FINAL REPORT — Lung Cancer Classification Project

Introduction

The goal of this project was to determine which machine learning model best predicts lung cancer from demographic and behavioral survey data. Our research question:**"Which machine learning model best predicts lung cancer from demographic and behavioral features?"**, was motivated by the growing need for accessible, low-cost screening methods that may help identify individuals at elevated risk. Traditional diagnostic procedures are expensive and often limited by healthcare access; however, survey-based indicators such as smoking history, coughing frequency, fatigue, and anxiety may capture early signals of health deterioration. By examining several supervised learning models, we aimed to understand both the predictive value of these features and the comparative effectiveness of modern classification algorithms.

Methods

The dataset used in this project was a publicly available lung cancer survey containing **309 samples and 15 features**, including both behavioral and demographic attributes. Before modeling, the data underwent standard preprocessing steps: categorical values were encoded numerically, column labels were cleaned, and all continuous variables were standardized when required for specific models. The data was then split into an 80% training set (247 samples) and a 20% test set (62 samples). Several classification algorithms were implemented, including penalized logistic regression, a Support Vector Machine (SVM), a Random Forest ensemble, and a neural network built with PyTorch. Each model included appropriate preprocessing—such as feature scaling for logistic regression and SVM, and no scaling for Random Forests—and hyperparameters were tuned using grid search when applicable. A five-fold cross-validation strategy was applied to the training set to compare models reliably, using accuracy, precision, recall, and F1-score as evaluation metrics.

Results

Across all models, the **SVM consistently demonstrated the strongest cross-validation performance**, indicating robust ability to generalize from the available survey data. The final SVM model was then retrained on the full training dataset and evaluated on the held-out test set. It achieved an accuracy of approximately **0.97 (0.9677).** The classification report showed strong performance across both classes: for individuals without lung cancer, precision and recall were both 0.88, while for individuals with lung cancer, precision and recall were both 0.98. The macro-averaged F1-score was

0.93, and the weighted F1-score matched the overall accuracy at 0.97. These results indicate that the **SVM not only generalized well but also performed particularly effectively on the majority class of lung cancer cases**, which aligns with the dataset's distribution and the model's strength in capturing non-linear decision boundaries.

Several notable patterns emerged from the comparative model evaluation. Logistic regression, despite its simplicity and interpretability, failed to capture the more complex, non-linear relationships present in the dataset. The Random Forest model performed well but showed slightly higher variance and did not achieve the same consistency across folds as the SVM. The neural network, although capable of modeling non-linearity, was limited by the small size of the dataset and therefore did not achieve superior performance. Ultimately, **the SVM's balance of flexibility and robustness made it the most effective model for this prediction task.**

Discussion

However, the project is not without limitations. **The dataset is relatively small**, which constrains the complexity of models and limits the reliability of generalizing these findings to broader populations. Many features are self-reported and may introduce bias or noise. Additionally, because the dataset contains a higher proportion of positive lung cancer cases, models may be **more prone to optimizing performance for the majority class.** Despite these limitations, the results demonstrate that behavioral and demographic factors can provide meaningful predictive insight into lung cancer status. Future work could benefit from **incorporating larger datasets,** assessing the contribution of individual features through interpretability methods such as SHAP values, and validating model performance across diverse samples.

In conclusion, by evaluating multiple machine learning models and comparing their ability to predict lung cancer from survey-based features, this project finds that the **Support Vector Machine achieves the strongest performance, with a 97% test accuracy and excellent precision and recall.** These results suggest that machine learning models, particularly SVMs, can be powerful tools for identifying at-risk individuals when diagnostic resources are limited, offering promising directions for early risk assessment and public health applications.