Lung Cancer Risk Prediction – Report

Thumma Hasini

Batch – 37

2303A52076

Key Findings

Class distribution (counts): Yes: 34364, No: 15636

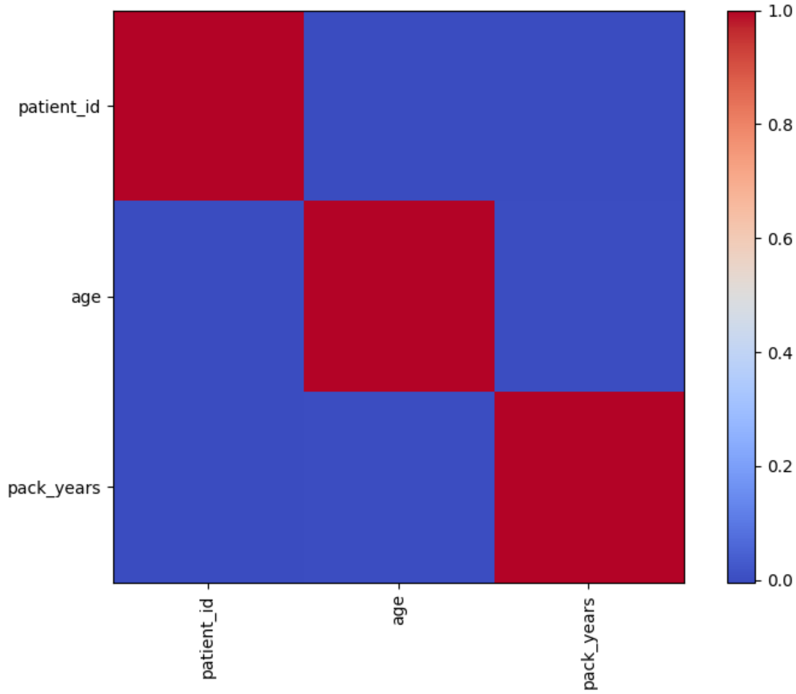
Top features by importance: pack_years (0.316); patient_id (0.228); age (0.227); radon_exposure (0.056); alcohol_consumption (0.046).

Best model by F1: 5 (F1=0.774).

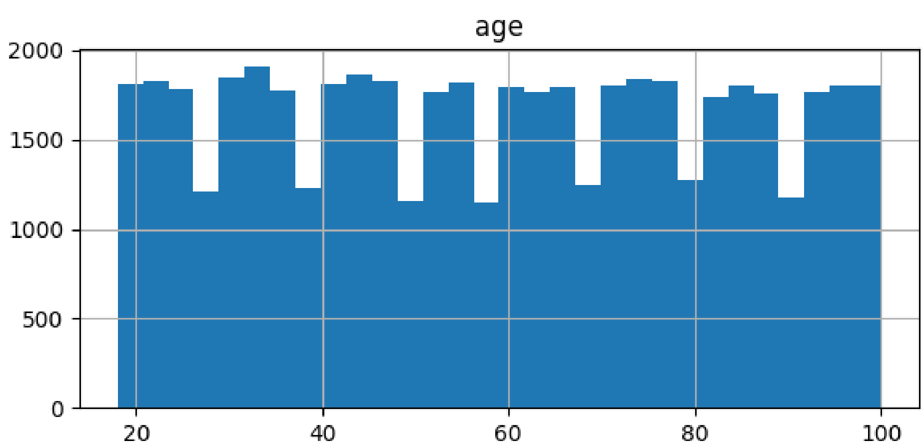
Comparison of Model Performance

Model	Accuracy	Precision	Recall	F1	ROC-AUC
0	0.673	0.820	0.671	0.738	0.741
1	0.629	0.755	0.680	0.716	0.598
2	0.686	0.790	0.739	0.764	0.733
3	0.669	0.840	0.641	0.727	0.750
4	0.633	0.781	0.648	0.708	0.670
5	0.701	0.805	0.746	0.774	0.752

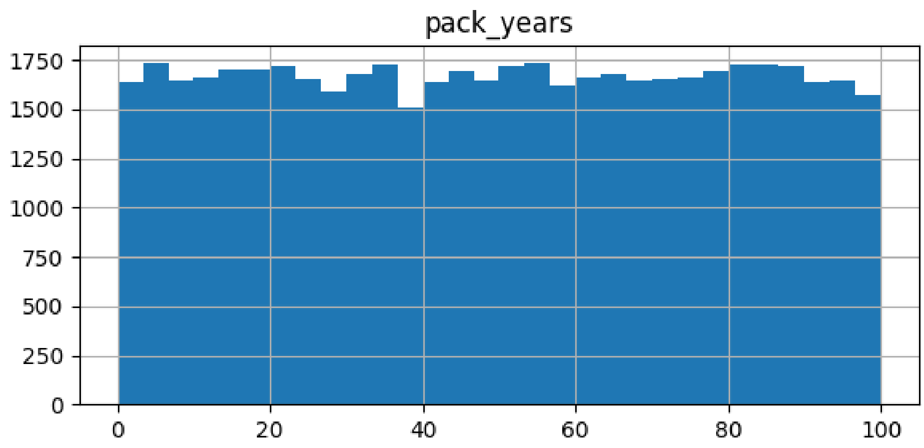
Visualizations



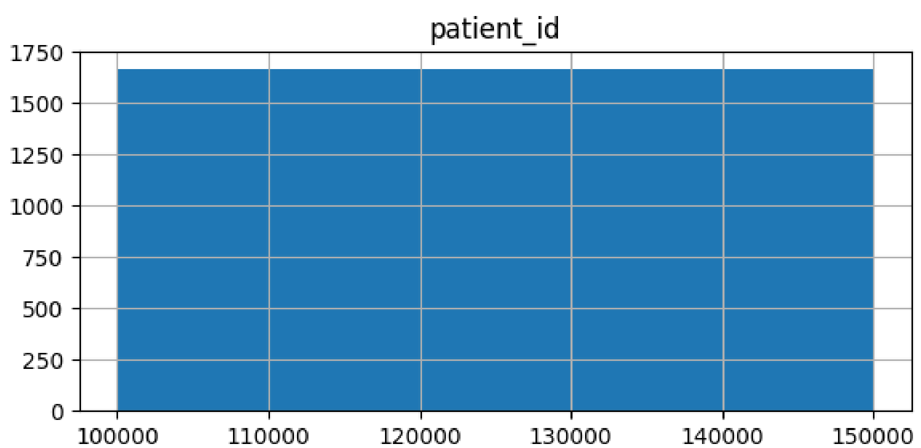
correlation_heatmap.png



hist_age.png



hist_pack_years.png



hist_patient_id.png

Insights from XAI Visualizations

Top predictors of lung cancer risk include:

- pack_years: importance 0.316

- patient_id: importance 0.228

- age: importance 0.227

- radon_exposure: importance 0.056

- alcohol_consumption: importance 0.046

- asbestos_exposure: importance 0.033

- copd_diagnosis: importance 0.030

- gender: importance 0.023

- family_history: importance 0.020

- secondhand_smoke_exposure: importance 0.019

Final Recommendations

1. Prioritize interpretability alongside performance for clinical deployment.

2. Tree-based models (Random Forest, Gradient Boosting) are recommended for their balance of accuracy and explainability.

3. Use SHAP explanations to provide per-patient insights on risk drivers.

4. Validate models on external datasets before real-world usage.