**Project Team:**

Sanjana Nalluru

Sunanda Chavalamudi

Anusha Mandepudi

**Introduction**

Lung cancer is a major health concern. Detecting it earlier and giving the treatment the treatment to the person if he/she is suffering from lung cancer is very much important. The project focuses on the prediction of lung cancer using the machine learning techniques K-Nearest Neighbors (KNN) and Logistic Regression Classifiers.

The main motivation behind my project is saving the lives of people by giving advance treatment to them. By earlier detection of the disease. By using the popular machine learning models, we can improve the efficiency and accuracy of lung cancer prediction.

In this project, we are going to take a dataset which is related to the information about the people. And depending on that we need to classify whether the person is suffering from lung cancer or not. Using K-Nearest Neighbors (KNN) and Logistic Regression Classifiers we can find the results that will be useful for earlier detection of lung cancer.

**Related Work**

There are several other studies that are used for lung cancer prediction. There are various techniques that they have used in their work. similar work is done, but they have used different machine learning algorithms.

**References:**

Krishnaiah et al. (1) have examined the use of data mining techniques like Rule-based, Decision tree, Naïve Bayes, and ANN on healthcare data. They developed a prototype lung cancer disease prediction system using various data mining classification techniques. Among the various techniques applied, the Naïve Bayes model was the most effective model.

Patil et al. (2) proposed a new technique for building a prediction model using a decision tree algorithm. The model aimed to enhance lung cancer's performance with better speed and accuracy.

Mehedi Masud et al. (3) have used a machine learning approach for diagnosing Lung and Colon Cancer using a Deep Learning based Classification Framework. They have used CNN on the data and made categories on them.

**Methods**

**Collection of data**

We intend to use a pre-existing dataset that is publicly available on the internet. After considering various factors like relevance and quality, we have chosen to use a pre-existing dataset from Kaggle. Kaggle is a very trusted and popular platform that provides a wide range of datasets that can be accessed and downloaded.

The data mining pipeline involved several steps like exploration of data, cleaning and preprocessing the data, feature selection, training the models using KNN and Logistic Regression techniques, and finally evaluating the models by finding its accuracy.

**Data mining Pipeline**

**Importing libraries**

Libraries used are NumPy, Pandas, Matplotlib, and Seaborn. Where NumPy is used for all the numerical operations on the dataset. Pandas is used for handling the data. Matplotlib is used for plotting the dataset used for visualization. And Seaborn is used for statistical visualization of the data.

**Importing the dataset**

Next step is reading the dataset. The input given is a CSV file, which has the independent variables denoted with X, and dependent variable is denoted with Y. the iloc() function in the pandas library is used to take the data from the dataset. The index location can be given using the row and column numbers. The last column of the dataset I have taken is dependent variable. The dependent variable is to classify whether the person is suffering from Lung cancer or not. The Lung cancer dataset is imported and the path is copied to read the csv file. The independent and dependent variables are set to the respective X and Y variables. The head() method is used to display the first few roads of the dataset. The dataset displayed the first 5 records from the dataset. The head() method is used from the pandas library.

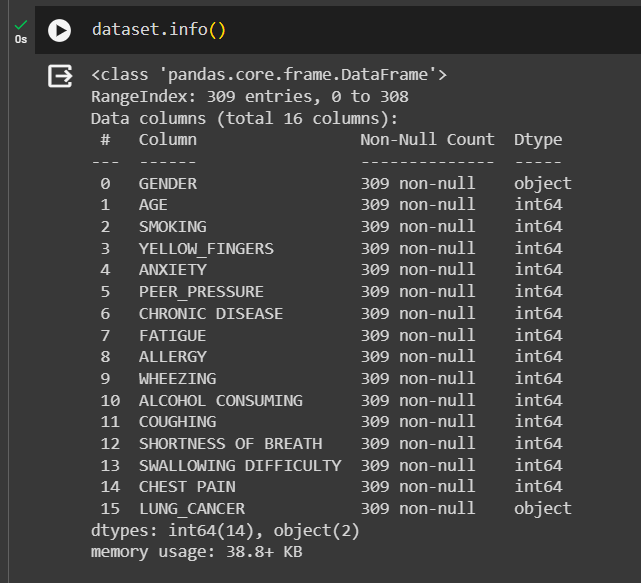


**Data Preprocessing**

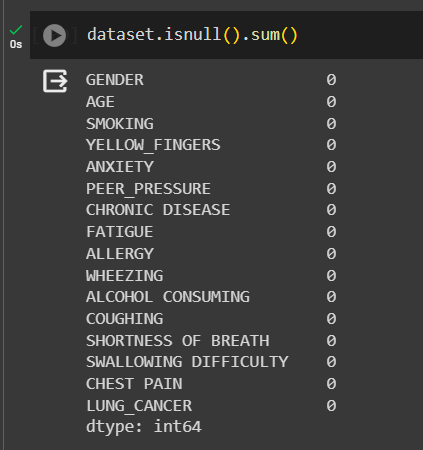
**Handling the missing data**

I have used the info() method to check for the datatypes of all the attributes. The information of the dataset can be known by this.

Then I have inspected the dataset, using the head() method I got some idea about the CSV file, regarding what are the attributes that are present in the dataset. Then I have checked for the number of null values for all the values for each of the attributes present in the dataset. For this I have used .isnull().sum() method.



Here in the given picture,it gives a clear view if there are any missing values in the records.



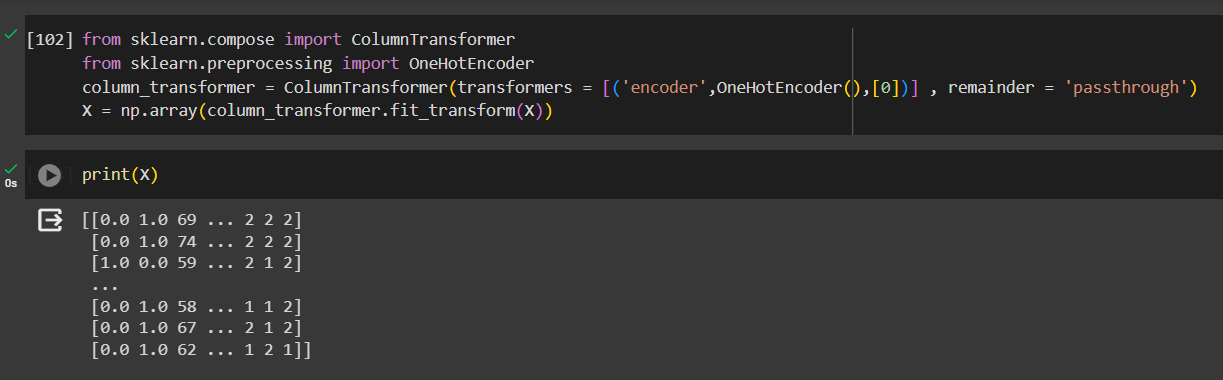
If there are any missing value sin the dataset it can be replaced with the mean of all the values.

After checking that I made a conclusion, there are no null values for all the values for the attributes present in the dataset.

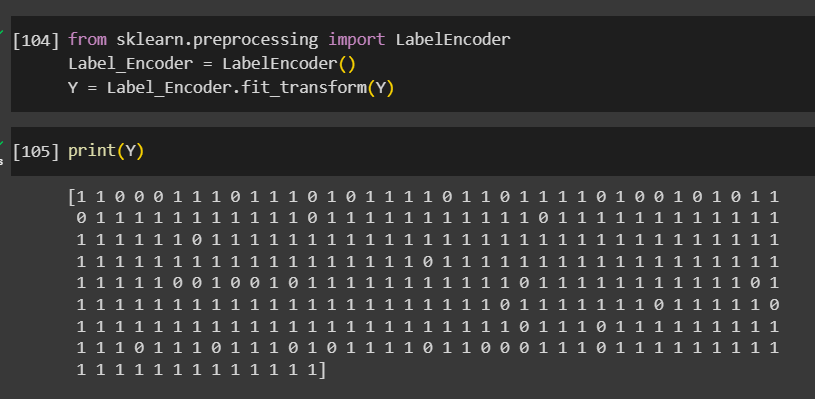
**Encoding the categorical data**

There are two attributes AGE and LUNG\_CANCER that are categorical data. Therefore, I have encoded the categorical data to convert it into numerical data. So that the model can establish the correlation between the independent variables and the dependent variable. It is possible if the data present for all the attributes is numerical data.

From the scikit library I have used the two classes ColumnTransformer and OneHotEncoder for encoding the categorical data of the independent variables.



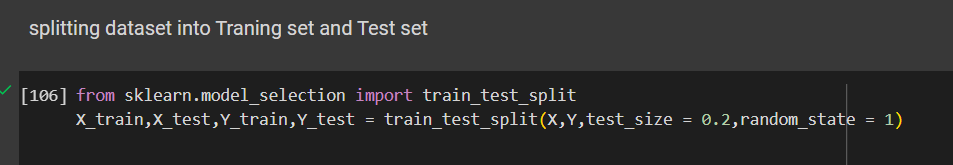
After that LabelEncoder is used for encoding the categorical data of the dependent variable into the numerical data.



Then the fit\_transform method is used to fit the scalar data and transform the data into fitted parameters.

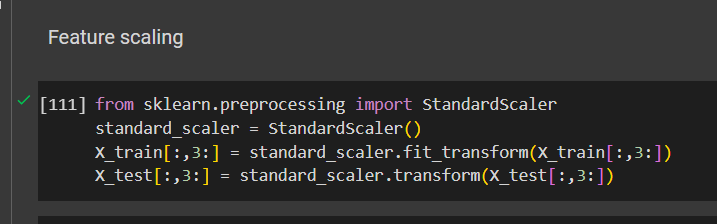
**Splitting dataset into Training set and Test Set**

From the model\_selection module of the sklearn library. The train\_test\_split class I used for splitting the dataset into training set and test set. The test\_size is set to 0.2. that is 80 percent of the dataset is taken as training set and 20 percent of the dataset is taken as test set. I have printed the train and test values of the independent and dependent variables. Therefore, we can see that the dataset is encoded and has been split into training set and test set.



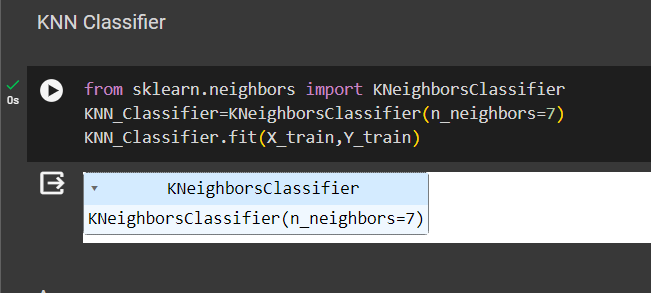
**Feature scaling**

I have used the StandardScaler class from the sklearn library for the process of standardization of all the values for the attributes in the training set and test set. fit\_transform is used to combine the fit and transform methods to single operation. This can be used for training the data as well and standardizing the values.



**K-Nearest Neighbours**

The KNN algorithm is used for the purpose of classification and in this model preparation. It memorizes the instances during its training. And it that knowledge it will classify and predict the output. The n\_neighbors is used to determine the number of neighbours that we are going to consider during the training. In this model I have set the number of n\_neighbors value to 7. The class label can be established by knowing the majority class among the 7 nearest neighbors.



If a new data point comes into testing of which class it belongs, it is put into the nearest class which can be calculated using the Euclidean distance.

I have imported the KNeighborsClassifier class from the neighbors module which is present in the sklearn library. The model is trained with 7 neighbors. I have used the fit method to train the data.

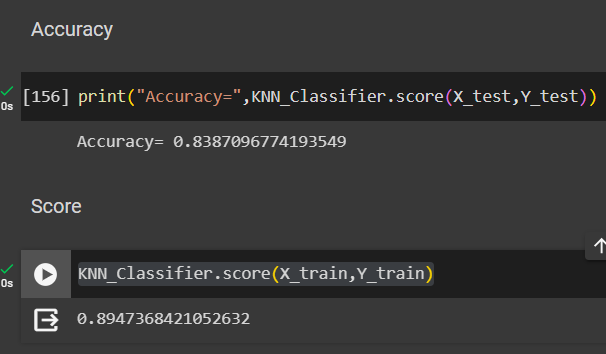
The KNN algorithm will learn the correlation among the variables and gain some knowledge from the training data.

**Model Evaluation**

Once, the model is developed, it is evaluated. We will evaluate the model by primarily focusing on its accuracy. The dataset will be divided into two sets: training data and testing data. Testing data will be used to evaluate the model. If the accuracy of the model is 90 or above, this means that the model is stated as suitable for this topic. If possible, we will use an independent testing dataset, which will not be a part of the original data collection to test the performance of the model.

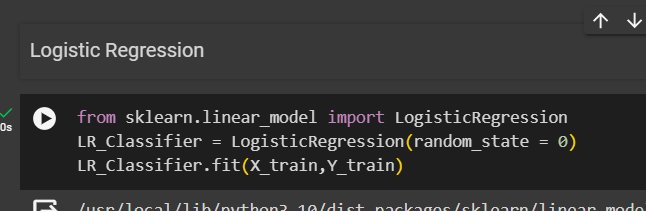
The accuracy of the model is also printed. It will give the overall correctness of the model.

For evaluating the accuracy of the training data I have used the score() method which takes X\_train and Y\_train as its parameters.

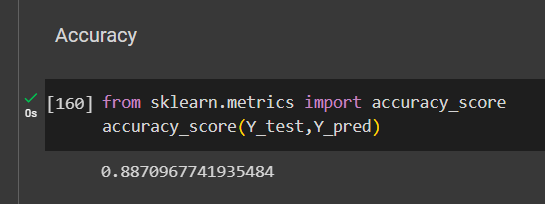


**Logistic Regression**

I have used logistic regression for binary classification whether the person is suffering from lung cancer or not. It is a very popular statistical method used for classification problems. It uses the logistic function to determine the test result between 0 and 1. Therefore, it is used for probabilistic classification. Where the decision is taken whether the point belongs to which class among the two. The decision boundary is set at 0.5.



The predict method is used to make the predictions for the test data. And finally, the accuracy of the model is also calculated.



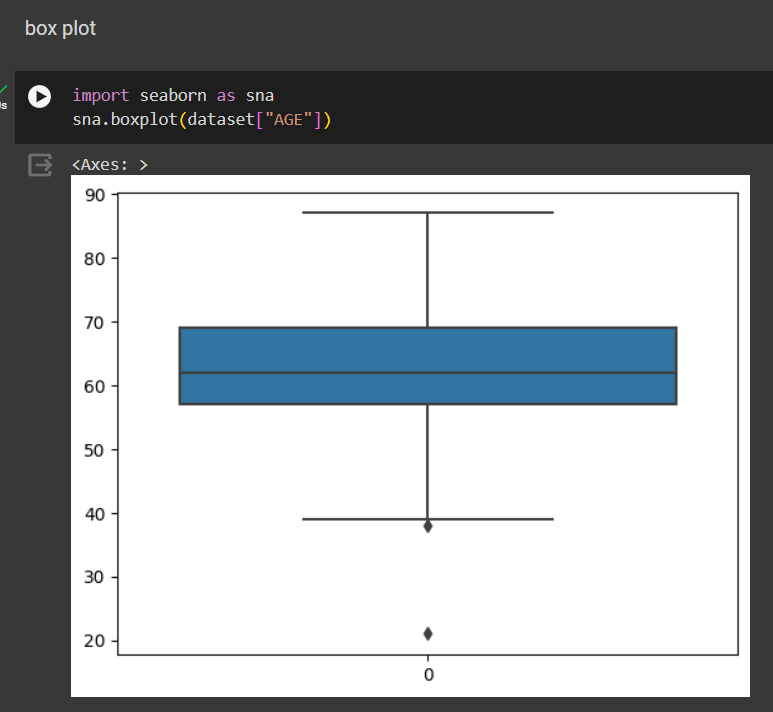
**Software Used**

All the work related to the project is done in Google colab. I have used python programming language in my project. I have used scikit-learn library for using all the classes that are required for splitting the data and generating the models. And seaborn and matplotlib libraries have been used for data visualization.

**Results and Discussion**

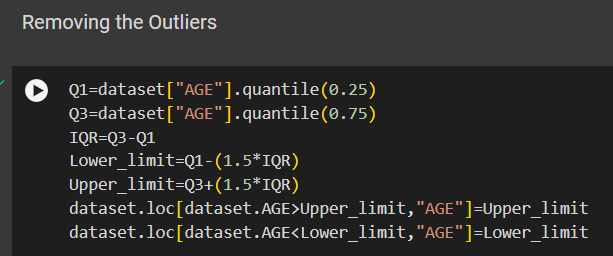
**Box plot**

Here the seaborn library is used for statistical visualization of the data. The box plot is used for the process of visualization on the “AGE” attribute. We can see some outliers in the box plot. In the next step we can remove that.

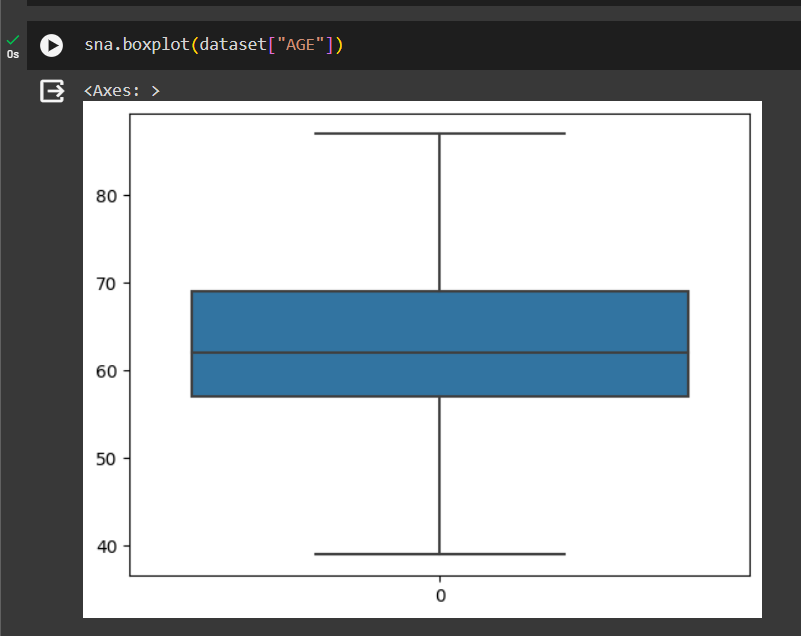


**Removing the Outliers**

we can remove the outliers by calculating the quartiles and Inter Quartile range (IQR). The values outside of the Lower and Upper bound are replaced with the Lower and Upper limits.



Finally, after removing the outliers the box plot is again.



**Conclusion**

By this project, I have learned a lot of things regarding machine learning. Our study gave us a detailed note on how the machine learning algorithms like KNN and Logistic Regression will work on predicting the lung cancer.

Though we have not included all other factors that may affect these predictions. Also, our model is not very accurate about the results of classifying the categories. The prediction is done only on limited data that is available from the datasets. In real world scenarios, it will be different as there will huge amount of data.

In future, I will work more on other machine learning models and compare the accuracy based on the evaluation performed on various problems and select the best model that can be suitable for the problem.

**Data and Software Availability**

The dataset is taken from Kaggle, which is available at:

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

GitHub link for the project:

**References**

[1] Rehman, A. U., Saleem, R. M., Shafi, Z., Imran, M., Pradhan, M., & Alzoubi, H. M. (2022, February). Analysis of Income on the Basis of Occupation using Data Mining. In 2022 International Conference on Business Analytics for Technology and Security (ICBATS) (pp. 1-4). IEEE.

[2] Bekena, S. M. (2017). Using decision tree classifier to predict income levels.

[3] N. Chakrabarty and S. Biswas, "A Statistical Approach to Adult Census Income Level Prediction," 2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN), Greater Noida, India, 2018, pp. 207-212, doi: 10.1109/ICACCCN.2018.8748528.