47413158 Aiden Lowe-Yip COMP2200 Portfolio 4 reflection

**Reflective Report on Portfolio 4**

**Introduction**

This report aims to reflect on the work I have done throughout the period of doing Portfolio 4. Whilst working on this report I have significantly developed my skills in handling and analysing data.

**Choice of Dataset**

Initially, the finding the right dataset seemed challenging due to the me having no clue on what dataset to analyse. However, one dataset caught my eye. I remember by tutor mentioning what real life applications machine learning has. As soon as I remembered this, I decided on the dataset of malignant cancer predictions named 'Wisconsin Breast Cancer Diagnosis'. Additionally, the dataset had clear goals set out in which I was able to distinguish if a patient had malignant or benign tumors.

**Identifying the Problem**

The main problem was whether a model can predict or diagnose breast cancer based on selected features. It involved identifying whether a tumor is malignant (cancerous) or benign (non-cancerous) using selected features such as radius, texture, perimeter, and area of the tumor cells. The goal of this portfolio is to build a model that can predict the diagnosis based on these features effectively.

**Challenges**

The challenges I faced not only extended to writing and evaluating code, but time management. Whilst working on this portfolio, I realised that with all of my other assignments from different units that I had to mange my time better. I approached this challenge by setting a timer at regular intervals so I could evenly distribute time to my work. Additionally, When evaluating models, I struggled with understanding what to do when getting errors after training the K-Nearest Neighbors.

**Machine Learning Models**

After loading the dataset and cleaning the data. I removed any columns that did not correlate to the predictions I wanted. Additionally, I changed features based on if they were numeric or categorical so that no errors would compute. Whilst working on this portfolio, I considered several machine learning models including Logistic Regression, K-Nearest Neighbors (KNN), Decision Tree Classifier and Random Forest Classifier. I chose these models because of thier effectiveness for binary classification problems. I then evaluated these models using accuracy scores by comparing their performances to see which model was the most effective.

**Conclusion**

Based on the model accuracy scores, the Random Forest Classifier had the highest accuracy amongst all the models that were tested. The results were within my expectations, where more complex models with the ability to analyse complex data performed better. This Portolio task reinforced my ability to select features and machine learning models. Furthermore, this portfolio has improved my understanding of real life machine learning applications in a healthcare system and improved my ability to approaching other tasks involving predictive modelling and data analysis.