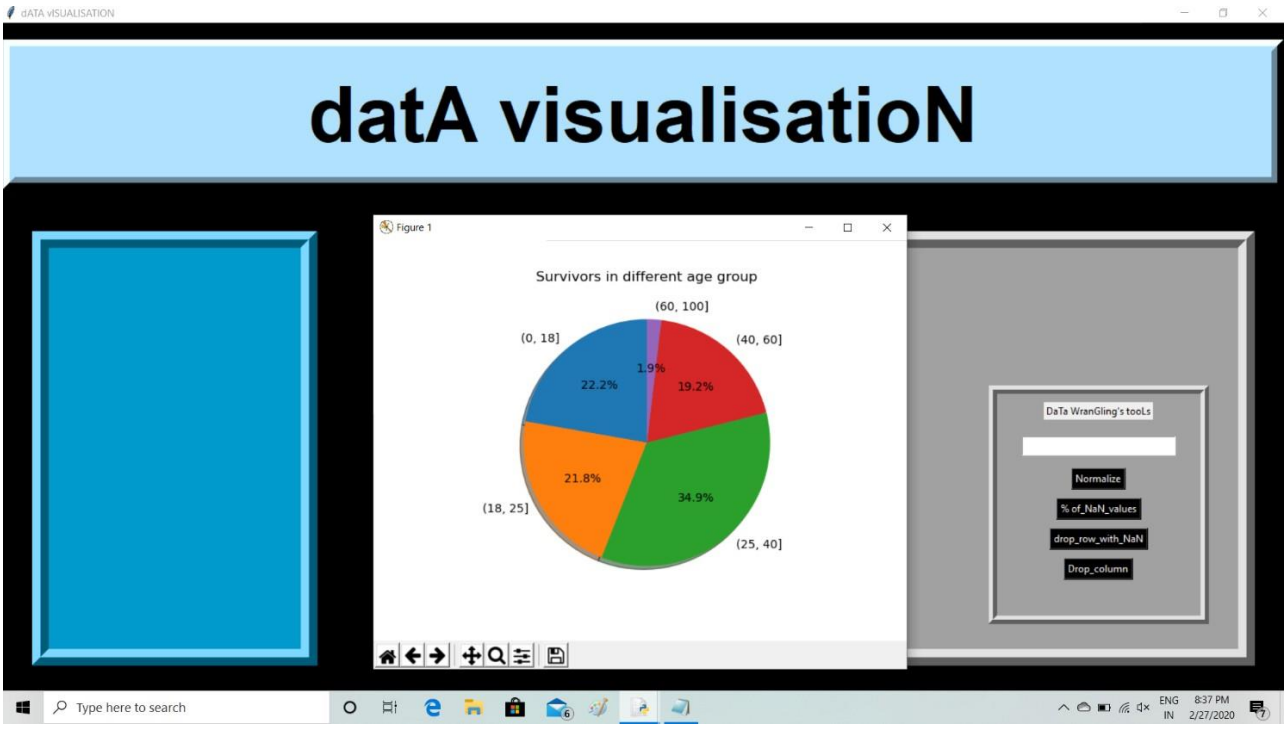
-------classwise survival vs non-survival of child passengers---------



----------passenger in different age group----------



----------survivor in different age group----------



Data Visualization:

- Data visualization is the discipline of trying to understand data by placing it in a visual context.
- Python offers multiple great graphing libraries that come packed with lots of different features.

- Can be done with the help of Matplotlib.

Matplot.lib:

- Matplot.lib is the most popular python plotting library.
- It is a low-level library with a Mat lab like interface which offers lots of freedom at the cost of having to write more code.
- Matplot.lib is specifically good for creating basic graphs like line charts, bar charts, histograms and many more.
- Can be imported as Import matplot.lib pie plot as plat.
- Different plots in matplot.lib:
    - ✓ Scatter plot
    - ✓ Plot
    - ✓ Histogram
    - ✓ Bar chart

Conclusion

In **"Data Visualisation of Titanic(.csv) Data-Set"** project, we were asked to perform data wrangling ,normalization and data visualisations.

During the process of project preparation I and my whole team put tremendous effort in developing a dynamic tool for analysing a given dataset. In our project we incorporated following stuffs:

- Functionality of adding any dataset using file_ dialog _menu.
- Plotting a different types of graphs using option menu.
- It has provision of dynamically dropping column, rows etc.
- It also includes buttons to display dataset in tabular format.

We accomplish above mentioned functionality using several modules like tkinter, matplotlib, re, numpy, pandas, tkintertable etc along with various build in methods like drop(),grid(),pack() ,plot() etc.

Though we tried to make it as user friendly as it can be but some portion still needs improvements such as graph plots to make it a complete data analytic tools. Hence we posted this project on Github(https://github.com/Imrankhan2712/Data-Visualisation/blob/master/pull.py) for future improvisation .

References:

- https://www.w3schools.in/python-tutorial/gui-programming/tkinter
- https://matplotlib.org