21341053-MD FOYSAL KABIR

Lung Cancer Disease Prediction

Introduction

Lung cancer is a leading cause of cancer-related deaths worldwide, emphasizing the need for early detection and accurate prediction. We chose to explore Lung Cancer Prediction Using Machine Learning due to its potential to improve patient outcomes. By leveraging machine learning techniques, we can analyze vast amounts of medical data to develop comprehensive predictive models. Early detection enables timely intervention, personalized treatment, and ultimately, better patient outcomes. Through this research, we aim to contribute to the development of more effective tools for lung cancer diagnosis and management.

Dataset Description

Source - https://www.kaggle.com/datasets/mysarahmadbhat/lung-cancer

It is taken from kaggle website

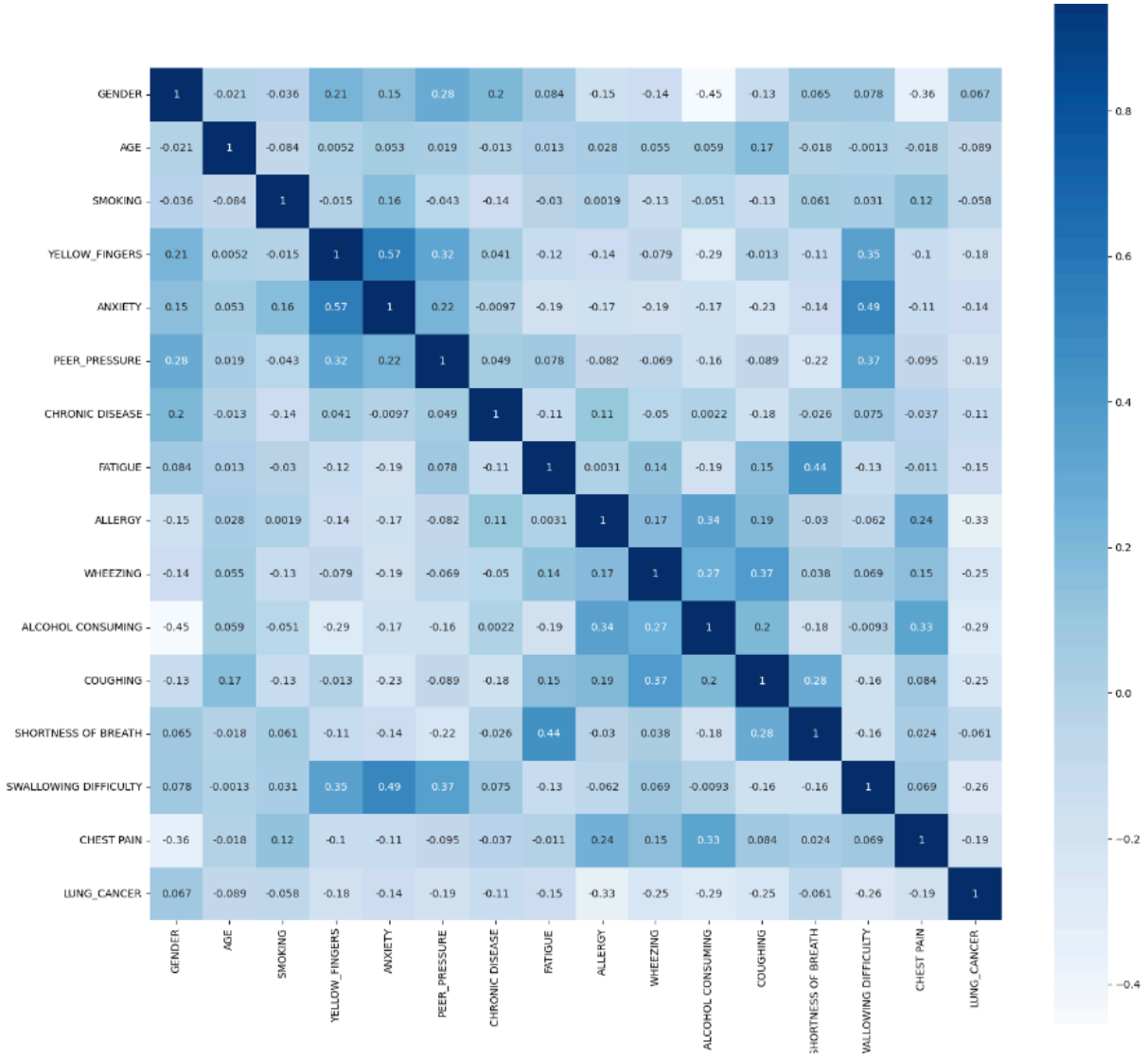
It has sixteen feature

It is a classification problem because it has categorical outputs such as Yes and NO.

There are three hundred and nine data points .

The features is in categorical type

The heatmap :



Dataset Preprocessing

Categorical Values

According to the dataset Gender feature and Lung cancer feature only have categorical value as M,F and Yes ,No .To maintain a balance and proper understanding of the dataset

we have changed the value of feature Gender (M,F) to (1,2) ,Lung Cancer feature (Yes,No) to (1,2) as well.

```
lung_data.GENDER = lung_data.GENDER.map({"M":1,"F":2})
lung_data.LUNG_CANCER = lung_data.LUNG_CANCER.map({"YES":1,"NO":2})
```

NULL Values

```
lung_data.isnull().sum()
```

```
GENDER                    0
AGE                       0
SMOKING                   0
YELLOW_FINGERS            0
ANXIETY                   0
PEER_PRESSURE             0
CHRONIC DISEASE           0
FATIGUE                   0
ALLERGY                   0
WHEEZING                  0
ALCOHOL CONSUMING         0
COUGHING                  0
SHORTNESS OF BREATH       0
SWALLOWING DIFFICULTY     0
CHEST PAIN                0
LUNG_CANCER               0
dtype: int64
```

There were 0 null values to work with .

Data splitting

```
#Splitting the Dataset: Training and Testing
from sklearn.model_selection import train_test_split
x_train,x_test,y_train,y_test = train_test_split(x,y,test_size=1/3,random_state=0)
```

Model training and testing: *Logistic regression* is the process of modeling probabilities of a specific outcome given input variables. The most common logistic regression models a binary outcome that can take two values such as healthy/not healthy, yes/no, true/false, and so on. Logistic regression is used to predict the categorical dependent variable. It's used when the prediction is categorical, for example, yes or no, true or false, 0 or 1.For instance, insurance companies decide whether or not to approve a new policy based on a driver's history, credit history and other such factors.

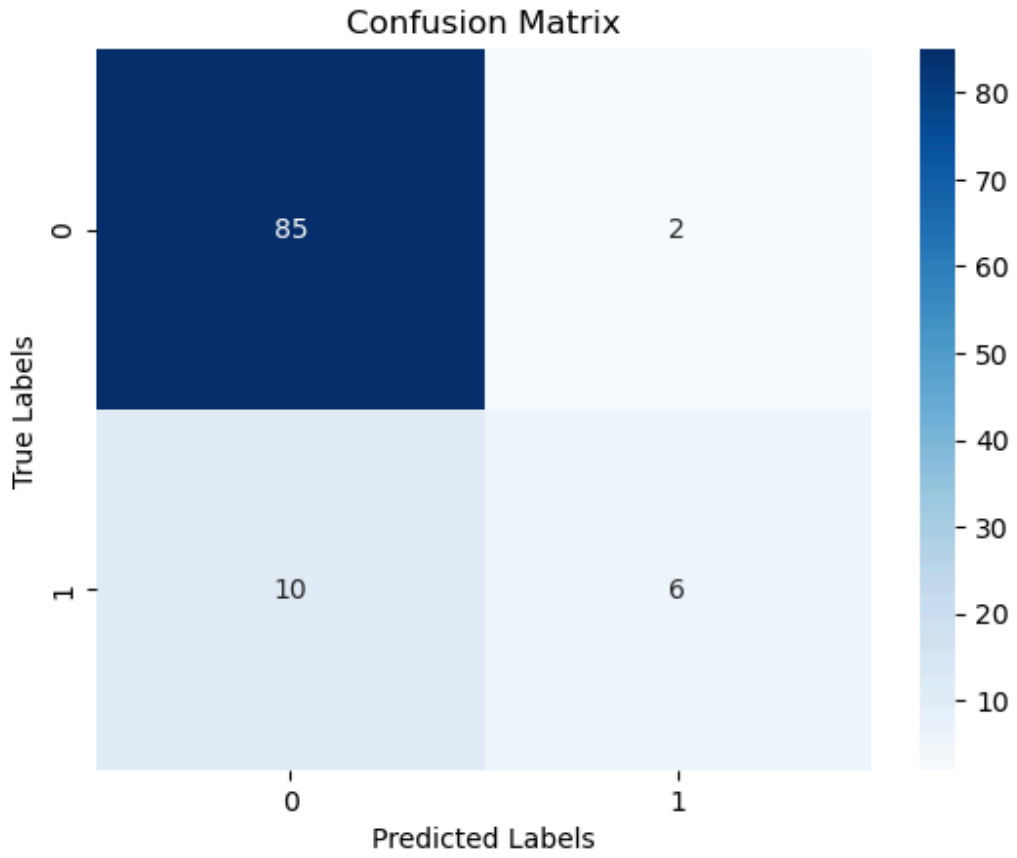
Knn classifies the new data points based on the similarity measure of the earlier stored data points. For example, if we have a dataset of tomatoes and bananas. KNN will store similar measures like shape and color. When a new object comes it will check its similarity with the color (red or yellow) and shape.

Svm support vector machine is also very fast for both categorical and quantitative datasets.

Comparison Analysis

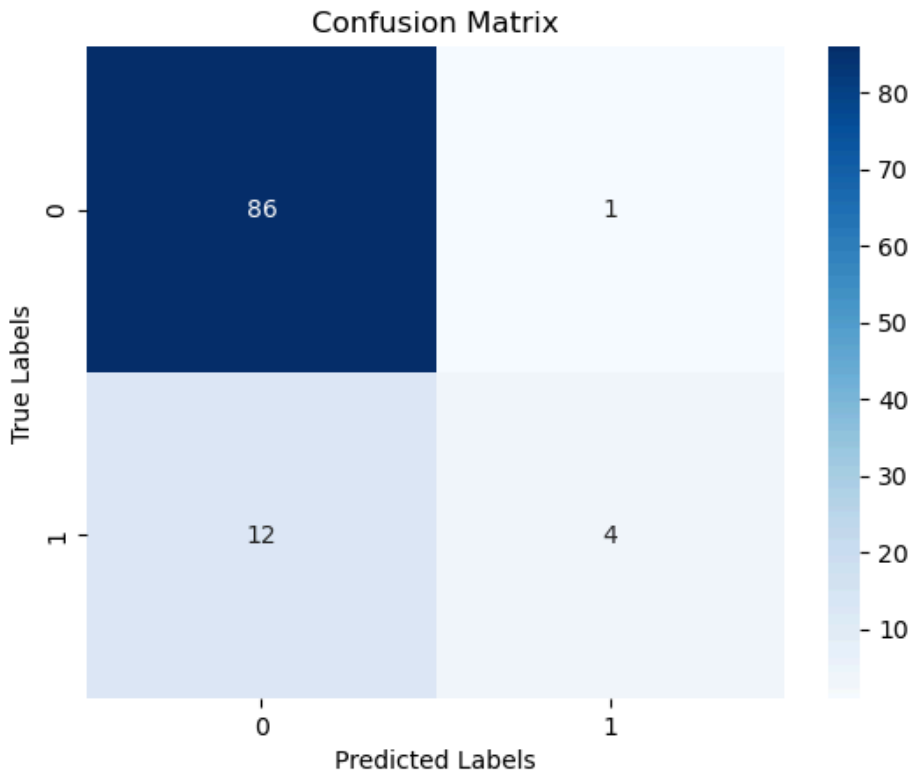
For Logistic regression we get

```
Accuracy: 0.883495145631068
Precision: 0.8947368421052632
Recall: 0.9770114942528736
F1 score: 0.9340659340659342
```

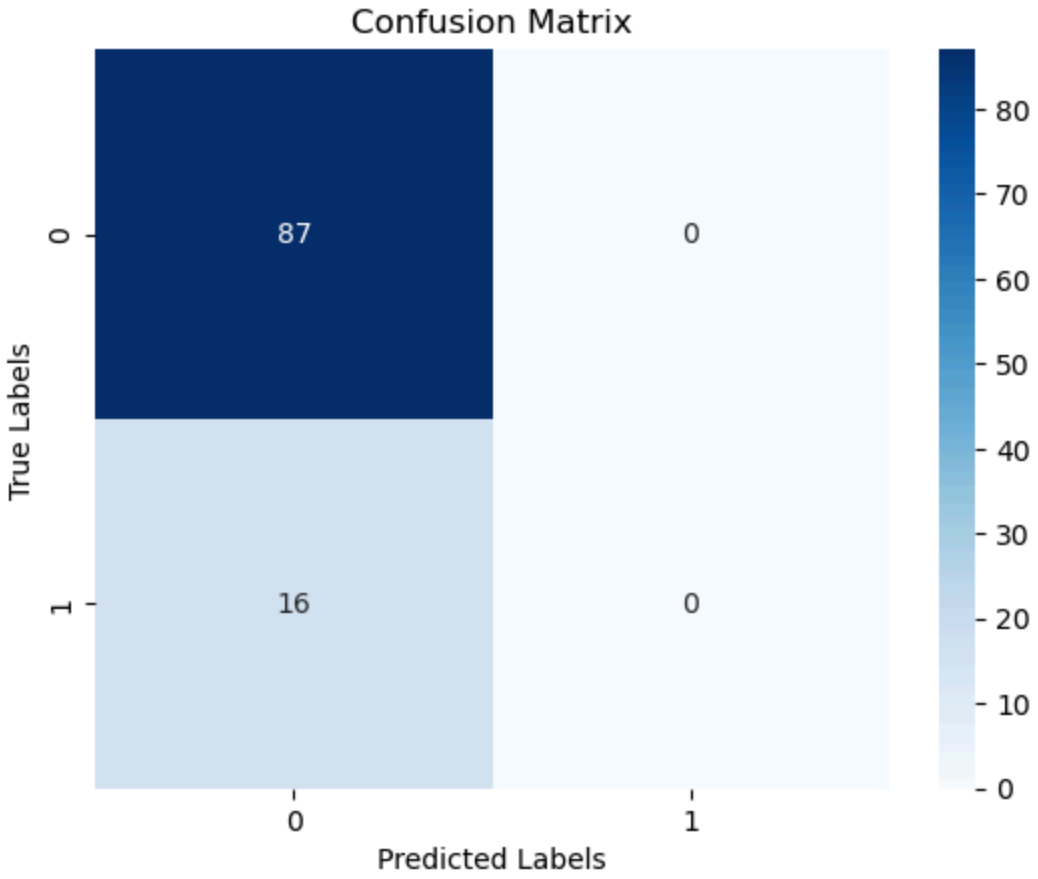


For Knn we get

```
Accuracy: 0.8737864077669902
Precision: 0.8775510204081632
Recall: 0.9885057471264368
F1 score: 0.9297297297297297
```



For Svm we get ,




```
Accuracy: 0.8446601941747572
Precision: 0.8446601941747572
Recall: 1.0
F1 score: 0.9157894736842105
```

The results are quite ok but we could have done better .

Conclusion

The model is working quite well but not as efficiently as it should be , to make it more efficient the data pre processing part should have more focus rather than applying

machine learning models .Especially on the dependent variables . Overall the precision and accuracy is up to the mark and definitely could have done better . Final result of this project which has three machine learning models can predict lung cancer and in the future with proper nurture it can precisely predict cancer for better future prediction and awareness to the society. Further improvement must be needed to work better in this project where data preprocessing will be more efficient .