Reflective Report

Mai Vy (Vivian) Nguyen | 47554029

1. Learning Process and Using Notebooks

In the initial half of COMP6200 (Weeks 1-7), I was systematically introduced to the foundational principles of Data Science. Beginning with a comprehensive understanding of Data Science, I adeptly learned Python programming and explored Jupyter Notebooks, an essential tool in this domain. The weeks that followed helped me to better comprehend data handling as I became proficient in descriptive statistics and understood the intricacies of data structures.

Privacy in data analytics emerged as a pivotal consideration in Week 3. Subsequently, I got a deeper insight into data visualization, acquiring the ability to discern patterns and relationships, crucially differentiating between correlation and causation. By Week 5, I approached predictive modelling through regression to make data-informed predictions.

In the latter half (Weeks 8-13), the coursework delved deeper into specialized machine learning techniques. From the broad domains of supervised and unsupervised learning, I transitioned to specific models: Naive Bayes, Neural Networks, and Decision Trees. Each week equipped me with nuanced methodologies, refining my data analysis capabilities.

2. Progress and Future Interests

Reflecting on my journey from Week 1 to 13, I've seen exponential growth in both my theoretical knowledge and practical application skills. Initially, my grasp was limited to basic Python scripting and rudimentary data handling. By Week 13, though, I felt confident enough to talk about more complex themes and even go to guest lectures on specialist topics.

My interest has been particularly piqued in the realms of Neural Networks and Decision Tree Models. In the future, I'm eager to delve deeper into these areas, potentially exploring hybrid models that combine the strengths of multiple techniques. Given the versatility and applicability of Data Science, I envision myself leveraging the skills acquired in this course to present business solutions in the consulting industry.

3. Portfolio 4

3.1. Dataset Selection

Relevance and Significance: The decision to select a dataset about "Heart Failure" was not merely a random choice but stemmed from the pressing concern heart diseases pose worldwide. By choosing this dataset, I was diving deep into an area where accurate predictions could significantly impact lives by facilitating timely interventions.

Diverse Data Attributes: Another compelling reason for this selection was the dataset's composition. With a mix of categorical and numerical variables, it provided a comprehensive landscape to explore different data preprocessing techniques, handle varied data types, and evaluate their influence on prediction models.

3.2. Problem Identification

Classifying Heart Diseases: The core problem targeted was predicting the likelihood of heart disease based on multiple clinical and physiological parameters. It was framed as a binary classification task, considering the dataset's nature and the need to delineate patients into two distinct groups: those at risk and those not.

Real-world Implications: Such a problem has palpable real-world implications. The ability to accurately predict heart disease could lead to early and more effective interventions, enhancing patient outcomes and even potentially saving lives.

3.3. Model Choices

Logistic Regression: Logistic regression was a natural choice for a starting point. Its strength in binary classification tasks is well-known, and its simplicity makes it easy to implement and understand. Moreover, in medical contexts, the ability of logistic regression to provide odds ratios – offering a clear, interpretable metric on how different features influence the outcome – is invaluable.

KNN Model: While logistic regression offers a parametric approach, the decision to also incorporate the K-Nearest Neighbors (KNN) model was driven by the desire to explore non-parametric techniques. KNN's principle, where predictions are based on 'neighborhood' data points, was intuitive for this dataset. The underlying assumption is straightforward: patients with similar physiological and clinical profiles might exhibit similar health outcomes.

Hyper Parameter Optimization: The inclusion of hyperparameter optimization, particularly for the KNN model, showcased the importance of fine-tuning. The significant difference in performance pre and post-optimization was a testament to how critical this step is in the model-building process.

3.4. Insights

Feature Importance and Consistency with Medical Knowledge: It was rewarding to see the models highlight features like Sex, ChestPainType, FastingBS, ExerciseAngina, and ST_Slope as significant predictors. These align with existing medical literature, reaffirming the model's credibility. It underscores that when machine learning models are trained appropriately, their findings often resonate with established domain knowledge.

Evaluating Beyond Accuracy: While achieving an accuracy of around 86% was promising, the deeper insights lay in the ROC curve and the AUC metric. An AUC-ROC of 0.9051 suggested a strong model capability to differentiate between positive and negative classes. It was a crucial reminder that in domains like healthcare, relying solely on accuracy might be misleading. It would be needed to evaluate model performance holistically, considering metrics that give a clearer picture of how well the model distinguishes between classes.

Potential Impacts and Limitations: The models' findings and their potential use in real-world scenarios underscore a bigger theme: the implications of false negatives in medical diagnosis. While the models performed commendably, it's imperative to evaluate and minimize false negatives, where patients with potential heart disease are predicted as healthy.