Reflective Report

Beatrice Pamittan 47558865

Report the process of solving problems and learning to use notebooks

First, there needs to be an understanding of the data, objectives, and a defined problem. The next step would be to dive into the data and choose a suitable dataset to solve the problem. Once this has been established, exploratory data analysis should be m conducted. This is where we would explore the data types, size of the dataset, features, and overall quality of the dataset we are working with. Next, we can proceed with data cleaning and processing. In this step, we identify any missing values and decide whether to drop these or impute their values where possible. We also transform the data by removing outliers and inconsistencies. In addition to that, we can also transform the features by encoding them, converting them into dummy variables and the like. After this, we can do the data visualization to see the distribution of the data and features in the dataset. Moreover, we can further analyze the data by doing univariate, bivariate, and multivariate analysis. Visualization will also help understand patterns, relationships, correlations, and trends to gain further insights.

At this stage, we would have established what kind of dataset we are dealing with and the problem to solve. This would then guide us in choosing which machine learning model to use to train our data based on the problem and characteristics at hand. If necessary, we also split and train the data. Once the model has been set up and run, we can now conduct model evaluations. Evaluating a model involves the use of metrics such as accuracy, recall, precision and more. Based on the metrics we obtain, we can further iterate on the models by narrowing down the features and further optimizing the models. Given these metrics, proper analysis and documentation of the results should be written down clearly to share with stakeholders and others that would like to understand the notebook. Lastly, if more than one model was used to evaluate the problem, there should be a comparison between the accuracy and other results of the model to determine which would be the best model for the scenario or problem.

Regarding learning to use notebooks, I had to be familiar with Jupyter Notebooks specifically and its capabilities and the syntax for Markdowns and Python code. It was always good to have references at hand for basics of Markdown codes and symbols needed as well as the syntax of libraries such as pandas, numpy, scikitlearn and more. It was good that we were initially provided with Portfolios that had a guide on how to go about them as it set the standard for the portfolios that we needed to churn out ourselves by Portfolio 3 and 4. I would say that it definitely took practice to be able to get to the level of comfortability I am at now with creating notebooks for data science projects moving forward.

How have you progressed from the start of the unit, what are you interested in doing with this in the future?

From the start of the unit, I had very little knowledge on data science. Most of what I knew back then came from high level topics or articles that I would see on Linkedin or on other forms of social media. I would hear of what types of machine learning models there were such as

supervised or unsupervised but had no idea what they meant. This unit has definitely helped me build and gain the knowledge I needed for Data Science. It was a very rigorous unit with numerous assessments involved but I see the value in each and every one of them. We had to put our learnings into practice and that was what the weekly practicals and 4 portfolios were for. I definitely learned a lot throughout this semester from what data science really is, to the methodology, and on to the different supervised and unsupervised machine learning models. It initially seemed daunting to learn a new coding language and embark on multiple portfolios with it but I have seen that it was a great experience for me as a whole. It tested my determination and persistence, especially during times when I encountered some difficulties. In the future, I would like to continue to create more projects in my spare time to continue building on my skills and so that I get to improve overtime as well. Some topics off the top of my head that I would like to explore further would be in the field of supply chain management, customer relationship management, and retail datasets.

Portfolio 4 Discussion

Why you choose the dataset you have used for your portfolio

Portfolio 4 was an interesting assessment as we had the freedom to choose our dataset and the models we would use. I chose the Breast Cancer Diagnostic dataset from the University of California Irvine because cancer unfortunately runs in the family and I have a few family members who have lost their battle to breast cancer and other forms of cancer. I wanted to learn more about what are the attributes that doctors in the field look at to determine that a cell is benign or malignant and what would be the more significant features that would predict if a cell was indeed cancerous or not.

How do you identify the problem you target to solve in your portfolio?

Identifying the problem was quite straightforward in this case, as given a set of nuclei measurement features, the target variable in my chosen data set would be the diagnosis. In other cases, I would approach identifying the problem by first identifying who are the potential stakeholders for this certain dataset and identify what their needs would be. To scope down on the problem, it would be helpful to identify the value and impact of the problem to solve. This is so that the study or project to be conducted would bring substantial value to the stakeholder.

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

The machine learning models I have chosen for portfolio 4 are Logistic Regression and K-Nearest Neighbors (KNN). These are both supervised machine learning models as the target variable can be located within the dataset. Logistic Regression is great for predicting categorical dependent variables. In my portfolio's case, the diagnosis was converted to binary variables. Logistic Regression is useful in identifying relationships between features as well as it is also computationally efficient and can handle large datasets. The other model I chose was KNN because it is sensitive to local patterns which would be a good option as samples in the dataset that have the same classification would be near each other. KNN is also able to handle data that is imbalanced. In addition to that, KNN was also chosen for its adaptability and simplicity.

Can you well explain the insights or conclusions you draw from your study? Is the result consistent with your intuitive expectation?

The study used two models, namely Logistic Regression and K-Nearest Neighbors (KNN). The model that emerged with the highest accuracy score is Logistic Regression - RFE with an accuracy of approximately 96.49%. This suggests that the Logistic Regression model with Recursive Feature Elimination (RFE) performs the best in terms of accurately predicting the classes for the given dataset. The model with the highest ROC curve area is also Logistic Regression - RFE with an ROC curve area of approximately 0.9969. The high ROC curve area indicates that Logistic Regression with RFE has a superior ability to discriminate between the positive and negative classes. Considering both accuracy and ROC performance, Logistic Regression - RFE stands out as the best-performing model for the given classification task. It demonstrates the highest accuracy and the largest area under the ROC curve, indicating strong predictive capabilities and effective class separation. Therefore, Logistic Regression with Recursive Feature Elimination (RFE) is recommended as the optimal model for this particular dataset based on the provided metrics. I initially thought that the KNN model would yield the higher accuracy among the two but the accuracy was not also very far off from each other. Amongst the 30 features, I already had a feeling from the start that not each and every feature would be very significant, hence why I conducted RFE to narrow down the features and it did increase the accuracy once it was trimmed down.