Group Members: Quentin Shin (ttvy6kv), Taylor Decina (mvu2ab), Keeley Messick (vmv6mp)

DS 3001

3/31/2025

02: Methods

- What is an observation in your study?

The observation in this study is the risk of developing thyroid cancer in relation to various independent variables. The development of the disease serves as the dependent variable; age, gender, country, ethnicity, family history, radiation exposure, iodine deficiency, smoking, obesity, diabetes, TSH level, T4 level, and nodule size serve as the independent variables. These independent variables were assessed on a scale of low, medium, and high risk factors.

- Are you doing supervised or unsupervised learning? Classification or regression?

We are conducting supervised learning as our training data is labeled and we have a baseline for what we think outputs should be given scientific research on thyroid cancer predictors. It's also classification given we will be predicting the risk on a scale of low, medium, and high, rather than continuous outcomes.

- What models or algorithms do you plan to use in your analysis? How?

We plan to use line graphs, histograms, and correlation matrices in our analysis using our variables and outcome.

- How will you know if your approach "works"? What does success mean?

If a patient exhibits "at risk" traits (i.e., family history for disease and abnormal T4 levels) the algorithm recognizes that this would be a high risk. Success means a correlation between thyroid cancer and an independent variable is found (?).

- What are weaknesses that you anticipate being an issue? How will you deal with them if they come up? If your approach fails, what might you learn from this unfortunate outcome?

There are still some potential issues with this data. One being that our dependent variable is in terms of levels (low, medium, high) so we will have to visually accommodate for that in our charts. Additionally, some of the data is not common knowledge, like T4_Level, so that will require explanations to the common user to allow for better understanding. A final challenge might be needing to change the categorical variables to binary. There are six variables that are "Yes" and "No" answers and it would be easier to display these numerically in our graphs. If our approach fails we will understand that not everything is perfect the first time around and that we can continue to do trial and error with our plan.

You should address the following topics in the text, as appropriate:
- Feature Engineering: How will you prepare the data specifically for your analysis? For example, are there many variables that should be one-hot encoded? Do you have many correlated numeric variables, for which PCA might be a useful tool?

To prepare the data specifically for our analysis, we will first need to change the categorical variables to binary. These will be done for six of our variables. There will be a few variables that need to be one-hot encoded like gender, countries so that the model will treat the categories independently and not assume a relationship between them. There are a good amount of numerical variables that could be correlated, so PCA might be useful to identify the most prominent factors that contribute to the risk of developing thyroid cancer.

- Results: How will you communicate or present your results? This might be a table of regression coefficients, a confusion matrix, or comparisons of metrics like R^2 and RMSE or accuracy and sensitivity/specificity. This is how you illustrate why your plan succeeded or explain why it failed.

We plan to present our results primarily through a comparison of metrics, mostly using R^2 and RMSE to illustrate the outcome of our plan.