LUNG CANCER PREDICTION

LIBRARIES USED:-

1. Numpy :- used to perform mathematical operations on the data.
2. Pandas :-used to load and make changes in the dataset.
3. Matplotlib.pyplot :- used to make plots, graphs and represent the data visually.
4. Seaborn :- same as mathplotlib.pyplot except the fact that it is much better at visually representing the data.

A black and white chart

AI-generated content may be incorrect.

1. Warning:- used to import warnings and can also be used to filter the warnings.

DATA ACQUISITION:-

Data obtained from Kaggle.

DATA CLEANUP AND PREPROCESSING:-

Dropping the duplicates. Using labelencoding to convert columns from object to integer type and changing values to 0&1 notation for easier classification and understanding.

Dropping columns without much relation to the target column.

FEATURE SELECTION:-

Finding the correlation between the various features and making a new feature using them.

Then splitting the independent and dependent variables.

TARGET DISTRIBUTION AND SAMPLE IMBALANCE HANDLING:-

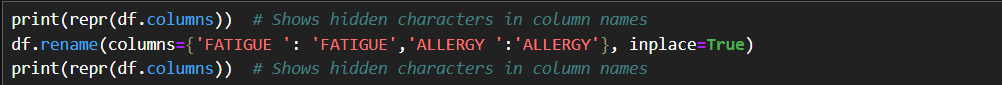
We use adaptive synthetic sampling. **ADASYN (Adaptive Synthetic Sampling) technique** o handle **imbalanced data** by generating synthetic samples for the minority class.

MODEL SELECTION:-

We make use of 10 different models and test them on the basis of their accuracy.

THINGS LEARNED :-

1. If we perform a method on the dataset while assigning the method execution to the dataset using an equal to we change the dataset into a method.
2. Column names in a csv file might contain special characters like a blank space so always check for it



1. We can use tab inside parenthesis for drop down suggestion box.
2. Although setting random state to zero is not mandatory for a p[rediction model but it is preferred because

A screenshot of a computer

AI-generated content may be incorrect.

1. If module isn’t found it can be installed using the [!pip install (module name)] command.

A screenshot of a black screen

AI-generated content may be incorrect.