Reflective Report on Portfolio 4

This is the Reflective report which covers process of solving problems using data. How I approach to the problem Solving by solving it in the Jupyter Notebook?

I will discuss over the progress of my unit on how I approached it and how I will use the learnings in the future. Then, I will discuss over the two points of discussion related to portfolio 4.

Problem Solving and Learning to Use Jupyter Notebooks

Initially, After choosing this unit I was in a dilemma on how to approach the unit. After attending the few practical workshops, I got an approach on how to lead on the fundamentals of data science. I found the Jupyter Notebook quite productive, as we were able to run the python 3 code snippet wise. The Jupyter notebook is good for documentation which makes it easier for the user to read. Jupyter Notebook records the entire workflow and it can be edited and run easily. All the machine learning libraries had good support over the Jupyter Notebook which makes it super-flexible.

During each portfolio I was learning the entire process of data Science. It all starts with the Problem Definition, Understand the business context of E-Commerce to understand the context. Datasets were provided in the portfolios which I uploaded over the Jupyter environment.

The biggest part of the entire data processing is data Cleaning and Pre-processing. In this I Learned about the handling missing data and how to detect and deal with outliers. How to handle the categorical variables by using encoders and doing normalisation and standardisation to the numerical variable. New relevant features can be made as per the context of the data.

The Exploratory Data Analysis section of the portfolio helped me in getting the insights related to the problem statement. Data Visualisation and Exploring the Data with patterns, trends and correlations. Feature selection step helped in selecting the relevant and important features related to the dependent variables.

Throughout this unit I become much confident in using Jupyter notebook and applying problem Solving through data science.

Progress And Future Interest.

With strong foundational concepts of data science, now I am quite confident in approaching the advance concepts like Deep Learning and natural language processing. I would like to Explore these topics and in future I would love to apply to the datasets that we received in portfolios.

Applying the data science based problem solving I would like to participate in Kaggle Data science Competition. I would like to enhance my profile over the Kaggle. I would like to be in top 10 submissions to any Kaggle Competition and keep my name and the Macquarie University name to the top.

Applying these skills on a practical ground is very crucial I would love to join start up company in which I can apply the learnt skills and apply with real datasets on any cloud environment.

Why you choose the dataset you have used for your portfolio 4 ?

In the portfolio 4, the choice of data set plays an important role as the covid data has a great dependency on the lives. Seeing which country performs better can help the government decide over the policies and share the information around the world.

Dataset provides proper demographics related to the problem. Date wise analysis and the visualisation from the data were clean .

Data had 223 countries and the date from 2019 to 2022 which is quite appropriate to the times. The size of the dataset is quite appropriate which shows there's enough information related to the Covid 19 predictions. The dataset has more than 98252 rows and 8 features of the same.

I consider the COVID-19 epidemic to be one of the most significant occurrences of our time, and I am passionate about utilizing data analytics to aid in understanding and mitigating this problem. I am certain that the knowledge and expertise I obtained from working with this dataset would be useful in my next profession as a data scientist.

The reason to choose your machine models in portfolio 4 and why these models are suitable for solving the problem you have raised ?

In the Portfolio 4, I have used K means algorithm. It helped me in diversification of clusters of countries which performed well during covid times. It’s a simple algorithm to read and applying through scikit makes it easy on Jupyter notebook.

The Elbow method helps in optimizing the model by judging the right K which in turns result the most appropriate number of cluster. The Dendo-diagram also helps in clarity. K means Roustness and clear visualisation has helped in selecting the countries that are performing well and countries which need to work over it's policies. ### It helps in the convergence of all the features.

In prediction of confirmed cases Initially i started with linear regression as the trend was not linear i applied Polynominal regression to improve the accuracy of the model

Conclusion to Portfolio 4

It appears that the growth of confirmed and death cases has slowed. which is a great sign. Hope that continues for a short while. Just as the USA occurred to be that epicenter for a brief while, no other nation should emerge as the new epicenter of COVID-19. The number of confirmed cases will increase dramatically if a new country becomes the epicenter.