Introduction

Utilizing machine learning techniques, such as KNN, can assist in decision-making in public health care. With a data set with features of both healthy individuals and lung cancer patients with different severities, the two research questions (RQ) in this study are (1) Can we predict the presence of lung cancer to assist doctors in flagging patients potentially at risk of lung cancer? (2) Can we predict the severity of lung cancer among cancer patients to support clinician diagnoses?

For RQ 1, general practitioners and family doctors typically do not have access to such a wide breadth of data, as seen in the study that we pulled this data from. It would be useful for doctors to have support in knowing which patients are most at risk in order to prioritize whom they screen/investigate further. RQ 2 is also important because predicting the severity of lung cancer based on symptoms and clinical manifestations can support clinical findings in patients undergoing cancer screening or diagnosis and provide supplementary materials for doctors' decision-making.

Data Engineering Process

We first imported, read the dataset, and checked for missing values. There are two pair plots: one focuses on healthy individuals and patients with cancer, corresponding to RQ 1, and the other looks at the different severities of cancer patients, relating to RQ 2. The pair plots for question 1 reveal that, for most features, the pair scatter plots show that healthy individuals and patients with cancer are naturally distinct. However, for the group of cancer patients in RQ 2, there is no obvious distinction among different severity levels. We also made correlation tables and heatmaps for variables in RQ 1 and 2, respectively. In the RQ1 heatmap, the least correlated features to cancer presence were age, gender, and air pollution, however air pollution had a moderate correlation with other features. In the RQ 2 heatmap, we found that all variables are reasonably correlated with severity except for age and gender.

The variables chosen for each research question are primarily based on the correlational data in the table and heatmap, as well as logistics of including different features in the context of each question. RQ 1 features are alcohol usage, obesity, smoking, and genetic risk. RQ 2 features are alcohol usage, genetic risk, air pollution, lung disease, obesity, smoking, passive smoking, chest pain, and coughing blood. For RQ 1, we created a new binary variable, cancer presence, where 0 indicates healthy people and 1 indicates cancer patients. For RQ 2, we created a new dataset by filtering out observations with cancer severity 3, which was coded as healthy in the original data set. This reduced our sample size to 1000 and severity levels 0 to 2, where 0 indicates low, 1 indicates moderate, and 2 indicates high. Finally, for both questions, we split the dataset into an 80% training set and a 20% test set and used a scaler to standardize the features to ensure more accurate results in KNN.

Analysis

The learning technique employed is K-nearest neighbour (KNN). In the dataset, we have cancer presence or cancer severity as outcomes, which could be used as labels in supervised learning. We used Euclidean distance as the features use similar scales with no outliers or sparsity.

For RQ 1, we plotted different values of K against the accuracy and class 1 precision of our mode over multiple random states of the train/test split. This is because, in this context, flagging positive patients is more important than negative patients, as this tool simply recommends investigation by a doctor and is not a replacement for a diagnosis. Average accuracy and class 1 precision have a local maximum at K = 20. To avoid overfitting, we excluded lower values of K.

For RQ 2, we focused on the overall accuracy of the model to find the optimal K. We plotted K against the error rate of our model on the test set. Because there was a decrease in the error rate plot starting at K = 20, our optimal K was chosen to be 20 to avoid overfitting and ensure accuracy simultaneously.
After determining the optimal K, we performed KNN for RQ 1 and 2, using the corresponding datasets and K values, respectively. We then created a confusion matrix and classification report.

Findings

In RQ 1, precision was 0.85 and 0.84, and recall was 0.93 and 0.68 for presence of cancer (class 1) and absence of cancer (class 0) respectively. The train accuracy was 0.89 and the test accuracy was 0.85. The model performed well on both classes, with a slight bias towards class 1, given its higher recall. The test accuracy is close to the training accuracy, indicating that the model generalizes well to unseen data. There may be room for improvement, particularly in maximizing recall for class 0, however in the context of the intended purpose of this model, it may not be necessary as we are primarily looking to flag potential positive patients in order to assist a doctor who will then investigate further.
In RQ 2, for each of the severity levels 0, 1, and 2, we obtained high precisions of 1, 0.90, and 0.98, respectively. The recalls were 0.89, 1.00, and 0.96 for each level. The training accuracy was 0.96, and the test accuracy was 0.95. The model performed best on severity 2 as it had the most balanced precision and recall. Based on the available features, our model excelled at predicting severity for patients with severe cases but encountered difficulties distinguishing between those with low or moderate cancer levels.
For secondary findings, we also conducted clustering analysis using unsupervised learning to see if using all features as metrics, the data tend to separate into cancer and healthy patients groups. Since K-means is only suitable for numerical data, we employed K-modes, which uses dissimilarity metrics and mode-based clustering for categorical features. We performed K-modes on both the original dataset and on a constructed balanced dataset, due to the disproportions between cancer and healthy patients in the original dataset. Both results indicate that more than half of cancer patients and approximately all healthy patients lay in the same cluster. This might be due to including all features leading to high dimensionality problems and presence of unaccounted confounding variables. Based on observations from exploratory analysis, healthy and cancer patients naturally separate from each other, which is suitable for discriminant analysis. We then further applied Quadratic Discriminant Analysis including all features, which gave us high precision of over 0.90.

Conclusion

The model created for RQ 1 has the potential to assist doctors in clinical decision making as it helps identify patients who may have/be at risk for lung cancer. While the model can be further tuned for maximum performance, it functions well as a proof of concept and satisfies the requirement of initially flagging patients for review. For RQ 2, our research suggests that our model is particularly useful for identifying lung cancer cases with the highest severity. However, it should be used with caution in patients with lower or moderate disease severity, as its efficacy may be less reliable in these instances.

Individual Contributions: Feifan was responsible for exploratory analysis and clustering, Myron was responsible for RQ 1, and Yutong was responsible for RQ 2. All team members contributed to the code, report, and presentation.

Code and Presentation: Link to code on GitHub: https://github.com/Yutong-Lu/Datathon-1. Link to presentation on google slides: https://bit.ly/3LCbHur