# Datathon 1: Analysis of Lung Cancer and BMI Risk Factors

## Introduction:

In Canada, lung cancer is a leading cause of cancer-related death, taking away the life of over 21,000 people per year (Statistics Canada, 2022). On the other hand, obesity, which is measured with a body mass index (BMI) above 30.0, increases the risk of health conditions such as cardiovascular diseases and adds financial burden to the healthcare system (Wharton et al., 2020; Statistics Canada, 2019). Lung cancer and obesity are both diseases with multiple behavioural, environmental, and hereditary factors. Thus the goal of this report is to analyse and visualise a lung cancer dataset from Ethiopia and a BMI dataset from Canada and predict the possible severity of cancer and rate of obesity through the use of linear regressions and K-nearest-neighbours (KNN).

## Data Engineering Process

Due to our intent to use both linear regression and KNN as methods to predict cancer severity and BMI levels, we utilised descriptive statistics, histograms, and statistical tests to check for the distributions and scale of the datasets. Since KNN is a distance based non-parametric model, the accuracy of the model is sensitive to scaling of the data, as variables with larger ranges will dominate the variables with smaller ranges.

We first checked the descriptive statistics and histograms of both datasets to get a more thorough view of the data's distribution and magnitude. Next, we tested the normality of the data through the wilks Shapiro test to ensure that a regression would still be a robust prediction model. From this, we can then determine the efficacy of KNN and linear regressions as possible models. Finally, we used a correlation matrix to determine the relationship among the different variables and to remove any variables with 0 correlation. This would reduce the dataset's multidimensionality and improve model performance.

## Analysis

The descriptive statistics for the BMI dataset showed a large disparity in the magnitude of the factors, which would necessitate scaling or normalisation for KNN to be effective. When testing for normality, only height and weight were normally distributed. The histograms further confirmed the distributions not being Gaussian. Finally, when looking at the correlation matrix, no factors other than height and weight seem to be correlated to BMI. However, this may be due to BMI being directly calculated through the height and weight of the patients, which introduces multicollinearity and would overpower any other possible correlation (supplementary figure 1). Due to the large range of scales and the non normal distribution among the different factors, the use of both linear regression or KNN would result in inaccurate models. As such, we decided to

not explore the BMI data further. Due to the data being mostly categorical for the Cancer dataset, we did not consider a linear regression. Furthermore, we used minMaxScaler instead of standardScaler since the dataset does not follow a normal distribution and the predictors are all ordinal with similar ranges. Since minMaxScaler does not rely on the normality assumption, it allows for better scaling of the data than standardScaler. When looking at the correlation matrix, age and gender had no correlation with the rest of the variables, Therefore, age and gender were removed to improve accuracy and model performance (supplementary figure 2).

## Findings

Based on the elbow method, K values of 1 to 5 resulted in the lowest errors, so the model was fixed at K = 2 . When testing across multiple different random states, the confusion matrices would never return false negatives but would return some positives across the predicted cancer rates on actually healthy people (although the rate of false positives decreased as the random state increased)(supplementary figure 3). The recall values were always above 0.95, with most around 1.00. However, the precision scores would tend to be slightly lower (however still above 90%). This overall resulted in very high f1 scores of at least 0.97 across all models regardless of the starting random states. These results imply that the model can very accurately cluster the cancer severities based on the 4 groups and that they each are very defined in their factors. However, the lower precision to recall rates also imply that there are some crossovers between healthy people and people who have cancer where some healthy people would experience similar levels of risk factors as those with cancer. The confusion matrix supports this since a few actually healthy people would always be predicted to have some level of lung cancer.

## Conclusion

Based on our KNN model, health practitioners can confidently utilise these risk factors to predict the cancer severity of a patient. Moreover, since the model had a low rate of false negatives, they can be confident that they would not miss some possible cancer diagnoses within the population. However, due to the weaker precision of the model, there may be some within the population who, although healthy, would flag as possibly having lung cancer. This may increase the possibility of additional diagnostics being done on healthy patients, which can cause additional healthcare stress on populations with weaker public health systems such as the ones found in lower- and middle-income countries like Ethiopia (the source of this dataset). These results also imply that there may be some unknown factors that would reduce the prevalence of cancer among people who live with high risk factors. Overall, this model would be a great benefit to use in the Ethiopian population to reduce lung cancer related burden of disease and can be used as justification for further research on factors that prevent lung cancer in Ethiopia.

# References

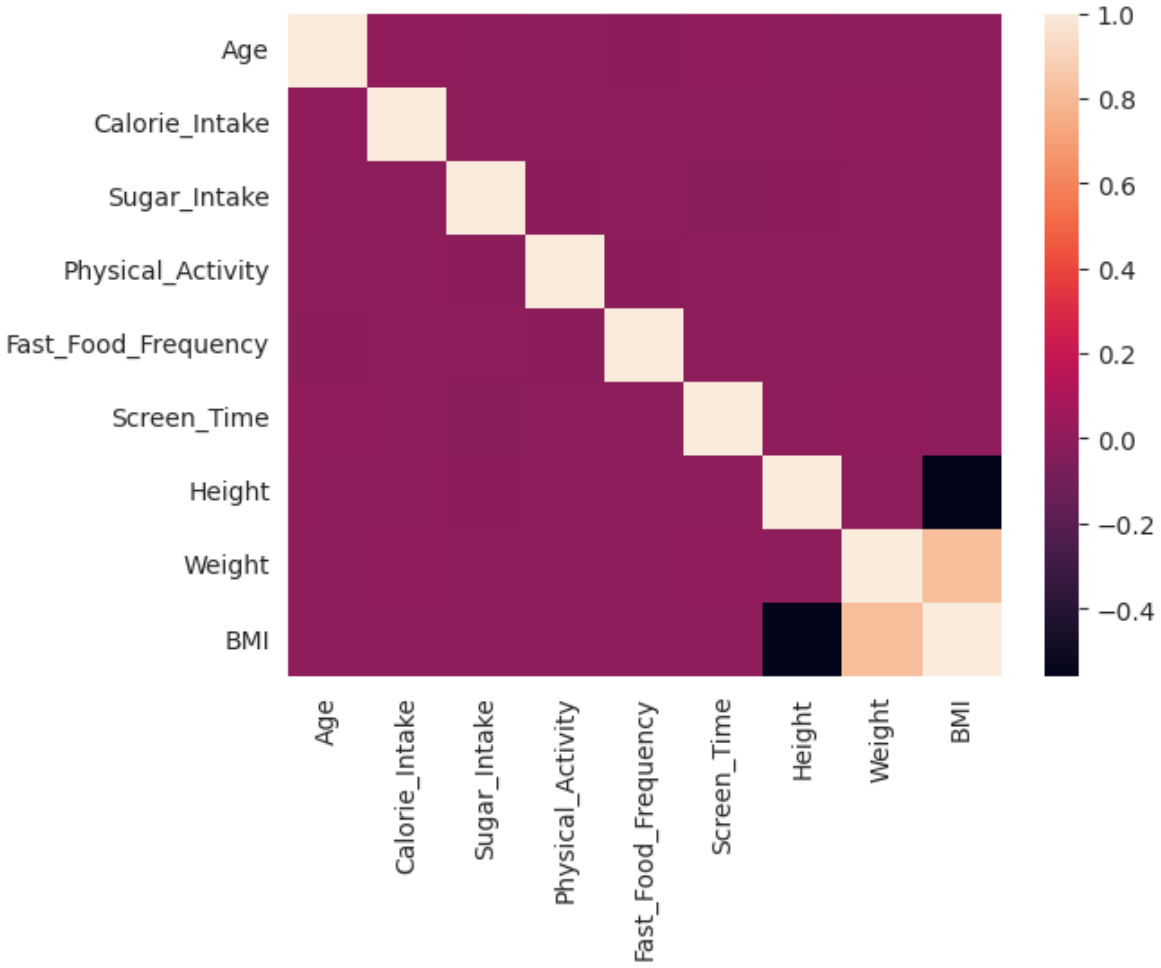
Government of Canada, S. C. (2022, January 4). *Lung cancer is the leading cause of cancer death in Canada*. https://www.statcan.gc.ca/o1/en/plus/238-lung-cancer-leading-cause-cancer-death-canada

Government of Canada, S. C. (2019, June 25). *Overweight and obese adults, 2018*. https://www150.statcan.gc.ca/n1/pub/82-625-x/2019001/article/00005-eng.htm

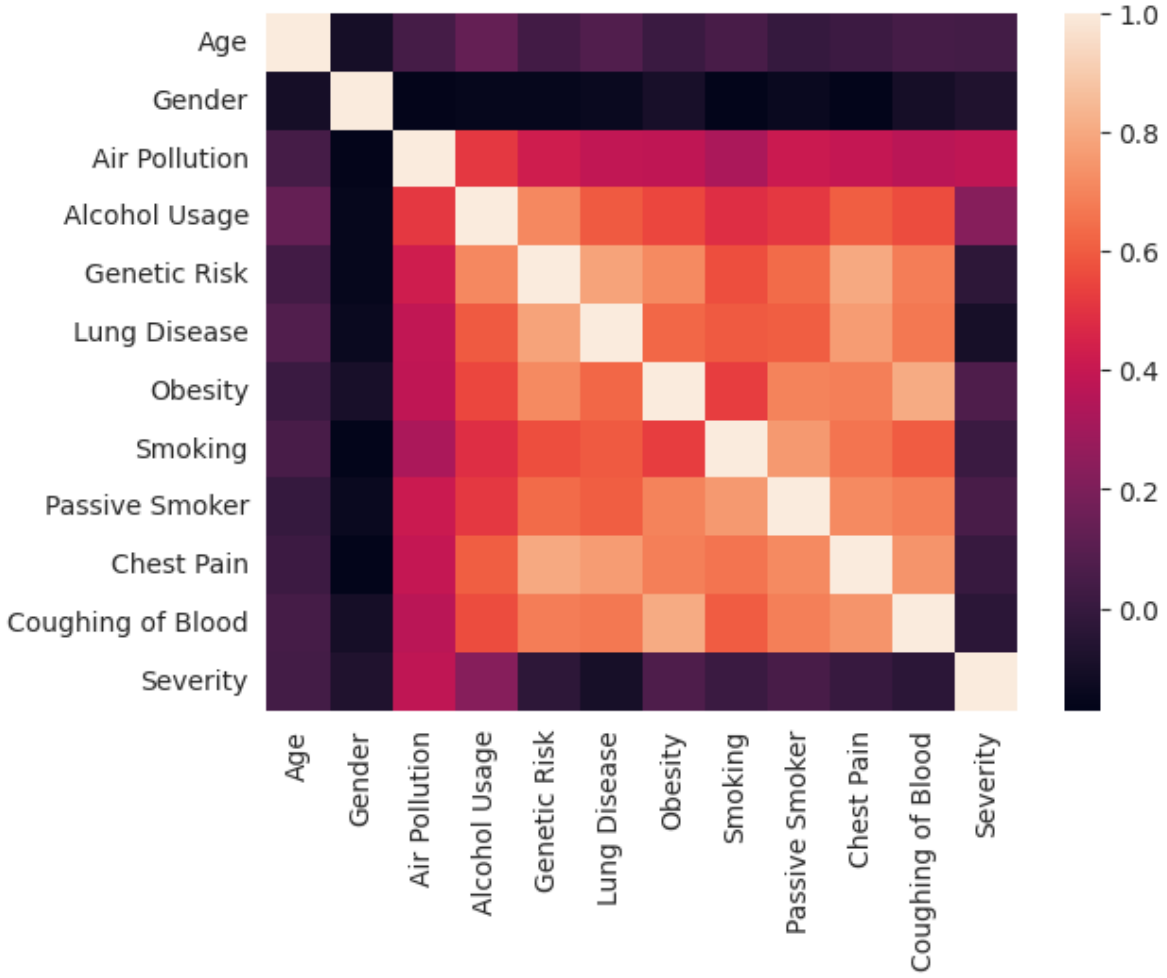
Wharton, S., Lau, D. C. W., Vallis, M., Sharma, A. M., Biertho, L., Campbell-Scherer, D., Adamo, K., Alberga, A., Bell, R., Boulé, N., Boyling, E., Brown, J., Calam, B., Clarke, C., Crowshoe, L., Divalentino, D., Forhan, M., Freedhoff, Y., Gagner, M., … Wicklum, S. (2020). Obesity in adults: a clinical practice guideline. *Canadian Medical Association Journal*, *192*(31), E875–E891. https://doi.org/10.1503/cmaj.191707

# Appendix

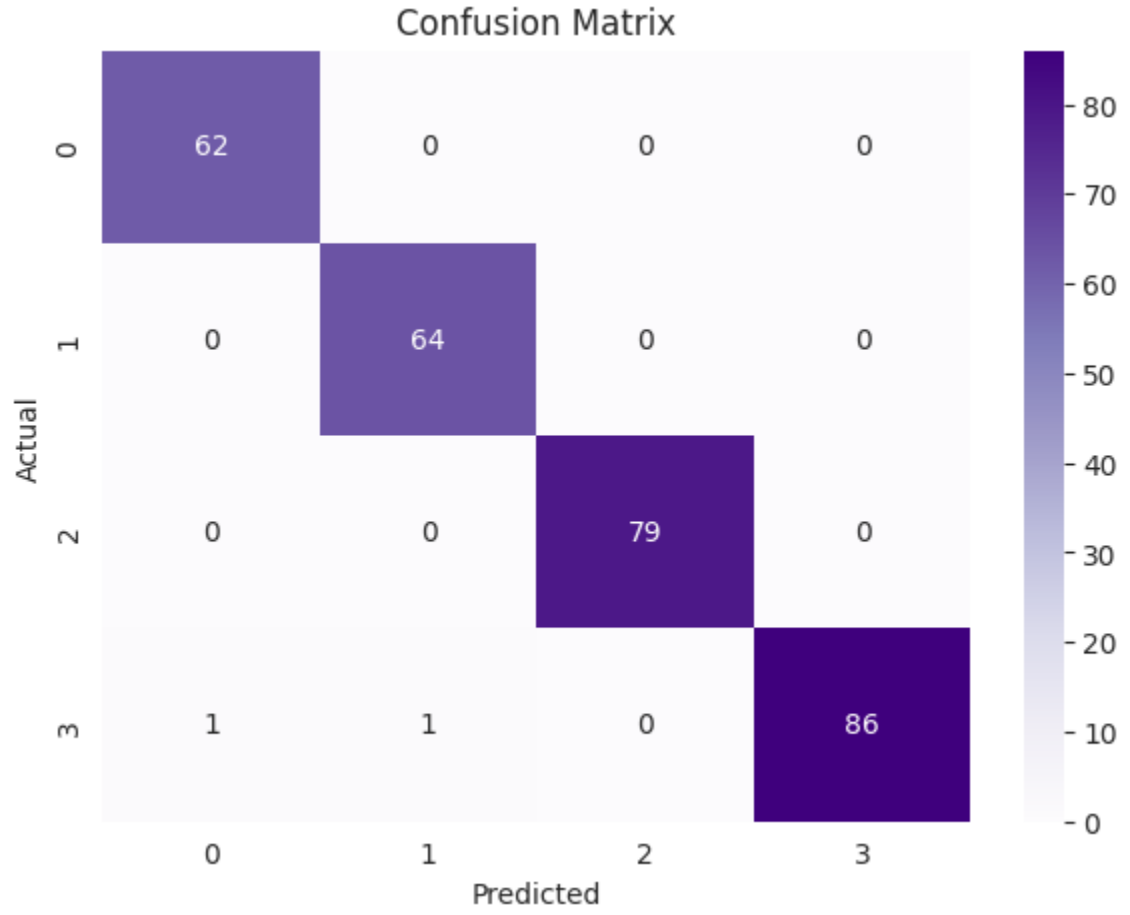
Supplementary Figure 1: Correlation Matrix for BMI Data



Supplementary Figure 2: Correlation Matrix for Lung Cancer Data



Supplementary figure 3: Confusion Matrix for Lung Cancer Data




## Individual Contributions

Kinna Zhao : data preprocessing and visualisation, exploratory data analysis
Priyonto Saha:  exploratory data analysis, model design and analysis
Yacine Marouf: model analysis, writing report+presentation

## Github Repository with Presentation and Jupiter Notebook

https://github.com/P-Saha/k-NN_and_Clustering_Analysis