Data Science II: Modeling And Data Analysis

Report

Lung Cancer Detection

Diana Sargsyan Sycamore Herlihy Jennifer Kimball Joan Albert Olivia Deckers

Supervisor
Oren Mangoubi, Ph.D.

Contents

1 Abstract

Predicting the presence of lung cancer early has the potential to save many lives. There are many different predictors that can be used in this process, including allergy, alcohol consumption, and fatigue. In this paper, we make use of various machine learning models, including Decision Trees, Random Forests, Support Vector Classifier, and K-Nearest Neighbors Classifier to predict lung cancer diagnoses in adult patients, eventually achieving a maximum accuracy of 96.7

2 Introduction

An effective cancer prediction system would help people understand their cancer risk and perhaps convince them to change some of the behaviors that are putting them at higher risk. A machine learning model that predicts the presence of lung cancer is a low cost solution that has the potential to save lives by catching the disease early. For our analysis, we will use Python libraries to fit a variety of different models to the data. We will then choose the best model based primarily off of the accuracy of the model's predictions. This analysis will help us discover which features are the most helpful in correctly predicting lung cancer and whether we can accurately predict the presence of lung cancer.

3 Data and Preprocessing

3.1 Dataset

We decided to use **Lung Cancer** dataset from **Kaggle**, having 16 attributes and 284 instances We need a lot of imports for our computations, however the most important ones are the following:

matplotlib
pandas
tree
numpy
SVC
train_test_split
StandardScaler
KNeighborsClassifier
RandomForestClassifier
svm

3.2 Analysis and description of the dataset

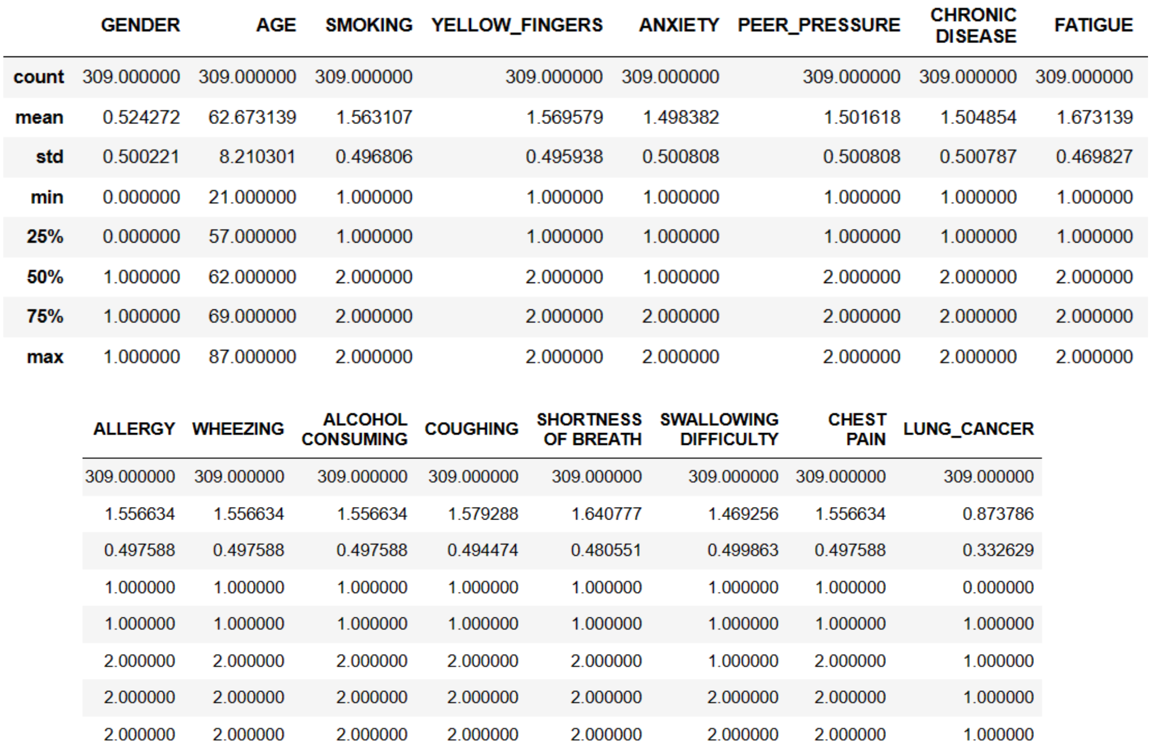
Here are the columns of the dataset:

Gender: M(male), F(female)
Age: Integer age of the patient.
Smoking: YES=2, NO=1.
Yellow fingers: YES=2, NO=1.
Anxiety: YES=2, NO=1.
Peer_pressure: YES=2, NO=1.
Chronic Disease: YES=2, NO=1.
Fatigue: YES=2, NO=1.
Allergy: YES=2, NO=1.
Wheezing: YES=2, NO=1.
Alcohol: YES=2, NO=1.
Coughing: YES=2, NO=1.
Shortness of Breath: YES=2, NO=1.
Swallowing Difficulty: YES=2, NO=1.
Chest pain: YES=2, NO=1.
Lung Cancer: YES, NO.

Here is some general information on the dataset:

```
class ’pandas.core.frame.DataFrame’
RangeIndex: 309 entries, 0 to 308
Data columns (total 16 columns):
Column Non-Null Count Dtype
— —— ————— ——
0 GENDER 309 non-null int64
1 AGE 309 non-null int64
2 SMOKING 309 non-null int64
3 YELLOW_FINGERS 309 non-null int64
4 ANXIETY 309 non-null int64
5 PEER_PRESSURE 309 non-null int64
6 CHRONIC DISEASE 309 non-null int64
7 FATIGUE 309 non-null int64
8 ALLERGY 309 non-null int64
9 WHEEZING 309 non-null int64
10 ALCOHOL CONSUMING 309 non-null int64
11 COUGHING 309 non-null int64
12 SHORTNESS OF BREATH 309 non-null int64
13 SWALLOWING DIFFICULTY 309 non-null int64
14 CHEST PAIN 309 non-null int64
15 LUNG_CANCER 309 non-null int64
dtypes: int64(16)
```

memory usage: 38.8 KB



	GENDER	AGE	SMOKING	YELLOW_FINGERS	ANXIETY	PEER_PRESSURE	CHRONIC DISEASE	FATIGUE	ALLERGY	WHEEZING	ALCOHOL CONSUMING	COUGHING	SHORTNESS OF BREATH	SWALLOWING DIFFICULTY	CHEST PAIN	LUNG_CANCER
count	309.000000	309.000000	309.000000	309.000000	309.000000	309.000000	309.000000	309.000000	309.000000	309.000000	309.000000	309.000000	309.000000	309.000000	309.000000	309.000000
mean	0.524272	62.673139	1.563107	1.569579	1.498382	1.501618	1.504854	1.673139	1.556634	1.556634	1.556634	1.579288	1.640777	1.469256	1.556634	0.873786
std	0.500221	8.210301	0.496806	0.495938	0.500808	0.500808	0.500787	0.469827	0.497588	0.497588	0.497588	0.494474	0.480551	0.499863	0.497588	0.332629
min	0.000000	21.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	0.000000
25%	0.000000	57.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000
50%	1.000000	62.000000	2.000000	2.000000	1.000000	2.000000	2.000000	2.000000	2.000000	2.000000	2.000000	2.000000	2.000000	1.000000	2.000000	1.000000
75%	1.000000	69.000000	2.000000	2.000000	2.000000	2.000000	2.000000	2.000000	2.000000	2.000000	2.000000	2.000000	2.000000	2.000000	2.000000	1.000000
max	1.000000	87.000000	2.000000	2.000000	2.000000	2.000000	2.000000	2.000000	2.000000	2.000000	2.000000	2.000000	2.000000	2.000000	2.000000	1.000000

Figure 1: **Description of the dataset**

Here we can observe the following:

AGE - The dataset contains mostly elderly individuals, with an average and median age of around 62 years, but some younger individuals are also included, with a minimum age of 21.

SMOKING - The dataset is composed mostly of smokers, as indicated by the mean value of the smoking variable, which should be 1.5 if perfectly balanced since values are either 1 or 2.

YELLOW_FINGERS - People with yellow fingers are also prevalent in the dataset, as indicated by the mean value of the yellow fingers variable.

ANXIETY - The dataset is balanced in terms of anxiety.

PEER_PRESSURE - The dataset is balanced in terms of peer pressure.

CHRONIC_DISEASE - The dataset is balanced in terms of chronic disease.

FATIGUE - The majority of individuals in the dataset show signs of fatigue.

ALLERGY - The majority of individuals in the dataset show signs of allergy.

WHEEZING - The majority of individuals in the dataset show signs of wheezing symptom.

ALCOHOL CONSUMING - The majority of individuals in the dataset consume alcohol.

COUGHING - The majority of individuals in the dataset show coughing symptom.

SHORTNESS OF BREATH - The majority of individuals in the dataset have shortness of breath.

SWALLOWING DIFFICULTY - Most individuals in the dataset do not have difficulty swallowing.

CHEST PAIN - The dataset is mostly comprised of people with chest pain.

3.3 Data Pre-processing

Pre-processing mainly consisted of cleaning up the data and filtering the dataset to only show rows that matched a certain criteria. The steps we followed are:

1. Encoding the categorical variables LUNG_CANCER and GENDER. For GENDER we will have 1 for 'M' which stands for Male, and 0 for 'F' which stands for Female. For LUNG_CANCER we will have 1 for 'YES' and 0 for 'NO'.

2. Separating independent and dependent attributes.

3. Changing values of other predictor columns from 2 to 1 and from 1 to 0.

4. Splitting independent and dependent attributes into training and testing data with the 80/20 rule. The shape of X_train is (247, 15) and the shape of X_test is (62, 15).

5. Scaling the AGE column with StandardScaler, scaling the data to a unit variance.

4 Set-Up, Create, Fit and Predict Models

Our models are:

- **SVC** with Polynomial, RBF, Sigmoid, and Linear kernels.

  This algorithm categorizes the data by finding an optimal hyperplane that separates the two categories of data that lies between two other planes that each go through data points on either side.

- **K-NearestNeighborsClassifier** with a range of 1 to 25 neighbors.

  This is an algorithm that determines which category a data point is in based on the categories that the K nearest neighbors are in.

- **DecisionTreeClassifier** with random state 1.

  This is an algorithm that iteratively divides the data set into smaller prediction regions based on values of the predictor variables.

- **RandomForestClassifier** with 50 estimators and random state 1.

  This is a group of decision trees that each give a class prediction, and the category predicted the most by all of the trees is the prediction for the data point being tested.

4.1 Visualizations

Now, using graphs, we will present our observations.
First, the heatmap shows that there is some multicollinearity present in the data. (see Figure 2)

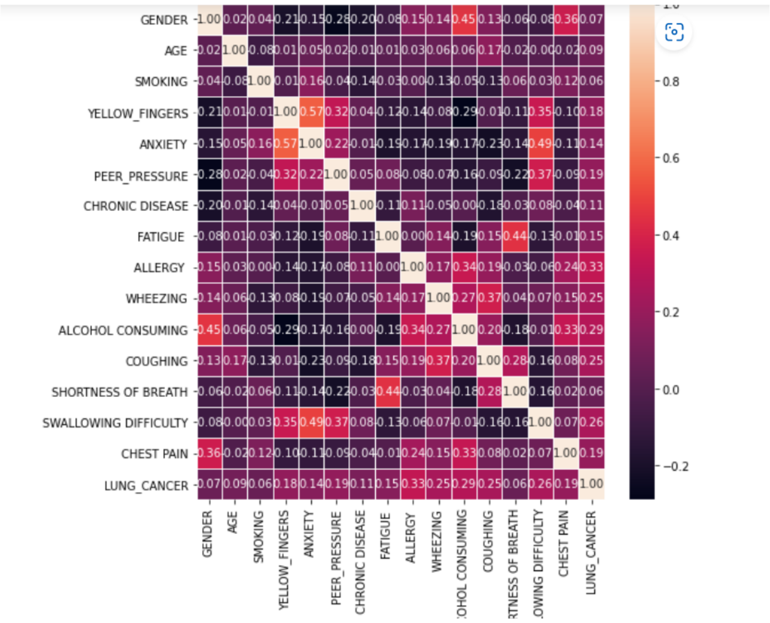


Figure 2: **Heatmap**

We have built KNN with number of neighbors ranging from [1, 24] to see which has the highest accuracy. According to the graph and our calculations, the highest accuracy is when number of neighbors is 12, with 96.7% accuracy. (see Figure 3)

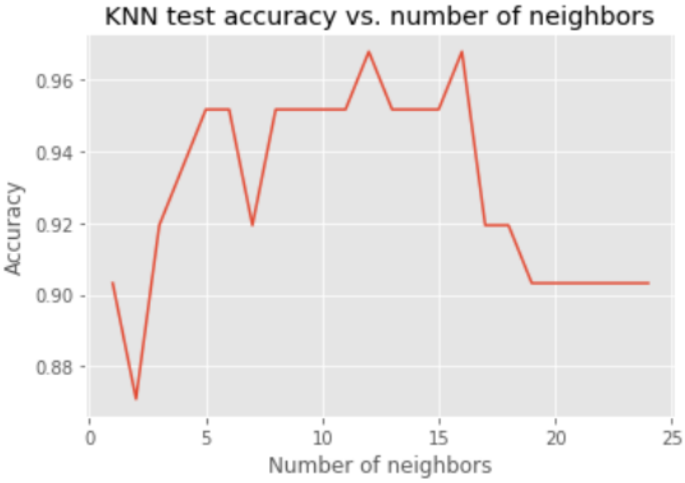


Figure 3: **KNN**

Next, we have a confusion matrix, which shows that in 53 cases out of 54 our model predicts correctly that the patient has lung cancer. It is important to have very few false negatives, a case where a patient is unaware of their risk for lung cancer, and the matrix shows that this only occurred in 1 instance. (see Figure 4)

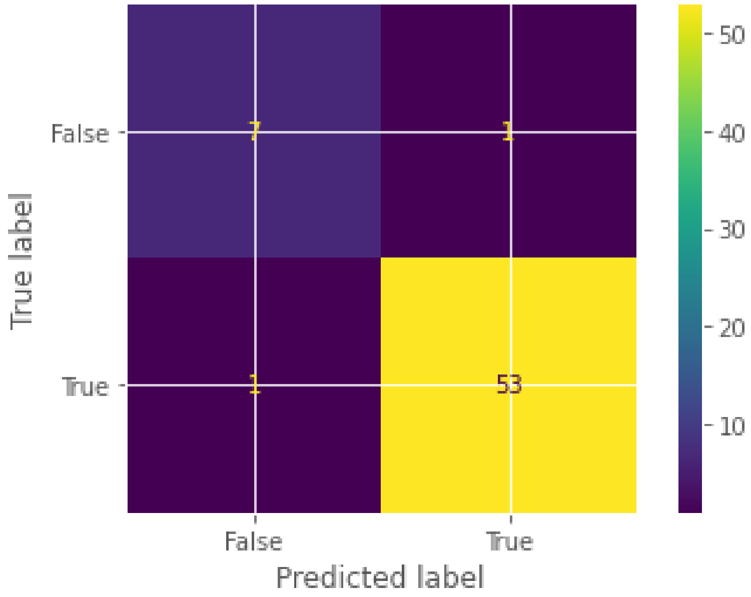
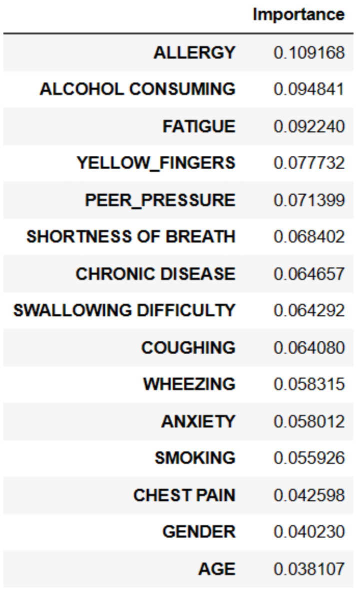


Figure 4: **Confusion matrix for KNN, k=12**

The next graph displays feature importances. It shows that allergy, alcohol consumption, and fatigue are the top 3 attributes that assist in detecting lung cancer. While it is true that patients cannot control whether or not they have an allergy, they can control their alcohol consumption. The patient also does not have direct control over feeling fatigued, but knowing that fatigue is one of the best predictors will help catch some instances of the disease earlier. It is also interesting to note that smoking, which is commonly thought of as the best predictor for lung cancer, is not in the top five. The presence of yellow fingers is the fourth most important feature, however, and this is usually caused by chemical stains that result from prolonged cigarette use. (see Figure 5)



	Importance
ALLERGY	0.109168
ALCOHOL CONSUMING	0.094841
FATIGUE	0.092240
YELLOW_FINGERS	0.077732
PEER_PRESSURE	0.071399
SHORTNESS OF BREATH	0.068402
CHRONIC DISEASE	0.064657
SWALLOWING DIFFICULTY	0.064292
COUGHING	0.064080
WHEEZING	0.058315
ANXIETY	0.058012
SMOKING	0.055926
CHEST PAIN	0.042598
GENDER	0.040230
AGE	0.038107

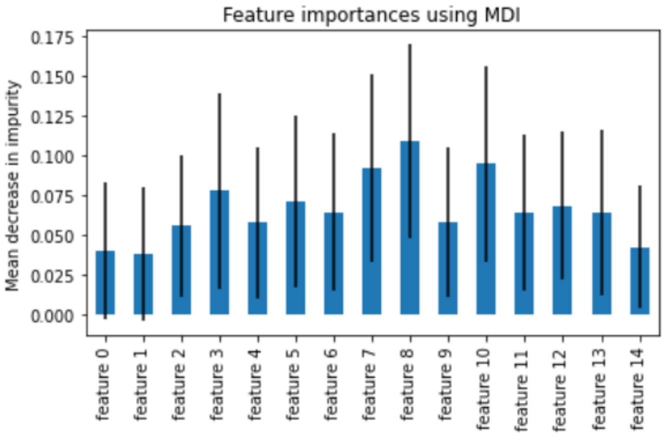


Figure 5: **Feature Importances**

Our last observation is that the most common age of people having lung cancer ranges from 50-70 based on the mean of the histogram. (see Figure 6)

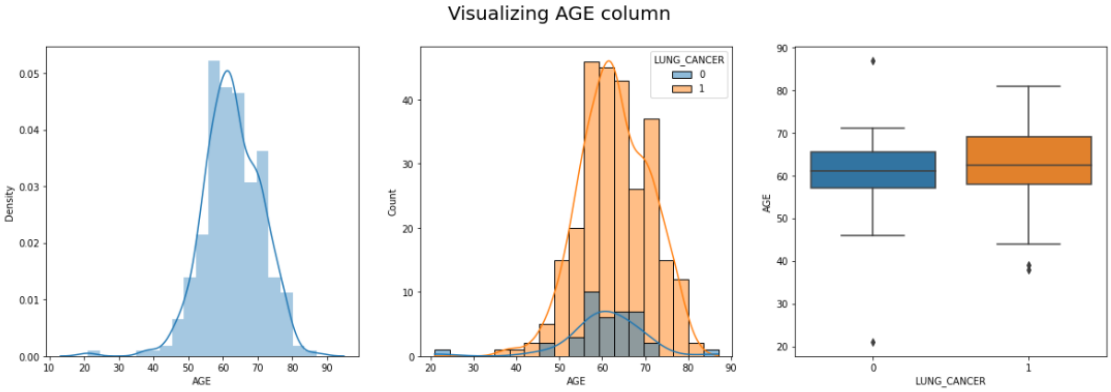
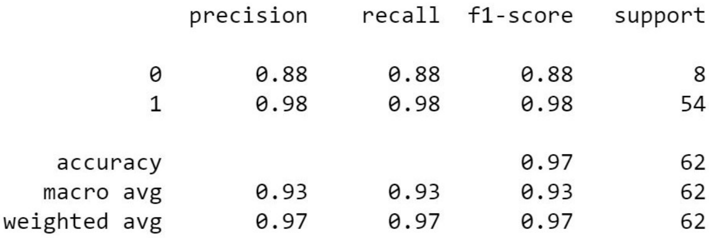


Figure 6: **Age and lung cancer**

5 Results

Here are the accuracies that each of our models reported:
SVC with Polynomial kernel: 87%
SVC with RBF kernel: 88.7%
SVC with Sigmoid kernel: 87.1%
SVC with Linear kernel with default learning rate: 93.5%
SVC with Linear kernel with 0.25 learning rate: 93.5%
KNN with 12 neighbors: 96.7%
Decision Trees: 90.3%
Random forests: 93.5%

We chose the **KNN** model with **k=12**, as it had the highest accuracy (96.7%) amongst all models that were tested. This is important because we would not to provide patients with incorrect diagnoses We tested k values from 1-24, and k=12 had the highest accuracy rate. Additionally, KNN is a good choice because it works very well with small data sets, numerical data types, and high-dimensional data sets where the distance between two points is easily calculated.



```
              precision    recall  f1-score   support

           0       0.88      0.88      0.88         8
           1       0.98      0.98      0.98        54

    accuracy                           0.97        62
   macro avg       0.93      0.93      0.93        62
weighted avg       0.97      0.97      0.97        62
```

Figure 7: **KNN Model Results with k=12**