LUNG CANCER PREDICTION USING

Introduction

Lung cancer is the major cause of cancer-related death in this generation, and it is expected to remain so for the foreseeable future. It is feasible to treat lung cancer if the symptoms of the disease are detected early. It is possible to construct a sustainable prototype model for the treatment of lung cancer using the current developments in computational intelligence without negatively impacting the environment. Because it will reduce the number of resources squandered as well as the amount of work necessary to complete manual tasks, it will save both time and money. A machine learning model based on random forest was used to optimize the detection process from the lung cancer dataset. Using a random forest classifier, lung cancer patients are classified based on their symptoms at the same time as the Python programming language is utilized to further the model implementation. The effectiveness of our model was evaluated in terms of several different criteria. As a result of the favorable findings of this research, hospitals can deliver better healthcare to their patients. Patients with lung cancer can be diagnosed early and treated accordingly thus increasing chances of survival. The proposed method gets a 94.9% accuracy rate when comparing the existing methods.

Data Collection and Preprocessing:

Data has been collected from Kaggle datasets and the patient dataset was used for lung cancer prediction and to explore the steps involved in cleaning and preprocessing the data. Many data-cleaning techniques and data-preprocessing methods were used.

Exploratory Data Analysis:

Selection of Machine Learning Algorithms

Many analyses were done to select the machine-learning algorithms for lung cancer

prediction. Many insights were gained into the strengths and limitations of different algorithms.

Training and Testing of Models

Train and test data were chosen and applied to the chosen machine learning models

using the dataset. Explored strategies to optimize model performance and accuracy.

Models used:

- Support Vector classifier
- Random Forest Classifier
- Logistic Regression
- Decision Tree
- Gradient boost
- Gaussian naive base

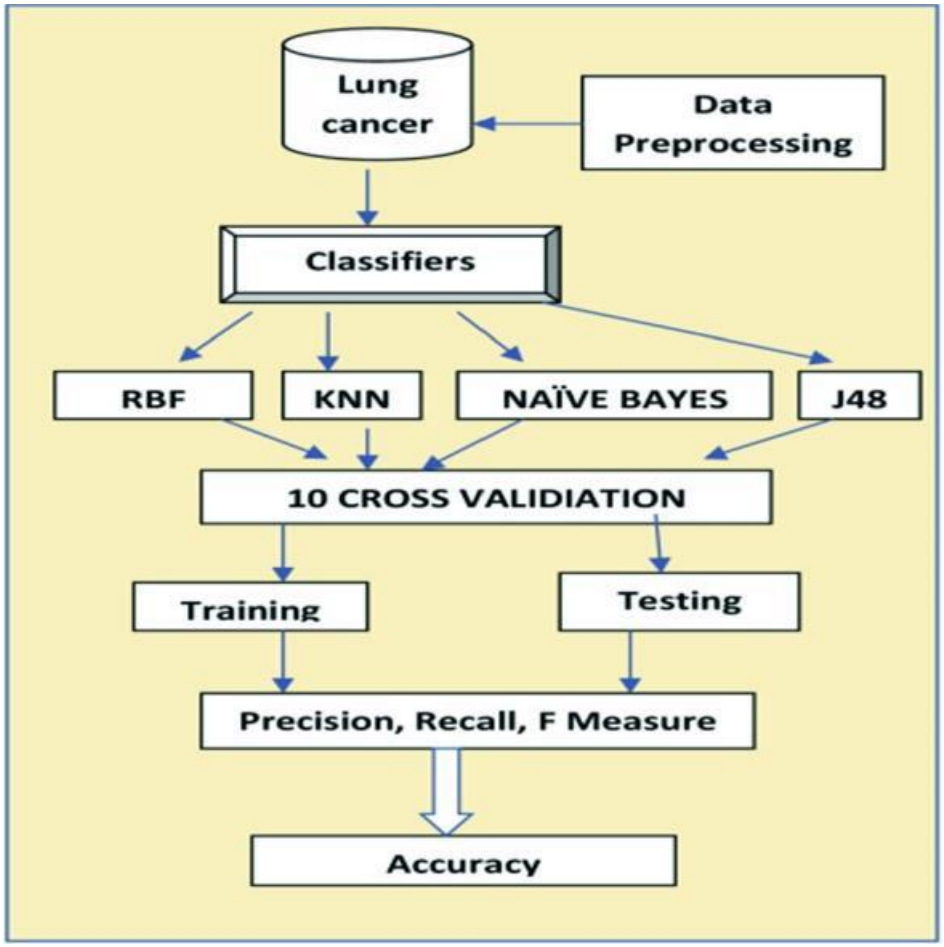
Feature Selection and Engineering:

Selection of Relevant Features

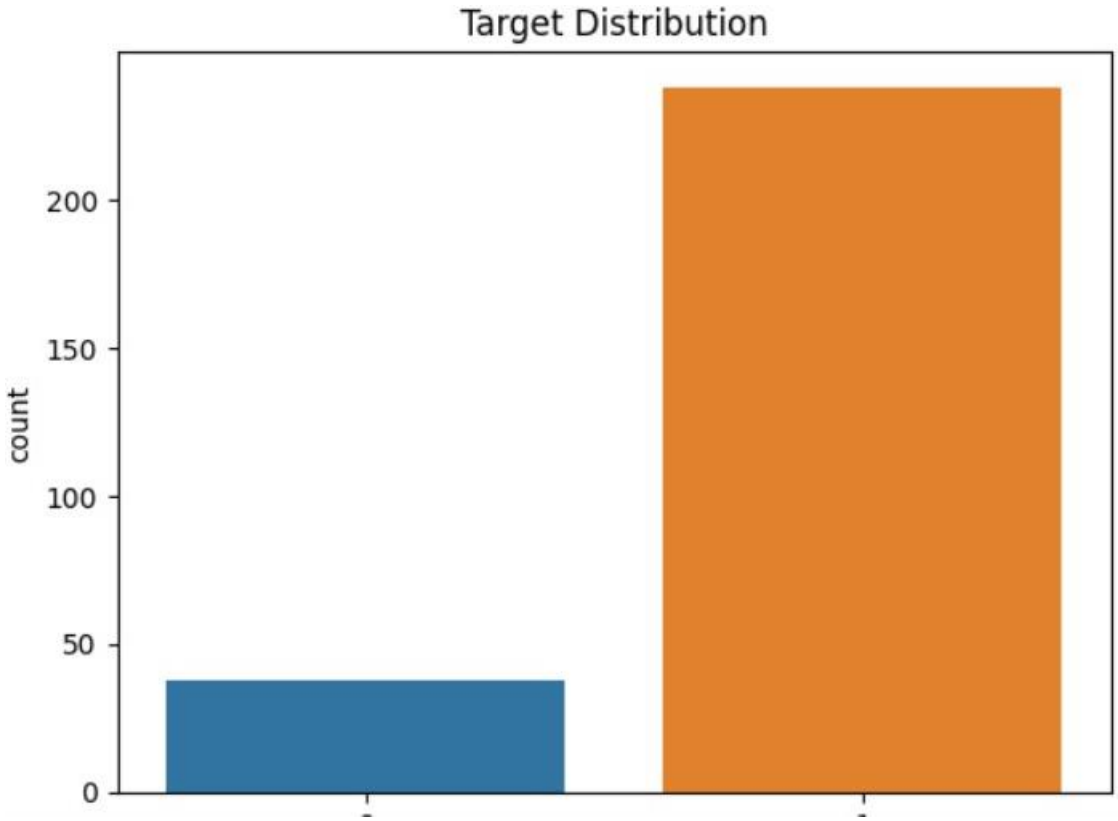
The most relevant features for accurate lung cancer prediction were selected. Explored different feature selection methods and understood their impact on model performance, by using a correlation matrix we found that the features 'Yellow Finger' and 'Anxiety ' had more correlation among themselves and also with the target variable.

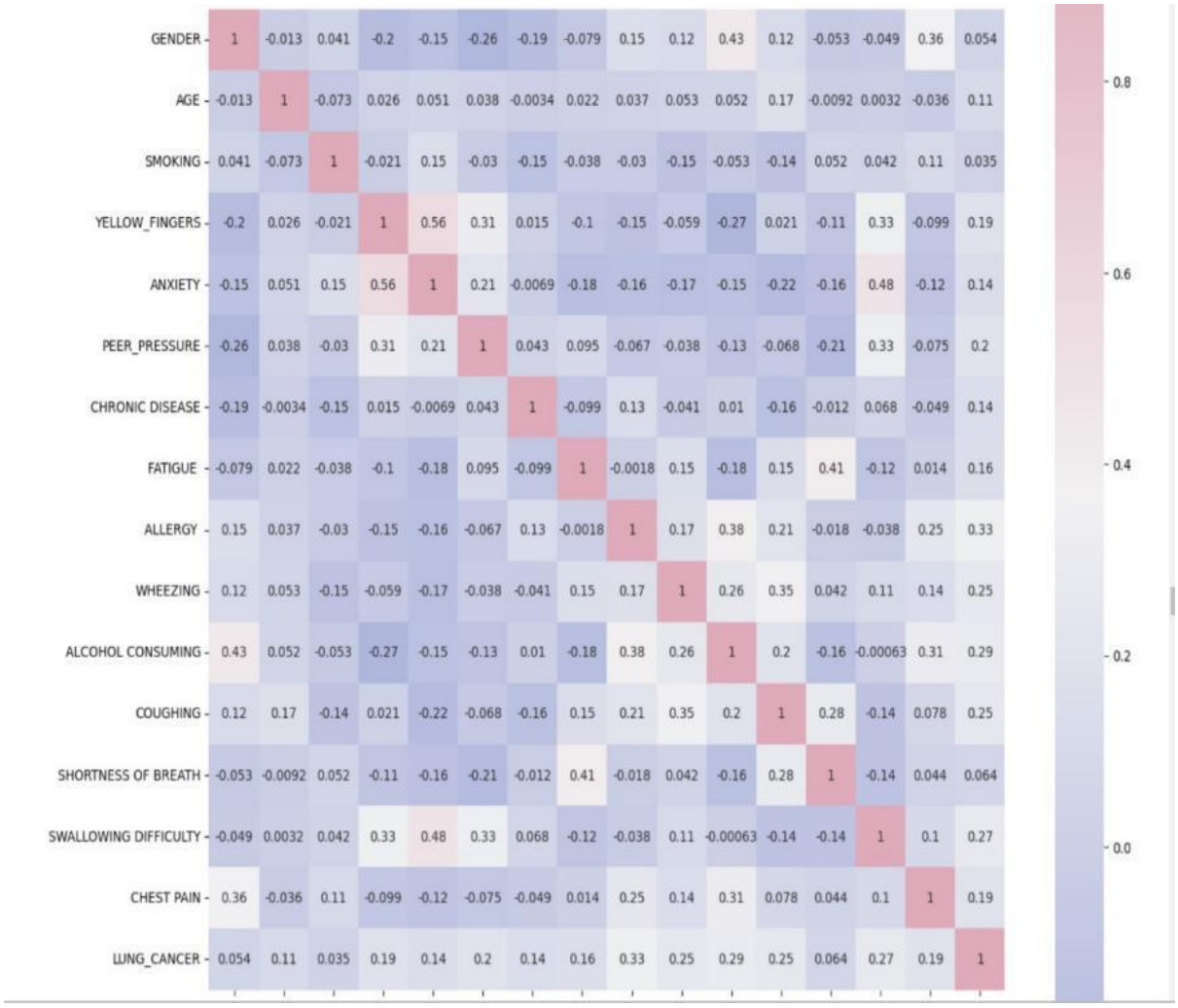
Feature Engineering and Transformation

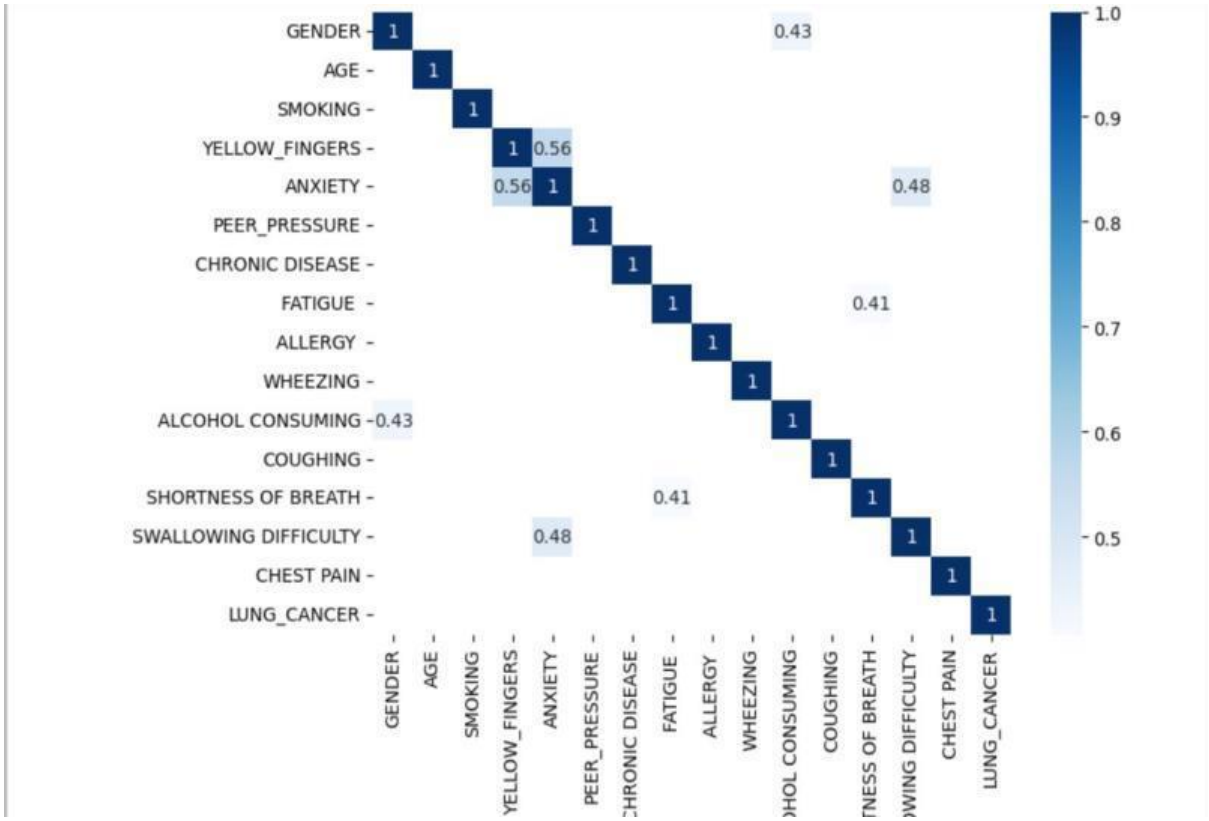
The art of feature engineering was discovered and the techniques used to transform the dataset were performed. A new feature was added combining the two other featured and exploring how feature engineering can enhance the predictive power of machine learning models.


Lung cancer
Data Preprocessing
Classifiers
RBF
KNN
NAÏVE BAYES
J48
10 CROSS VALIDIATION
Training
Testing
Precision, Recall, F Measure
Accuracy

OUTPUTS:


Target Distribution
count
200
150
100
50
0


GENDER 1 -0.013 0.041 -0.2 -0.15 -0.26 -0.19 -0.079 0.15 0.12 0.43 0.12 -0.053 -0.049 0.36 0.054
AGE -0.013 1 -0.073 0.026 0.051 0.038 -0.0034 0.022 0.037 0.053 0.052 0.17 -0.0092 0.0032 -0.036 0.11
SMOKING 0.041 -0.073 1 -0.021 0.15 -0.03 -0.15 -0.038 -0.03 -0.15 -0.053 -0.14 0.052 0.042 0.11 0.035
YELLOW_FINGERS -0.2 0.026 -0.021 1 0.56 0.31 0.015 -0.1 -0.15 -0.059 -0.27 0.021 -0.11 0.33 -0.099 0.19
ANXIETY -0.15 0.051 0.15 0.56 1 0.21 -0.0069 -0.18 -0.16 -0.17 -0.15 -0.22 -0.16 0.48 -0.12 0.14
PEER_PRESSURE -0.26 0.038 -0.03 0.31 0.21 1 0.043 0.095 -0.067 -0.038 -0.13 -0.068 -0.21 0.33 -0.075 0.2
CHRONIC DISEASE -0.19 -0.0034 -0.15 0.015 -0.0069 0.043 1 -0.099 0.13 -0.041 0.01 -0.16 -0.012 0.068 -0.049 0.14
FATIGUE -0.079 0.022 -0.038 -0.1 -0.18 0.095 -0.099 1 -0.0018 0.15 -0.18 0.15 0.41 -0.12 0.014 0.16
ALLERGY 0.15 0.037 -0.03 -0.15 -0.16 -0.067 0.13 -0.0018 1 0.17 0.38 0.21 -0.018 -0.038 0.25 0.33
WHEEZING 0.12 0.053 -0.15 -0.059 -0.17 -0.038 -0.041 0.15 0.17 1 0.26 0.35 0.042 0.11 0.14 0.25
ALCOHOL CONSUMING 0.43 0.052 -0.053 -0.27 -0.15 -0.13 0.01 -0.18 0.38 0.26 1 0.2 -0.16 -0.00063 0.31 0.29
COUGHING 0.12 0.17 -0.14 0.021 -0.22 -0.068 -0.16 0.15 0.21 0.35 0.2 1 0.28 -0.14 0.078 0.25
SHORTNESS OF BREATH -0.053 -0.0092 0.052 -0.11 -0.16 -0.21 -0.012 0.41 -0.018 0.042 -0.16 0.28 1 -0.14 0.044 0.064
SWALLOWING DIFFICULTY -0.049 0.0032 0.042 0.33 0.48 0.33 0.068 -0.12 -0.038 0.11 -0.00063 -0.14 -0.14 1 0.1 0.27
CHEST PAIN 0.36 -0.036 0.11 -0.099 -0.12 -0.075 -0.049 0.014 0.25 0.14 0.31 0.078 0.044 0.1 1 0.19
LUNG_CANCER 0.054 0.11 0.035 0.19 0.14 0.2 0.14 0.16 0.33 0.25 0.29 0.25 0.064 0.27 0.19 1
0.8
0.6
0.4
0.2
0.0




```
print("Logistic regression models' average accuracy:", np.mean(lr_model_scores))
print("Decision tree models' average accuracy:", np.mean(dt_model_scores))
print("Gaussian naive bayes models' average accuracy:", np.mean(gnb_model_scores))
print("Support Vector Classifier models' average accuracy:", np.mean(svc_model_scores))
print("Random forest models' average accuracy:", np.mean(rf_model_scores))
print("Gradient boost models' average accuracy:", np.mean(gb_model_scores))
```

```
Logistic regression models' average accuracy: 0.9288120567375886

Decision tree models' average accuracy: 0.9227393617021278

Gaussian naive bayes models' average accuracy: 0.8870124113475178

Support Vector Classifier models' average accuracy: 0.9476063829787235

Random forest models' average accuracy: 0.9498670212765958

Gradient boost models' average accuracy: 0.947695035460993
```

So the Stratified K-Fold cross validation is showing Random Forest model gives the most accuracy of 94.9%, and also other models like Gradient Boost, Support Vector Classifier gives almost same accuracies, while Gaussian Naive Bayes model gives the least accuarcy of 88.7%.

ACCURACY:

➢ Logistic regression models' average accuracy: 0.9288120567375886

➢ Decision tree models' average accuracy: 0.9227393617021278

➢ Gaussian naive Bayes models' average accuracy: 0.8870124113475178

➢ Support Vector Classifier models' average accuracy: 0.9476063829787235

➢ Random forest models' average accuracy: 0.9498670212765958

➢ Gradient boost models' average accuracy: 0.947695035460993

So, the Stratified K-Fold cross-validation is showing Random Forest model gives the highest accuracy of 94.9%, and also other models like Gradient Boost, and Support Vector Classifier give almost the same accuracies, while the Gaussian Naive Bayes model gives the lowest accuracy of 88.7%.

CONCLUSION:

Through rigorous evaluation, diverse machine learning models exhibited high accuracy, precision, and recall rates in predicting lung cancer based on patient data.