Study of the Prediction of Lung Cancer using Machine Learning

Cristian DEGNI, Enrico SBUTTONI

Motivation

What problem are you tackling?

The problem we are addressing is the prediction of whether an individual has lung cancer based on clinical and lifestyle data. Lung cancer is one of the most aggressive and fatal types of cancer, and early and accurate diagnosis can significantly impact patient outcomes by enabling timely intervention and treatment planning.

Detecting lung cancer using clinical, lifestyle, and other diagnostic information allows for earlier interventions, potentially improving survival rates and quality of life for patients. This research aims to leverage machine learning models to provide more precise and reliable predictions of lung cancer presence, thereby supporting patients, families, and healthcare professionals in making critical decisions about diagnostics and care pathways.

What is the setting you are considering?

In this study, we utilized two distinct datasets to analyze the capability of a machine learning model to diagnose the presence of lung cancer. The first dataset, a smaller one with approximately 300 records, has been extensively studied in the past, with previous research reporting very high diagnostic accuracy. However, the limited size of this dataset highlighted a potential issue of overfitting, where the model performs exceptionally well on the specific dataset but has reduced generalizability. The second dataset, a larger one with approximately 3300 records, had not been used in similar studies before, providing a unique opportunity to explore and improve results by leveraging new machine learning techniques.

To balance specificity and generalizability, we selected the entire first dataset and a subset of the second, resulting in a combined total of 1500 records. This approach allowed us to work with a medium-sized dataset that is more representative of generic data. The aim of the study is not only to replicate the high performance achieved on the first dataset but also to improve results on the second, previously unexplored dataset, contributing to the advancement of knowledge through innovative techniques.

Method

1. Random Forest

Random Forests are ideal for handling diverse clinical data, including categorical and continuous variables, while providing feature importance scores to identify key factors influencing the likelihood of having lung cancer. The model will be trained on labeled patient data, using feature importance to determine the variables most associated with the presence of lung cancer.

Preliminary experiments

0.1 Data Encoding and Visualization

In the preprocessing phase, we applied *Label Encoding* to categorical variables, such as `GENDER`. This step is crucial because many machine learning models, including Random Forests, cannot process non-numerical data directly. Encoding transforms categories into numerical representations while preserving the relationships and allowing the model to work seamlessly.

Feature Visualizations: Two sets of plots were created to gain insights into the data:

- **Feature Distributions**: Continuous features, such as `AGE`, were visualized using histograms with density curves, while categorical features (e.g., `SMOKING`, `ANXIETY`) were displayed using count plots. This helps identify trends and imbalances in the dataset.
- **Correlation Heatmap**: A heatmap shows the correlation between features, providing insights into multicollinearity and the potential importance of predictors. For example, `AGE` and `SMOKING` display low correlations, indicating they provide distinct information.

0.2 Random Forest Model

We trained a *Random Forest Classifier*, an ensemble method that constructs multiple decision trees during training and averages their predictions for robustness. Random Forests are particularly effective in handling mixed types of features (numerical and categorical) and automatically capturing non-linear relationships without requiring feature scaling.

0.3 Model Training and Evaluation Results

Training and Evaluation: The Random Forest model was successfully trained, and its performance was evaluated on both the validation and test sets. Below are the detailed results:

Validation Set Results:

- **Accuracy:** 0.70
- **Classification Report:**

```
              precision    recall  f1-score   support

          NO       0.56      0.37      0.45        73
         YES       0.74      0.86      0.80       152

    accuracy                           0.70       225
   macro avg       0.65      0.62      0.62       225
weighted avg       0.68      0.70      0.68       225
```

Test Set Results:

- **Accuracy:** 0.67
- **Classification Report:**

```
              precision    recall  f1-score   support

          NO       0.55      0.31      0.39        78
         YES       0.70      0.86      0.77       147

    accuracy                           0.67       225
   macro avg       0.62      0.59      0.58       225
weighted avg       0.65      0.67      0.64       225
```

Observations:

- The model performs better in predicting the `YES` class, achieving higher precision, recall, and F1-score compared to the `NO` class. This suggests the model is better at identifying positive cases of lung cancer.
- The relatively lower recall for the `NO` class indicates that the model struggles to identify negative cases, possibly due to class imbalance.
- The overall accuracy is satisfactory, with 0.70 on the validation set and 0.67 on the test set, showing consistent generalization performance.

Next steps

Future work could include hyperparameter tuning, feature importance analysis, and applying other ensemble methods to further enhance predictive performance. We will also use other methods like:

1. Support Vector Machines (SVM)

SVMs are ideal for binary classification and can handle non-linear relationships in patient data. An SVM with an RBF kernel will classify patients based on clinical and lifestyle features.

2. XGBoost (Extreme Gradient Boosting)

XGBoost is highly accurate and effective for structured data, capturing complex patterns in clinical data. The model will be trained and tuned to identify key factors influencing lung cancer diagnosis.

Division of the tasks

We worked on the machine learning algorithm, with Enrico focusing on the testing phase, evaluating its performance and ensuring its robustness. Cristian, on the other hand, was responsible for searching and selecting the dataset, as well as performing the necessary preprocessing tasks.

Although our tasks were distinct, most of the activities were carried out together and in sync, with both of us contributing to a collaborative and harmonious workflow.