Capstone 3 Final Report

As a doctor, having an idea of whether a disease will reoccur or not in their patient is very important and difficult to know at times because not every medical case will be the same, even if the disease itself is the same. Data analysis can provide insight on specific cases that would have a lower or higher chance of recurrence, specifically in thyroid cancer. Thyroid cancer continues to rise in the population, but if identified at an early stage, it can be very curable. Through data analysis, thyroid cancer can potentially be identified earlier via clinicopathologic feature observation within patients.

To begin, we need to collect quality and sufficient data, which can be found in this CSV file provided by Kaggle: https://www.kaggle.com/datasets/abuchionwuegbusi/thyroid-cancer-recurrence-prediction. Additionally, we need to identify which parts of the data will be useful to the problem we are trying to solve. This can be done through data wrangling. Once this step is complete, we can continue to the next step of exploratory data analysis where we can identify relationships between characteristics of thyroid cancer patients and see which of these characteristics affects the state of the patient. Finally, we can develop a machine learning model that predicts the state a patient is in using their physical characteristics and any additional crucial information.

The deliverables in this project include a GitHub repository containing all the Jupyter notebooks I will use for each data analysis step of the project, a slide deck explaining the important findings from the data set, and a project report explaining the overall project.

To begin, I took the given data set and performed data wrangling and data analysis. Initially, the data set contained all categorical columns, except one. To change this, I changed the string values into numerical values so that I could use the data for further analysis and also to

create my model later on. Once all these values were changed, I then changed the data type of each of the columns so that further analysis could be done.

Next, I used a heat map and box plot to identify any trends within the dataset. Correlation wise, there were quite a few characteristics that correlated with one another. One example being “Risk” and “Stage” or “Risk” and “Adenopathy”. The box plot showed a wide range of ages and all other characteristic columns showed a very small range since the values were transformed from categorical to numerical values.

Finally, I began pre-processing the data by creating dummy variables from the features “Age” and “Gender”. Next, I did the train test split technique to check how the model would perform with a new data set. I used the random forest classifier as my model to predict the classes for the test values. Here we got an accuracy of 0.935 and a f1-score of 0.935. Further, I used a confusion matrix to review the model performance and to see where the misclassification happened in terms of determining if the cancer recurs or not. 6% of the true label values that were “yes” predicted inaccurately and 7% of the true label values that were “no” predicted inaccurately. To finish up the model analysis, I used a bar graph to show the importance of each variable in determining the prediction of recurrence. The bar graph shows “Response” (almost 100 relative importance) and “Risk” (almost 60 relative importance) as the two variables that were of most importance out of all the variables.

To further my research, I would test out additional models, such as the linear regression model, to see if there is a higher accuracy when classifying the test values. I would also research additional characteristics that contain numerical values because the given data set only contained categorical data. Although I was able to change it into numerical values, I think it would also be

helpful to see values of greater variability to further help with the prediction of thyroid cancer recurrence.