Lung Cancer data detection and analysis

Payel Moishal (pm3009@hw.ac.uk) - Mohammed Bilal (bm3009@hw.ac.uk)

- Abdullah Allami (aa4030@hw.ac.uk) - Husam Isied (hi2005@hw.ac.uk)

- Mohammed Zanella (mkz4000@hw.ac.uk)

Git Link: https://github.com/mkz4000/Mon_Dubai_PG_11

R1 Introduction

This study explores the prediction of lung cancer severity levels (low, medium, high) using a comprehensive dataset containing patient demographics, lifestyle factors, environmental exposures, genetic predispositions, medical symptoms, and CT scan images. The investigation is guided by the following hypotheses:

Hypothesis

Predicting lung cancer levels can be effectively achieved using supervised learning models by leveraging patterns in the features.

Model performance hypothesis: "Complex models like MLP and CNN will outperform simpler models like Naïve Bayes and linear regression due to the nonlinear relationships in the data."

Clustering will reveal distinct subgroups among patients, potentially corresponding to different severity levels

Datasets

The main dataset includes 25 features spanning demographics, lifestyle, environmental factors, genetic predispositions, and symptoms, with a multi-class target variable (low, medium, and high cancer severity).

Additionally, a CT scan image dataset with 1,190 slices from 110 cases (normal, benign, malignant) was included for deep learning analysis. Key steps in data analysis include:

R2 Data Analysis

The work done in data analysis addressed quite some points, here's a summary:

1. **Introduction and Dataset Exploration:**
    - We started by importing basic libraries and exploring the dataset, which is related to cancer patients and provided a link to an external Kaggle dataset source.
    - Initial exploratory data analysis including descriptive statistics and visualizations to understand the data distribution and patterns.

2. **Data Preprocessing:**
    - Performed data preprocessing steps such as scaling, cleaning to prepare the data for analysis.
    - Dividing data into classes based on a target variable.
3. **Correlation Analysis:**
    - Calculating correlation matrices for different classes (e.g., low, medium, and high classes) of the target variable.
    - Identifying top correlated features for each class, creating datasets containing the most significant features.
4. **Feature Selection:**

Datasets with subsets of features were created based on the top correlated features across different classes:

- Top 6 Features Dataset: Combined the top 2 correlated features from each class.
- Top 15 Features Dataset: Merged the top 5 features from each class.

Key Takeaways and Findings

1. **Correlation values before and after scaling:**

The categorical features have different scales. For example: “Air Pollution” (1 min - 8 max) while “Genetic Risk” (1 min - 7 max) while “Fatigue” (1 min - 9 max) and so on. When we calculated the correlation values, we tried to unify the scales of the categories but surprisingly the correlation values stayed the same

2. **Class labels Correlation:**

We found each class label has different correlations with the dataset features, for example the less value for “Dust Allergy” and “Obesity” the more likely the Cancer is still in Low level while the more values of these features the more likely the Cancer reached High level. Also some of the features have high correlation ranking in the original dataset but ranked differently when correlation considered separately per class, for example “Alcohol use” and “Balanced Diet” among the top6 in the original dataset but not in the 3 combined datasets of top2 features, on the other hand “Smoking” is not among the top6 in the original dataset but it is in the 3 combined datasets of top2 features.

3. **Correlation-focused Analysis:**

Our core analysis centered around finding relationships between variables to choose features that could contribute most to understanding or predicting the target outcome.

4. **Data Organization:**

By creating smaller datasets with relevant features, the analysis sets a foundation for building more efficient machine learning models or conducting targeted analysis.

R3 Clustering

Determining Optimal Clusters:

- **Elbow Method:** Plotted the inertia to find the point where it decreases sharply, indicating the optimal number of clusters.
- **Silhouette Scores:** Calculated silhouette scores for different numbers of clusters (2 to 20) to evaluate clustering quality.
    - Best silhouette score found: **0.37** for **12 clusters.**

K-Means Implementation:

- Chosen **10 clusters** (based on Elbow or silhouette analysis).
- Achieved: **Silhouette Score:** 0.35 **Inertia:** 7048.59

Visualization:

- Used PCA to reduce data dimensions to 2 components for plotting.
- Visualized clusters with scatter plots, observing separations and overlaps.
- Performed silhouette analysis to assess within-cluster consistency.

Cluster Profiling:

    - Cluster profile 0: Young adults with relatively lower risk factors (e.g., genetic risk, air pollution).
    - Cluster profile 1: Patients with high exposure to occupational hazards, genetic risk, and lifestyle-related factors.
    - Cluster profile 2: Patients with high air pollution and chronic lung diseases.
    - Others: Various profiles defined by combinations of age, environmental exposure, and symptoms.

R4 Classification insights with DecisionTrees

The main objective of this part is to try to discover which group of features has more significance in determining the cancer level by utilizing Decision Trees.

we are going to divide our features into: Genetic and Medical History Features, Environmental Features, and Behavioral Features

And build decision trees of each and compare the results to determine which group of features give better results in predicting cancer levels. The idea is not to reach 100% prediction accuracy as a must, but to figure out more details that help to confirm the common conceptions and challenge the hypothesis about lung cancer reasons.

At the end we will build a decision tree that includes all the features in general and check the prediction results

Below are the selected features of each group:

- Genetic and Medical History: Genetic Risk, Chronic Lung Disease, and Obesity
- Environmental Features: Air Pollution, OccuPational Hazards, Passive Smoker
- Behavioral Features: Alcohol use, Balanced Diet, Smoking

DecisionTrees Summary and Remarks

- When looking into predictions from a feature category perspective, it seems that behavioral factors are more deterministic than genetic and environmental causes (at least this is what the data shows!).
- notice how passive smoker feature is at root of the environmental factors tree and more significant than air pollution when it relates to lung cancer
- notice how obesity is at root of the genetic and medical history factors tree and more significant than genetic factor feature itself when it relates to lung cancer
- we decided to consider the behavioral elements the more prevailing factor with relation to answering questions on the level of cancer predictions, this because the decision tree give more accuracy then the other factors, and it's branching is simpler and more interpretable
- as per the whole features decision tree, we utilized Randomized search to figure which hyper parameter to set and which feature are more important in classifying our data and reached accuracy of 100%

Experimenting Basic Classifiers

The basic idea is we present different binary classifiers and compare their results. We choose to predict the High level cancer category so this practice will help us determine the conditions that best discover the cases with High level of cancers. We will display different measurement metrics including sensitivity, so if we manage to have a very accurate classifier with high sensitivity, we can be sure that if our prediction says it's not High level

diagnosis then for sure it's not High level, which will give more trust with the patients and hope to get treatment.

Summary

- KNN Classifier: Achieved perfect accuracy and ROC-AUC. However, KNN's overfitting potential should be evaluated on larger, more varied datasets.
- Logistic Regression: Also perfect performance, likely due to data suitability for linear separability.
- Naive Bayes: Slightly lower accuracy but retains a strong ROC-AUC, showcasing robustness in probabilistic classification.
- SGD Classifier: Performed exceptionally well, matching logistic regression with its flexible linear model.

R5 Neural Networks

The MLP was evaluated with different activation functions (ReLU, tan h, logistic). and varying numbers of iterations:

1. **Performance with 500 Iterations**:
    - Achieved **100% accuracy** across all activation functions, indicating full convergence and optimal learning.
2. **Performance with 25 Iterations**:
    - Accuracy varied with activation functions: **Tanh**: 89.5% **Logistic**: 86.25% **ReLU**: 82.25%
    - Demonstrates that while MLP can generalize well with limited training, its performance depends on the chosen activation function and training duration.

With sufficient training (500 iterations), MLP achieved perfect results, validating its ability to model complex, nonlinear relationships. Even with fewer iterations, MLP remained competitive, particularly with tan h, underscoring its robustness.

Convolutional Neural Network (CNN) Analysis

The CNN was applied to the IQ-OTH/NCCD lung cancer dataset. Key results are as follows:

- Achieved **100% accuracy** on the test set, indicating perfect classification performance.

- CNN's ability to capture spatial hierarchies in image data was pivotal, leveraging its convolutional layers to distinguish subtle differences across classes.

CNN's strong performance highlights its strength in medical imaging tasks, validating its utility in classifying lung cancer severity based on CT scans. While this result is promising, further validation on external datasets and analysis to assess potential overfitting are recommended to ensure the model's robustness in real-world applications.

Perceptron Analysis

The Perceptron classifier was evaluated using a one-against-all approach:

- Achieved **88.5% accuracy** overall, with strong performance for the high severity class (F1-score: 0.97).
- Struggled to distinguish low and medium severity cases, as reflected in a lower recall (0.71) for low severity.

While effective for simpler tasks, the Perceptron struggled with overlapping features between low and medium severity levels, highlighting its limitations compared to more advanced neural networks like MLP and CNN.

Conclusion

The results validate our hypotheses:

Predictive Hypothesis: All supervised learning models, including simple classifiers (Logistic Regression, KNN) and complex neural networks (MLP, CNN), achieved high accuracy, proving that lung cancer severity levels can be effectively predicted using patterns in the dataset features.

Model Performance Hypothesis: Complex models like CNN and MLP achieved perfect accuracy, outperforming simpler models like Naïve Bayes and Perceptron. This validates the hypothesis that advanced models can better capture the nonlinear relationships in the data.

Clustering Hypothesis: While not discussed here, clustering provided insights into patient subgroups, complementing the supervised learning approaches.