CSE422 : ARTIFICIAL INTELLIGENCE
Project Report
Project Title :Lung Cancer Prediction

Name	ID
Omor Bin Amjad Chowdhury	23241085
Asif Imtiaz Chowdhury	21241050
Rayhan Sharif Sadif	21101063

Submitted To:

Sumaiya Akter
Lecturer
Brac University

Swattic Ghose
Lecturer
Brac University

Table of Contents

INTRODUCTION

Lung Cancer Prediction is an artificial intelligence (AI) project that aims to predict the risk of lung cancer in individuals based on various factors and datasets. The project's objectives, motivation, and problems it aims to solve can be summarized as follows:

Objectives: The project aims to develop an AI tool that can accurately predict the risk of lung cancer for individuals with or without a significant smoking history, based on analyses of low-dose computed tomography (LDCT) scans from patients in the U.S. and Taiwan. This tool can help improve early detection, diagnosis, and treatment of lung cancer.

Motivation: Lung cancer is the leading cause of cancer death in the United States and around the world. The development of an AI tool that can accurately predict lung cancer risk can help reduce the number of lung cancer deaths by identifying individuals at high risk and providing timely interventions.
Problems: The project faces several challenges, such as the heterogeneity of lung cancer and the need for large and diverse datasets to train and validate the AI models. Additionally, AI tools currently cannot replace physicians in clinical trials and have limitations in terms of decision-making.

Solutions: AI has the potential to help treat lung cancer in various aspects, including detection, diagnosis, decision-making, and prognosis prediction. By integrating and analyzing large datasets, AI models can strengthen various aspects of lung cancer therapy, such as early detection, auxiliary diagnosis, prognosis prediction, and immunotherapy practice.

In summary, our project aims to develop an AI tool that can accurately predict lung cancer risk, helping to improve early detection, diagnosis, and treatment. The project is motivated by the high mortality rate associated with lung cancer and aims to address the challenges posed by the heterogeneity of lung cancer and the need for large datasets.

Dataset Description:

Dataset Link: https://www.kaggle.com/datasets/rishidamarla/cancer-patients-data

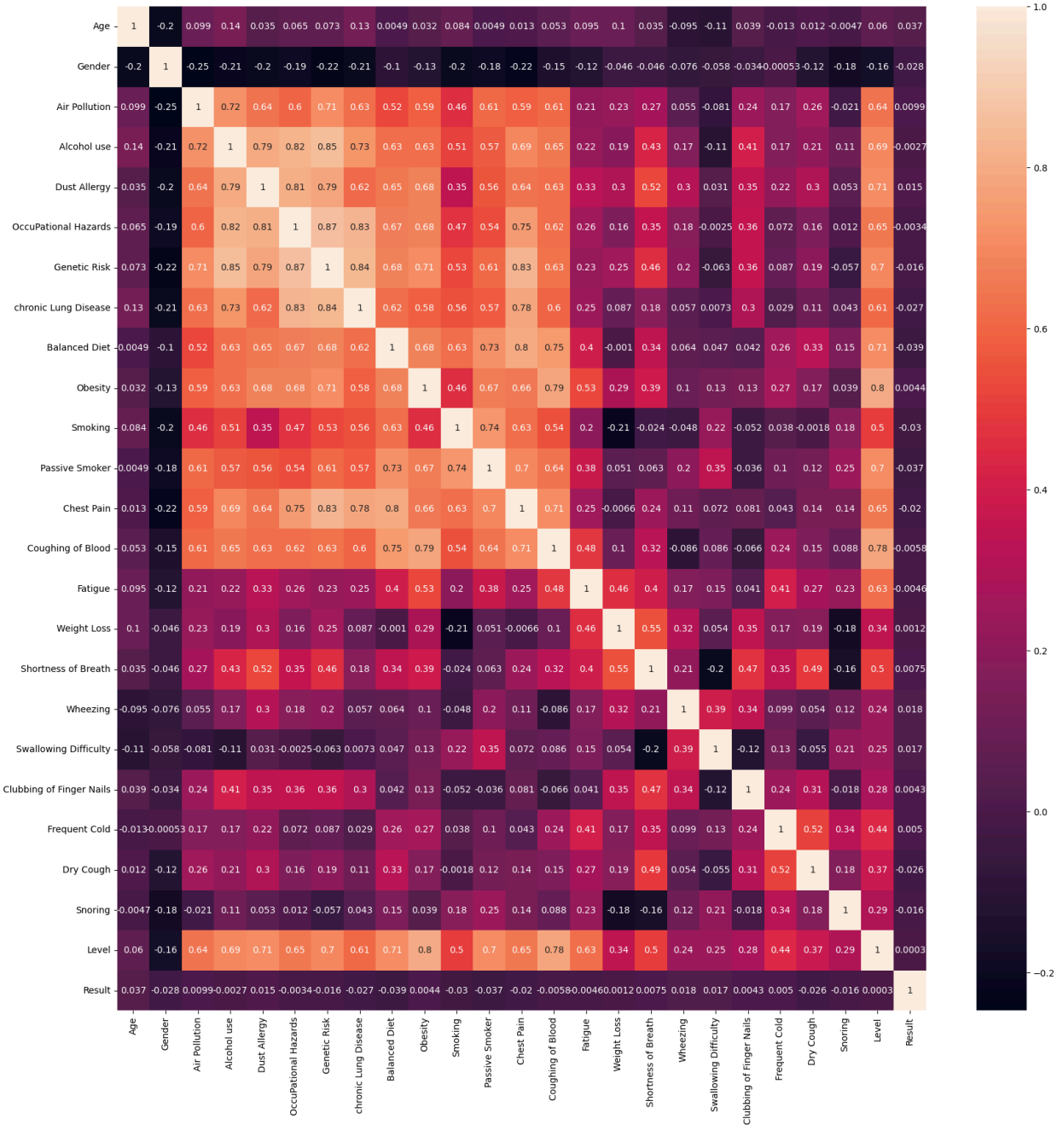
Task Type: The project focuses on a regression task, specifically predicting Lung Cancer Prediction of each person.
Number of Data Points: 1000
No of Features: 28 Features

Qualitative: Gender ,Level and Age

Quantitative: All others

Category: Classification problem as the risk level falls in one of three categories: Low, Medium or High.

Correlation of Features:


Age 1 -0.2 0.099 0.14 0.035 0.065 0.073 0.13 0.0049 0.032 0.084 0.0049 0.013 0.053 0.095 0.1 0.035 -0.095 -0.11 0.039 -0.013 0.012 -0.0047 0.06 0.037
Gender -0.2 1 -0.25 -0.21 -0.2 -0.19 -0.22 -0.21 -0.1 -0.13 -0.2 -0.18 -0.22 -0.15 -0.12 -0.046 -0.046 -0.076 -0.058 -0.034 -0.00053 -0.12 -0.18 -0.16 -0.028
Air Pollution 0.099 -0.25 1 0.72 0.64 0.6 0.71 0.63 0.52 0.59 0.46 0.61 0.59 0.61 0.21 0.23 0.27 0.055 -0.081 0.24 0.17 0.26 -0.021 0.64 0.0099
Alcohol use 0.14 -0.21 0.72 1 0.79 0.82 0.85 0.73 0.63 0.63 0.51 0.57 0.69 0.65 0.22 0.19 0.43 0.17 -0.11 0.41 0.17 0.21 0.11 0.69 -0.0027
Dust Allergy 0.035 -0.2 0.64 0.79 1 0.81 0.79 0.62 0.65 0.68 0.35 0.56 0.64 0.63 0.33 0.3 0.52 0.3 0.031 0.35 0.22 0.3 0.053 0.71 0.015
OccuPational Hazards 0.065 -0.19 0.6 0.82 0.81 1 0.87 0.83 0.67 0.68 0.47 0.54 0.75 0.62 0.26 0.16 0.35 0.18 -0.0025 0.36 0.072 0.16 0.012 0.65 -0.0034
Genetic Risk 0.073 -0.22 0.71 0.85 0.79 0.87 1 0.84 0.68 0.71 0.53 0.61 0.83 0.63 0.23 0.25 0.46 0.2 -0.063 0.36 0.087 0.19 -0.057 0.7 -0.016
chronic Lung Disease 0.13 -0.21 0.63 0.73 0.62 0.83 0.84 1 0.62 0.58 0.56 0.57 0.78 0.6 0.25 0.087 0.18 0.057 0.0073 0.3 0.029 0.11 0.043 0.61 -0.027
Balanced Diet 0.0049 -0.1 0.52 0.63 0.65 0.67 0.68 0.62 1 0.68 0.63 0.73 0.8 0.75 0.4 -0.001 0.34 0.064 0.047 0.042 0.26 0.33 0.15 0.71 -0.039
Obesity 0.032 -0.13 0.59 0.63 0.68 0.68 0.71 0.58 0.68 1 0.46 0.67 0.66 0.79 0.53 0.29 0.39 0.1 0.13 0.13 0.27 0.17 0.039 0.8 0.0044
Smoking 0.084 -0.2 0.46 0.51 0.35 0.47 0.53 0.56 0.63 0.46 1 0.74 0.63 0.54 0.2 -0.21 -0.024 -0.048 0.22 -0.052 0.038 -0.0018 0.18 0.5 -0.03
Passive Smoker 0.0049 -0.18 0.61 0.57 0.56 0.54 0.61 0.57 0.73 0.67 0.74 1 0.7 0.64 0.38 0.051 0.063 0.2 0.35 -0.036 0.1 0.12 0.25 0.7 -0.037
Chest Pain 0.013 -0.22 0.59 0.69 0.64 0.75 0.83 0.78 0.8 0.66 0.63 0.7 1 0.71 0.25 -0.0066 0.24 0.11 0.072 0.081 0.043 0.14 0.14 0.65 -0.02
Coughing of Blood 0.053 -0.15 0.61 0.65 0.63 0.62 0.63 0.6 0.75 0.79 0.54 0.64 0.71 1 0.48 0.1 0.32 -0.086 0.086 -0.066 0.24 0.15 0.088 0.78 -0.0058
Fatigue 0.095 -0.12 0.21 0.22 0.33 0.26 0.23 0.25 0.4 0.53 0.2 0.38 0.25 0.48 1 0.46 0.4 0.17 0.15 0.041 0.41 0.27 0.23 0.63 -0.0046
Weight Loss 0.1 -0.046 0.23 0.19 0.3 0.16 0.25 0.087 -0.001 0.29 -0.21 0.051 -0.0066 0.1 0.46 1 0.55 0.32 0.054 0.35 0.17 0.19 -0.18 0.34 0.0012
Shortness of Breath 0.035 -0.046 0.27 0.43 0.52 0.35 0.46 0.18 0.34 0.39 -0.024 0.063 0.24 0.32 0.4 0.55 1 0.21 -0.2 0.47 0.35 0.49 -0.16 0.5 0.0075
Wheezing -0.095 -0.076 0.055 0.17 0.3 0.18 0.2 0.057 0.064 0.1 -0.048 0.2 0.11 -0.086 0.17 0.32 0.21 1 0.39 0.34 0.099 0.054 0.12 0.24 0.018
Swallowing Difficulty -0.11 -0.058 -0.081 -0.11 0.031 -0.0025 -0.063 0.0073 0.047 0.13 0.22 0.35 0.072 0.086 0.15 0.054 -0.2 0.39 1 -0.12 0.13 -0.055 0.21 0.25 0.017
Clubbing of Finger Nails 0.039 -0.034 0.24 0.41 0.35 0.36 0.36 0.3 0.042 0.13 -0.052 -0.036 0.081 -0.066 0.041 0.35 0.47 0.34 -0.12 1 0.24 0.31 -0.018 0.28 0.0043
Frequent Cold -0.013 -0.00053 0.17 0.17 0.22 0.072 0.087 0.029 0.26 0.27 0.038 0.1 0.043 0.24 0.41 0.17 0.35 0.099 0.13 0.24 1 0.52 0.34 0.44 0.005
Dry Cough 0.012 -0.12 0.26 0.21 0.3 0.16 0.19 0.11 0.33 0.17 -0.0018 0.12 0.14 0.15 0.27 0.19 0.49 0.054 -0.055 0.31 0.52 1 0.18 0.37 -0.026
Snoring -0.0047 -0.18 -0.021 0.11 0.053 0.012 -0.057 0.043 0.15 0.039 0.18 0.25 0.14 0.088 0.23 -0.18 -0.16 0.12 0.21 -0.018 0.34 0.18 1 0.29 -0.016
Level 0.06 -0.16 0.64 0.69 0.71 0.65 0.7 0.61 0.71 0.8 0.5 0.7 0.65 0.78 0.63 0.34 0.5 0.24 0.25 0.28 0.44 0.37 0.29 1 0.0003
Result 0.037 -0.028 0.0099 -0.0027 0.015 -0.0034 -0.016 -0.027 -0.039 0.0044 -0.03 -0.037 -0.02 -0.0058 -0.0046 0.0012 0.0075 0.018 0.017 0.0043 0.005 -0.026 -0.016 0.0003 1
Age Gender Air Pollution Alcohol use Dust Allergy OccuPational Hazards Genetic Risk chronic Lung Disease Balanced Diet Obesity Smoking Passive Smoker Chest Pain Coughing of Blood Fatigue Weight Loss Shortness of Breath Wheezing Swallowing Difficulty Clubbing of Finger Nails Frequent Cold Dry Cough Snoring Level Result
1.0 0.8 0.6 0.4 0.2 0.0 −0.2

Dataset Pre-processing:

Several steps were taken during the preprocessing phase to prepare the dataset for effective machine learning modeling. These steps are essential for ensuring data quality and usability, as well as improving model performance. The following preprocessing procedures were carried out:

- Duplicate columns like "Smoking" and "Swallowing Difficulty" were removed.

- Columns like "Patient id" which do not contribute to the results were dropped.

- Missing values were handled through imputation using the "SimpleImputer" from `sklearn`. The mean strategy was employed.

- Categorical features like "Gender" and "Level" were encoded with appropriate values.

- Less informative features, such as "Genetic Risk" and "Occupational Hazard", were dropped using correlation analysis, visualized using Seaborn.

Feature Scaling:

Data Loading and Exploration:

It includes data exploration using functions like info(), head(), and visualization using Seaborn and Matplotlib for creating count plots, strip plots, histograms, and heatmap.

Data Preprocessing:

The dataset undergoes preprocessing steps, including handling missing values (SimpleImputer), mapping categorical values to numerical ones, and dropping unnecessary columns.

Data Scaling:

Min-Max scaling is applied to the features using MinMaxScaler from scikit-learn.

Model Training and Evaluation:

Machine learning models such as Decision Tree Classifier, Gaussian Naive Bayes, and Random Forest Classifier are trained and evaluated.

The models are trained both with and without feature scaling to observe the impact on performance.

Hyperparameter Tuning:

Grid Search Cross-Validation (GridSearchCV) is employed for hyperparameter tuning in the Support Vector Classifier (SVC).

Model Evaluation Metrics:

The code calculates and prints various evaluation metrics, including accuracy, confusion matrix, and classification report, to assess the models' performance.

Data Visualization:

Seaborn's heatmap is utilized to visualize confusion matrices, aiding in understanding the model's performance.

Machine Learning Models:

Decision Tree Classifier, Gaussian Naive Bayes, Random Forest Classifier, and Support Vector Classifier are used as machine learning models.

These features collectively contribute to the development, training, and evaluation of machine learning models for lung cancer prediction.

Dataset Splitting:

The splitting of the data into training and testing sets is an important step in preparing the dataset for machine learning modeling. This split enables the model to be evaluated on unseen data, providing an estimate of its performance on real-world data. The following are the steps and details for splitting the dataset in this project:

Defining Features and Target Variable:

The features (X) and the target variable (y) were defined. 'X' includes all the columns except for the target variable

Splitting Ratio:

The dataset was divided into training (70%) and test (30%) sets, with two variants for each: one scaled using "MinMaxScaler" and the other unscaled. The data to be split was selected randomly.. A 70-30 split is commonly used in machine learning as it provides a substantial amount of data for learning while retaining enough data to test and validate the models effectively.

Random Split with Fixed Random State:

The train_test_split function from scikit-learn was used for this purpose. A random_state was set to ensure reproducibility of the results.

Result of the Split:

The training set consisted of samples (features and target), and the testing set comprised samples.

This methodical approach to dataset splitting ensures that the models are trained on a diverse set of data points and are tested on different samples to gauge their generalization ability effectively.

Model Training and Evaluation:

The project included training and testing three distinct machine learning models: decision trees,naive bayes,random forest,support vector machine and grid search. Each model was chosen for its distinct characteristics and ability to address regression issues. The following explains the training and evaluation process for each model:

Decision Tree:

Training: Decision Trees are trained by recursively partitioning the dataset based on feature values. The algorithm selects the best feature at each node to split the data, aiming to maximize information gain or Gini impurity reduction.

Testing: During testing, new data traverses the tree, following the learned decision rules at each node until it reaches a leaf. The predicted class at the leaf is assigned to the input data.

Naive Bayes:

Training: Naive Bayes is a probabilistic model based on Bayes' theorem. It assumes that features are conditionally independent given the class. The model calculates class probabilities and conditional probabilities of features given the class during training.

Testing: In testing, the model uses Bayes' theorem to calculate the probability of each class given the input features. The class with the highest probability is assigned to the input data.

Random Forest:

Training: Random Forest is an ensemble model that builds multiple decision trees during training. Each tree is trained on a random subset of the data and a random subset of features to promote diversity. The final prediction is made by averaging or voting among the individual trees.

Testing: During testing, the input data passes through each tree in the forest, and the final prediction is determined by aggregating the predictions of all trees.

Support Vector Machine (SVM):

Training: SVM aims to find a hyperplane that best separates data points of different classes. During training, it identifies the optimal hyperplane by maximizing the margin between classes. SVM can handle linear and non-linear separation using different kernels.

Testing: In testing, the input data are mapped into the feature space, and their position relative to the hyperplane is used to determine the class label.

Grid Search:

Training and Testing: Grid Search is not a model but a hyperparameter tuning technique. It systematically searches through a predefined hyperparameter grid to find the combination that yields the best model performance. It involves training and testing the model with various hyperparameter values, allowing for optimization.

Evaluation Process:

Each model was trained on the training set and then used to make predictions on the test set.

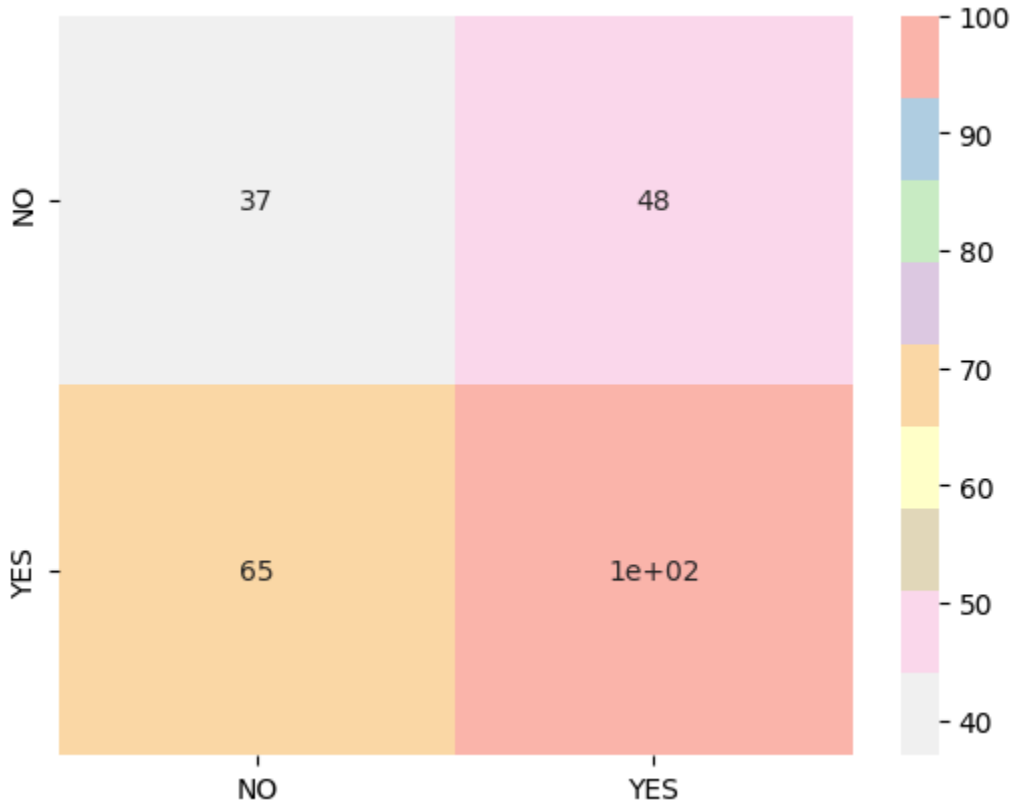
The evaluation focused on training and testing before and after scaling. This allowes for an in-depth comparison of how each model performs in predicting and understanding the strengths and weaknesses of each approach in the context of the given dataset.

Model Selection and Comparison Analysis

The project involved a thorough comparison and analysis of three machine learning models:Decision Tree,Naive Bayes,Random Forest.The initial performance of these models was evaluated based on Without Scaling Training and testing accuracy of the model and with Scaling Training and testing accuracy of the model.

The results were as follows:

Decision Tree:



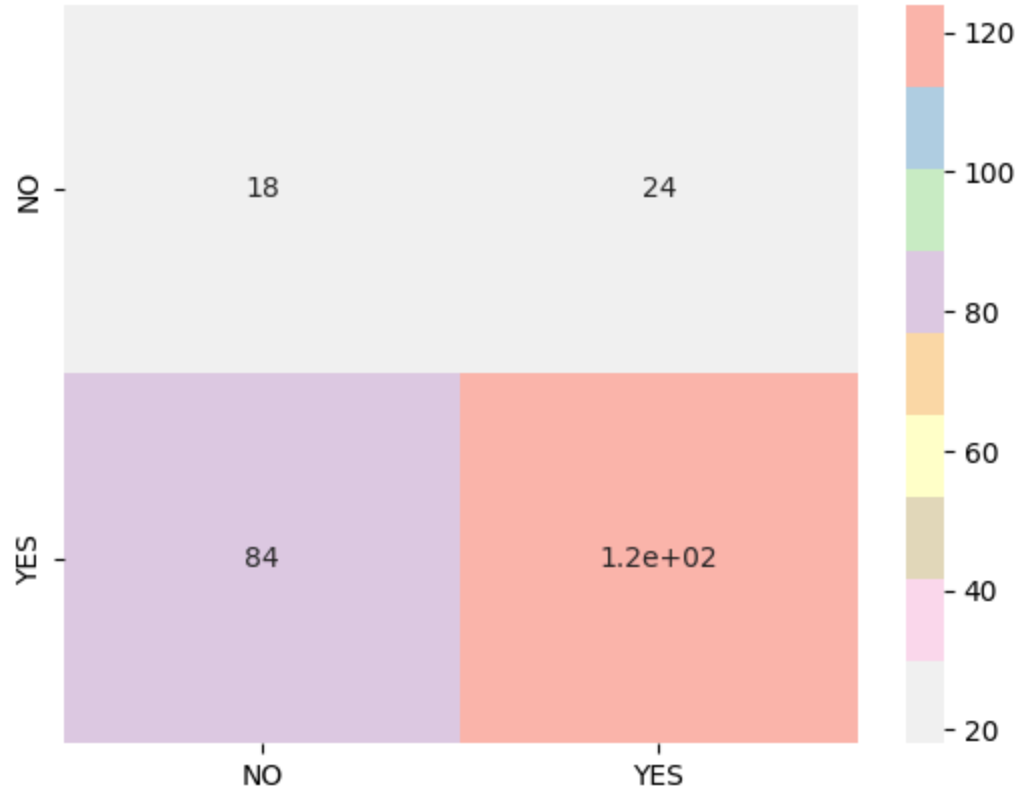
Without Scaling Training accuracy of the model is 0.7453333333333333
Without Scaling Testing accuracy of the model is 0.492
With Scaling Training accuracy of the model is 0.7453333333333333
With Scaling Testing accuracy of the model is 0.548

```
[ 63 100]]
              precision    recall  f1-score   support

           0       0.44      0.36      0.40       102
           1       0.61      0.68      0.64       148

    accuracy                           0.55       250
   macro avg       0.52      0.52      0.52       250
weighted avg       0.54      0.55      0.54       250
```

Naive Bayes:

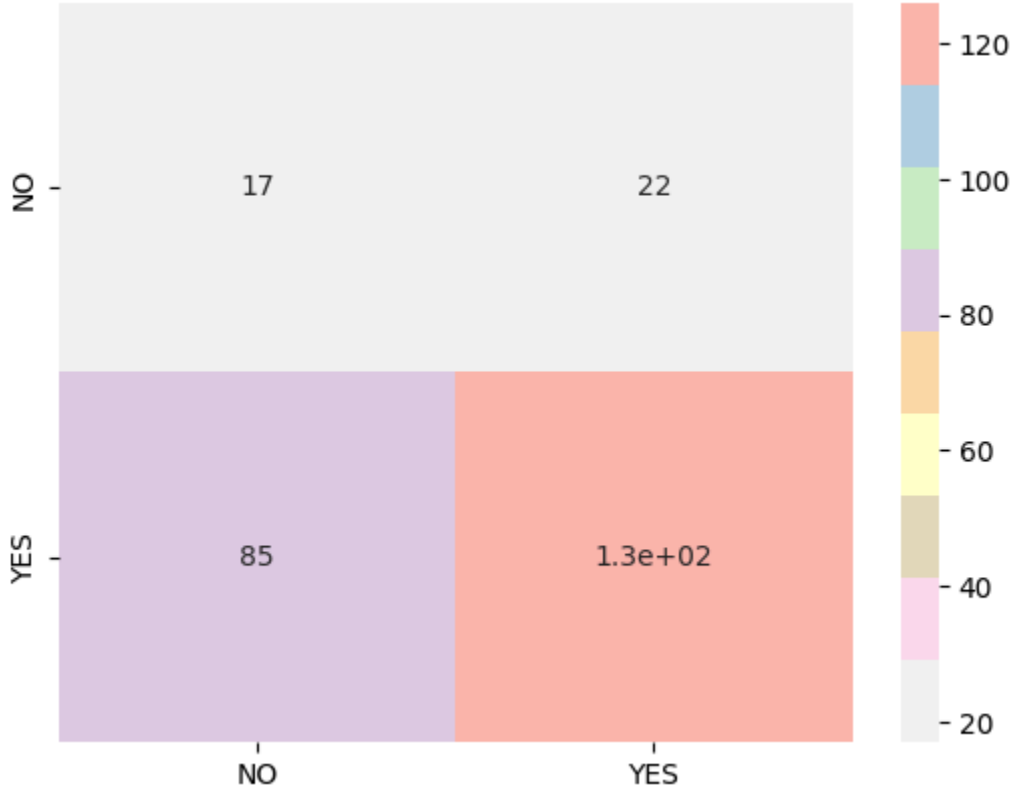


Without Scaling Training accuracy of the model is 0.60
Without Scaling Testing accuracy of the model is 0.57
With Scaling Training accuracy of the model is 0.60
With Scaling Testing accuracy of the model is 0.57

[84 124]]

	precision	recall	f1-score	support
0	0.43	0.18	0.25	102
1	0.60	0.84	0.70	148
accuracy			0.57	250
macro avg	0.51	0.51	0.47	250
weighted avg	0.53	0.57	0.51	250

Random Forest:



Without Scaling Training accuracy of the model is 0.74
Without Scaling Testing accuracy of the model is 0.52
With Scaling Training accuracy of the model is 0.75
With Scaling Testing accuracy of the model is 0.57

	precision	recall	f1-score	support
0	0.44	0.17	0.24	102
1	0.60	0.85	0.70	148
accuracy			0.57	250
macro avg	0.52	0.51	0.47	250
weighted avg	0.53	0.57	0.51	250

Outcome from Model Comparison:

Based on these metrics, the Random Forest model significantly outperformed the other two. The random forest model proved to be the most robust and consistent performer, both before and after scaling. Naive bayes show no change before and after scaling and the decision tree improves its testing accuracy but there is no change in its training accuracy.

Conclusion

This project, which centered on lung cancer prediction.It successfully navigated through the various stages of a machine learning project, from data preprocessing to model training and evaluation. The journey resulted in insightful findings that not only shed light on the capabilities of various machine learning models, but also highlighted the importance of careful feature selection and model evaluation.

Finally, this project not only met its goal of predicting product selling prices, but it also provided a thorough understanding of the complexities involved in machine learning projects, from data preparation to model evaluation and selection. These discoveries pave the way for further analysis and continuous improvement in the field of predictive modeling.