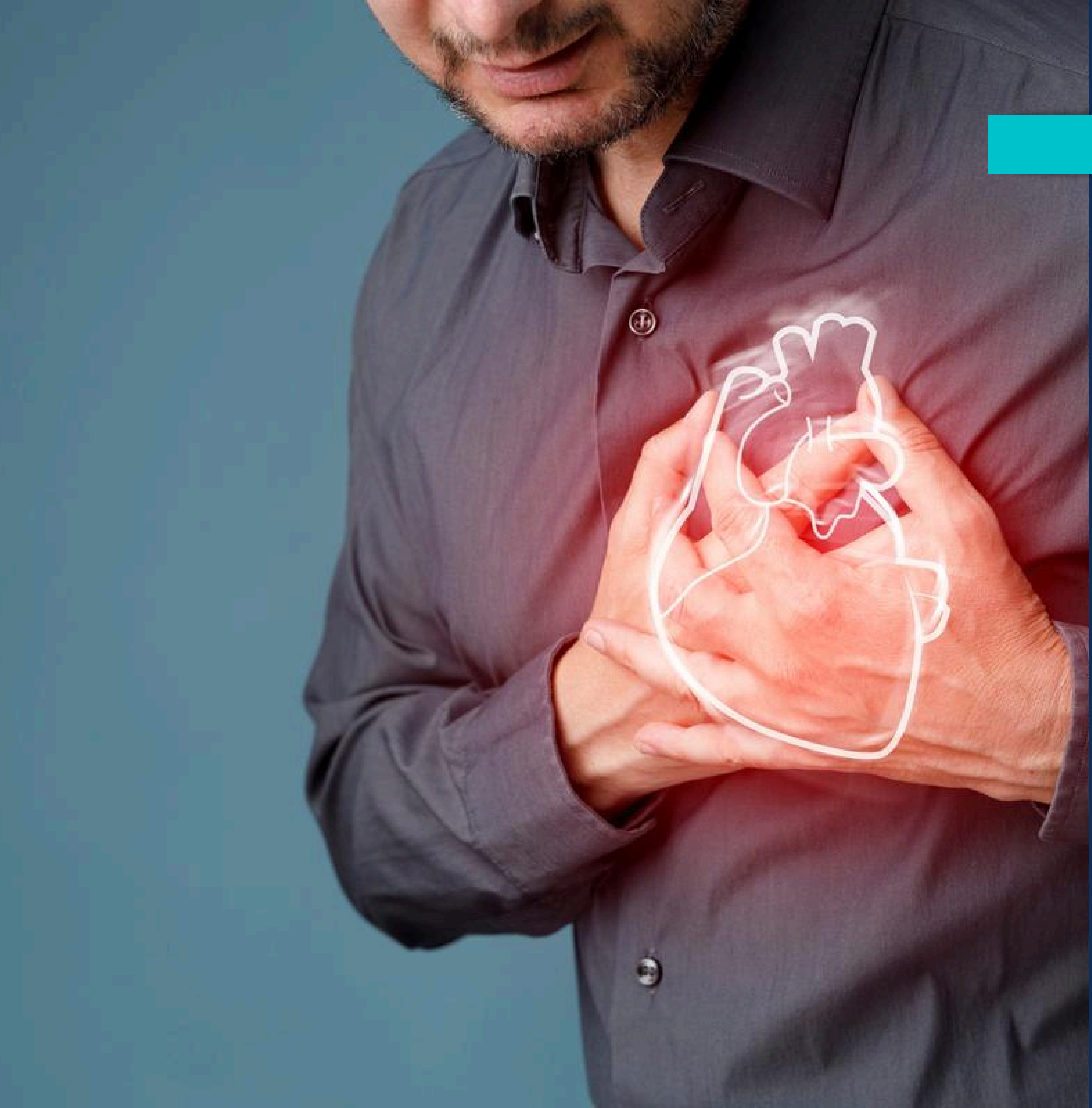
ANALÍSIS PREDICTIVO
FINAL EXAM

Heart Disease Prediction

Thea Boge



Business Case

Heart disease is a leading cause of mortality, however early detection drastically improves outcomes.

Objective

- Predict likelihood of heart disease based on routine clinical measurements
- Support earlier diagnosis and more efficient triage
- Provide doctors with a probability-based decision-support tool

Dataset Summary

raw_merged_heart_dataset.csv from Kaggle

- 2181 rows, 14 features
- Binary target: presence of heart disease
  - target (1 = heart disease)
  - Patients without heart disease: 1099
  - Patients with heart disease: 1082

→ balanced dataset

Cleaning:

- Columns standardized (lowercase, trimmed)
- Missing-value placeholders replaced with NaN
- All object-like numeric columns converted to float
- Define numerical and categorical columns

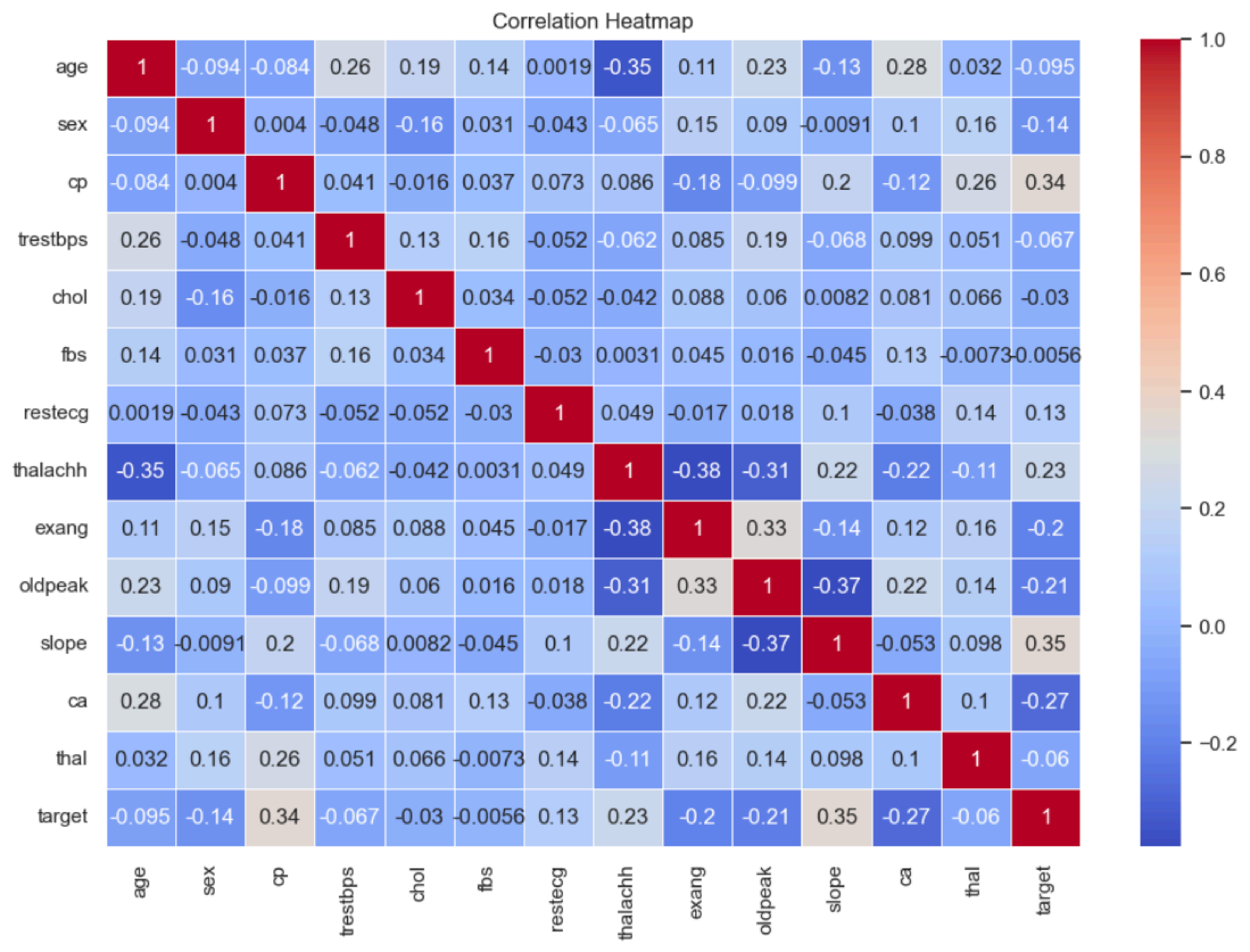
Column	Description	Data Type	Feature Type
age	Age in years.	int64	Numerical
sex	Biological sex (0 = female, 1 = male).	int64	Categorical
cp	Chest pain type (0–3): typical, atypical, non-anginal, asymptomatic.	int64	Categorical
trestbps	Resting blood pressure (mm Hg).	float64	Numerical
chol	Serum cholesterol (mg/dL).	float64	Numerical
fbs	Fasting blood sugar >120 mg/dL (1 = true, 0 = false).	float64	Categorical
restecg	Resting ECG results (0–2).	float64	Categorical
thalachh	Maximum heart rate achieved during exercise.	float64	Numerical
exang	Exercise-induced angina (1 = yes, 0 = no).	float64	Categorical
oldpeak	ST depression induced by exercise relative to rest.	float64	Numerical
slope	Slope of the ST segment during peak exercise (0–2).	float64	Categorical
ca	Number of major vessels (0–3) visible under fluoroscopy.	float64	Categorical
thal	Thalassemia status (1–3): normal, fixed defect, reversible defect.	float64	Categorical
target	Heart disease diagnosis (1 = disease, 0 = no disease).	int64	Target

Exploratory Data Analysis

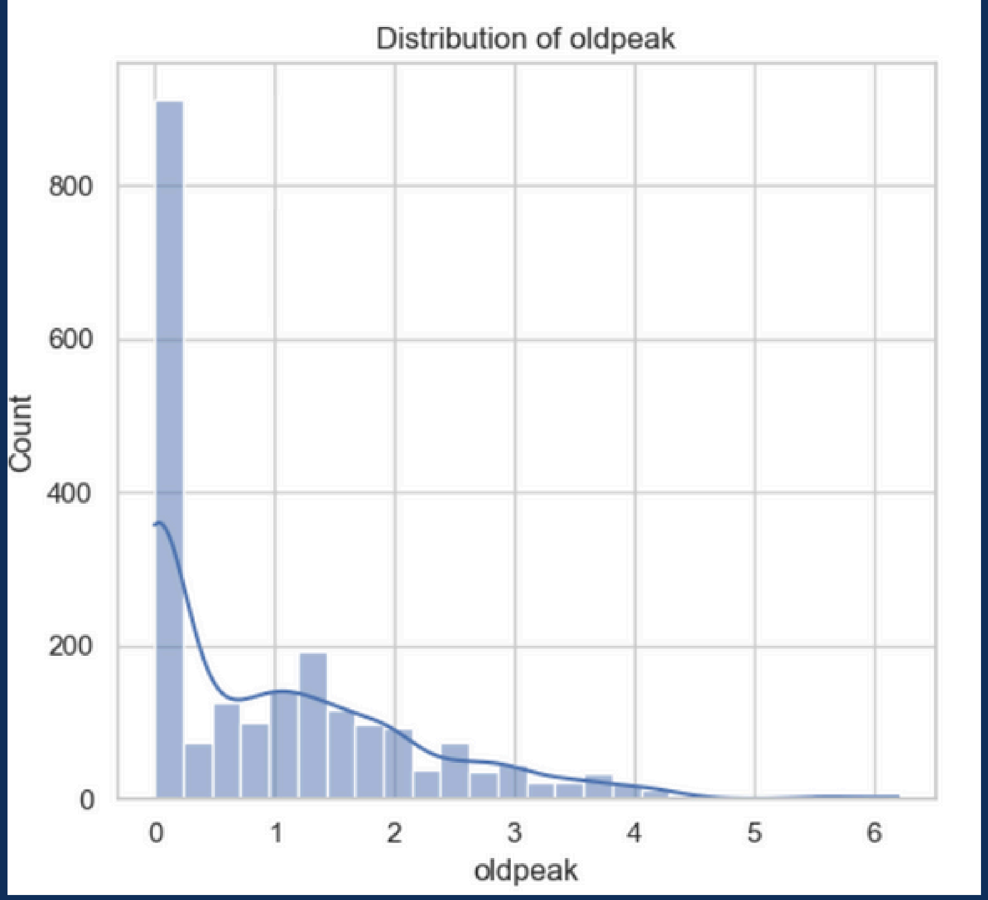
Weak linear correlations between features
→ suggests non-linear patterns

Target also shows weak correlations
→ linear models expected to underperform

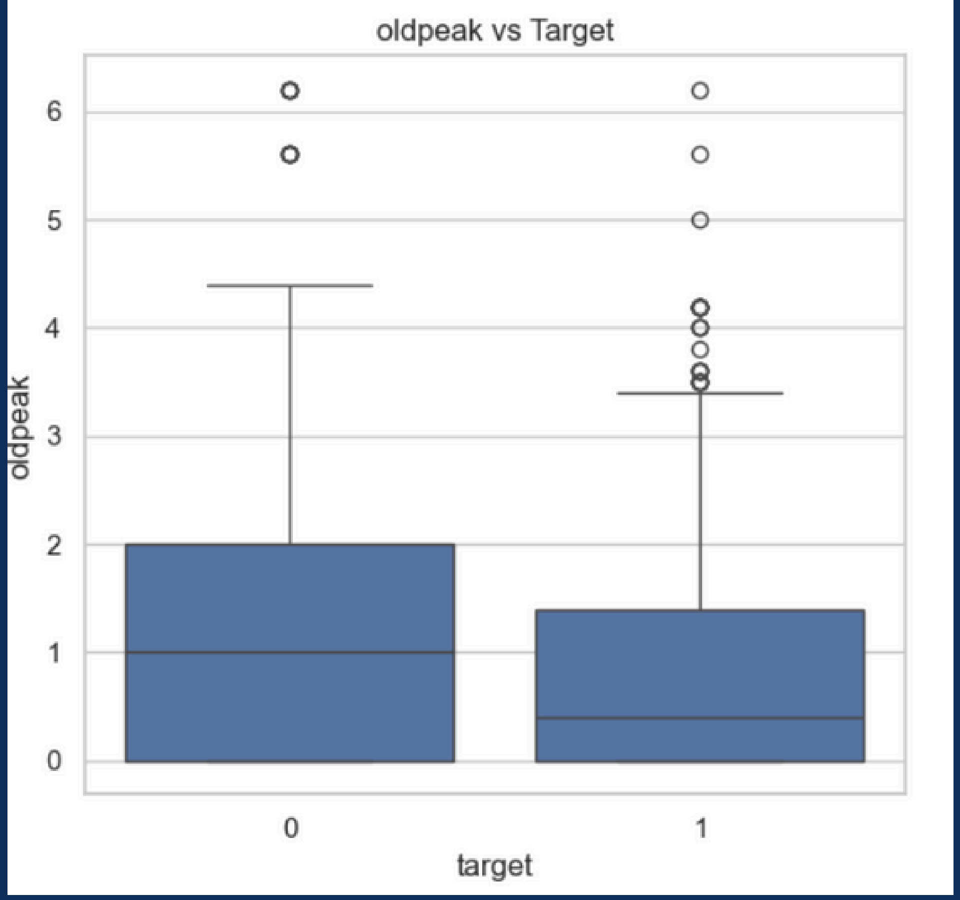
Tree-based models handle these relationships better



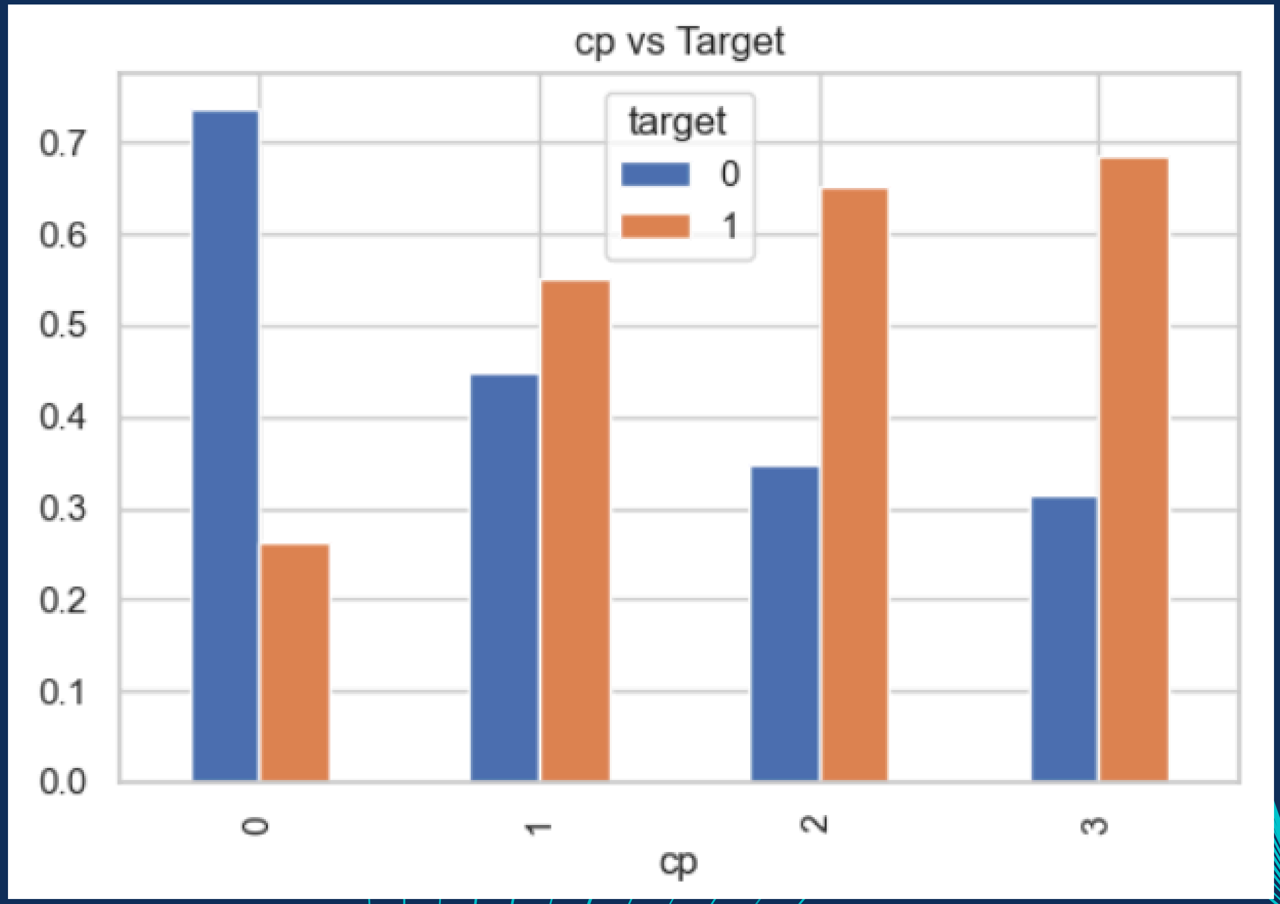
Exploratory Data Analysis



Skewed distributions → trees handle without scaling



Outliers
→ trees robust to outliers; linear models are sensitive



Categorical differences → tree models split effectively on categorical features

→ Tree-based models are ideal

Preprocessing

Missing data:

- Low overall missingness
- Three variables have 8–13% missing

Baseline Preprocessing (preprocess())

- Numerical: median imputation + standard scaling
- Categorical: most-frequent imputation + one-hot encoding

Pipelines ensure every step is applied in the correct order and prevent data leakage

	Missing Count	Percent
age	0	0.000000
sex	0	0.000000
cp	0	0.000000
trestbps	1	0.045851
chol	23	1.054562
fbs	8	0.366804
restecg	1	0.045851
thalachh	1	0.045851
exang	1	0.045851
oldpeak	0	0.000000
slope	190	8.711600
ca	291	13.342503
thal	266	12.196240
target	0	0.000000

Baseline Model

DecisionTreeClassifier

- Accuracy Score: 0.906
- Stronger due to non-linear splits

```
Baseline Accuracy: 0.9061784897025171
              precision    recall  f1-score   support

           0       0.92      0.89      0.91       220
           1       0.89      0.92      0.91       217

    accuracy                           0.91       437
   macro avg       0.91      0.91      0.91       437
weighted avg       0.91      0.91      0.91       437

Confusion Matrix:
[[196  24]
 [ 17 200]]
```

LogisticRegression

- Accuracy Score : 0.764
- Struggles with non-linear clinical relationships

```
Linear Baseline Accuracy: 0.7368421052631579
              precision    recall  f1-score   support

           0       0.75      0.71      0.73       220
           1       0.72      0.76      0.74       217

    accuracy                           0.74       437
   macro avg       0.74      0.74      0.74       437
weighted avg       0.74      0.74      0.74       437

Confusion Matrix:
[[156  64]
 [ 51 166]]
```

→ Tree-based model handles nonlinear patterns better than linear models

Feature Engineering

preprocessor_lightgbm()

→ LightGBM handles missing values and categorical splits natively

transform()

Clinical Ratios

- Cholesterol / age
- Resting BP / age

Heart Rate Reserve

- MaxHR reserve and % max HR → captures cardiovascular fitness

Log Transform

- Log of oldpeak → reduces skew

Interaction Terms

- cp × exang
- slope × oldpeak

LightGBM:

- Accuracy Score: 0.924

LightGBM + Feature Engineering:

- Accuracy Score: 0.917

Parameter Tuning

objective()

- Builds a LightGBM model with trial-suggested hyperparameters
- Returns the accuracy score → this is what Optuna tries to maximize

optuna.create_study(direction="maximize")

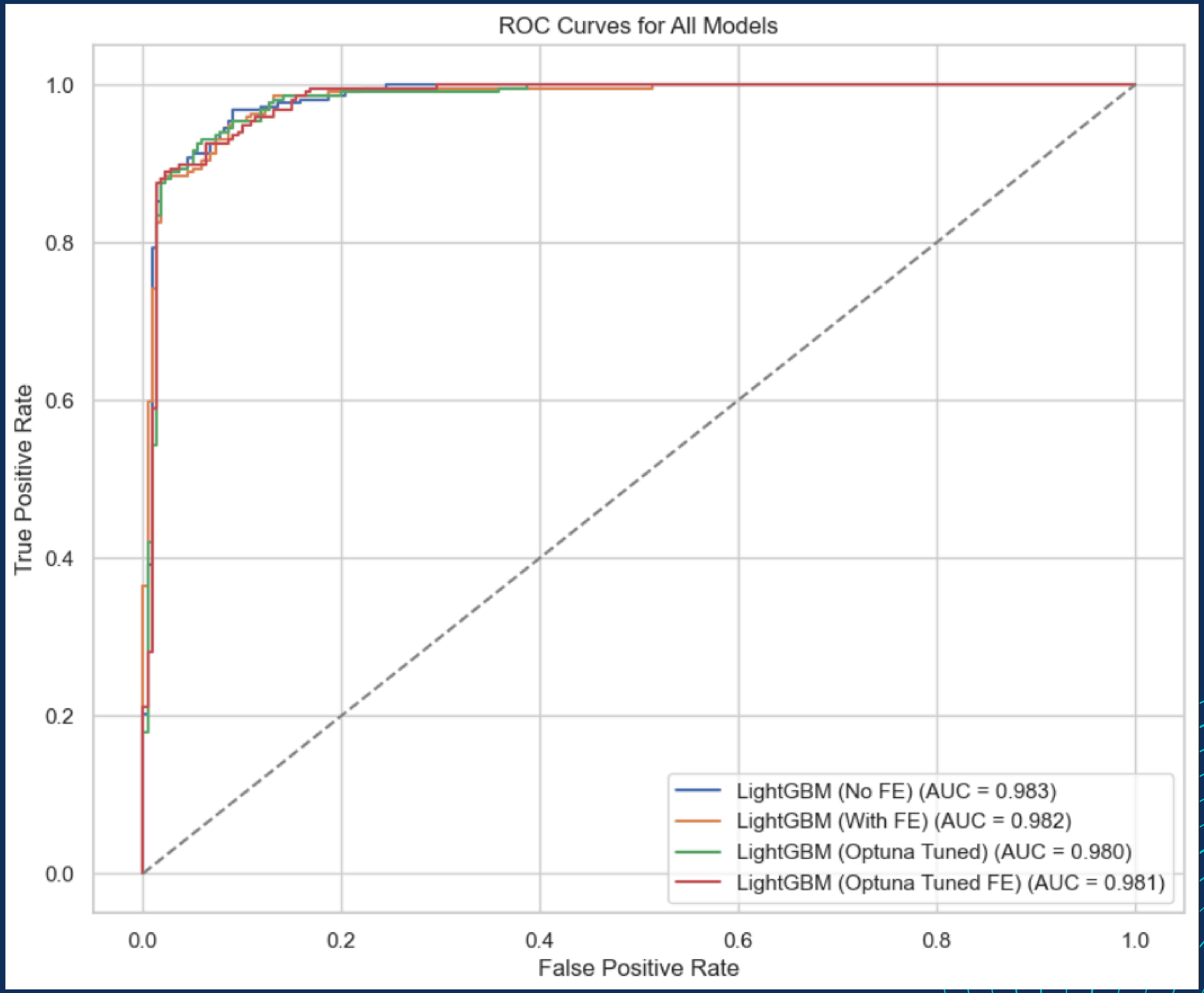
Finds the parameter set that gives highest accuracy

learning_rate
num_leaves
n_estimators
max_depth
min_child_samples

LightGBM + Feature Engineering + Optuna:
Accuracy Score: 0.917

LightGBM + Optuna:
Accuracy Score: 0.936

Model Evaluation




```
Tuned Accuracy: 0.9359267734553776
              precision    recall  f1-score   support

           0       0.93      0.95      0.94       220
           1       0.94      0.93      0.93       217

    accuracy                           0.94       437
   macro avg       0.94      0.94      0.94       437
weighted avg       0.94      0.94      0.94       437

Confusion Matrix:
[[208  12]
 [ 16 201]]
```

Limitations and Improvements

Limitations

- Small dataset
- Limited features
- Minimal impact from feature engineering
- Tuning gives only minor improvements

Improvements

- Use cross-validation
- Collect more data
- Try CatBoost
- SHAP for model interpretability

Sources

Kaggle Dataset:
https://www.kaggle.com/datasets/mfarhaannazirkhan/heart-dataset?select=raw_merged_heart_dataset.csv

Introduction to LightGBM:
https://sefiks.com/2018/10/13/a-gentle-introduction-to-lightgbm-for-applied-machine-learning/

How to handle missing features:
https://jimmy-wang-gen-ai.medium.com/how-do-xgboost-lightgbm-and-catboost-handle-missing-features-e541da94d528

¡Gracias por su atención!