FACULTY OF COMPUTING

SECB3203-01

PROGRAMMING FOR BIOINFORMATICS

TITLE:

Prediction of Lung Cancer using Recursive Feature Elimination

LECTURER:

DR. NIES HUI WEN

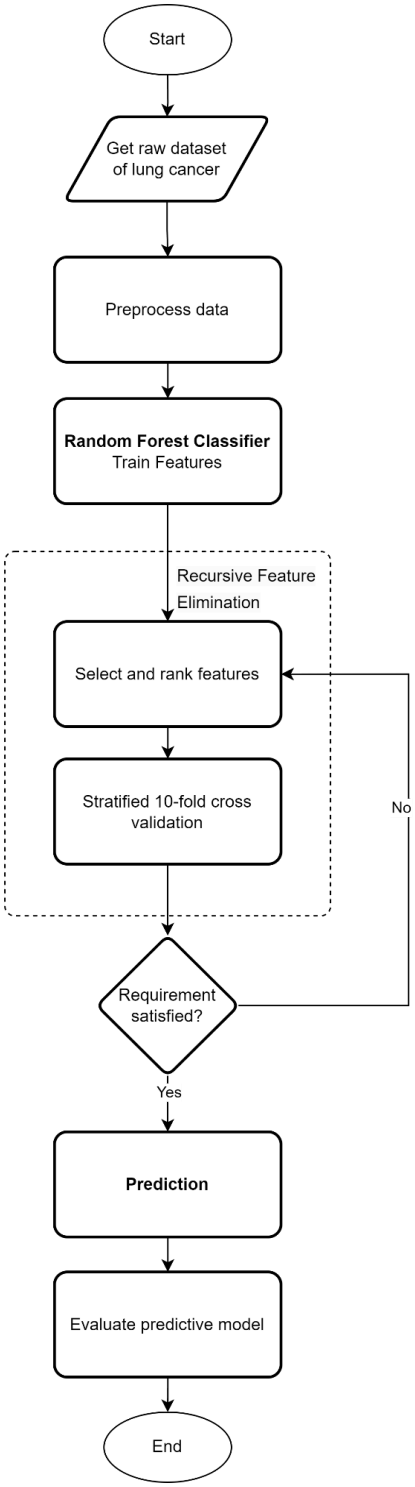
GROUP 4

GROUP MEMBERS:

NAME	MATRIC NUMBER
LEE RONG XIAN	A21EC0043
LU QI YAN	A21EC0049

Discussion

Throughout the project, we are following closely the flowchart of our project as a guide. The flowchart of the proposed approach is shown below.



1.0 Get raw dataset of the lung cancer

The dataset we use in this project can be found at https://www.kaggle.com/datasets/mysarahmadbhat/lung-cancer.

The raw dataset consists of 309 rows and 16 columns. The 16 features are as below:

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

2.0 Data Preprocessing

The first step in data processing is to handle the duplicated data. A total of 33 duplicated data in the dataset are identified and dropped. This is because the duplicates are taking up unnecessary storage space and may lead to inconsistencies during the model training. By removing the duplicated data, we are able to ensure a higher accuracy of the data analysis result.

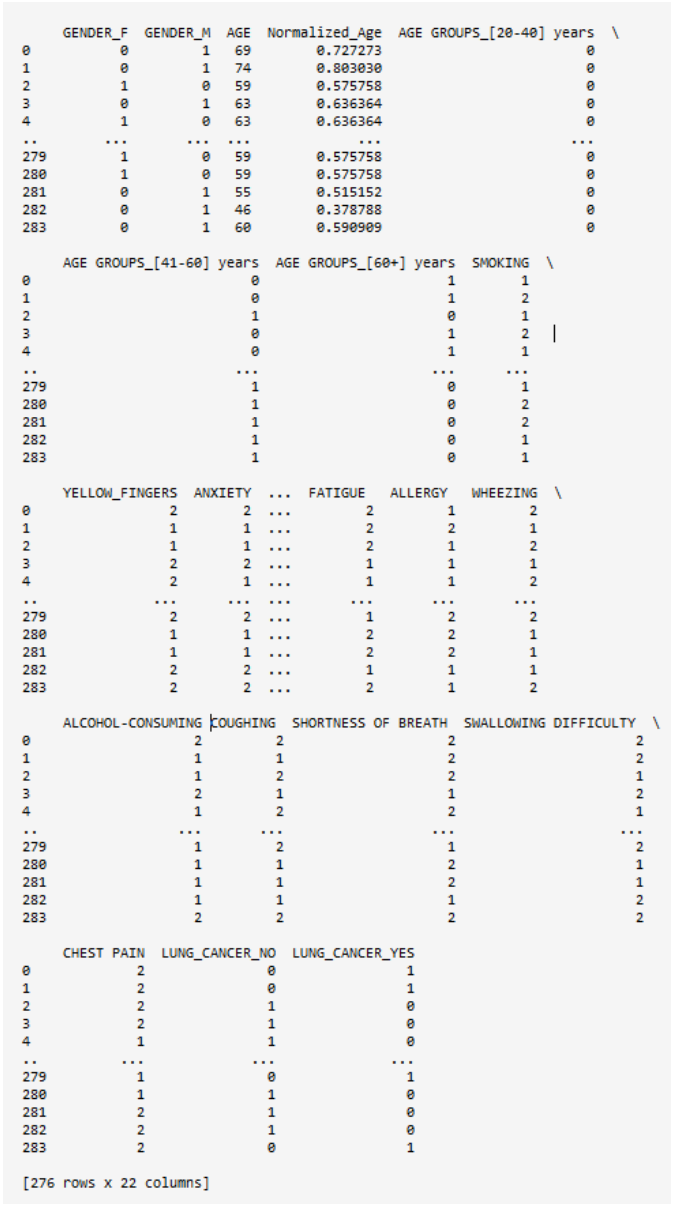
```
Total duplicate values:  33
    GENDER  AGE  SMOKING  YELLOW_FINGERS  ANXIETY  PEER_PRESSURE  \
0        M   69        1               2        2              1
1        M   74        2               1        1              1
2        F   59        1               1        1              2
3        M   63        2               2        2              1
4        F   63        1               2        1              1
..     ...  ...      ...             ...      ...            ...
279      F   59        1               2        2              2
280      F   59        2               1        1              1
281      M   55        2               1        1              1
282      M   46        1               2        2              1
283      M   60        1               2        2              1
```

The second step is to handle the missing values as the null values would result in data inconsistencies. However, there are no null values in this dataset. Next, we do data normalization on the “Age” feature. The method used is Min-Max scaling which transforms the data to a range

between 0 and 1(change integer to floating point datatype), making it easier to interpret the relative importance of values.

```
0      0.727273
1      0.803030
2      0.575758
3      0.636364
4      0.636364
         ...
279    0.575758
280    0.575758
281    0.515152
282    0.378788
283    0.590909
Name: Normalized_Age, Length: 276, dtype: float64
```

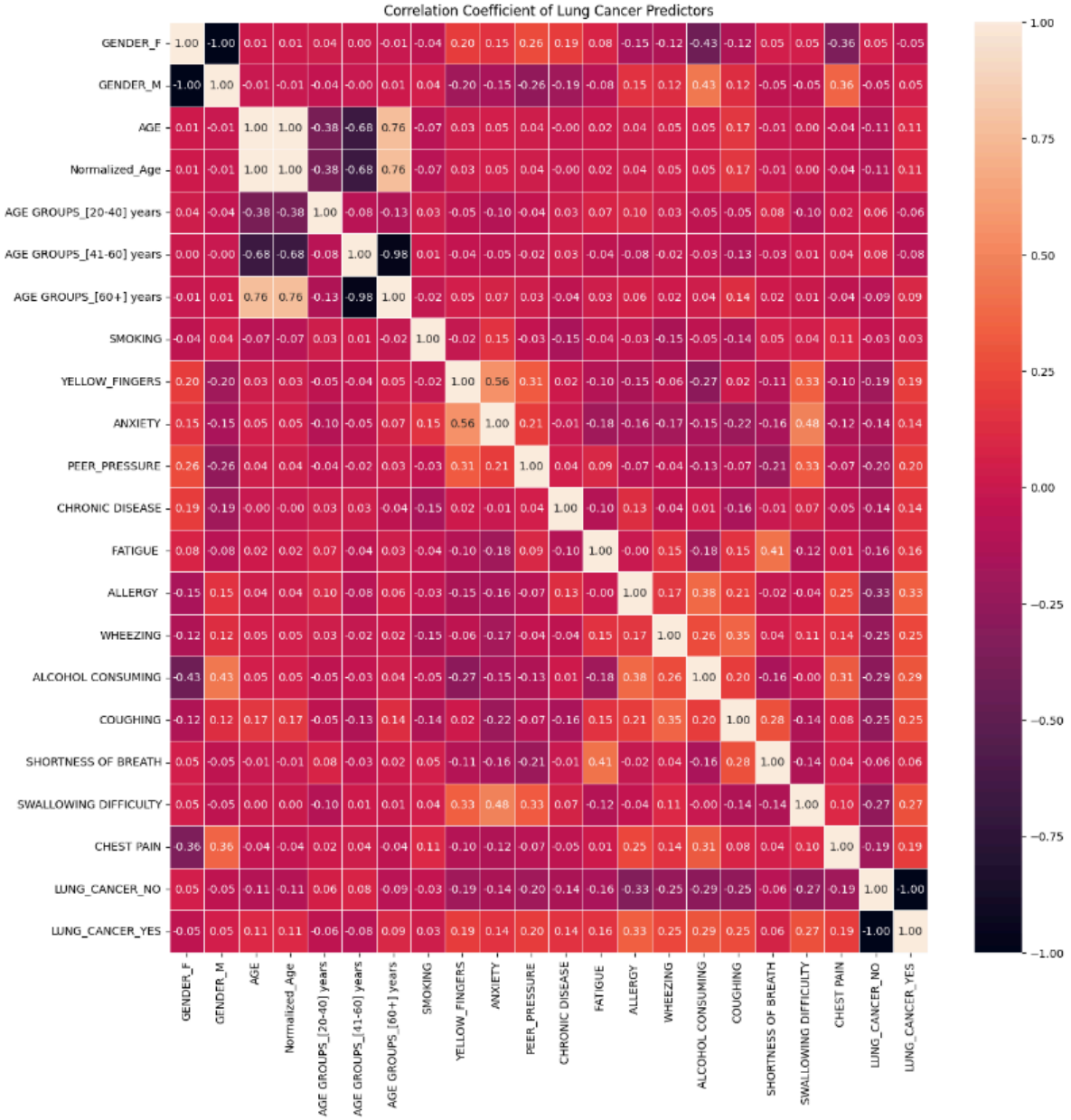
Then, we do data binning where we group the “age” feature into 3 categories which are “20-40 years old”, “41-60 years old” and “above 60 years old”. It is shown that the continuous data value which is “age” is grouped into discrete intervals. This is to reduce the complexity of data and make it more manageable and easier to analyze. Lastly, we create the indicator variables for our dataset whereby we convert the categorial data into numerical values. The indicator variable is set for 3 attributes namely, Gender, Age Groups, and Lung Cancer. As we can visualize in the result below, the boolean values (True and False) are converted into integer values (1 and 0) for the 3 features: Gender, Age Groups, and Lung Cancer. Besides, it is shown that the continuous numerical data which is “Age” and “Normalized Age” remain the same. Setting indicator variables is to ease the data analysis process and to ensure a clearer representation of the dataset.



```
     GENDER_F  GENDER_M  AGE  Normalized_Age  AGE GROUPS_[20-40] years  \
0           0         1   69        0.727273                         0
1           0         1   74        0.803030                         0
2           1         0   59        0.575758                         0
3           0         1   63        0.636364                         0
4           1         0   63        0.636364                         0
..        ...       ...  ...             ...                       ...
279         1         0   59        0.575758                         0
280         1         0   59        0.575758                         0
281         0         1   55        0.515152                         0
282         0         1   46        0.378788                         0
283         0         1   60        0.590909                         0

     AGE GROUPS_[41-60] years  AGE GROUPS_[60+] years  SMOKING  \
0                           0                       1        1
1                           0                       1        2
2                           1                       0        1
3                           0                       1        2
4                           0                       1        1
..                        ...                     ...      ...
279                         1                       0        1
280                         1                       0        2
281                         1                       0        2
282                         1                       0        1
283                         1                       0        1

     YELLOW_FINGERS  ANXIETY  ...  FATIGUE   ALLERGY   WHEEZING  \
0                 2        2  ...         2         1         2
1                 1        1  ...         2         2         1
2                 1        1  ...         2         1         2
3                 2        2  ...         1         1         1
4                 2        1  ...         1         1         2
..              ...      ...  ...       ...       ...       ...
279               2        2  ...         1         2         2
280               1        1  ...         2         2         1
281               1        1  ...         2         2         1
282               2        2  ...         1         1         1
283               2        2  ...         2         1         2

     ALCOHOL-CONSUMING  COUGHING  SHORTNESS OF BREATH  SWALLOWING DIFFICULTY  \
0                    2         2                    2                      2
1                    1         1                    2                      2
2                    1         2                    2                      1
3                    2         1                    1                      2
4                    1         2                    2                      1
..                 ...       ...                  ...                    ...
279                  1         2                    1                      2
280                  1         1                    2                      1
281                  1         1                    2                      1
282                  1         1                    1                      2
283                  2         2                    2                      2

     CHEST PAIN  LUNG_CANCER_NO  LUNG_CANCER_YES
0             2               0                1
1             2               0                1
2             2               1                0
3             2               1                0
4             1               1                0
..          ...             ...              ...
279           1               0                1
280           1               1                0
281           2               1                0
282           2               1                0
283           2               0                1

[276 rows x 22 columns]
```

In a nutshell, we have obtained 276 rows and 22 columns after data preprocessing.

Next, we produced a heatmap heatmap to show the predictors' correlation coefficient. This is for us to visualize the strength of associations between data variables in a way clearer.



3.0 Model Development

Firstly, we identify the features that are not required for future analysis. These features include AGE, NORMALIZED_AGE, and LUNG_CANCER. Hence, these columns are dropped. This is because AGE, NORMALIZED_AGE and AGE GROUPS represent the same information, keeping them might cause the data to be redundant whereas LUNG_CANCER is not a relevant

predictor or we can say the target variable for the analysis. Then, we are using the 15 features below to train the model.

```
Index(['GENDER', 'AGE GROUPS', 'SMOKING', 'YELLOW_FINGERS', 'ANXIETY',
       'PEER_PRESSURE', 'CHRONIC DISEASE', 'FATIGUE ', 'ALLERGY ', 'WHEEZING',
       'ALCOHOL CONSUMING', 'COUGHING', 'SHORTNESS OF BREATH',
       'SWALLOWING DIFFICULTY', 'CHEST PAIN'],
      dtype='object')
```

In this case, we split the dataset into a train set (70%) and a test set (30%) because this contributes to the highest accuracy of our predictive model.

3.1 Select and rank features

We use RandomForestClassifier as the estimator to carry out the recursive feature elimination. The number of features to select is set to 10 because it provides enough data for training and testing and it also may result in a reliable performance in our model. The eliminated features are represented by false whereas 10 relevant features are represented by true.

	columns	Kept
0	GENDER	False
1	AGE GROUPS	True
2	SMOKING	False
3	YELLOW_FINGERS	True
4	ANXIETY	True
5	PEER_PRESSURE	True
6	CHRONIC DISEASE	True
7	FATIGUE	True
8	ALLERGY	True
9	WHEEZING	False
10	ALCOHOL CONSUMING	True
11	COUGHING	False
12	SHORTNESS OF BREATH	True
13	SWALLOWING DIFFICULTY	False
14	CHEST PAIN	True

Then, the features are ranked. This helps us to identify the most relevant features that cause lung cancer.

	columns	Kept
0	GENDER	3
1	AGE GROUPS	1
2	SMOKING	5
3	YELLOW_FINGERS	1
4	ANXIETY	1
5	PEER_PRESSURE	1
6	CHRONIC DISEASE	1
7	FATIGUE	1
8	ALLERGY	1
9	WHEEZING	2
10	ALCOHOL CONSUMING	1
11	COUGHING	6
12	SHORTNESS OF BREATH	1
13	SWALLOWING DIFFICULTY	4
14	CHEST PAIN	1

3.2 Stratified 10-fold cross-validation

Then, we evaluate the model performance by using the Stratified 10-fold cross-validation. The evaluation result is as below.

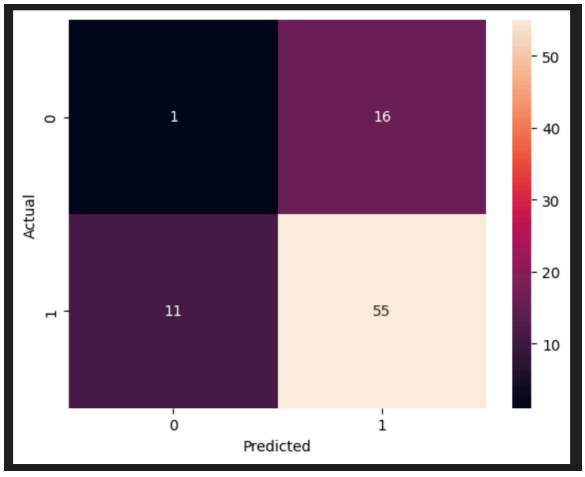
Metrics	**Score**
Mean score	0.89
Accuracy	0.94
Precision	0.96
Recall	0.97

In this project, the reason that a train set (70%) and a test set (30%) are used to train the model is that they contribute to the highest accuracy of our predictive model. Therefore, it is proven in the table below:

Train set	Test set	Accuracy
0.10	0.90	0.89
0.20	0.80	0.84
0.30	0.70	0.89
0.40	0.60	0.90
0.50	0.50	0.91
0.60	0.40	0.93
0.70	0.30	0.94
0.80	0.20	0.91
0.90	0.10	0.89

4.0 Predictive model evaluation

In this project, a confusion matrix is applied to evaluate the predictive model. It can help to show how many predictions are correct and how many are incorrect. The result is shown in the figure below. From the figure, we can see that there is 1 in the True Negative region, 16 in the False Positive region, 11 in the False Negative region, and 55 in the True Positive region.



The scores of the key metrics of the confusion matrix are shown below.

Metrics	Score
Accuracy	0.67
Precision	0.77
Recall	0.83

To wrap up the discussion, it is proven that the project successfully achieved its objectives by developing an accurate lung cancer prediction model, applying Recursive Feature Elimination (RFE) to systematically eliminate irrelevant features, and demonstrating the effectiveness of RFE in selecting the most relevant features for enhanced predictive performance.

5.0 Reflection

This project requires us to understand the dataset that we have chosen and to understand the algorithm that we have chosen to do. From this, we learned how to find suitable datasets online. At the same time, it is important to find a related thesis or scholarly paper that can help us to understand the algorithms. By doing this, we can fully understand the algorithms that we choose which is the algorithm for Recursive Features Elimination. After understanding and referring to the sources that we found, we learned to modify the algorithm based on our requirements. In this project, what we do is play around with the parameters in the algorithm so that we can get the highest accuracy with the specific value of the parameters. In summary, this research has given us practical experience and insightful knowledge in the field of machine learning, with a particular emphasis on the application of recursive feature elimination (RFE). Through this project, we have improved our coding abilities and become more skilled at creating algorithms and composing code specifically for the RFE method.

Feedback from Dr Sharin Hazlin Binti Huspi

FEEDBACK: Overall, the group has done a good job in learning and applying what they need to do for the project. It was easy to guide them as they were able to understand what they needed to do. They were also interested in the area, thus they were able to understand the process and develop the algorithm. However they still need to learn to discuss their findings. I think there are still areas to learn in doing the discussion analysis of the results.			
CLIENT NAME	Sharin Hazlin Huspi		
SIGNATURE	Shmie	**DATE**	11/02/2024