**COM7003-24/25-Artificial Intelligence**

**Lung Cancer Prediction Survey**

**Module Leader:** Dr. Xin Lu

**Student Name:** Safa Dana

**Student ID:** 2404831

Contents

[**Introduction** 3](#_Toc192178535)

[**Details of the Dataset** 4](#_Toc192178536)

[**Import necessary libraries** 5](#_Toc192178537)

[**Take a look at the data frame specs** 6](#_Toc192178538)

[**EDA** 8](#_Toc192178539)

[**Data Visualization** 8](#_Toc192178540)

[**Age Distribution** 8](#_Toc192178541)

[**Gender distribution** 9](#_Toc192178542)

[**Distribution based on Alcohol consumption** 10](#_Toc192178543)

[**Mutual Information score** 10](#_Toc192178544)

[**Data Cleaning and Preparation** 11](#_Toc192178545)

[**Data Preprocessing** 13](#_Toc192178546)

[**Train/Test split** 13](#_Toc192178547)

[**Machine Learning Algorithms** 14](#_Toc192178548)

[**Naïve Bayes** 14](#_Toc192178549)

[**KNN** 15](#_Toc192178550)

[**Random Forest** 17](#_Toc192178551)

[**Neural Network** 18](#_Toc192178552)

[**Model Comparison** 20](#_Toc192178553)

[**ROC Curve** 21](#_Toc192178554)

[**Conclusion** 23](#_Toc192178555)

[**References** 24](#_Toc192178556)

# **Introduction**

Lung cancer is one of the leading causes of cancer-related deaths worldwide. Early detection of lung cancer can improve treatment outcomes and increase the chances of survival. Patients diagnosed with lung cancer through early CT screening have a 20-year lung cancer-specific survival rate of 81% (Lokaj, 2023). With advancements in technology and data science, machine learning models can help predict the likelihood of lung cancer based on medical data.

This report presents a project focused on building a predictive model for lung cancer using a Jupyter notebook. The main goal of this project is to analyze medical data, apply machine learning techniques, and evaluate the model's performance in predicting lung cancer. The report covers data preprocessing, model selection, evaluation metrics, and the final results of the prediction models.

# **Details of the Dataset**

The dataset used in this project is named **"**Lung Cancer**"**, sourced from Kaggle (Bhat, 2022). It contains medical data related to lung cancer diagnosis, with the goal of predicting whether a patient has lung cancer based on various attributes. The link for the dataset is: <https://www.kaggle.com/datasets/mysarahmadbhat/lung-cancer>

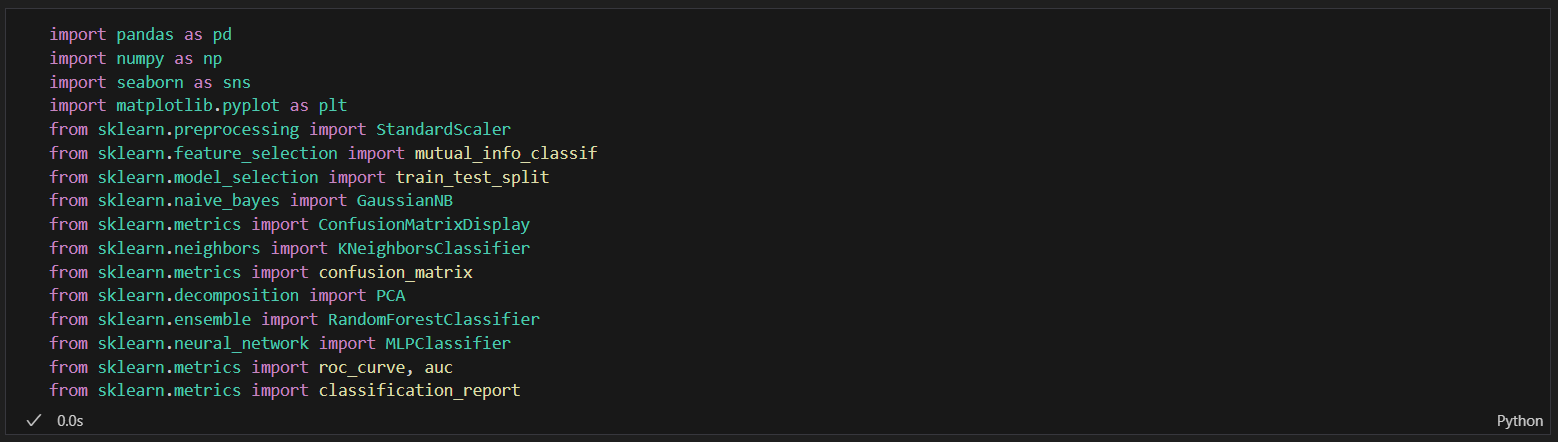
The dataset includes the following key features:

* **Age**: The age of the patient (**Integer**).
* **Gender**: The gender of the patient (**M** or **F**).
* **Smoking**: Indicates whether the patient is a smoker (**1** for **No**, **2** for **Yes**).
* **Yellow fingers**: Whether the patient's fingers are yellow, which can be a sign of heavy smoking (**1** for **No**, **2** for **Yes**).
* **Anxiety**: Whether the patient has anxiety (**1** for **No**, **2** for **Yes**).
* **Peer pressure**: Whether the patient is influenced by peer pressure (**1** for **No**, **2** for **Yes**).
* **Chronic disease**: Indicates if the patient has a chronic disease (**1** for **No**, **2** for **Yes**).
* **Fatigue**: Whether the patient often feels tired (**1** for **No**, **2** for **Yes**).
* **Allergy**: Whether the patient has allergies (**1** for **No**, **2** for **Yes**).
* **Wheezing**: Indicates if the patient experiences wheezing (**1** for **No**, **2** for **Yes**).
* **Alcohol consumption**: Whether the patient consumes alcohol (**1** for **No**, **2** for **Yes**).
* **Coughing**: Whether the patient has a persistent cough (**1** for **No**, **2** for **Yes**).
* **Shortness of breath**: Whether the patient experiences difficulty breathing (**1** for **No**, **2** for **Yes**).
* **Swallowing difficulty**: Whether the patient has trouble swallowing (**1** for **No**, **2** for **Yes**).
* **Chest pain**: Whether the patient has chest pain (**1** for **No**, **2** for **Yes**).
* **Lung Cancer**: The target variable indicating if the patient has lung cancer (**Yes** or **No**).

The dataset contains **309 rows** and **16 columns**, with the **Lung Cancer** column as the target variable for prediction. The data includes both categorical and binary features, which require preprocessing before applying machine learning models.

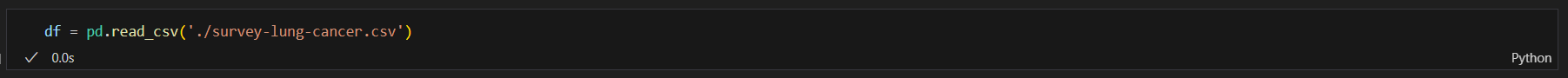
# **Import necessary libraries**

First, important libraries must be imported to the project.



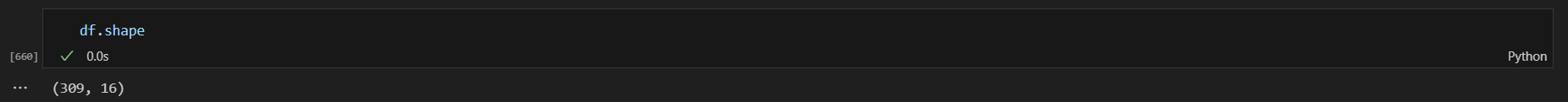
The first step after importing the libraries is to load the dataset into memory as a Data Frame.

A screenshot of a computer

AI-generated content may be incorrect.

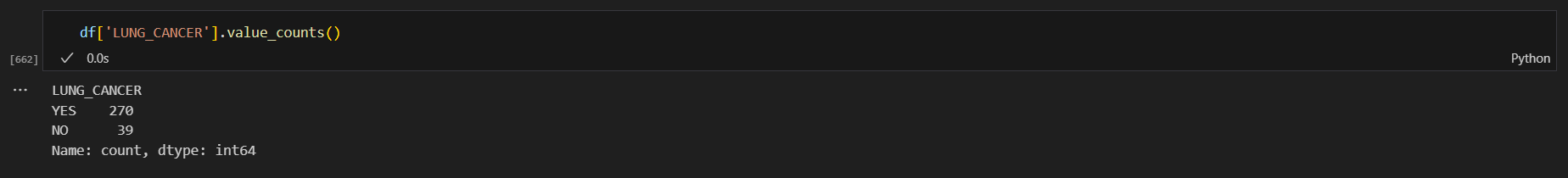
# **Take a look at the data frame specs**

df.shape returns a tuple representing the number of rows and columns in a pandas DataFrame, in the format (rows, columns).



df.info() provides a quick summary of a pandas DataFrame, showing the index, column names, non-null counts, and data types of each column.A black screen with a black border

AI-generated content may be incorrect. See how many people in the data frame suffer from Lung Cancer and how many don't.

 As we can see, 270 people in the data frame suffer from lung cancer and 39 people don't have such disease.

We should always check for null values in the Data frame because missing data can affect data analysis and machine learning models, leading to inaccurate results or errors during processing (Faruk, 2023). A black screen with a black border

AI-generated content may be incorrect. As we can see there is no null value in our data frame.

# **EDA**

EDA techniques help show data in pictures, making it easier to see patterns, unusual values, and connections between different pieces of data (Maitra, 2023).

## **Data Visualization**

Before cleaning or modifying the data, we take a look at the data frame by visualizing them through human-friendly plots to gain a better vision about what we are going to do (Yau, 2013).

### **Age Distribution**

A black screen with colorful text

AI-generated content may be incorrect.

A graph of patients with age distribution

AI-generated content may be incorrect.

The dark blue bars are for all patients, and the light blue bars are just for lung cancer patients. The curved lines show the overall age trends. Most lung cancer patients are between 55 and 75 years old, with the highest number around 60 years old. This helps see if certain ages are more at risk for lung cancer.

### **Gender distribution**

A black rectangular object with white text

AI-generated content may be incorrect.

A graph of different colored rectangular shapes

AI-generated content may be incorrect.

The blue bars are total male and females in the data frame. The red bars are male and females who suffer from the lung cancer.

### **Distribution based on Alcohol consumption**

A black screen with a black background

AI-generated content may be incorrect.

A graph of a bar chart

AI-generated content may be incorrect. The blue bars are total of people who consume alcohol. The red bars are people who consume alcohol and suffer from lung cancer.

### **Mutual Information score**

The mutual information score measures how much information two things share. Mutual information is commonly used in medical imaging to measure the relationship between variables. It is especially useful in image registration, where it helps align images by maximizing shared information (Maes et al., 1997).A black screen with many colorful lights

AI-generated content may be incorrect. A graph with different colored bars

AI-generated content may be incorrect.This means “Allergy” has the most effect on the outcome and “Chest Pain” has the least effect.

# **Data Cleaning and Preparation**

We already checked that we don’t have any null values to be concerned about in our dataset. Instead, we shall make some changes in the dataset to make it more ready for machine learning algorithms. Binary representation is desirable for its memory efficiency (Zhirong Wu, 2015). In our dataset we have “YES/NO”, “1/2” variables that can be converted to simple “0/1” as representatives for “False/True”. Also “Gender” which is defined as “M” for Male and “F” for Female can be mapped to “0” for Female and “1” for Male.

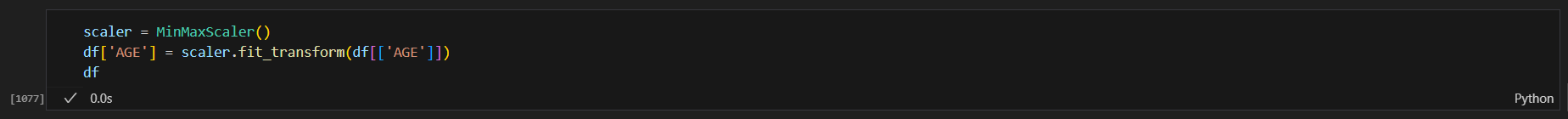
A black screen with a black background

AI-generated content may be incorrect.Now, all our binary features are changed to “0/1” for “False/True” or “Female/Male”

A screenshot of a black and white screen

AI-generated content may be incorrect.

Now we Normalize the “AGE”



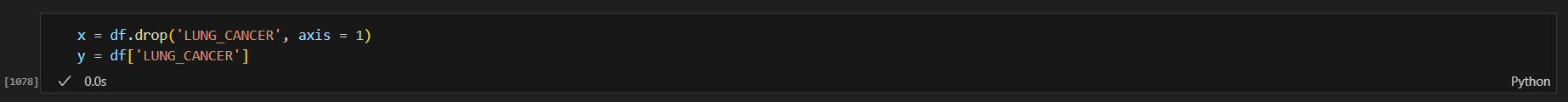
The output will be like this

A screenshot of a computer

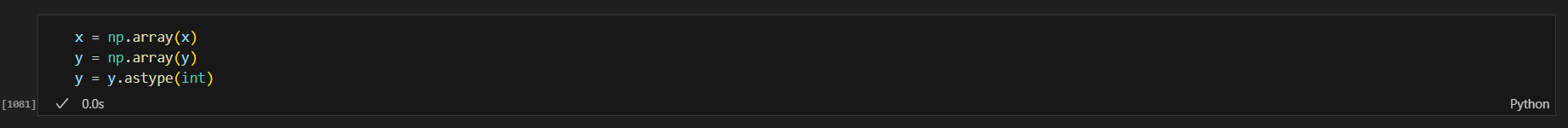
AI-generated content may be incorrect.

# **Data Preprocessing**

Dividing dataset to X for features and Y for answer

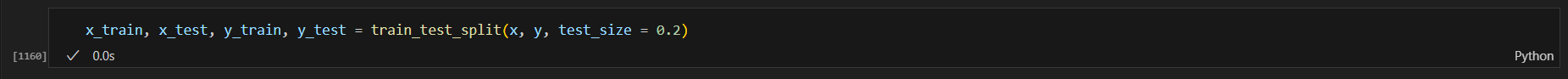
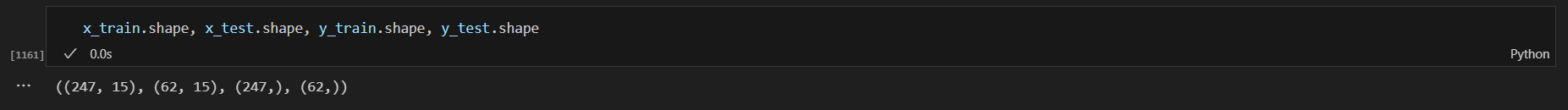


We convert X and Y to NumPy arrays to make them compatible with machine learning libraries like scikit-learn. NumPy arrays are faster and use less memory than lists (Oliphant & Oliphant, 2015). They also allow easy, vectorized operations on data without needing loops.



## **Train/Test split**

Now it’s time to split the data into two parts, one for training the AI model and the other for testing the model. Usually it’s split to %80 for training and %20 for testing. The 80/20 or 70/30 ratio for splitting data is used to ensure enough data for training while still having enough to test the model’s performance. These ratios are common because they balance effective training and testing the model on new data (Hastie, Tibshirani, & Friedman, 2009).

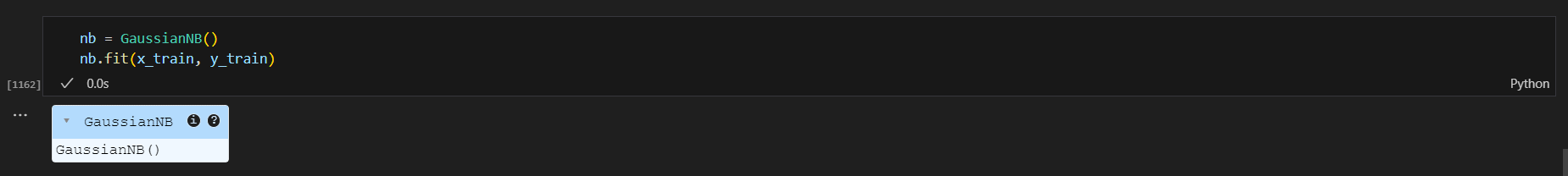
As we can see, 247 rows are designated for training and 62 rows are designated for testing the model

# **Machine Learning Algorithms**

We will try four different ML algorithms to predict Lung Cancer using our dataset. The algorithms are “Naïve Bayes”, “KNN”, “Random Forest” and “Neural Networks”. After training each model we will use a method called classification\_report() to evaluate the accuracy of the model. The classification\_report() gives a summary of a classification model's performance. It shows important details like precision, recall, and f1-score for each class. We will also show the Confusion Matrix for the model. A confusion matrix is a table that shows how well a classification model works. It compares the model’s predictions with the real answers, showing where it was right and where it was wrong. To show the confusion matrix we can use confusion\_matrix() method from scikit-learn library which gives a plain text as output, but we prefer to use ConfusionMatrixDisplay() class from the same library to show it as a plot.

## **Naïve Bayes**

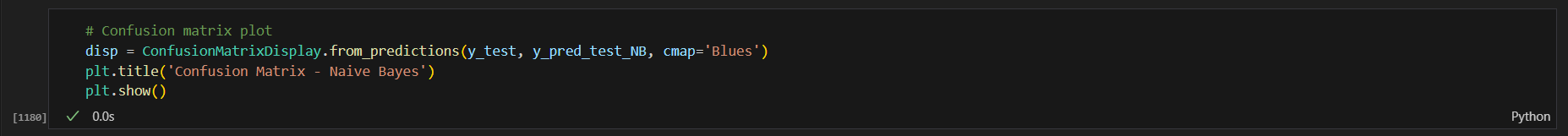
Train the model

****

The performance

A black rectangular object with a black rectangle

AI-generated content may be incorrect.

Confusion Matrix

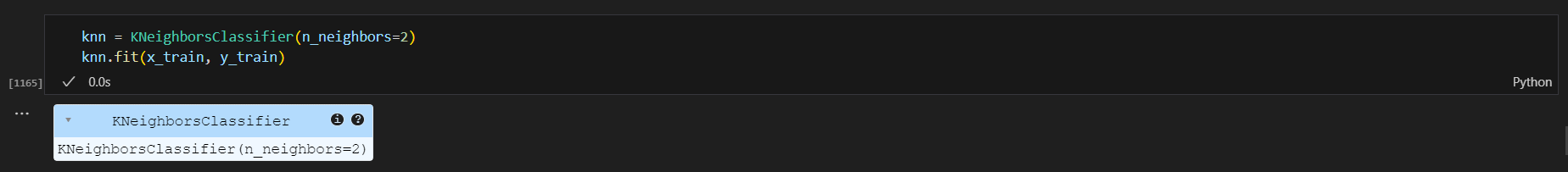
A diagram of a confused matrix

AI-generated content may be incorrect.

## **KNN**

K-Nearest Neighbors (KNN) is an algorithm that predicts a result by looking at the closest examples in the data and choosing the most common answer among them.

Train the model



Performance

A screenshot of a computer

AI-generated content may be incorrect.

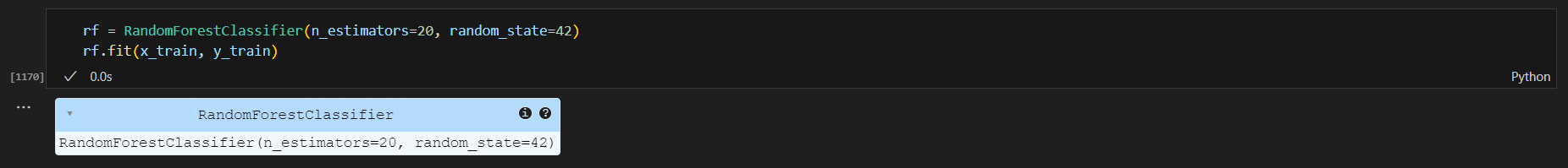
Confusion Matrix

A diagram of confusion matrix

AI-generated content may be incorrect.

## **Random Forest**

Random Forest is an algorithm that makes predictions by combining the results of many decision trees, improving accuracy and reducing errors.



Performance

A screenshot of a computer

AI-generated content may be incorrect.

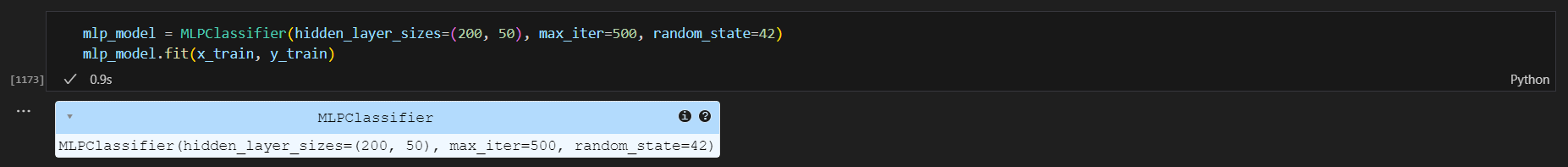
Confusion Matrix

A diagram of confusion matrix

AI-generated content may be incorrect.

## **Neural Network**

A Neural Network (Multi-Layer Perceptron) is an algorithm that mimics how the human brain works, using layers of connected nodes (neurons) to learn patterns and make predictions.



Performance

A screenshot of a computer

AI-generated content may be incorrect.

Confusion Matrix

A diagram of confusion matrix

AI-generated content may be incorrect.

# **Model Comparison**

To evaluate the performance of different machine learning models, I compared several algorithms, including Naive Bayes, K-Nearest Neighbors (KNN), Random Forest, and Neural Network. I used metrics like accuracy, precision, recall and F1-score to measure how well each model performed. By analyzing these results, I was able to identify which model had the best balance of accuracy and reliability for my classification task. This comparison helped me choose the most suitable model for my dataset.

A computer screen with colorful text

AI-generated content may be incorrect.

A graph of different colored bars

AI-generated content may be incorrect.

The results showed that all models had high accuracy, with performance metrics very close to each other. Among them, the KNN model performed slightly better than the others, demonstrating a small advantage in terms of accuracy and evaluation metrics. This suggests that while all models are effective, KNN may be the best choice for this classification task.

## **ROC Curve**

An ROC curve (Receiver Operating Characteristic curve) is a graph that shows how well a classification model can separate the positive and negative classes. It plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at different thresholds. The Area Under the Curve (AUC) tells us how good the model is; the closer to 1, the better the model. An AUC of 0.5 means the model is no better than random guessing.

A graph of a comparison of roc curves

AI-generated content may be incorrect.

# **Conclusion**

While KNN performed slightly better than other models in accuracy, precision, recall, and F1-score, its ROC AUC score was a bit lower than some of the other models. A model can perform well in accuracy, precision, recall, and F1-score, but still have a lower ROC AUC compared to others (Fawcett, 2006). This means that, although KNN is generally a strong performer, other models might do a better job at distinguishing between classes across different thresholds. Overall, KNN is still a good choice, but the ROC AUC suggests that other models might be more consistent in some situations.

# **References**

Bhat, M. A. (2022). *Lung Cancer*. Retrieved 02/05/2025 from <https://www.kaggle.com/datasets/mysarahmadbhat/lung-cancer>

Faruk, O. (2023). *Data preprocessing: Identifying and Handling Null Values, High and Low Cardinality, Leakage, and Multicollinearity*. Retrieved 03/05/2025 from <https://www.linkedin.com/pulse/data-preprocessing-identifying-handling-null-values-high-omar-faruk/>

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861-874. <https://doi.org/https://doi.org/10.1016/j.patrec.2005.10.010>

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics)*.

Lokaj, R. (2023). *Early CT Lung Cancer Screening Significantly Improves Survival in Lung Cancer*. Retrieved 03/05/2025 from <https://www.cancernetwork.com/view/early-ct-lung-cancer-screening-significantly-improves-survival-in-lung-cancer>

Maes, F., Collignon, A., Vandermeulen, D., Marchal, G., & Suetens, P. (1997). Multimodality image registration by maximization of mutual information. *IEEE Trans Med Imaging*, *16*(2), 187-198. <https://doi.org/10.1109/42.563664>

Maitra, M. (2023). *Importance and Impact of Exploratory Data Analysis in Data Science*. Retrieved 03/05/2025 from <https://dzone.com/articles/importance-and-impact-of-exploratory-data-analysis>

Oliphant, T. E., & Oliphant, T. E. (2015). *Guide to NumPy*. Continuum Press. <https://books.google.co.uk/books?id=g58ljgEACAAJ>

Yau, C. (2013). *R Tutorial with Bayesian Statistics Using OpenBUGS*.

Zhirong Wu, D. L., Xiaoou Tang. (2015). Adjustable Bounded Rectifiers: Towards Deep Binary Representations. 11. <https://www.mdpi.com/2076-3417/11/17/7825>