Lung Cancer Prediction- Minor Project

-By

Sanjay R

23m0002@iitb.ac.in

Krutanic-Machine Learning Course

Problem Statement

The objective of this project is to develop a machine learning model to predict the likelihood of lung cancer based on a set of given features. Lung cancer is one of the leading causes of cancer-related deaths globally, and early detection is crucial for effective treatment. By leveraging machine learning techniques, we aim to build a predictive model that can assist in the early diagnosis of lung cancer, potentially saving lives and improving patient outcomes.

Models Tested

- **Logistic Regression:** It is a linear, interpretable model that provides a straightforward baseline for binary classification tasks, making it ideal for quick implementation and initial analysis.
- **Random Forest:** An ensemble of decision trees, it handles non-linear relationships and is robust against overfitting, providing reliable predictions with feature importance insights.
- **Support Vector Machine:** Effective in high-dimensional spaces, SVM focuses on maximizing the margin between classes, making it powerful for complex classification problems with non-linear boundaries.
- **K-Nearest Neighbors:** A simple, instance-based algorithm, KNN is non-parametric and flexible, making predictions based on the majority class of nearest neighbors, suitable for datasets with complex structures.
- **Decision Tree:** A model that splits data into branches based on feature values, it's easy to interpret and visualizes decision-making, but can overfit without proper pruning or regularization.

Performance of models

Performance	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.9744	0.9733	1	0.9865
Random Forest	0.9744	0.9863	0.9863	0.9863
Support Vector Machine	0.9615	0.9730	0.9863	0.9796

K-Nearest Neighbours	0.9231	0.9589	0.9589	0.9589
Decision Tree	0.9615	0.9861	0.9726	0.9793

- **Logistic Regression** and **Random Forest** have the highest accuracy (0.9744), with the Random Forest model slightly outperforming in Precision, Recall, and F1 Score.
- **Support Vector Machine** shows a lower accuracy compared to Logistic Regression and Random Forest but has high Recall.
- **K-Nearest Neighbors** has the lowest accuracy and is less competitive compared to the other models.
- **Decision Tree** performs similarly to the Support Vector Machine but slightly better in Precision and F1 Score.

Overall, **Logistic Regression, Random Forest, Support Vector Machine and Decision Tree** perform closely well, but the Random Forest model offers the best balance of Precision, Recall, and F1 Score, making it a strong choice for the prediction task.

Challenges faced:

- The need to encode categorical variables (like gender) before feeding them into the models increases preprocessing complexity.
- Models like SVM and KNN are sensitive to the scale of the data, requiring normalization or standardization. Ensuring consistent preprocessing across all models adds complexity.
- Storing large models, particularly Random Forests with many trees, can consume significant memory, which can be a challenge depending on the computational resources available. As the data used was comparatively low dimensional, the computation was not highly time-consuming.