CS 5593 995-999 - Data Mining - Proposal

1) Lung Cancer Analysis

2) Names and Email Address of Authors

Brady Carden - brady.a.carden-1@ou.edu
Charles Cornelison - charles.c.cornelison-1@ou.edu
Isabella Cruz - bella.v.cruz@ou.edu

3) Category and Objectives of the Project

Category

- Our group has chosen option b for the project.

Objectives

- Identify individuals with lung cancer risk
- Cluster individuals into various groups of risk
- Identify the most significant factors contributing to lung cancer
- (I'm not really sure what these are supposed to be)

4) The Significance of the Project

As medicinal and health sciences have grown more advanced, our ability to combat illness and chronic disease has rapidly increased. However, there remain several conditions, such as cancer, which seem to be an inevitability in many people's lives. Even with ideal health, a clean environment, and well maintained relationships and stress levels, a diagnoses of cancer may be not be avoidable, in large part due to the genetic nature of the disease. There are various types of cancer, some worse than others, but all are undesirable due to their deleterious health effects, medical treatment costs, and effects on general well being. Lung caner, specifically, is the "leading cause of cancer deaths worldwide among both men and women [Esposito, 2010]." Though cancer may not be entirely preventable, the best way to treat it is via the power of early detection. "When cancer is detected at the earliest stages, treatment is more effective and survival drastically improves [Crosby, 2022]." Unfortunately, some 50% of cancers are detected in an advanced stage [Crosby, 2022]. With the growth of Big Data, the medical world has realized the usefulness of technologies such as artificial intelligence, machine, learning, and data mining, and some have already been applied to lung cancer prediction [Pontes, 2021]. Other realms of data science have proven powerful: "ML also offers potential for early detection of cancers by scalably synthesizing trends across patients over a potentially distant time horizon (Bertsimas, 2020)."

For various reasons, predicting lung cancer is difficult. Smoking cigarettes is known to be one of the main causes of lung cancer, but "There is no correlation between lung cancer and the number of packs smoked per year due to the complex interplay between smoking and environmental and genetic factors. [Siddiqui, 2023]" For this reason, applying data mining algorithms could prove to extremely beneficial, especially with the large attribute counts like in the dataset we will use. Found on Kaggle, the data set is called Lung Cancer prediction. Each of the 1,000 entries contains 26 attributes, 2 of which are used for identification, 1 for risk level (low, medium, high), with the remaining 23 representing various biological, environmental, medical history, and habitual factors. Below are 10 of the 23 attributes:

1) Gender: 1 or 2 representing male or female
2) Air Pollution: Integers 1 - 8 representing patient exposure to air pollution
3) Alcohol Use: Integers 1 - 8 representing the level of alcohol use
4) Occupational Hazard: Integers 1 - 8 representing level of occupational hazard
5) Genetic Risk: Integers 1 - 7 representing level of genetic risk
6) Chronic Lung Disease: Integers 1 - 7 representing the history of lung disease in the patient
7) Smoking: Integers 1 - 8 representing level of smoking history
8) Chest Pain: Integers 1 - 9 representing level of chest pain

9) Obesity: Integers 1 - 7 representing the level of obesity
10) Shortness of Breath: Integers 1 - 9 representing the level of shortness of breath

As provided on Kaggle, the data is from a study published in the journal Nature Medicine which looked at data from over 462,000 people in China who were followed for an average of roughly 6 years. The total file size is 61.47 kB. Outside of age, which is numeric, the attributes are categorical according to the data set.

Potential data mining questions (??)
- Group individuals with similar risk (clustering)
- Determine whether someone should take action / determine risk level (classification)
- What attributes and trends seem to have similar outcomes (association analysis)

5) Implementation, Research Methodology, and Time Table

Proposal requirements want this to be very specific so we can just write down random timelines and random assignments. If we're cool with the assignments and timelines I put here then cool. We can change them if need be, probably even after we submit this proposal.

Algorithm Implementation: 9/13/2024 - 10/13/2024
Note: Algorithms will be implemented in Python (or whatever you want)

Brady
- Clustering algorithm.
- The final product will be able to cluster the entries based on various distance metrics.

Charles
- Classification algorithm.
- The final product will be able to classify each entry into its provided risk level with a high level of accuracy.

Isabella
- Association analysis algorithm.
- The final product with analyze associations among entries, attributes, and outcomes (what is association analysis)

UI Implementation: 9/26/2024 - 11/1/2024

UI work will be headed up by Brady. Currently, the plan is to build a MacOS app using SwiftUI. If this proves to not work as intended the group will shift to a web dev based application. The finished product will allow a user to interface with our algorithms and use the app to produce results after adding their own entry.

Putting it All Together: 11/2/2024 - 12/5/2024

This time will be used to compile all work to make a report and video for our final submission. The final product will be up to par with the project guidelines provided on canvas

6) References

Bertsimas, D., & Wilberg, H. (2020, October 15). *Machine learning in oncology: Methods, applications, and challenges*. ASCO Publications. https://ascopubs.org/doi/full/10.1200/CCI.20.00072 (pp. 6)

Crosby, D., Bhatia, S., Brindle, K., Coussens, L., Dive, C., Emberton, M., Esener, S., Fitzgerald, R., Gambhir, S., & Balasubramanian, S. (2022, March 18). *Early detection of cancer*. Science. https://www.science.org/doi/10.1126/science.aay9040 (pp. 1)

Esposito, L., Conti, D., Ailavajhala, R., Khalil, N., & Giordano, A. (2010, November). *Lung cancer: Are we up to the challenge?*. Current genomics. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3048313 (pp. 1)

Pontes, B., Núñez, F., Rubio, C., Moreno, A., Nepomuceno, I., Moreno, J., Cacicedo, J., Praena-Fernandez, J.

M., Rodriguez, G. A. E., Parra, C., León, B. D. D., Del Campo, E. R., Couñago, F., Riquelme, J., & Guerra, J. L. L. (2021, December 30). *A data mining based clinical decision support system for survival in Lung cancer*. Reports of practical oncology and radiotherapy : journal of Greatpoland Cancer Center in Poznan and Polish Society of Radiation Oncology. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8726446 (pp. 2)

Siddiqui, F., Vaqar, S., & Siddiqui, A. (2023, May 8). *Lung cancer*. StatPearls [Internet]. https://www.ncbi.nlm.nih.gov/books/NBK482357 (pp. 2)