## **Name: Alana St.Bernard**

## **Part 4: Short-Answer Questions (Upload to Blackboard)**

Answer the following based on your work in **Part 3**:

1. Please provide the link to your public **GitHub repository**.

https://github.com/AlanaStB/BINF-5507/tree/main/Assignments/Assignment2

**Regression Models:**

1. Explain how ElasticNet regularization balances L1 and L2 penalties. Why might this be advantageous for certain datasets?

L1 penalty (Lasso) adds a penalty term is proportional to the absolute values of the coefficients and so prevents the model from having extremely large coefficients. It also drives the coefficients that it does not deem useful to zero and in doing so performs some measure of feature selection. L2 adds a penalty that is proportional to the square of the coefficients. It shrinks them without them becoming zero. In doing, this all the features are retained. Elastic net combines the squared approach and the absolute approach along with a hyperparameter to tune them. It can preform feature selection and coefficient shrinkage together. It is useful when there are a lot of highly correlated features and a complex dataset like that in healthcare.

1. What does the heatmap of R² and RMSE reveal about the relationship between alpha, l1\_ratio, and model performance? How did you determine the optimal configuration?

In general, alpha controls the regularization and L1\_ratio, the ratio of L1 and L2 penalties. The heatmap of the R2 and RMSE should reveal the effect of varying these parameters on the performance of the model. In my heatmap, a higher L1\_ratio correlating to a higher RMSE but a lower R2. The effect of alpha is the higher the alpha the higher the R2 and the lower the RMSE. The best combination of alpha and L1\_ratio is that where the R2 is highest and the RMSE is the lowest.

**Classification Models:**

1. When comparing logistic regression and k-NN, what evaluation metric(s) did you prioritize, and why?

Accuracy was used to evaluate as the categories are roughly balanced with a difference of less than 25% between them.

1. What insights did you gain from comparing AUROC and AUPRC curves for the top-performing models? Which model would you recommend, and under what circumstances?

The AUROC tells us how well the model can assign the two categories. Both models preform extremely well with values of 0.90 for the logistic regression and 0.92 for K-NN. These are both perfection classifiers. The AURPC tells us about the trade between precision and recalls. It shows for both models that precision can be maintained with compromising sensitivity. For both the value is 0.9. Therefore, I believe either model would be good but I would go with K-NNs as AUROC is slightly higher.

#### **Critical Thinking (BONUS)**

1. How do different solvers (e.g., liblinear, saga) affect the behavior of logistic regression models? Which solver worked best in your experiments and why?

The choice of solver can be based on a number of considerations including if the data is multiclass, the penalty used and the speed. Liblinear is good with binary classes and is fastest for small datasets. Saga is best for large datasets. Based on accuracy ‘newton-cg’ worked best. This solver is great when the dataset is relatively balanced and has a moderate to high number of features.

1. For k-NN, what trade-offs arise from increasing or decreasing the value of n\_neighbors, and how does this impact model complexity?

Small value of n\_neighbours is better for detecting local patterns but is sensitive to noise. This model may be prone to overfitting. It is good to produce a higher complexity model.

Large values will produce a more generalized model that is less affected by noise but loses complexity and may miss small patterns