**Final Project Progress**

**Project Team:**

Sanjana Nalluru

Sunanda Chavalamudi

Anusha Mandepudi

**Completed Tasks:**

**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.

**Handling the missing data**

Then we have inspected the dataset, using the head() method we got some idea about the CSV file, regarding what are the attributes that are present in the dataset. Then we have checked for the number of null values for all the values for each of the attributes present in the dataset. For this, we have used .isnull().sum() method. If there are any missing values in the dataset it can be replaced with the mean of all the values.

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

**Encoding the categorical data**

We have used the info() method to check for the datatypes of all the attributes. There are two attributes AGE and LUNG\_CANCER that are categorical data. Therefore, We 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 we 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.

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

From the model\_selection module of the sklearn library. The train\_test\_split class we 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. We 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**

We 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.

**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 visualized where we can see the removal of the outliers.

**Data:**

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

**Challenges**

There are a few difficulties which we have faced in the project. One of them will be checking which columns will be suitable for my model. And then doing the feature extraction. After that, we faced a small problem in checking and removing the outliers. We did it by plotting a boxplot.

**Collaboration**

We have been communicating with each other through Discord and have been meeting in person whenever it is possible. All the work has been split among ourselves. One person has been focused on acquiring and pre-processing the data and the other two will focus on building the prediction model.

**Next steps**

The data will be further visualized by appropriate visualization. Then the data will used to build the model and will be trained. The testing data will be used to build the model, which will be evaluated then. We are planning to gain an accuracy of 90 or above.