



IT-326

# Lung Cancer Data Mining

Presented By:

**Alanoud Alhudhaif**

**Aleen Aldosari**

**Maraheb Alrashedi**

**Shatha AlSheddey**



# PROBLEM STATEMENT & OBJECTIVES

- Dataset of lung cancer patients from Kaggle.
- Includes demographic, lifestyle, exposure, and medical risk factors.
- Objectives
- Understand the relationship between **risk** factors and **lung cancer**.
- **Build** classification models to predict lung cancer.
- Cluster patients into **meaningful** risk groups.

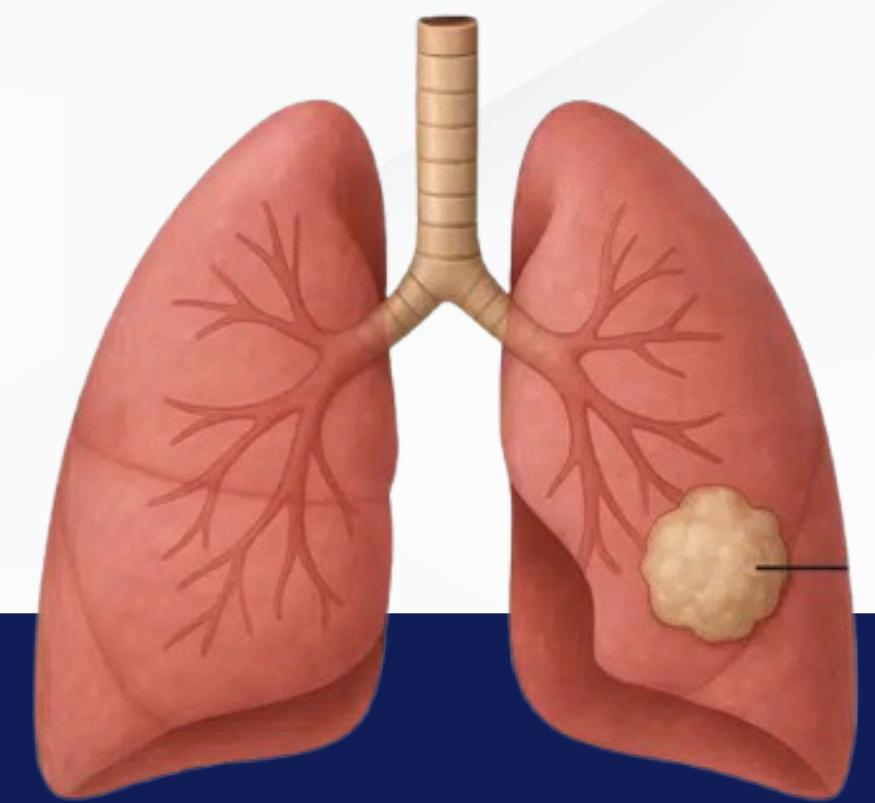


# DATASET DESCRIPTION

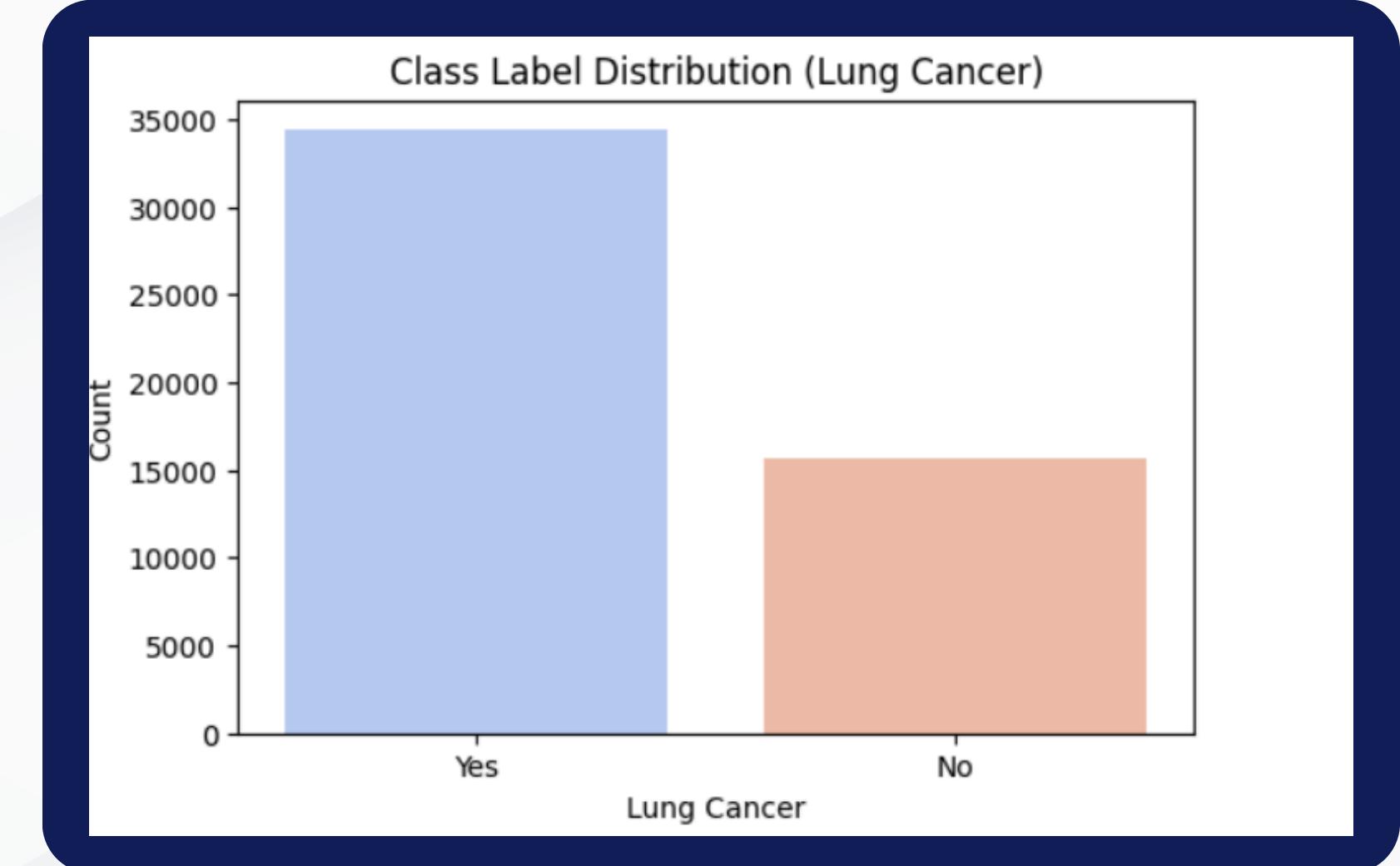
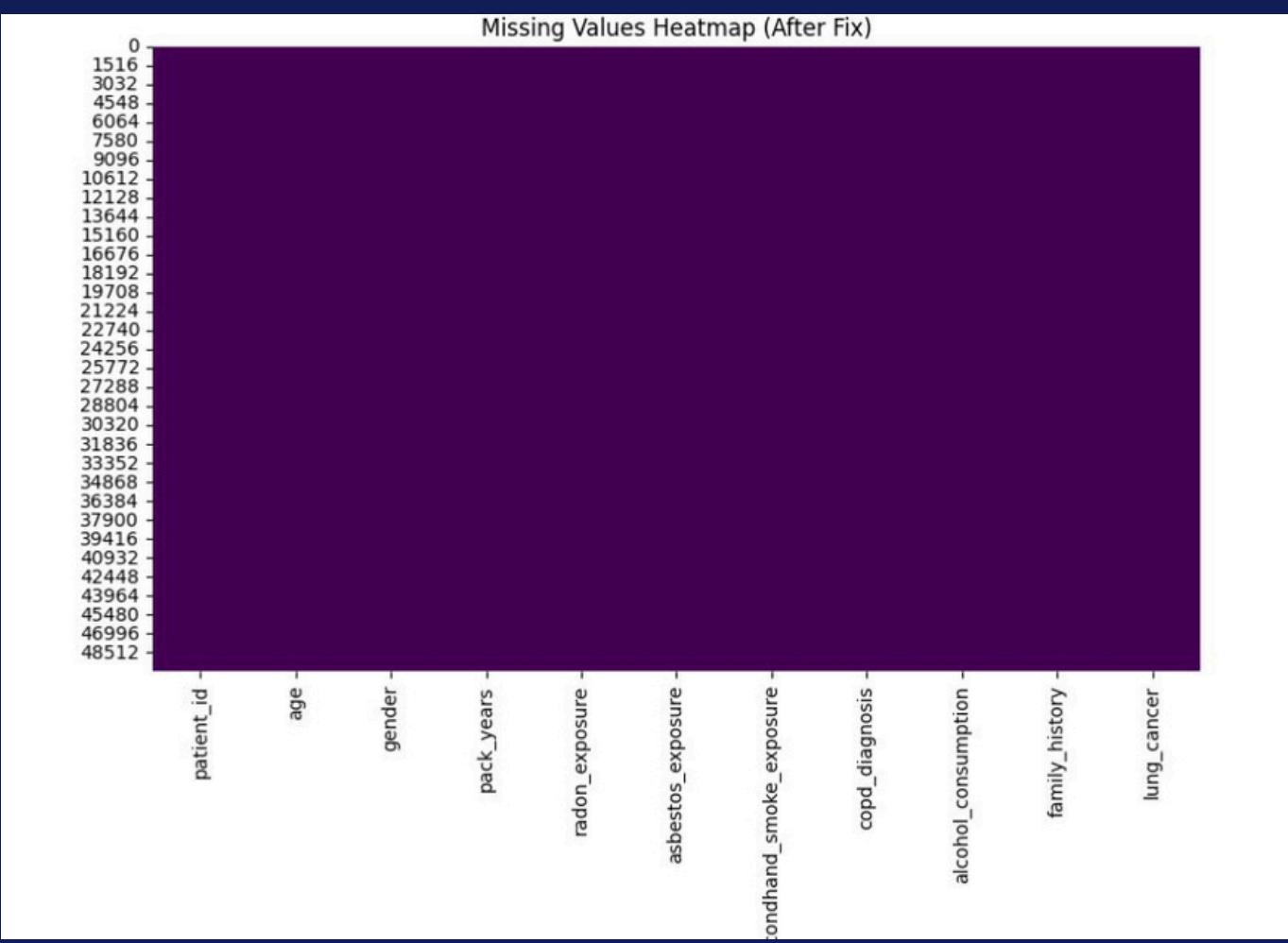
- **50,000** records, **11** attributes
- Target: **lung\_cancer** (Yes / No)

## EXAMPLE FEATURES:

- age, gender
- pack\_years (smoking history)
- radon\_exposure,
- asbestos\_exposure
- secondhand\_smoke\_exposure
- copd\_diagnosis,
- alcohol\_consumption,
- family\_history

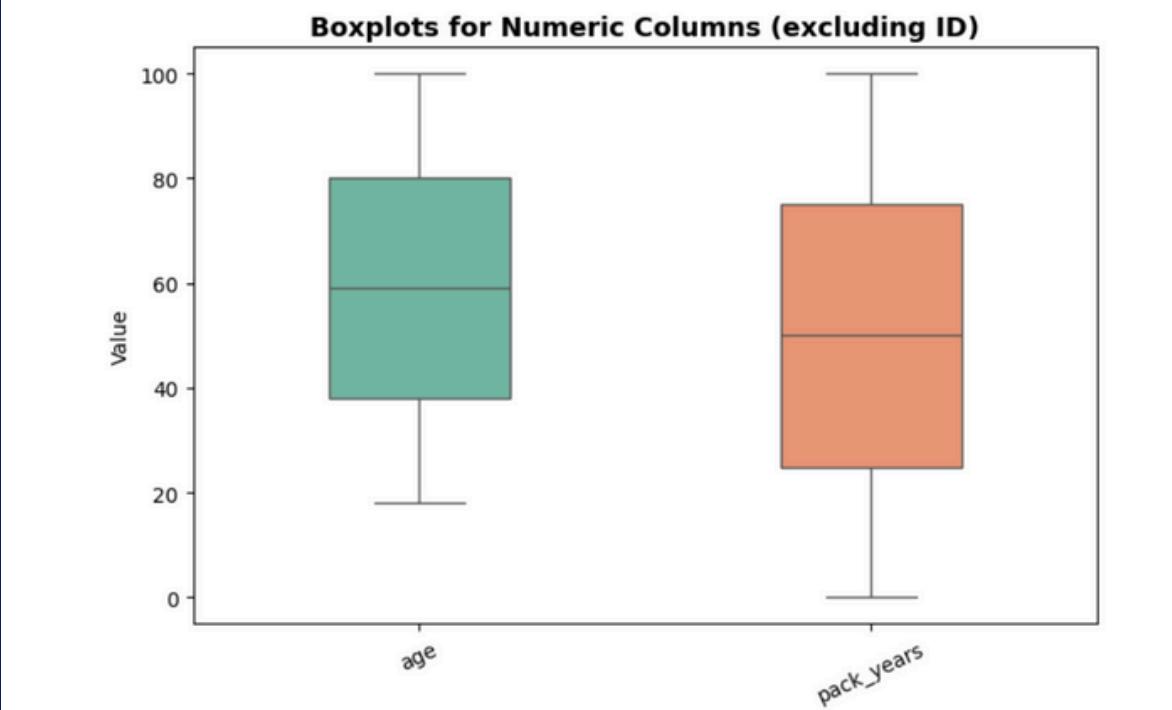
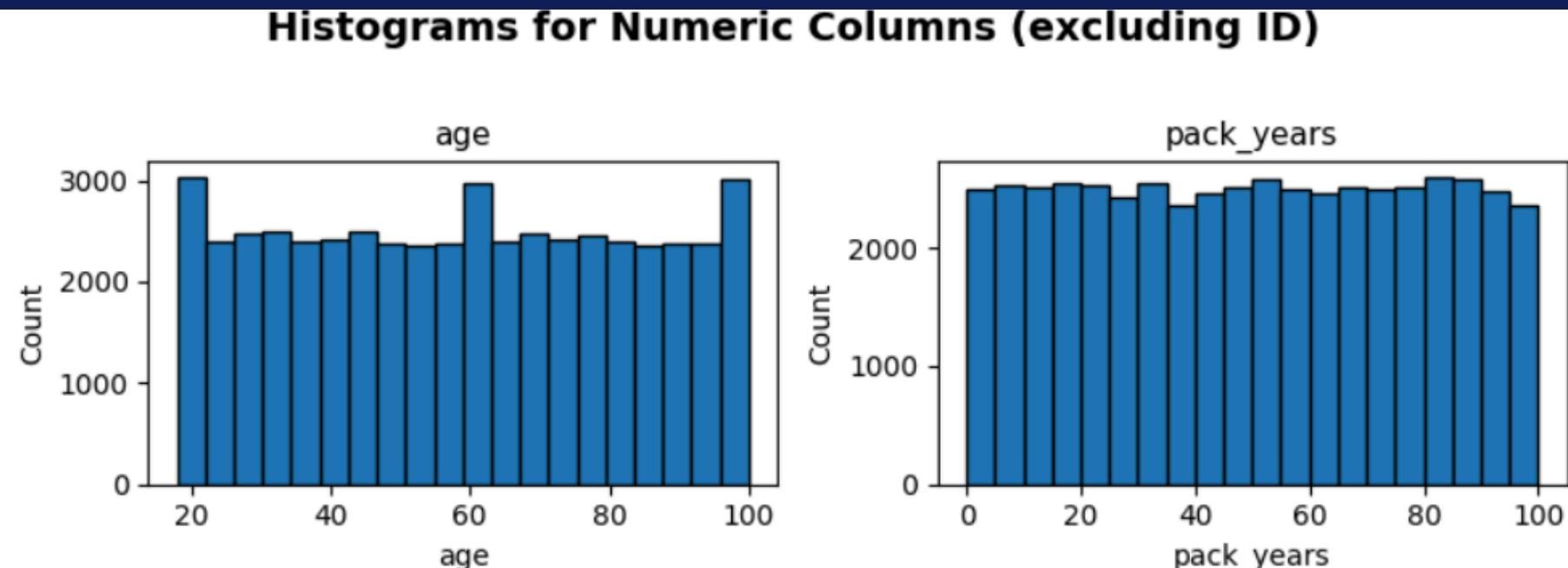


# EXPLORATORY DATA ANALYSIS (EDA)



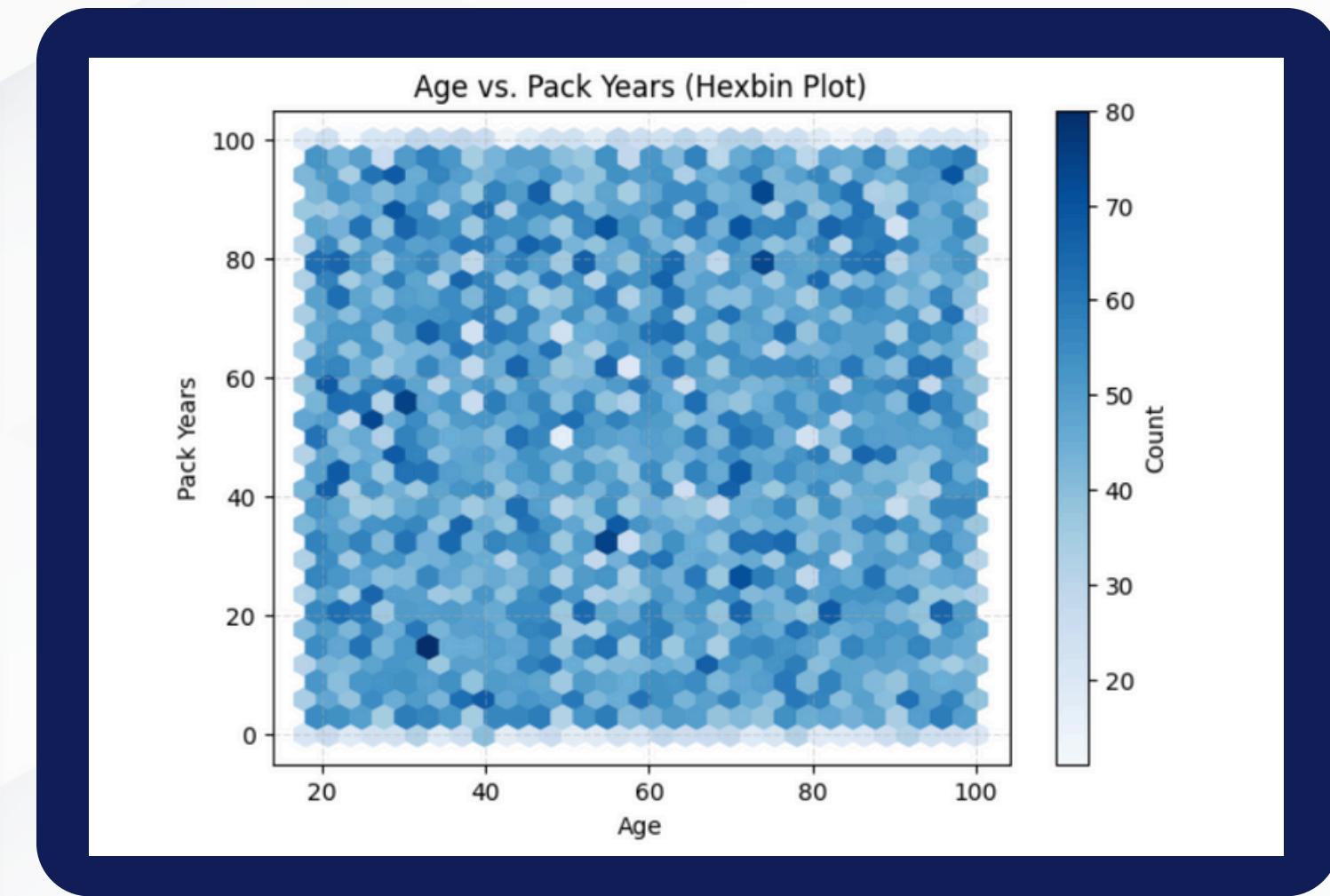
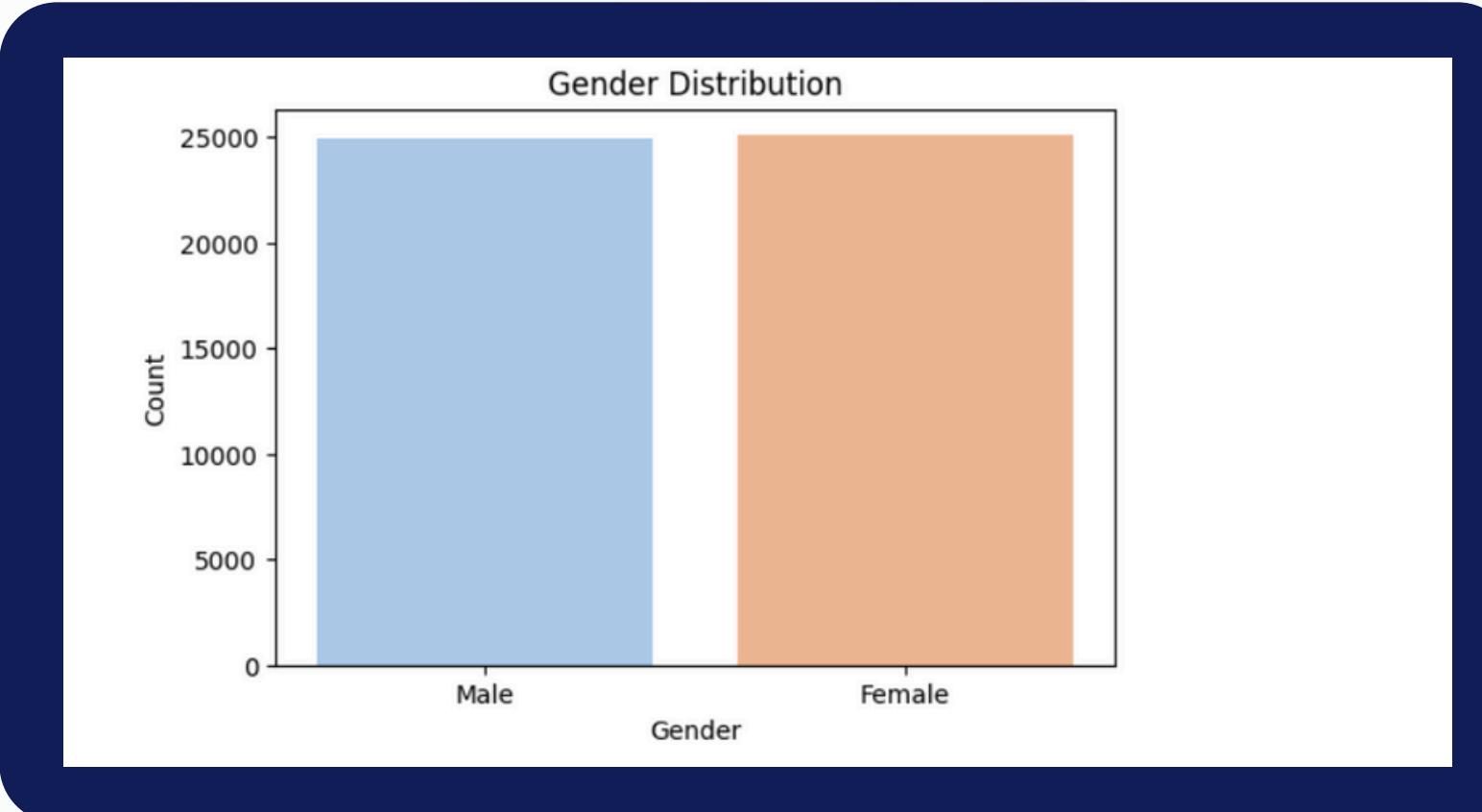
- **No missing** values detected across all features.
- Dataset is **imbalanced**: more “Yes” cases than “No”.

# EXPLORATORY DATA ANALYSIS (EDA)



- Age and **pack\_years** show wide, evenly distributed ranges.
- Both **age** and **pack\_years** show normal spread with **no significant outliers**.”

# EXPLORATORY DATA ANALYSIS (EDA)



- Balanced **gender** distribution (50%-50%).
- **Age** and **pack\_years** appear uniformly distributed, with no clear linear relationship. The density pattern indicates no strong correlation between the two variables.

# PRE-PROCESSING

Several preprocessing techniques were applied to **clean** and **prepare**

We applied **three** preprocessing techniques:

*Expert in financial  
management and strategy*

age	age
69	Senior
32	Young
89	Senior
78	Senior
38	Senior
100	Senior

*Before*      *After*

DISCRETIZATION

lung_cancer	family_hist	alcohol_cc	copd_diag	secondhar	asbestos_exposure	radon_exposure	pack_year	gender
No	No	Moderate	Yes	No	No	High	66.02524	Male
Yes	Yes	Moderate	Yes	Yes	No	High	12.7808	Female
Yes	No	None	Yes	Yes	Yes	Medium	0.408278	Female
Yes	No	Moderate	No	Yes	No	Low	44.06523	Female

*After*

age	gender	family_history	copd_diagnosis	asbestos_exposure	secondhand_smoke_exposure	alcohol_consumption
0	2	1	0	1	0	0
1	0	0	1	1	0	1
2	2	0	0	1	1	1
3	2	0	0	0	0	1
4	0	0	1	1	1	0

ENCODING

pack_year	pack_year
66.02524	0.660248
12.7808	0.127785
0.408278	0.004055
44.06523	0.44064
44.43244	0.444313
81.18055	0.811807
18.15675	0.181546
27.51112	0.275093

*Before*      *After*

NORMALIZATION

# CLASSIFICATION

- Classification is **supervised learning** which mean is need a class label to classify the objects.
- We trained our model to be able to predict if the patient has lung cancer or not using (**lung\_cancer**) class label
- build our model We used a decision tree algorithm which is a **recursive algorithm** produces a tree with a leaf nodes representing the final decisions.

# CLASSIFICATION TECHNIQUES

**1-Data splitting**  
**70/30, 80/20, and 90/10**

**2-Decision Tree Training**  
**Gini Index and Entropy.**

**3-Accuracy Evaluation**  
comparing predicted  
labels with actual labels

**4-Confusion Matrix**  
visualize correct vs.  
incorrect predictions

**5-Decision Tree Visualization**  
We visualized the structure of  
the Decision Tree

# DATA SPLITS

- **Training dataset:** for building the decision tree.
- **Testing dataset:** to evaluate the model.

01

90% Training  
10% Testing

0.6378  
0.6346

02

80% Training  
20% Testing

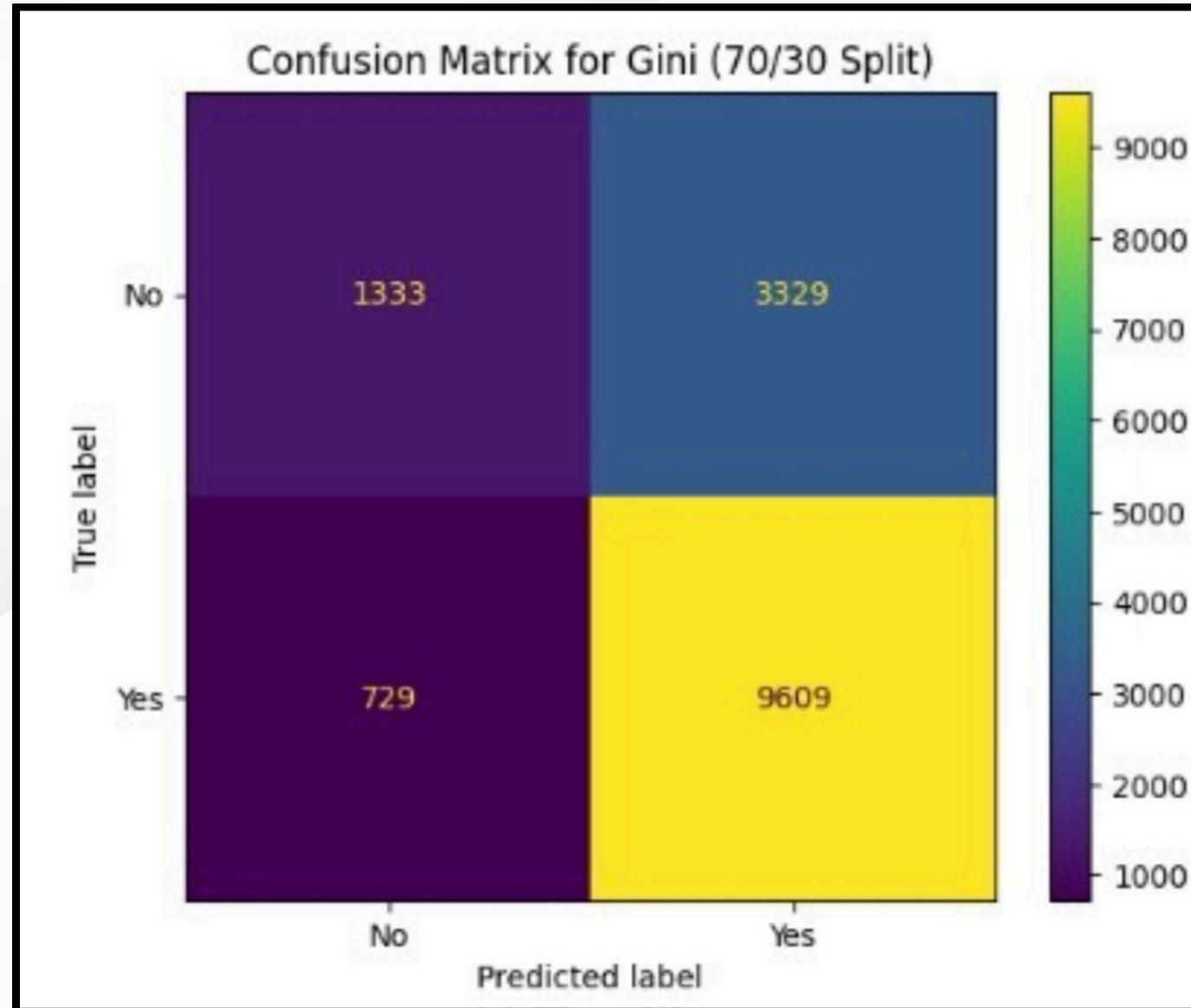
0.6488  
0.6487

03

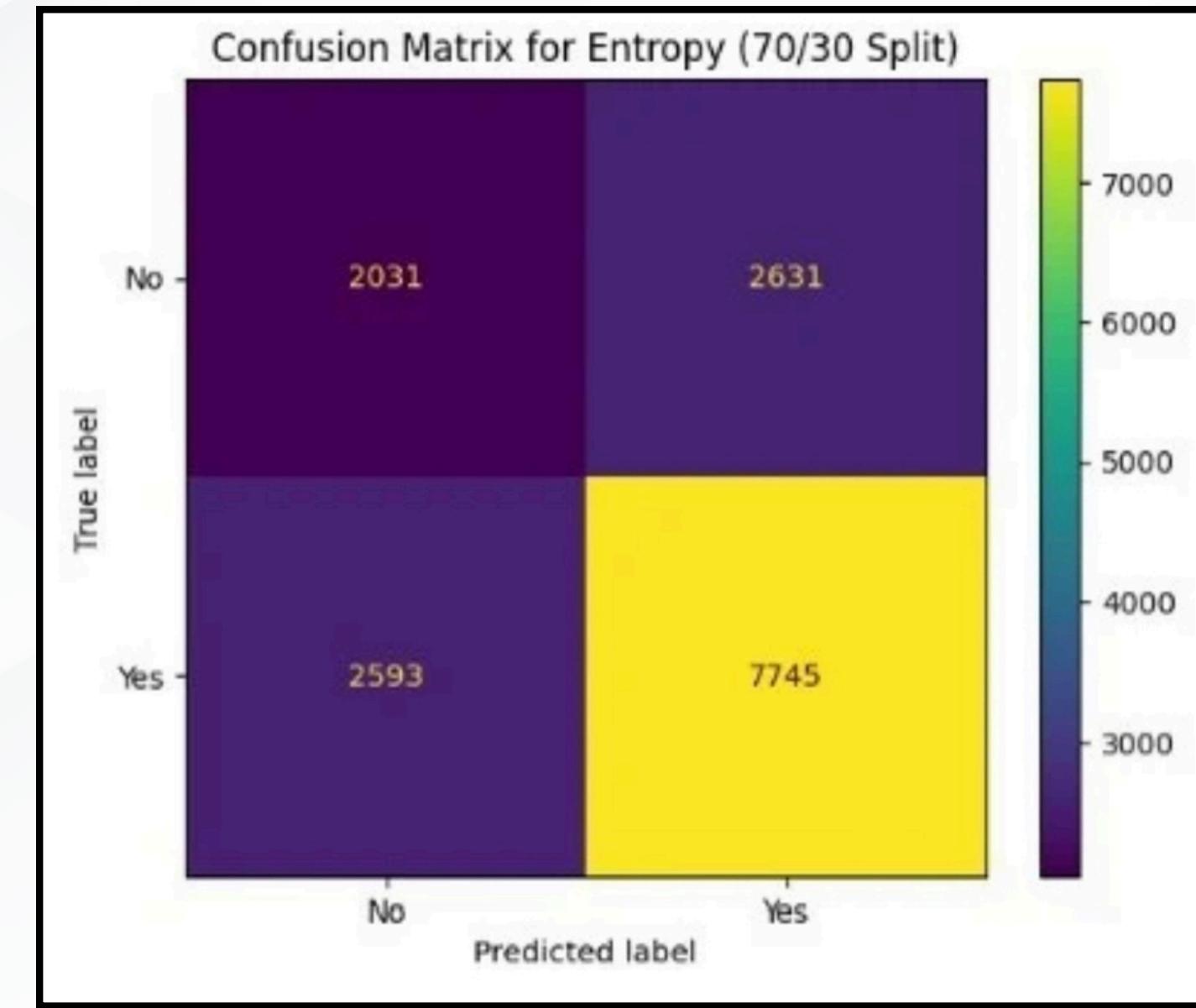
70% Training  
30% Testing

0.7294  
0.6517

# CONFUSION MATRIX



- True Positive (TP): 9609
- False Negative (FN): 729

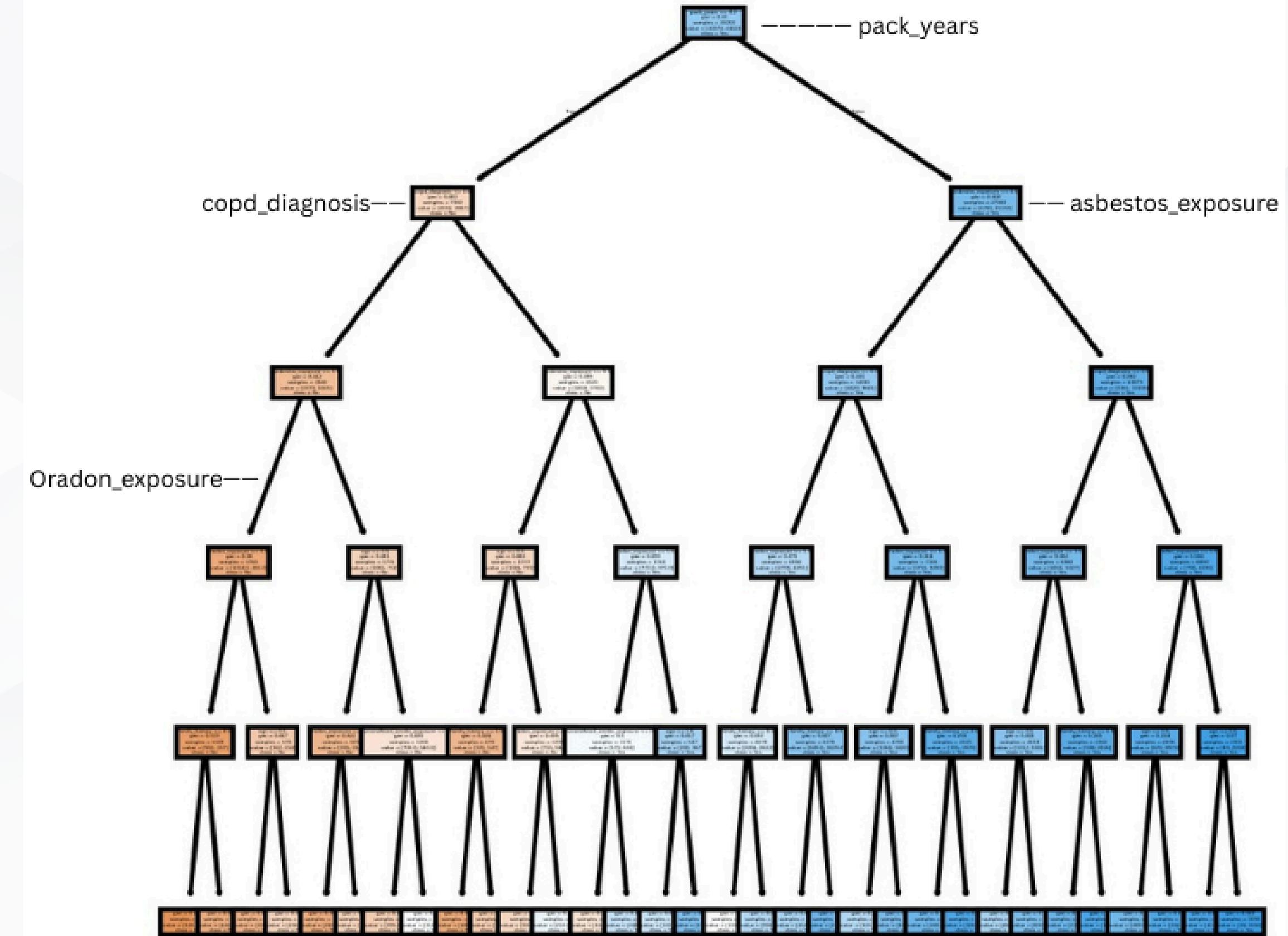


- True Positive (TP): 7745
- False Negative (FN): 2593

# DECIDECISION TREE

Final Decision Tree based on  
the best-performing model  
**(Gini, 70/30 split).**

This split produced the  
highest accuracy, and the tree  
shows the most important  
features influencing lung  
cancer prediction.



# CLUESTRING

- Clustering is an **unsupervised** learning technique that groups patients based on similarity in their risk-factor attributes.
- the model groups patients who share similar characteristics such as smoking level, radon exposure, asbestos exposure, and other medical risk factors.
- After forming these clusters, the grouping structure can help identify patterns among patients and support future analysis or prediction for new patients with similar risk profiles.

# CLUESTRING TECHNIQUES

## 1-Feature Selection

Remove non-useful  
columns

## 4- PCA Visualization

Reduce dimensions for  
3D visualization

## 2-Standardization

Scale all features to  
same range

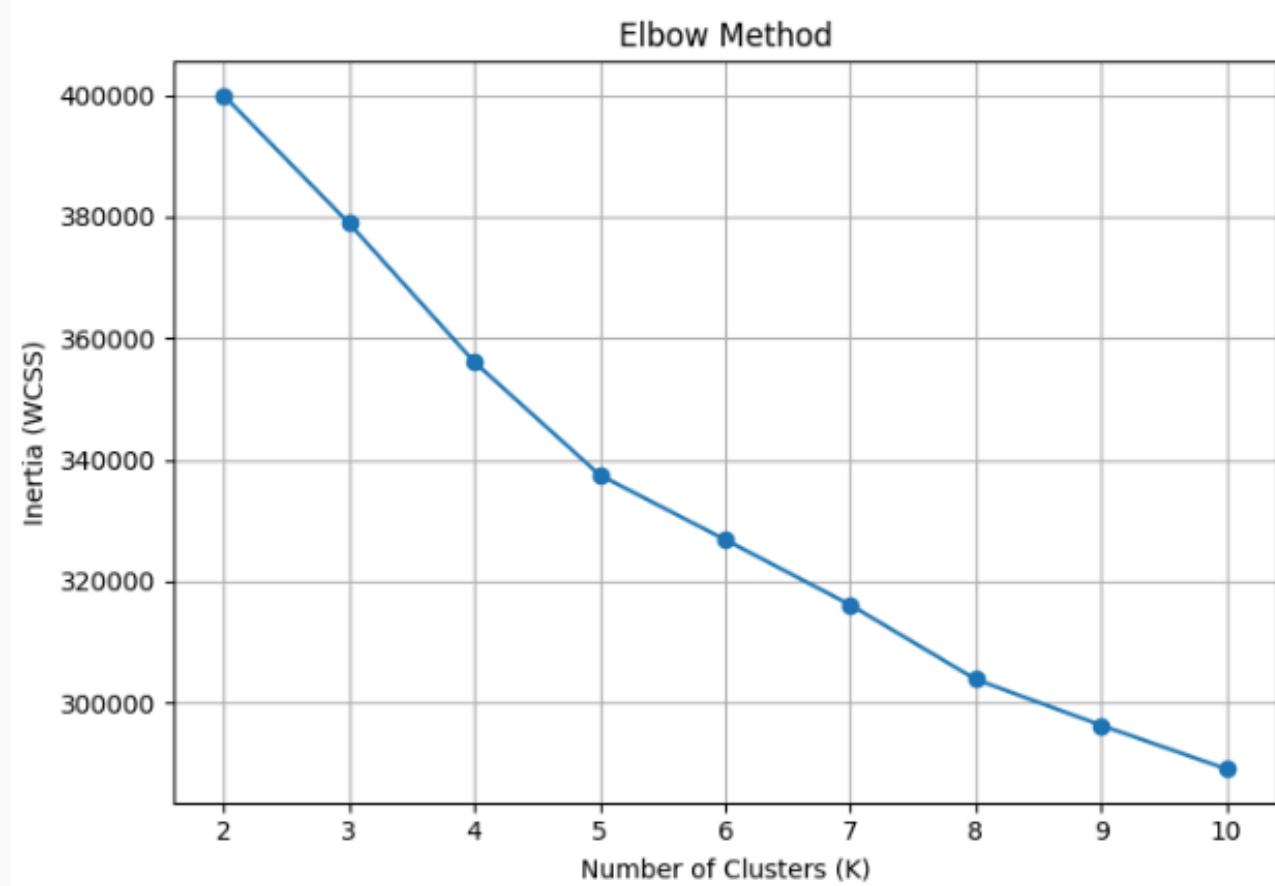
## 3- K-Means Clustering

Group patients based on  
similarity

# CLUESTRING

## Elbow & Silhouette Evaluation

- **Elbow** Method
- No sharp “elbow” point
- **K = 4** and **K = 6** show reasonable stability
- **K = 8** achieves the lowest WCSS → strongest compactness



- **Silhouette** Score
- **K = 8** gives the highest silhouette value
- Indicates the best separation between clusters
- Confirms **K = 8** as the optimal choice
- Conclusion: Both metrics support choosing **K = 8**.



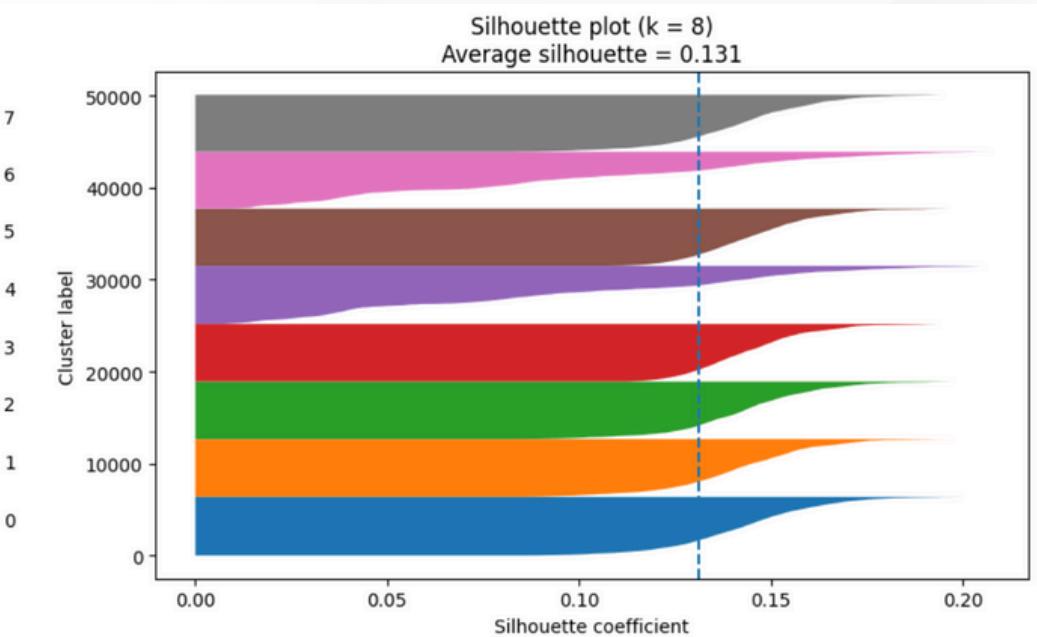
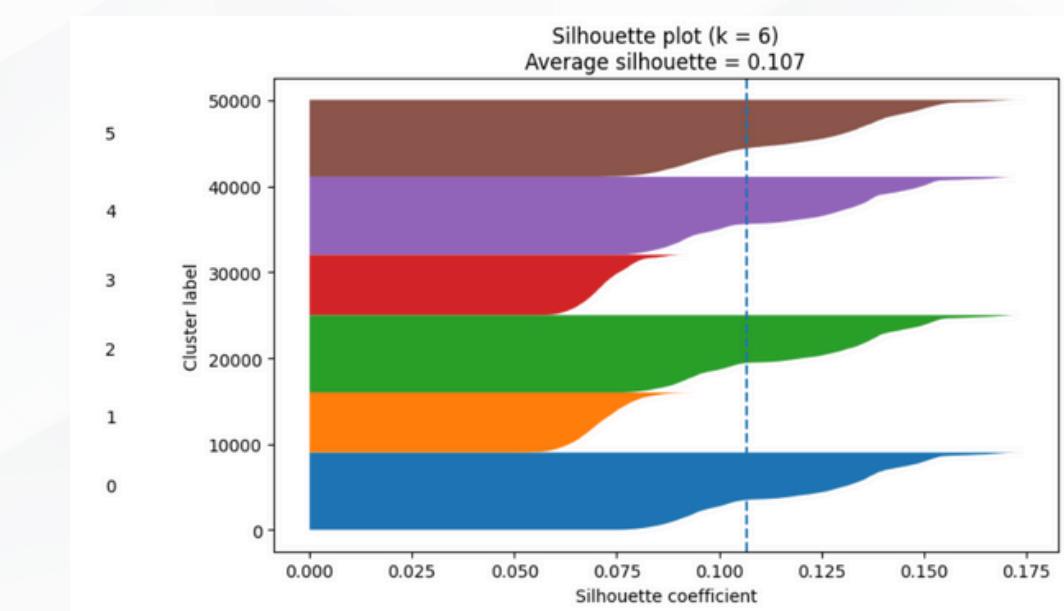
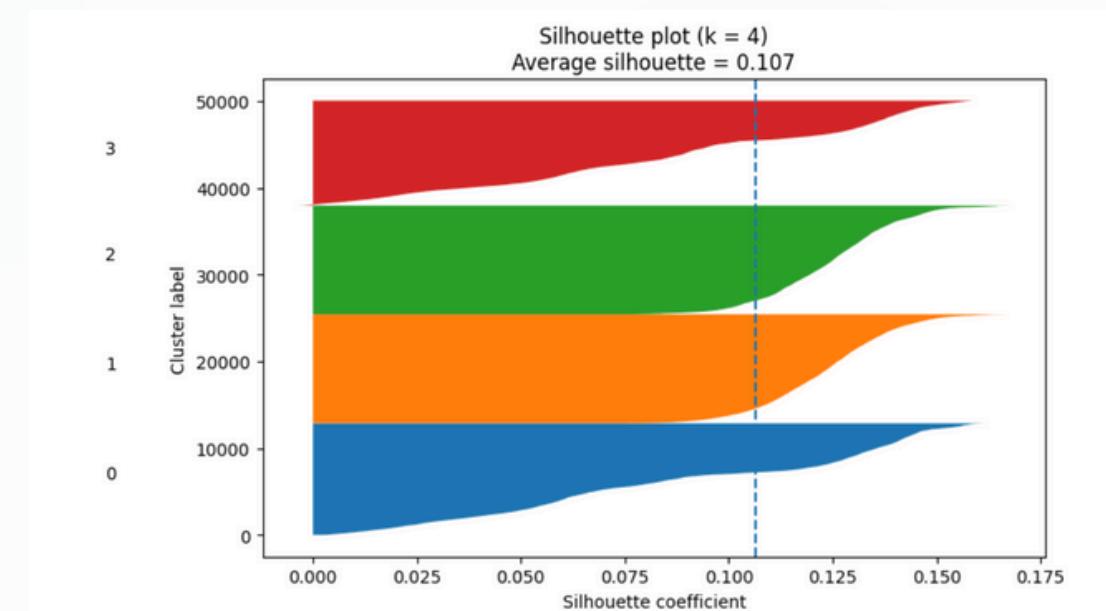
# CLUESTRING

## Silhouette Plots for Each K

silhouette analysis helps evaluate how well-separated and compact the clusters are.

Below are the silhouette plots for K = 4, K = 6, and K = 8.

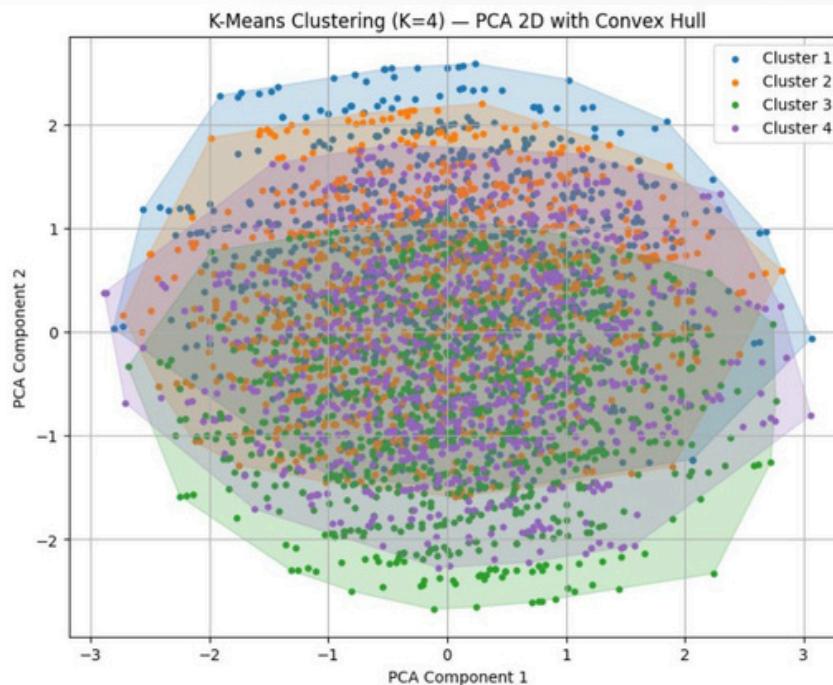
- K = 4: Average silhouette  $\approx 0.106$
- K = 6: Average silhouette  $\approx 0.107$
- K = 8: Average silhouette  $\approx 0.131$  (highest, best separation)



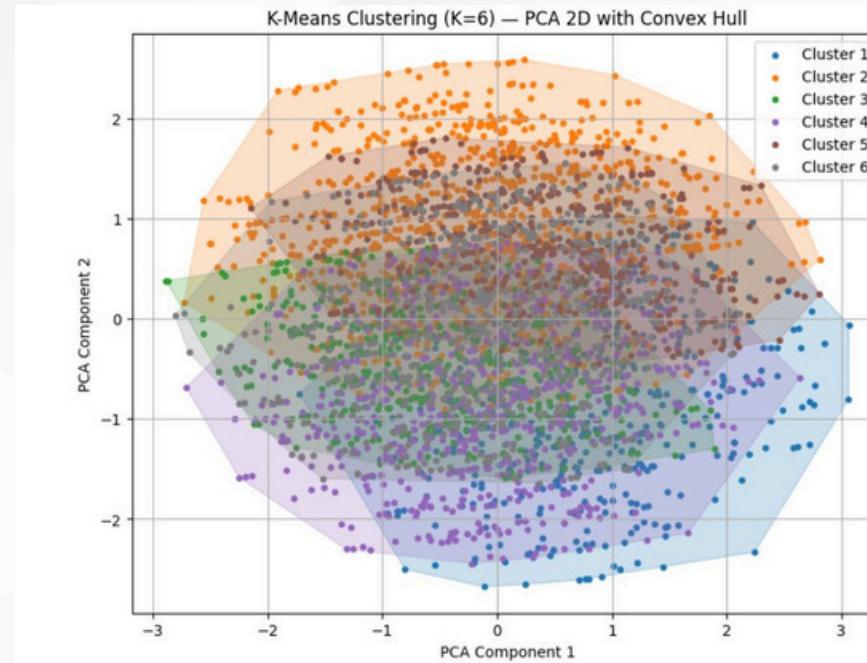
# CLUESTRING

## PCA CLUSTER VISUALIZATION

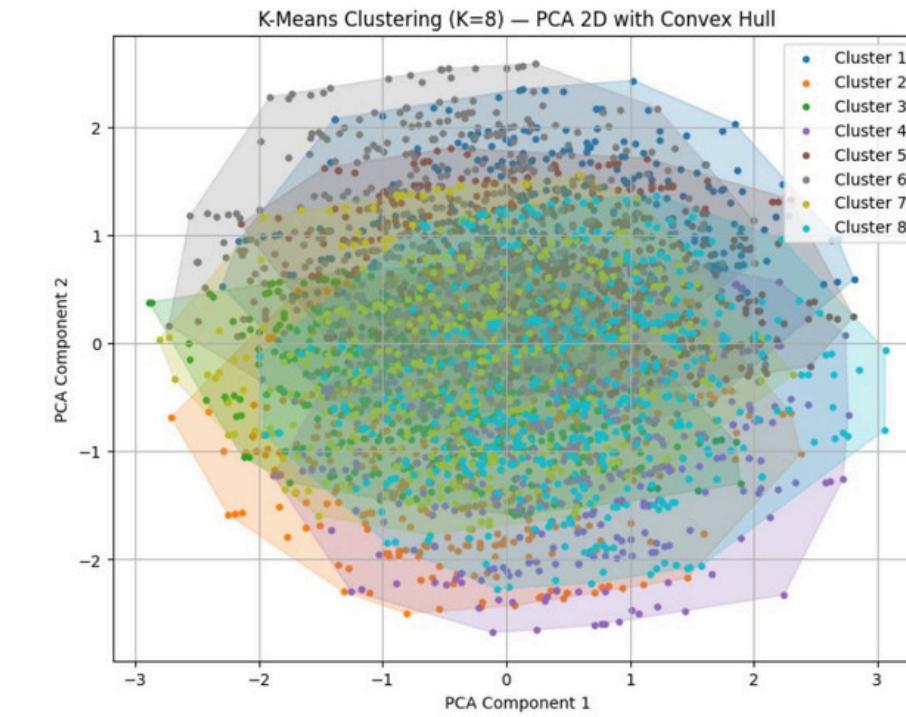
PCA (3D) WAS USED TO VISUALLY COMPARE CLUSTERS FOR K = 4, 6, AND 8.



K = 4  
VERY BROAD CLUSTERS  
SIGNIFICANT OVERLAP  
WEAK SEPARATION



K = 6  
SLIGHTLY MORE DEFINED STRUCTURE  
OVERLAP STILL PRESENT  
MODERATE IMPROVEMENT



K = 8 (MOST CLEAR)  
MOST DEFINED CLUSTER BOUNDARIES  
BETTER GROUPING AND LOWER OVERLAP  
CONVEX HULL VISUALIZATION SHOWS  
CLEARER SEGMENTATION

FINAL : K = 8 PROVIDES THE MOST STRUCTURED, COMPACT, AND INTERPRETABLE CLUSTERING.

# FINDINGS

- The **70/30 split** gave the highest accuracy across all models.
- Gini outperformed Entropy, achieving the best accuracy (0.729).
- Both models struggled more with predicting the “Yes” class, likely due to feature overlap.
- For clustering, K-means with K=8 produced the best separation (highest silhouette score), **PCA** visualizations confirmed.
- Overall, results show that multiple factors—not a single feature—drive lung cancer risk.
- Most influential features: COPD diagnosis and smoking-related exposures.



# THANK YOU

