BINF610 Interim Report: Breast Cancer Classification

For my final project, I have decided that I will be working with a breast cancer related dataset. The problem that I am trying to solve within this project is the classification of breast cancer as either benign (meaning non-cancerous) or malignant (cancerous) based on the 32 features that have been provided within the Breast Cancer Wisconsin (Diagnostic) data set. Within this dataset, there are 569 observations and 32 features, which include characteristics of cell nuclei extracted from breast mass samples. These breast mass samples were then computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. My goal is to develop a machine learning model that can accurately classify breast cancer cases, with the objective of achieving the best possible classification performance.

In order to address this problem, I plan to implement several machine learning techniques using Python as my primary programming language. The tools and libraries I plan to use include, but are not limited to NumPy, Pandas, Scikit-Learn, Matplotlib, and a couple of others. Specifically, I plan to create the following analyses:

- Random Forest Classifier: Within this final project, I will create a random forest classifier model. I will train the model on the labeled data from the dataset and tune its parameters to optimize its performance.
- Boosting: I will implement the use of boosting, a technique that we have learned combines multiple weak classifiers to create a strong classifier. I will use the AdaBoost algorithm to boost the performance of the Random Forest Classifier. I will also try other boosting methods as well.
- Hyperparameter Tuning: I will tune the hyperparameters of the random forest classifier model in addition to the AdaBoost model. I plan to incorporate the use of the grid search method to find the best combination of hyperparameter values that optimize the performance of the models.

In order to evaluate the performance of my models, I plan to use the following techniques to gage whether or not my models are performing well:

- ROC Curve: Given that I will be performing classification of benign vs malignant cases, I can make use of ROC curves. From this, I will be able to visualize the trade-off between true positive rates and false positive rates as a method to assess the model's classification performance.
- Cross Validation: In order to tell how well my models are performing, I will incorporate cross validation. Within these methods, just like we learned in class I will partition the dataset into multiple subsets and train the model on different combinations of training and testing data, in order to obtain more reliable results of the models performance.

- Classification Report and Confusion Matrix: Lastly, I will create a classification report and confusion matrix to assess the various models performances in terms of precision, recall, F1-Scores, and accuracy. Additionally, by creating a confusion matrix, I will be able to assess both type 1 and type 2 errors, which will serve to be very useful in the context of breast cancer diagnosis.

In the situation that I come across false negatives, I would like to adjust my model's weighting so that the "optimal" model has more false positives rather than false negatives, since I strongly believe that in this context it would be very valuable to have a model that is more cautious.

In conclusion, for this project, I plan to use various machine learning techniques such as random forest classifier, boosting, and hyperparameter tuning in order to eventually develop a classification model for breast cancer diagnosis using the Breast Cancer Wisconsin (Diagnostic) Data Set. I will then evaluate the performance of the models using ROC curves, cross validation, and classification reports with the support of a confusion matrix. It is my hope that by doing this work, this model could potentially be used for utility in clinical practice for breast cancer diagnosis.

Resources to be used:
https://archive.ics.uci.edu/ml/datasets/breast+cancer+wisconsin+(diagnostic)
https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data
https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html