ANALYSIS OF CLASSIFICATION OF LUNG CANCER USING THE DECISION TREE CLASSIFIER METHOD

Wendy Setiawan[*1], *Muhammad Faja Shandika*[2], *Amalia*[3], *Muhammad Radhi*[4], *Jepri Banjarnahor*[*5]
[1,2,3,4,5]*Indonesian Prima University*
Jl. Sampul No.3, Sei Putih Bar., Kec. Medan Petisah, Medan City, North Sumatra 20118
Email : jepribanjarnahor@unprimdn.ac.id*

ABSTRACT-The International Agency for Research on Cancer (IARC) revealed staggering figures, with 19.3 million global cancer cases and 10 million related deaths in that year. Cancer, characterized by abnormal cell growth, can potentially be dangerous with the ability to metastasize. Notably, lung cancer is often detected in an advanced stage due to a lack of awareness and comprehensive medical assessment. Lung cancer usually presents with a late-stagediagnosis. From 60% to 85% of individuals diagnosed with lung cancer show a lack of awarenessabout their condition. Early diagnosis using an accurate classification method can significantly increase the success of lung cancer diagnosis. To improve predictions, Decision Tree Classifier method was used in lung cancer classification, resulting in a significant increase in accuracy. This study achieved a good level of accuracy, with an accuracy value of 95.16% at a max_depth modeldepth of 15, and tested in 40 experimental iterations. These results are expected to provide hopefor progress in the classification of lung cancer.

Keywords: Lung, Cancer, Classification, Decision Tree

1. INTRODUCTION

The International Agency for Research on Cancer (IARC) published the latest edition of its Globocan 2020 report. The global cancer incidence reached 19.3 million cases, and cancer-related deaths reached 10 million in 2020. According to estimates provided by the International Agency for Research on Cancer (IARC), cancer deaths indicate that 1 in 8 men and 1 in 11 women die from this disease [1].

Cancer is a pathological condition characterized by the abnormal proliferation of body tissue cells, leading to cancer cell formation [2]. During its development, cancer cells have the potential to metastasize to distant anatomical locations, resulting in fatal outcomes [3][4]. Cancers are generally recognized by the general population as neoplasms, although not all neoplasms are malignant [5][6].

Lung cancer often presents with a late-stage diagnosis. From 60% to 85% of individuals diagnosed with lung cancer show a lack of awareness about their condition[7]. This phenomenon can be attributed to patients who perceive coughing and shortness of breath as typical bodily experiences, leading them to avoid seeking medical intervention [8]. In addition, many health professionals focus on symptom management without a comprehensive assessment [9].

Early diagnosis and accurate classification significantly improve lung cancer diagnosis and treatment success. Therefore, the selection of the appropriate classification method dramatically affects the results of the accuracy value. One of the classification methods that can be used in classifying lung cancer is the Decision Tree.

In the previous lung cancer classification study conducted by [10] using the decision tree method and support vector machine, it was found that the Support Vector Machine (SVM) method, when applied with Forward Selection and a data separation ratio of 80:20, showed the highest level of accuracy, with an accuracy rate of 62.3%. In comparison, the decision tree method only gets the highest accuracy value of 56.7%.

Based on the background presentation, it is necessary to develop a lung cancer classification method to get better accuracy in predicting lung cancer. Therefore researchers conducted a study entitled "Analysis of Lung Cancer Classification Using the Decision Tree Classifier Method."

2. RESEARCH METHODS

Classification is a data mining technique that assigns data to predefined categories. Classification is the process of identifying patterns representing data classes to predict the value of an unknown object class. We have to validate the training data to get the model. Using the test data, the level of accuracy and the built model are evaluated simultaneously. Classification makes it possible to predict the name or value of data objects [11].

Decision Trees (DTs) are non-parametric supervised learning methods for classification and regression. The goal is to create a model that predicts the value of the target variable by learning simple decision rules inferred from data features. A tree can be seen as a constant approach piece by piece [12][13].

In creating a Decision Tree algorithm, several steps must be followed. First of all, we need to collect the training data to use. This data is grouped into several different classes based on their characteristics. We are using pre-existing historical data as an example and calculating the values that will be the starting point of the decision tree. Once we have the training data, the next step is to estimate scores for each characteristic. We find out which factors are most influential in making decisions. After that, we will choose the part with the highest score as the root of our decision tree [14].

The Decision Tree method has advantages that include high interpretability, allows simple decision rules, and the ability to handle non-linear relationships in complex data. Grouping similar features also facilitates interpretation and strengthens the model's ability to capture the relationship between characteristics. Good scalability allows its use on large datasets with reliable performance, while its ability to manage various types of variables, take advantage of important information from features, and overcome outliers in data is a significant added value [15]

2.1 Research Steps

Steps or research flow are needed to help the research run and finish on time. We can see the steps or research flow in Figure 1 below:

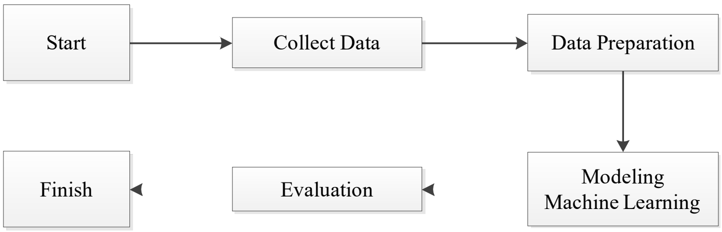


Figure 1. Research Steps

a. Collect Data
I am gathering information or facts from various sources or resources for analysis, understanding, or decision-making.
b. Data Preparation
Data Preparation is the step where the data that has been collected is processed and prepared before further analysis is carried out. This involves cleaning the data of errors, filling in missing values, changing the data format if necessary, and ensuring the data is ready for proper use.
c. Modeling Machine Learning
Modeling Machine Learning is creating and training computer models to learn from data and

make predictions or decisions based on patterns identified in the data.

d. Evaluation

Evaluation is an essential step in which we assess how good the results or models that have been made are. We use specific metrics to measure how accurate or suitable the model is at predicting or making decisions. Evaluation helps us ensure that our model is working correctly with the new data and also allows us to understand what kinds of errors can occur so we can correct them.

3. RESULTS AND DISCUSSION

3.1 Problem Analysis

Lung cancer classification analysis using the Decision Tree Classifier method is a complex effort that aims to investigate the effectiveness of this approach in overcoming essential problems in the medical world. This research addresses the complexities of early identification of lung cancer, a crucial issue because of its profound health implications. By utilizing the Decision Tree Classifier algorithm, which can map decisions based on data patterns, this study tries to build a model that can classify types of lung cancer based on a set of clinical features and symptoms.

3.2 Data Processing

In processing lung cancer classification data in this study, a data processing flow is needed so that the processed data obtains a good accuracy value for the model used. In Figure 2, you can see the flow of data processing in this study:

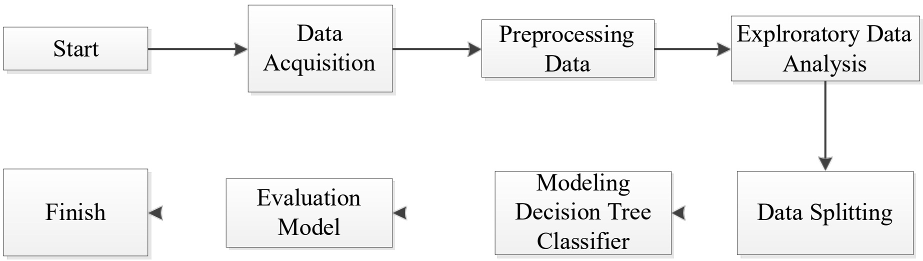
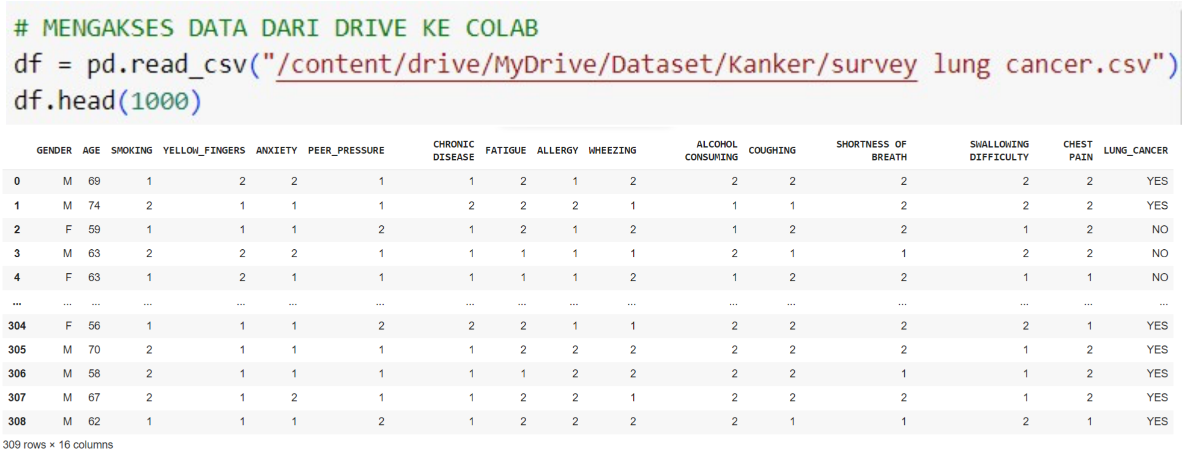


Figure 2 Data Processing Flow

3.2.1. Data Acquisition

Data acquisition is a fundamental procedure carried out by researchers to collect relevant data. In this study, the data used consisted of a dataset consisting of 309 rows and 16 columns explicitly focusing on patients diagnosed with lung cancer sourced from an open-source data provider [16]. As illustrated in Figure 3 presented below:



```
# MENGAKSES DATA DARI DRIVE KE COLAB
df = pd.read_csv("/content/drive/MyDrive/Dataset/Kanker/survey lung cancer.csv")
df.head(1000)
```

	GENDER	AGE	SMOKING	YELLOW_FINGERS	ANXIETY	PEER_PRESSURE	CHRONIC DISEASE	FATIGUE	ALLERGY	WHEEZING	ALCOHOL CONSUMING	COUGHING	SHORTNESS OF BREATH	SWALLOWING DIFFICULTY	CHEST PAIN	LUNG_CANCER
0	M	69	1	2	2	1	1	2	1	2	2	2	2	2	2	YES
1	M	74	2	1	1	1	2	2	2	1	1	1	2	2	2	YES
2	F	59	1	1	1	2	1	2	1	2	1	2	2	1	2	NO
3	M	63	2	2	2	1	1	1	1	1	2	1	1	2	2	NO
4	F	63	1	2	1	1	1	1	1	2	1	2	2	1	1	NO
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
304	F	56	1	1	1	2	2	2	1	1	2	2	2	2	1	YES
305	M	70	2	1	1	1	1	2	2	2	2	2	2	1	2	YES
306	M	58	2	1	1	1	1	1	2	2	2	2	1	1	2	YES
307	M	67	2	1	2	1	1	2	2	1	2	2	2	1	2	YES
308	M	62	1	1	1	2	1	2	2	2	2	1	1	2	1	YES

309 rows × 16 columns

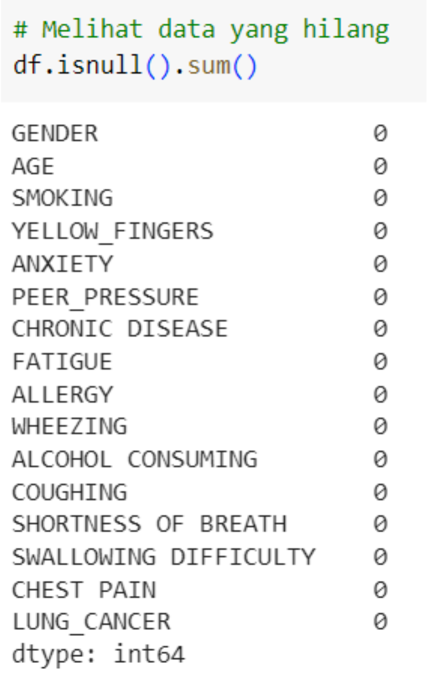
Figure 3 Lung Cancer Classification Datasheet

3.2.2. Preprocessing Data

The data obtained then proceeds to the preprocessing stage, where the data set is prepared in connection with developing and training machine learning models. At this point, the main goal is to improve the dataset's quality. The following procedure is carried out in the following way:

1. Missing Values

At this stage, the dataset is checked for any examples of missing data. After performing an analysis of missing data, it was determined that there were no missing data. So the process moves on to the next stage, which involves removing duplicate entries. Missing values can be seen in Figure 4 below:



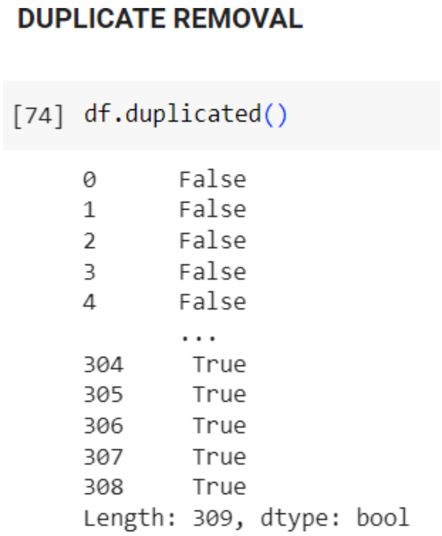
```
# Melihat data yang hilang
df.isnull().sum()
```

```
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
```

Figure 4 Missing Handling

2. Duplicate Removal

At this stage, data with identical or duplicate values undergoes an elimination process to eliminate the same data. Figure 5 below illustrates the deduplicate phases.



DUPLICATE REMOVAL

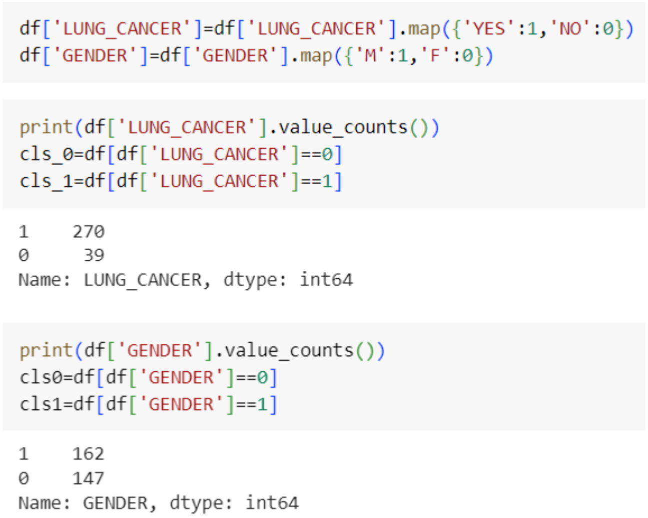
```
[74]  df.duplicated()
```

```
0      False
1      False
2      False
3      False
4      False
       ...
304     True
305     True
306     True
307     True
308     True
Length: 309, dtype: bool
```

Figure 5 Duplicate Removal

3. Featured Engineering

During the feature engineering stage, the data set is engineered into classes 0 and 1, selected from the GENDER and LUNG_CANCER columns, respectively. Class 0 represents a class that does not have lung cancer, while Class 1 represents one with cancer. Figure 6 below illustrates the stages of engineering features:



```
df['LUNG_CANCER']=df['LUNG_CANCER'].map({'YES':1,'NO':0})
df['GENDER']=df['GENDER'].map({'M':1,'F':0})
```

```
print(df['LUNG_CANCER'].value_counts())
cls_0=df[df['LUNG_CANCER']==0]
cls_1=df[df['LUNG_CANCER']==1]
```

```
1    270
0     39
Name: LUNG_CANCER, dtype: int64
```

```
print(df['GENDER'].value_counts())
cls0=df[df['GENDER']==0]
cls1=df[df['GENDER']==1]
```

```
1    162
0    147
Name: GENDER, dtype: int64
```

Figure 6 Future Engineering

3.2.3 Exploratory Data Analysis

This stage includes the initial phase of exploratory data analysis (EDA), which requires examination of the distribution pattern of lung cancer cases. The data exploration carried out is as follows:

1. Variable Correlation

Valuable correlation is the process of seeing the correlation of variable 1 with other variables. In this study, the variables used are Yellow_Fingers, Anxiety, Peer_Pressure, Chronic Disease, Fatigue, Allergy, Wheezing, Alcohol Consuming, Coughing, Shortness Of Breath, Swallowing Difficulty, Chest Pain, Lung_Cancer.

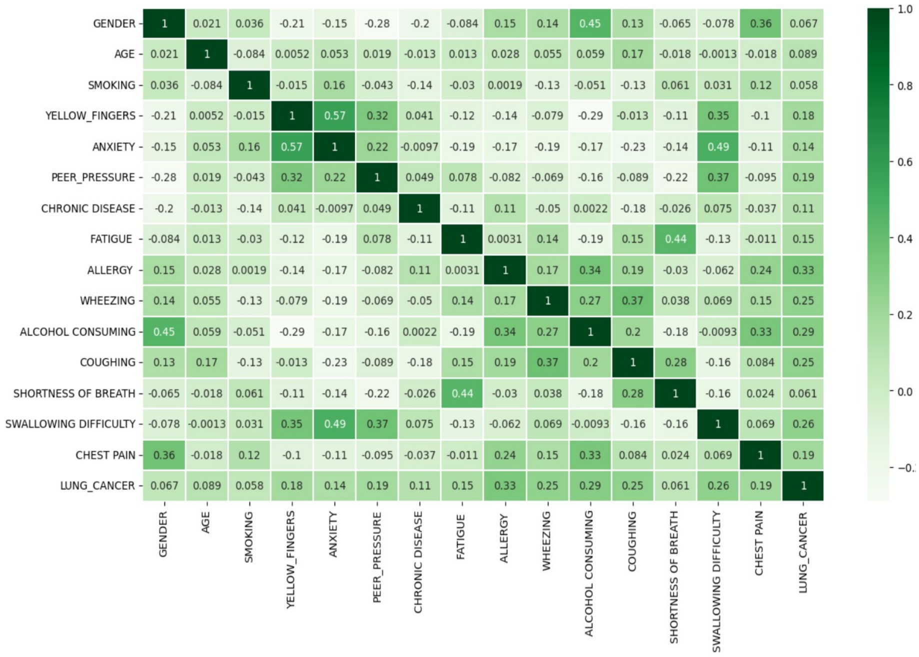


Figure 7 Correlation Data

In the context of this study, it was observed that darker colors indicate higher levels of correlation, whereas lighter colors indicate lower levels of correlation.

2. Age Column Visualization

For a more detailed understanding of the data, we visualized the age column with lung_cancer to see the distribution of the data on the age range of positive patients and those who were not favorable for lung cancer.

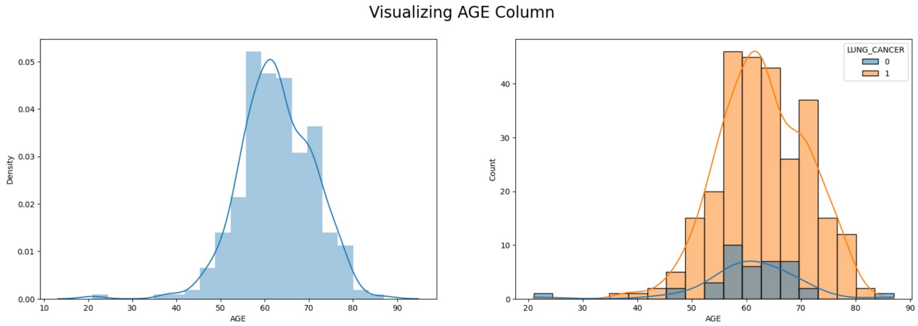


Figure 8 Visualization of Age and Lung Cancer Columns

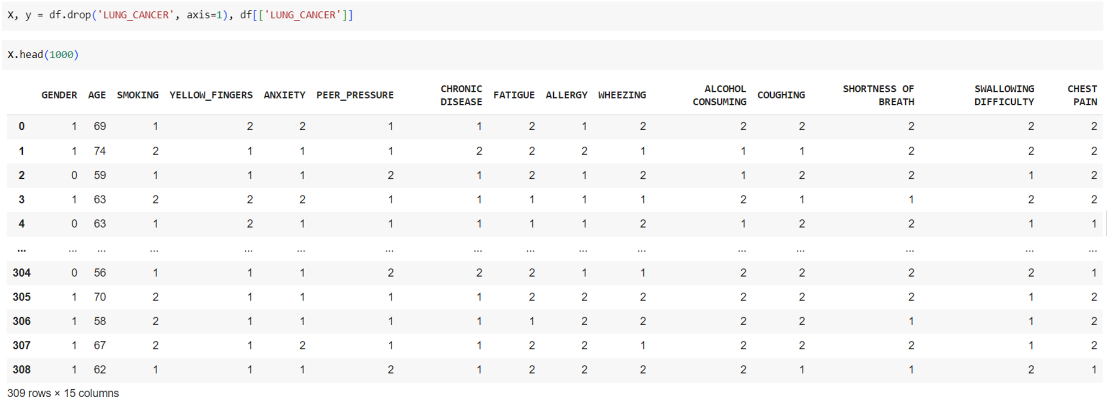
Based on Figure 8, we can see that in the dataset used, there are the most data from patients with a patient age range of 40 years to 50 years, while the most affected patients with lung cancer occur in patients aged 50 years to 80 years.

3.2.4. Data Splitting

The data splitting process involves dividing the dataset into two subsets: training data and testing data. The goal of data training is to enable the movement of a machine learning model by leveraging a predefined training data set while assessing the performance of a potentially flawed model using a different test data set.

1. Creating the Defining Variable (X)

In making the determining variable (X), 15 columns are used, which can be seen in Figure 9 below:



```
X, y = df.drop('LUNG_CANCER', axis=1), df[['LUNG_CANCER']]
```

```
X.head(1000)
```

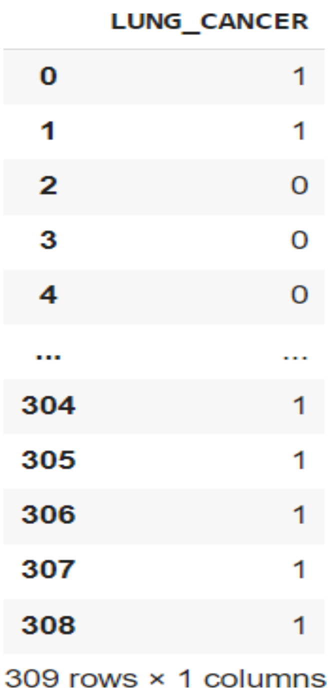
	GENDER	AGE	SMOKING	YELLOW_FINGERS	ANXIETY	PEER_PRESSURE	CHRONIC DISEASE	FATIGUE	ALLERGY	WHEEZING	ALCOHOL CONSUMING	COUGHING	SHORTNESS OF BREATH	SWALLOWING DIFFICULTY	CHEST PAIN
0	1	69	1	2	2	1	1	2	1	2	2	2	2	2	2
1	1	74	2	1	1	1	2	2	2	1	1	1	2	2	2
2	0	59	1	1	1	2	1	2	1	2	1	2	2	1	2
3	1	63	2	2	2	1	1	1	1	1	2	1	1	2	2
4	0	63	1	2	1	1	1	1	1	2	1	2	2	1	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
304	0	56	1	1	1	2	2	2	1	1	2	2	2	2	1
305	1	70	2	1	1	1	1	2	2	2	2	2	2	1	2
306	1	58	2	1	1	1	1	1	2	2	2	2	1	1	2
307	1	67	2	1	2	1	1	2	2	1	2	2	2	1	2
308	1	62	1	1	1	2	1	2	2	2	2	1	1	2	1

309 rows × 15 columns

Figure 9. Determinant Variable (X)

2. Create Target Variable (Y)

In making the target variable (Y), 1 column is used, which can be seen in Figure 10 below:



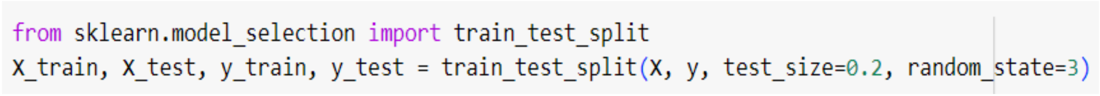
	LUNG_CANCER
0	1
1	1
2	0
3	0
4	0
...	...
304	1
305	1
306	1
307	1
308	1

309 rows × 1 columns

Figure 10 Target Variable (Y)

3. Data Splitting

The data is divided into two parts, namely 80% training data and 20% test data with random state 3 settings using the Skcitlearn library, which can be seen in Figure 11 below:

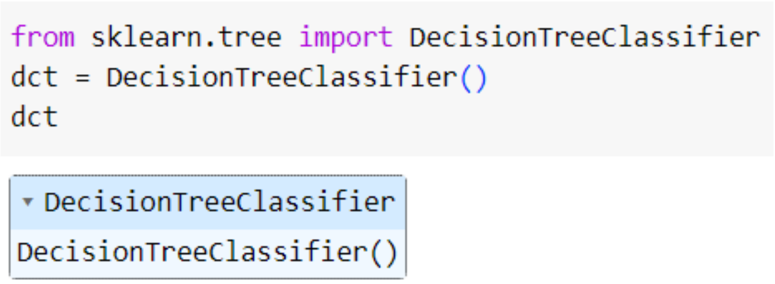


```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=3)
```

Figure 11 Splitting Data

3.2.5 Modeling Decision Tree Classifier

This study created a Decision Tree using the Decision Tree Classifier library, which you can see in Figure 12 below:

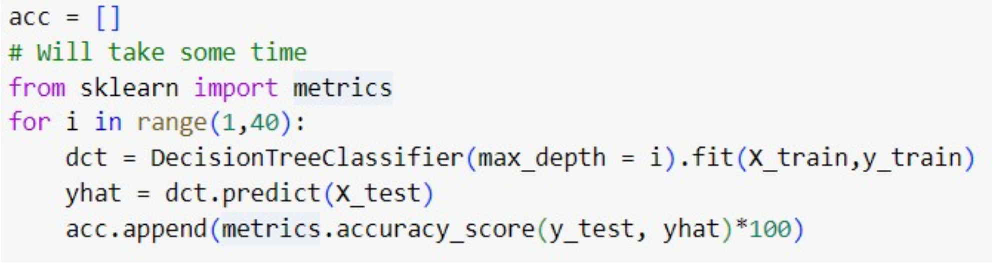


```
from sklearn.tree import DecisionTreeClassifier
dct = DecisionTreeClassifier()
dct
```

```
▾ DecisionTreeClassifier
DecisionTreeClassifier()
```

Figure 12 Import Library Decision Tree Classifier

After importing the library, the next step is setting the model configuration using max_depth. The max_depth setting is repeated 40 times to get the best accuracy. The max_depth location can be seen in Figure 13 below:



```
acc = []
# Will take some time
from sklearn import metrics
for i in range(1,40):
    dct = DecisionTreeClassifier(max_depth = i).fit(X_train,y_train)
    yhat = dct.predict(X_test)
    acc.append(metrics.accuracy_score(y_test, yhat)*100)
```

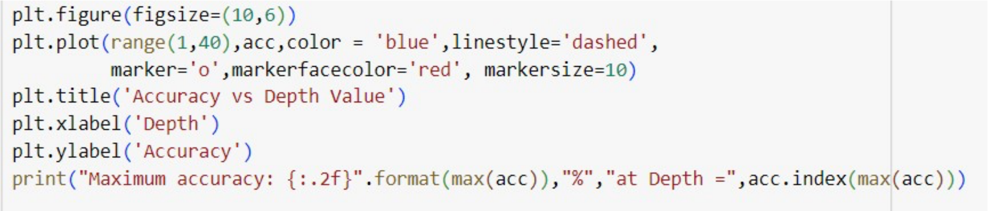
Figure 13 Setting Max_Depth Model

The "max depth" parameter in the decision tree refers to the maximum depth or maximum rate of the tree allowed to grow during the training process. In a decision tree, each level represents a decision node that divides data based on specific attributes or features. The tree depthindicates the number of stories from the root node (starting point) to the leaf nodes (final result or classification). From the max_dept experiment, the maximum accuracy value is 95.16% at the 15th depth, the accuracy value obtained from 40 trials can be seen in Table 1 below:

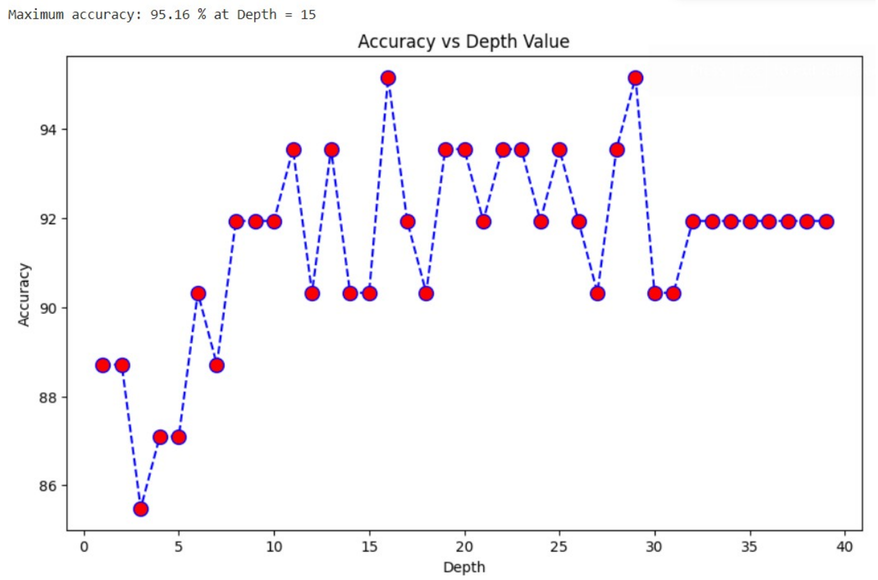
Table 1 Accuracy Value

depth	Accuracy Value
0	88.70967741935483
1	88.70967741935483
2	85.48387096774194
3	87.09677419354838
4	87.09677419354838
5	90.32258064516128
6	88.70967741935483
7	91.93548387096774
8	91.93548387096774
9	91.93548387096774
10	93.54838709677419
11	90.32258064516128
12	93.54838709677419
13	90.32258064516128
14	90.32258064516128
15	95.16129032258065
16	91.93548387096774
17	90.32258064516128
18	93.54838709677419
19	93.54838709677419
21	93.54838709677419
22	93.54838709677419
23	91.93548387096774
24	93.54838709677419
25	91.93548387096774
26	90.32258064516128
27	93.54838709677419
28	95.16129032258065
29	90.32258064516128
30	90.32258064516128
31	91.93548387096774
32	91.93548387096774
33	91.93548387096774
34	91.93548387096774
35	91.93548387096774
36	91.93548387096774
37	91.93548387096774
38	91.93548387096774

To make it easier to see the accuracy value of each depth, we visualize the accuracy value data using a line chart which we can see in Figure 14 below:



```
plt.figure(figsize=(10,6))
plt.plot(range(1,40),acc,color = 'blue',linestyle='dashed',
         marker='o',markerfacecolor='red', markersize=10)
plt.title('Accuracy vs Depth Value')
plt.xlabel('Depth')
plt.ylabel('Accuracy')
print("Maximum accuracy: {:.2f}".format(max(acc)),"%","at Depth =",acc.index(max(acc)))
```

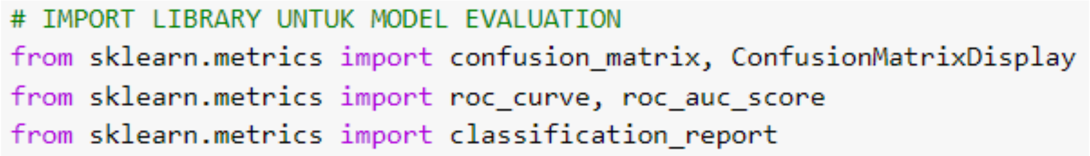


```
Maximum accuracy: 95.16 % at Depth = 15
```

Figure 14 Accuracy Value Line Diagram

3.2.6. Model Evaluation

To assess the success of the algorithm used in accurately identifying lung cancer cases, it is very important to examine the model evaluation of the Decision Tree Classifier algorithm approach. The library used for model evaluation is Scikit-Learn, as illustrated in Figure 15 below:



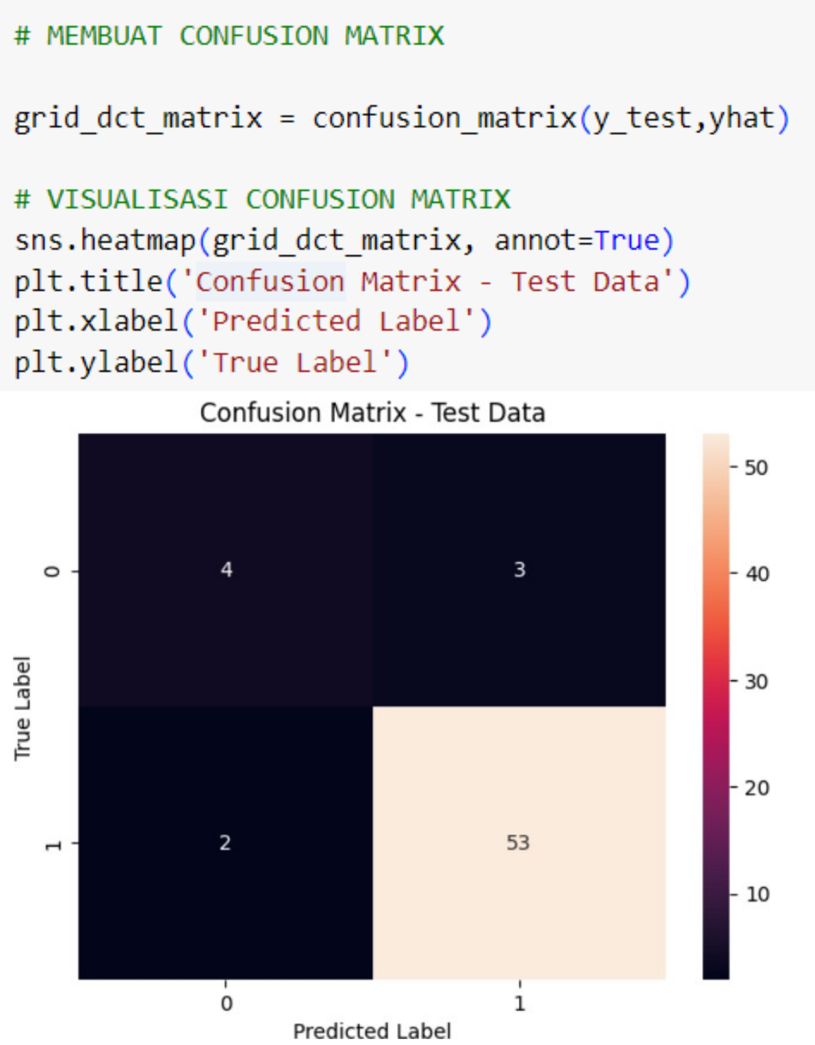
```
# IMPORT LIBRARY UNTUK MODEL EVALUATION
from sklearn.metrics import confusion_matrix, ConfusionMatrixDisplay
from sklearn.metrics import roc_curve, roc_auc_score
from sklearn.metrics import classification_report
```

Figure 15 Library Model Evaluation

The confusion matrix in the context of a decision tree is a table that is used to describe the performance of a classification model based on the predicted results of the actual data. The confusion matrix maps four possible outcomes from the classification process: true positive (TP), true negative (TN), false positive (FP), and false negative (FN).

True positive is when the model correctly predicts a class as positive, true negative is when the model correctly predicts a class as negative, false positive is when the model correctly predicts a class as positive, and false negative is when the model incorrectly predicts a class as unfavorable.

The confusion matrix is helpful in calculating evaluation metrics such as accuracy, precision, recall, and F1-score, which helps analyze how well the decision tree model classifies data and identifies the errors made by the model. By analyzing the confusion matrix, we can understand more deeply the performance of the decision tree model and take steps to improve and enhance the model.



```
# MEMBUAT CONFUSION MATRIX

grid_dct_matrix = confusion_matrix(y_test,yhat)

# VISUALISASI CONFUSION MATRIX
sns.heatmap(grid_dct_matrix, annot=True)
plt.title('Confusion Matrix - Test Data')
plt.xlabel('Predicted Label')
plt.ylabel('True Label')
```

Figure 16 Confusion Matrix decision tree algorithm

The results of the Confusion Matrix Decision Tree algorithm process can be seen in Figure 16. The results were accurate positive 4 times, true negative 3 times, false positive 2, and false negative 53 times

CONCLUSION

In this study, the application of the Decision Tree Classifier method to lung cancer datasets has yielded auspicious results. By achieving an accuracy rate of 95.16% in a configuration involving a max_depth of 15, as well as through 40 experimental iterations, this study shows that the model can correctly classify the types of lung cancer.

BIBLIOGRAPHY

[1] "GLOBOCAN 2020: DataNew Global Cancer | UICC." https://www-uicc-org.translate.goog/news/globocan-2020-new-global-cancer-data?_x_tr_sl=en&_x_tr_tl=id&_x_tr_hl=id&_x_tr_pto=tc (accessed Aug. 09, 2023).

[2] PPurnamawati, C. Tandrian, EM Sumbayak, and W. Kertadjaja, "Analysis of Primary Lung Cancer Incidence in Indonesia in 2014-2019 Purnamawati1," J. Kedokt. Meditek, vol. 27, no. 2, pp. 164–172, 2021, doi: 10.36452/jkdoktmeditek.v27i2.2066.

[3] N.Kamisi and R. Ratianingsih, "Analysis of the Stability of Mathematical Models for the Spread of Smoking Behavior with Risk Factors for Lung Cancer," J. Ilm. Matt. And Applied.,

vol. 19, no. 1, pp. 72–81, 2022, doi: 10.22487/2540766x.2022.v19.i1.15710.

[4] V. No, E. April, and EMIHandrina, “RISK FACTORS OF BREAST CANCER IN ASIA (WITH A META ANALYSIS STUDY APPROACH) ALFITA,” vol. 4, no. 3, pp. 304–312, 2022, [Online]. Available: https://jurnal.ensiklopediaku.org/ojs-2.4.8-3/index.php/ensiklopedia/article/view/501

[5] Juwita, N. Amalita, and MD Parma, "Risk Factors Affecting Lung Cancer Using Logistic Regression Analysis," UNPjoMath, vol. 4, no. 1, pp. 38–42, 2021.

[6] K.Lungs, "A Comparative Analysis of Filter Variations in Edge Detection Using the Canny Method for Ct-Scan Images," vol. 8, no. 2, pp. 77–81, 2023.

[7] E. Tiana and S.Wahyuni, "Results of Analysis of Data Mining Techniques with the Naive Bayes Method for Diagnosing Breast Cancer," J. Sist. Computer. and Inform., vol. 1, no. 2, p. 130, 2020, doi: 10.30865/json.v1i2.1766.

[8] MsNaezer, "Naïve Bayes Algorithm Performance Analysis and k-NN for Predicting Lung Cancer," J. Ilm. Computing, vol. 4, no. 1, pp. 88–100, 2023.

[9] L. Sari, A.Romadloni, and R. Listyaningrum, "Application of Data Mining in Lung Cancer Prediction Analysis Using the Random Forest Algorithm," Infotekmesin, vol. 14, no. 1, pp. 155–162, 2023, doi: 10.35970/infotekmachine.v14i1.1751.

[10] D.Septhya, K. Rahayu, S. Rabbani, and V. Fitria, “Implementation of Decision Tree Algorithm and Support Vector Machine for Lung Cancer Classification,” vol. 3, no. April, pp. 15–19, 2023.

[11] vvPermana, HN Fazri, MFN Athoilah, M. Robi, and R. Firmansyah, "Application of Data Mining in Lung Cancer Prediction Analysis Using the Random Forest Algorithm," J. Ilm. Tech. inform. and Commun., vol. 3, no. 2, pp. 27–41, 2023, doi: doi.org/10.55606/juitik.v3i2.472.

[12] “1.10. Decision Trees — scikit-learn 1.3.0 documentation.” https://scikit-learn.org/stable/modules/tree.html (accessed Aug. 09, 2023).

[13] E. Indra et al., “A Comparison of Heart Abnormalities Detection on ECG using KNN and Decision Tree,” InternetworkingIndonesia. J., vol. 12, no. 2, pp. 19–23, 2020.

[14] AROktavyani et al., "Comparison of the Naive Bayes, K-NN and Decision Tree Methods to the Healthcare Stroke Dataset", doi: 10.31284/p.snestik.2023.4067.

[15] A. Wahab, S.Samarinda, I. Lishania, R. Goejantoro, and YN Nasution, "Comparison of the Naive Bayes Classification Method and the Algorithmic Decision Tree Method (J48) in Stroke Patients at Abdul Wahab Sjahranie Hospital Samarinda Hospital," J.EXPONENTIAL, vol. 10, no. 2, 2019.

[16] “Lung Cancer | Kaggle." https://www.kaggle.com/datasets/mysarahmadbhat /lung-cancer (accessed Aug. 09, 2023).