Informatics and Software Engineering
p-ISSN/e-ISSN: xxxx-xxxx/xxxx-xxxx
Homepage: https://sanscientific.com/journal/index.php/ise
1(1) 8-14 (2023)
DOI: 10.58777/ise.v1i1.56

Lung Cancer EDA Classification Using the Decision Trees Method in Python

Aqila Darin Makkyah [1 *], Muhammad Faisal [2]
[1] UIN Maulana Malik Ibrahim Malang, Malang
[2] UIN Maulana Malik Ibrahim Malang, Malang

Received: June 12, 2023; Accepted: June 25, 2023

Abstract
Cancer is the second leading cause of death worldwide. In Indonesia, it is one of the diseases with a high mortality rate. Most patients are unaware of their lung cancer condition, resulting in delayed treatment. A prediction method with high accuracy is needed for the early detection of lung cancer. This study aims to classify lung cancer using the Decision Trees method and perform Exploratory Data Analysis (EDA) using a dataset obtained from Kaggle. The research achieved a high recall value for the positive class (Yes class) but a low recall for the negative class (No class). The study utilized the Decision Trees algorithm, known for its good performance. The dataset used includes clinical and demographic information of patients. By building a Decision Trees model, the research successfully classified lung cancer with good accuracy. The EDA results also provide insights into important factors in lung cancer classification. This study has the potential to contribute to the development of predictive models for lung cancer.

Keywords : Classification, lung cancer, Decision Trees, Explanatory data analysis

Abstrak
Kanker merupakan penyebab kematian tertinggi kedua di dunia. Di Indonesia termasuk penyakit dengan tingkat kematian yang tinggi. Sebagian besar penderita tidak mengetahui bahwa dirinya terkena kanker paru sehingga penanganan menjadi terlambat. Metode prediksi dengan tingkat akurasi yang tinggi diperlukan untuk mendeteksi secara dini kanker paru. Penelitian ini untuk melakukan klasifikasi kanker paru-paru menggunakan metode Decision Trees dan melakukan Analisis Data Eksploratori (EDA) menggunakan dataset yang diperoleh dari Kaggle. Penelitian tersebut menghasilkan nilai recall yang tinggi untuk kelas positif (kelas Yes) namun rendah untuk kelas negatif (kelas No). Penelitian ini dibuat dengan algoritma Decision Trees yang dikenal memilki performa yang baik. Dataset yang digunakan berisi informasi klinis dan demografis pasien. Dengan membangun model Decision Trees, penelitian ini berhasil mengklasifikasikan kanker paru-paru dengan akurasi yang baik. Hasil EDA juga memberikan wawasan tentang faktor-faktor penting dalam klasifikasi kanker paru-paru positif dan negatif. Penelitian ini berpotensi memberikan kontribusi dalam pengembangan model prediktif untuk kanker paru-paru.

Kata Kunci: klasifikasi, kanker paru-paru, Decision Trees, Penjelasan data analisis

*Corresponding author: Aqila Darin Makkyah (200605110100@student.uin-malang.ac.id)

1. Introduction

The lungs are one of the organs of the respiratory system, which functions as a place for exchanging carbon dioxide and oxygen in the blood. The problem that often occurs in the respiratory system is polluted air quality, so the air that is inhaled contains many bacteria that can attack the respiratory system, especially the lungs (Oktavianto, 2020). One of the diseases that attacks the respiratory system is lung cancer (Adiwijaya, 2018).

Cancer, according to the definition of the National Cancer Institute, is a genetic disease caused by changes in genes that control cell function, especially the function to grow and divide. Cancer is new cells that grow abnormally, then attack the contralateral body parts and spread to other organs (Charan & Parthiban, 2023). Lung cancer is one of the most common and deadly types of cancer worldwide. Lung cancer is one of the three diseases and the highest cause of death (Nasrullah, 2021). Several factors influence the development of lung cancer globally, including long-term exposure to cigarette smoke, genetic factors, radon gas, and air pollution (Wardani et al., 2023). This condition occurs when abnormal cells grow uncontrollably in the lung tissue. Lung cancer can develop quickly and spread to other body parts, making it difficult to treat in its advanced stages (Sari & Listyaningrum, 2023).

The main factor that causes lung cancer is smoking, either actively or passively. Long-term exposure to secondhand smoke contains various harmful chemicals that can damage lung cells and cause genetic changes, leading to cancer. Apart from smoking, environmental exposures such as air pollution, exposure to toxic chemicals such as asbestos, and a family history of lung cancer can also increase a person's risk of developing this disease.

Data from the Global Cancer Observatory WHO states the ten most deadly types of cancer globally. Lung cancer occupies the first position with 1,796,144 deaths, followed by colorectal cancer (935,173 deaths) and prostate cancer (375,304 deaths). In Indonesia, lung cancer has the highest incidence, with around 34,783 new cases and 30,843 deaths in 2020 (Rahman, 2020). The main factor for cancer in Indonesia is smoking; besides smoking directly, inhaled cigarette smoke also increases the risk of lung cancer. Other factors are genetics, a family history of cancer, drinking coffee more than 6 cups/day, chronic lung disease, alcohol consumption, consumption of fried or grilled meat, air pollution, and exposure to chemicals (Meiyanti et al., 2023).

Lung cancer symptoms can vary depending on the stage of the disease. However, common symptoms include a chronic cough that does not go away, shortness of breath, chest pain, unexplained weight loss, persistent tiredness, and hoarseness. Unfortunately, these symptoms often do not appear in the early stages of the disease, so lung cancer diagnosis often occurs at a more advanced stage when treatment becomes more difficult (Fardian, 2021). Research continues to improve understanding to address the challenges posed by lung cancer, of this disease, and to develop better early detection methods and more effective therapies (Cahyadie, 2016).

Research on lung cancer has been conducted (Wulandari & Perdana, 2022), that study uses the Naïve Bayes algorithm to predict whether a person has lung cancer (yes/no). Based on the algorithm's accuracy measurement, the recall percentage for the positive class was 98.77%, and the negative class was 66.67%. The precision value is 95.24%, and the accuracy level is 94.62%. This research produced a model with high recall for the positive class (class Yes) but low for the negative class (class No).

Researchers increase their understanding of the types of lung cancer and contribute to the field of data science; exploratory data analysis (EDA) was used in this study as an effective approach. One method that can be used in this analysis is Decision Trees or Decision Trees. This method allows us to classify a person as having lung cancer or not based on the features present, thereby assisting in a more precise diagnosis and treatment. In this context, this study aims to explore the application of the Decision Trees method in classifying lung cancer using the Python programming language. EDA on relevant lung cancer datasets and applies the Decision Trees method to classify a person as having lung cancer or not based on relevant features, such as age, sex, smoking history, histological type, and disease stage.

Implementation using Python provides an added advantage in performing this analysis. Python is a popular programming language in data science and machine learning. Powerful libraries and tools like NumPy, Pandas, and sci-kit-learn in Python allow us to load, process, and analyze data efficiently. Besides that, Python also provides powerful visualization tools such as Matplotlib and Seaborn, which help in the graphical representation of EDA results and interpretation of Decision Trees models. Through this research, it is hoped that it can increase our understanding of the classification of lung cancer using the Decision Trees method, as well as contribute to the prevention, early detection, and more appropriate treatment of this deadly disease.

This study aims to classify lung cancer using the Decision Trees method and perform Exploratory Data Analysis (EDA) using the dataset available on Kaggle. The dataset used is the "lung cancer.csv survey," which contains clinical and demographic information about the patient. The research phase begins with identifying the problem,

namely the classification of lung cancer based on the predictor variables in the dataset. Then, data collection was carried out from Kaggle sources. After that, data preprocessing, including the necessary data cleaning and transformation, is carried out before building the Decision Trees model.

2. Methods

The stages in this study can be described in Figure 1.

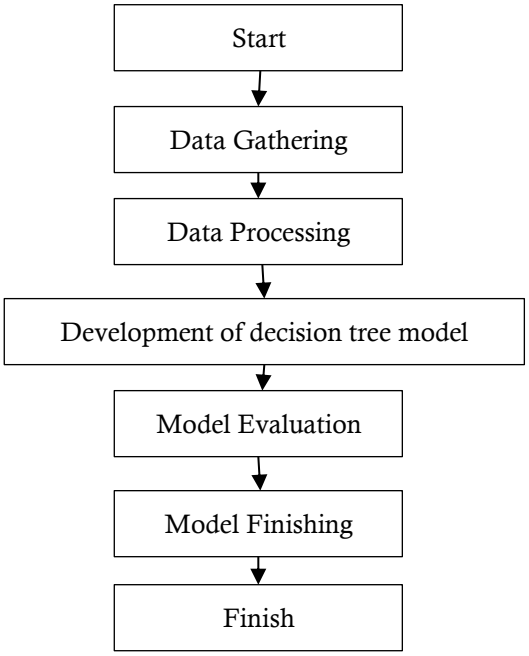

Figure 1. Research stages

2.1 Data Gathering

The dataset used in this study is the Kaggle dataset. This dataset contains information about what attributes are used as features to classify lung cancer EDA. This dataset is divided into two EDA lung cancer categories: YES and NO. It displays the number of positive and negative cancer patients based on an analysis of their habits, such as smoking, alcohol, allergies, and others.

2.2 Data Sharing and Data Preprocessing

At this stage, the data is divided into two parts, namely training data and testing data. The training data is used to train the algorithm in building a model, while the testing data measures the performance obtained from the training data. Data preprocessing is done by cleaning and transforming the collected data. This method may involve removing missing or outlier data, combining or separating variables, and normalizing or standardizing the data to make the results more accurate.

2.3 Decision Trees

This system uses a Decision Trees algorithm to help classify lung cancer, which can accurately predict whether the patient has lung cancer based on the

attributes in the dataset. The Decision Trees algorithm is a learning method usually used for classifying data, including in the context of EDA for Lung Cancer. The working principle of the Decision Trees algorithm is to build a model using a decision tree structure. Each tree will have a node that represents separation based on the attributes in the dataset to minimize misclassification in each node. After the decision tree is built, it can predict new data based on the given attributes.

2.4 Evaluation of Model Decision Trees

This research will be carried out by applying data mining to predict the occurrence of lung cancer. The data mining method used is Decision Trees. This stage involves evaluating the performance of the Decision Trees model that has been built. This effort can be done using evaluation metrics such as accuracy, precision, recall, F1-score, or area under the ROC curve (AUC-ROC). This evaluation helps determine how well the model can predict unseen

data.

2.5 Improvement of Model Decision Trees

If the model's performance does not meet expectations, this stage involves refining the Decision Trees model. This can involve adjusting model parameters, selecting better features, or using ensemble techniques such as Random Forest or Gradient Boosting to improve model performance and will involve interpreting the results from the Decision Trees model that has been built. This tree involves understanding the decision rules generated by decision trees and analyzing the importance of predictor variables in making decisions.

3 . Results and Discussion

Modeling and analysis using Python on Google Colab. Model Decision Trees generated using data train. Based on the summary model, it is known that the optimal number of variables considered to be broken down at each tree node is the optimal model from the Decision Trees in Figure 2, then calculates first and prints the importance level of each feature in the decision trees in Figure 3, and the final classification results are in Figure 4.

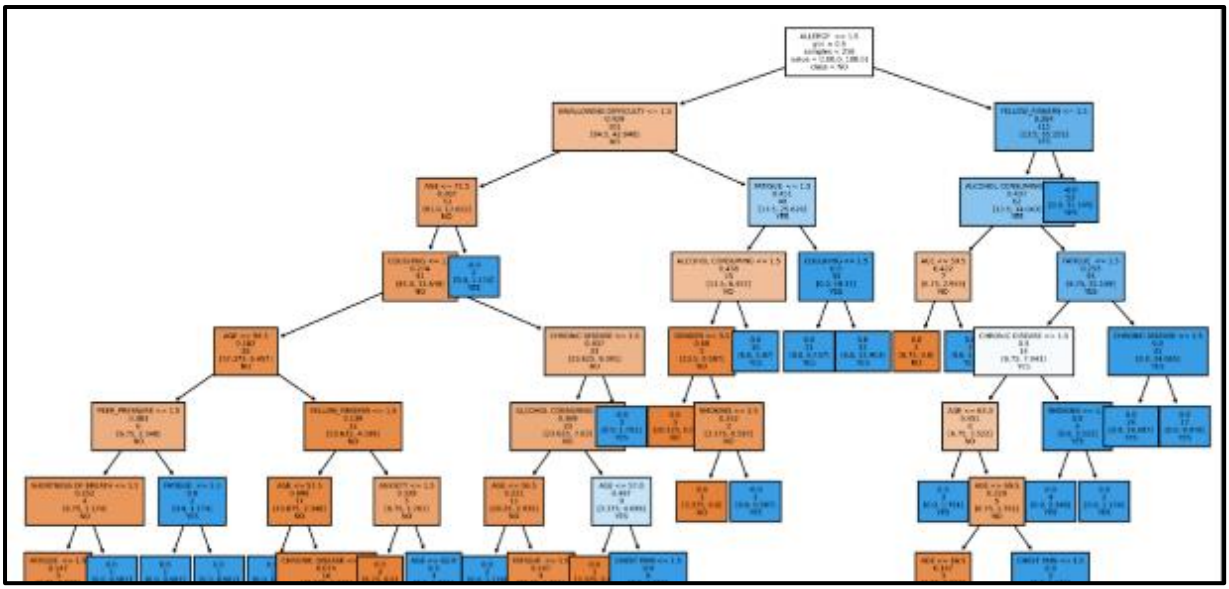

Figure 2. Modeling Decision Trees

Next, the Decision Trees model is built using the algorithms available in Python. Decision Trees algorithms such as ID3, C4.5, or CART can be used to build this model. During this process, optimal model parameters can be set to improve performance.

The results of the model performance evaluation show that the Decision Trees model has good accuracy in predicting lung cancer classification based on the predictor variables in the dataset. Evaluation metrics such as precision, recall, F1-score, or AUC-ROC can also be used to assess the performance of this model. Next, the Decision Trees model results are interpreted to understand the decision rules produced by the decision trees. This model helps identify the most important predictor variables in classifying lung cancer.

```
GENDER : 0.00117
AGE : 0.18560
SMOKING : 0.00935
YELLOW_FINGERS : 0.03161
ANXIETY : 0.03769
PEER_PRESSURE : 0.01387
CHRONIC DISEASE : 0.04033
FATIGUE  : 0.12825
ALLERGY  : 0.24928
WHEEZING : 0.00741
ALCOHOL CONSUMING : 0.13309
COUGHING : 0.01367
SHORTNESS OF BREATH : 0.02627
SWALLOWING DIFFICULTY : 0.12242
CHEST PAIN : 0.00000
```

Figure 3. Calculating and Printing the Importance Level of Each Feature in Decision Trees.

The function of Figure 4 helps to see the relative importance of each feature in the Decision Tree model. This level of importance indicates how much each feature contributes to making decisions on the model. By evaluating this level of importance, we can identify the features that have the most significant influence on prediction and further understand the factors that play a role in lung cancer classification.

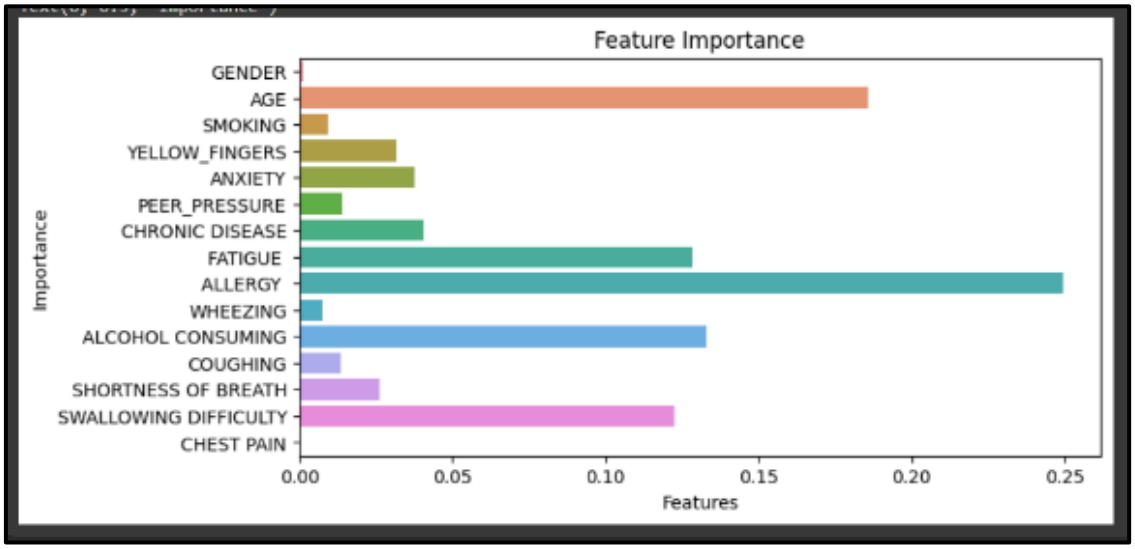

Figure 4. Final Result

Figure 4 is the final result for visually comparing the level of importance between the existing features to provide a clearer understanding of the most influential features in the classification of lung cancer. The final results can be seen in Figure 4.

4. Conclusion

The results of the model performance evaluation show that the Decision Trees model has good accuracy in predicting lung cancer classification based on the predictor variables in the dataset. Evaluation metrics such as precision, recall, F1-score, or AUC-ROC can also be used to assess the performance of this model. Next, the Decision Trees model results are interpreted to understand the decision rules produced by the decision trees. This result helps identify the most important predictor variables in classifying lung cancer. This study concludes that the Decision Trees method can effectively classify lung cancer based on the clinical and demographic data analyzed. The results of this study provide valuable insights into the development of lung cancer prediction models. They can be used for further research and development in this area. However, it should be remembered that these conclusions

are based on analysis using specific datasets found on Kaggle. Therefore, generalizing these findings to the general population or using different datasets must be exercised cautiously. In future research, it is possible to use a wider dataset and diversify methods to broaden knowledge about the classification of lung cancer with the Decision Trees method.

References

Adiwijaya, A. (2018). Deteksi Kanker Berdasarkan Klasifikasi Microarray Data. *Jurnal Media Informatika Budidarma*, *2*(4), 181. https://doi.org/10.30865/mib.v2i4.1043

B. Bawono and R. Wasono, "Perbandingan Metode Random Forest dan Naive Bayes," Jurnal Sains dan Sistem Informasi, vol. 3, no. 7, pp. 343–348, 2019, [Online]. Available: http://prosiding.unimus.ac.id

Cahyadie, R. C. R. (2016). *Hubungan kebiasaan merokok dengan kejadian kanker paru Di rsud ulin banjarmasin*. http://repository.unism.ac.id/395/

Charan, N., & Parthiban, S. (2023). *Logistic Regression over Decision Trees For Lung Cancer Detection To Increase Accuracy*. *10*, 2944–2953.

D. Dablain, B. Krawczyk, and N. v. Chawla, "DeepSMOTE: Fusing Deep Learning and SMOTE for Imbalanced Data," IEEE Trans Neural Netw Learn Syst, pp. 1–14, 2022, doi: 10.1109/TNNLS.2021.3136503.

D. H. Depari et al., "Perbandingan Model Decision Tree , Naive Bayes dan Random Forest untuk Prediksi Klasifikasi Penyakit Jantung," vol. 4221, pp. 239–248, 2022.

Fardian, A. I., & Riana, D. (2021). Prediksi Harapan Hidup Pasien Kanker Paru-Paru Pasca Operasi Bedah Thoraks Menggunakan Boosted Neural Network Dan Smote. *Jurnal Infomedia: Teknik Informatika, Multimedia, & Jaringan*, *6*(1), 9–15. http://archive.ics.uci.edu/ml/datasets .

Maiyanti, S. I., Zayanti, D. A., Andriani, Y., Suprihatin, B., Desiani, A., Salsabila, A., & Marselina, N. C. (2023). *Perbandingan Klasifikasi Penyakit Kanker Paru-paru menggunakan Support Vector Machine dan K-Nearest Neighbor*. *18*(1), 54–62.

Nasrullah, A. H. (2021). Implementasi Algoritma Decision Tree Untuk Klasifikasi Produk Laris. *Jurnal Ilmiah Ilmu Komputer*, *7*(2), 45–51. https://doi.org/10.35329/jiik.v7i2.203

Oktavianto, H., & Handri, R. P. (2020). Analisis Klasifikasi Kanker Payudara Menggunakan Algoritma Naive Bayes. *INFORMAL: Informatics Journal*, *4*(3), 117. https://doi.org/10.19184/isj.v4i3.14170

Purba, W., Wardani, S., Lumbantoruan, D. F., Celia, F., Silalahi, I., & Edison, T. L. (2023). *Optimization Of Lung Cancer Classification Method Using Eda-Based Machine Learning*. *6*(2), 43–50.

Rahman, C. A., & Kudus, A. (2022). Penggunaan Metode K Nearest Neighborhood untuk Imputasi Data Tersensor Kanan pada Pasien Kanker Paru-Paru Sel Kecil. *Bandung Conference Series: Statistics*, *2*(2), 441–448. https://doi.org/10.29313/bcss.v2i2.4615

Ramadani and B. H. Hayadi, "Perbandingan Metode Naive Bayes Dan Random Forest Untuk Menentukan Prestasi Belajar Siswa Pada Jurusan RPL (Studi Kasus SMK Swasta Siti Banun Sigambal)," Journal Computer Science and Information Technology(JCoInT) Program Studi Teknologi Informasi, no. 2, p. 2022, 2022, [Online]. Available: http://jurnal.ulb.ac.id/index.php/JCoInT/index

Sari, L., Romadloni, A., & Listyaningrum, R. (2023). Penerapan Data Mining dalam Analisis Prediksi Kanker Paru Menggunakan Algoritma Random Forest. *Infotekmesin*, *14*(01), 155–162. https://doi.org/10.35970/infotekmesin.v14i1.1751

S. Amaliah and M. Nusrang, "Penerapan Metode Random Forest Untuk Klasifikasi Varian Minuman Kopi Di Kedai Kopi Konijiwa Bantaeng," Variansi: Journal of Statistic and Its Application on Teaching and Research, vol. 4, no. 2, pp. 121–127, 2022, doi: 10.35580/variansiunm31.