Housing Price Prediction

Group: 4

Advisor: Prof. Khoat Than

Subject: Introduction to Artificial Intelligence

contents

1. Problem statement
2. Exploratory data analysis
3. Approaches & Evaluation
4. Discussion

Problem Statement

What affects house prices?

People believe:

- The square foot area
- The number of bedrooms

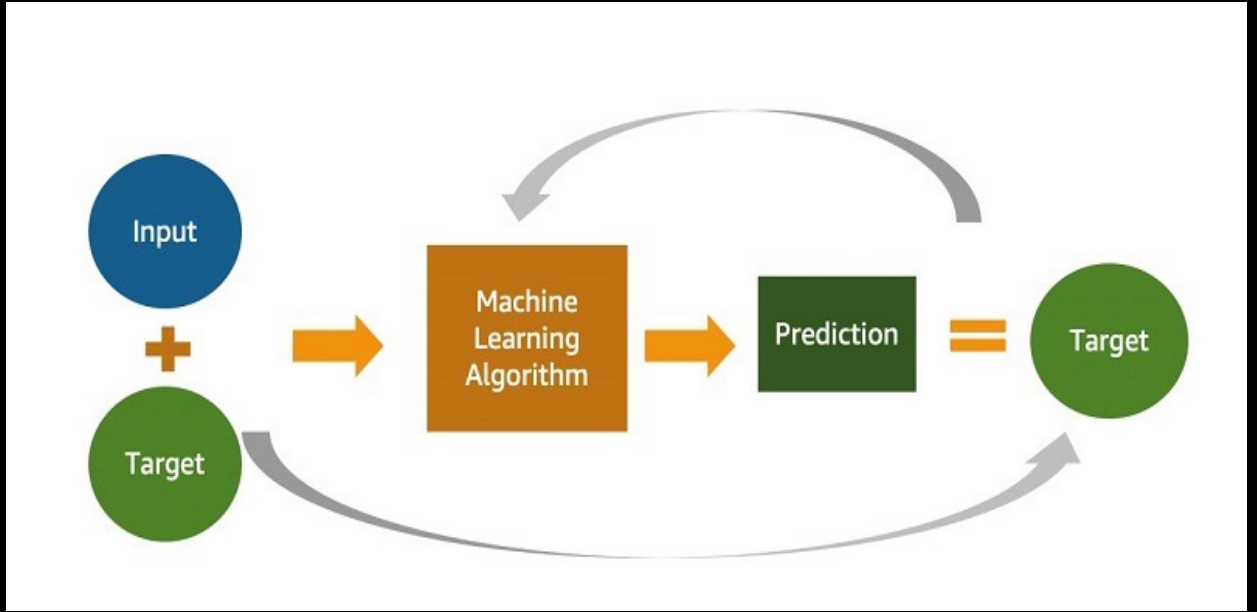
Truth:

- Area outside the house
- Type of housing

How to predict house prices?

Artificial Intelligence

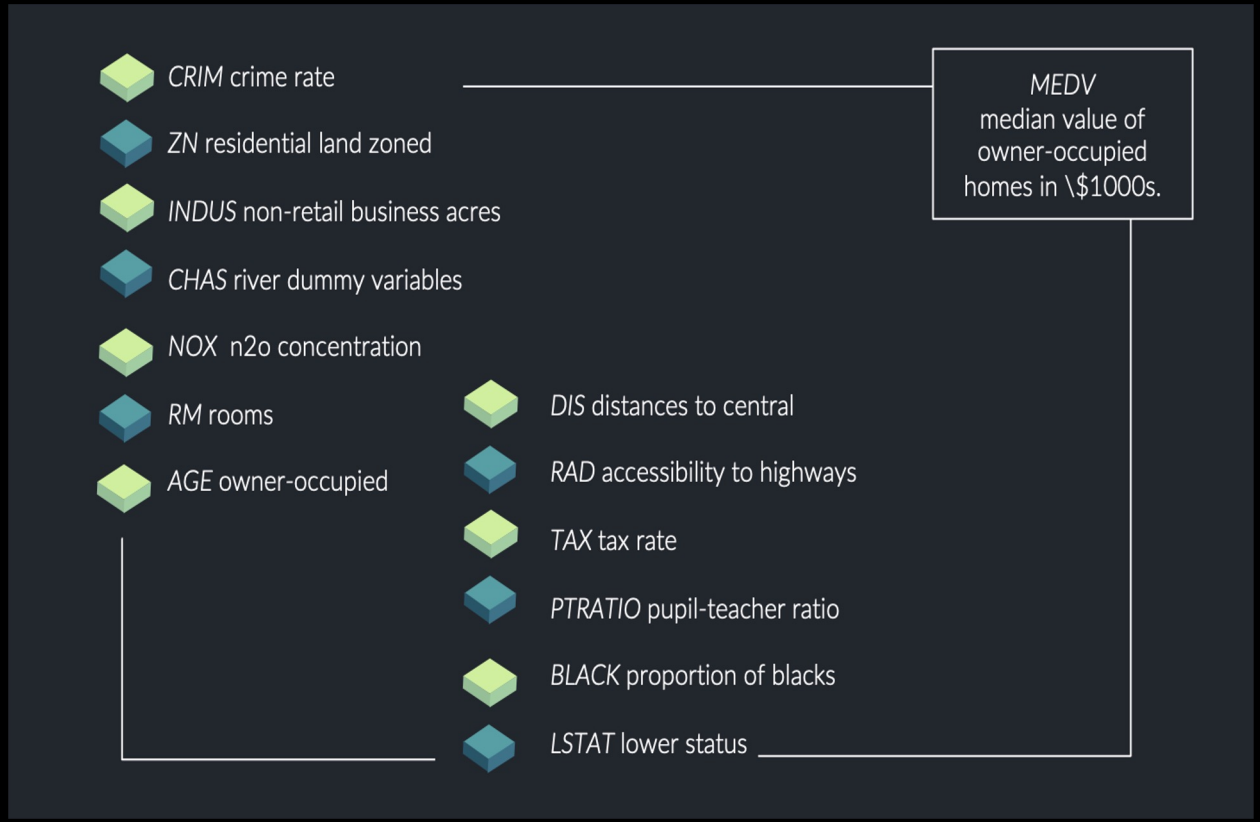
- Machine learning
- Deep learning



Datasets: Boston Housing Prices

506 instances
Missing Attribute Values: 0

Features: 13 numeric/ categorical features
Target: MEDV (median value)



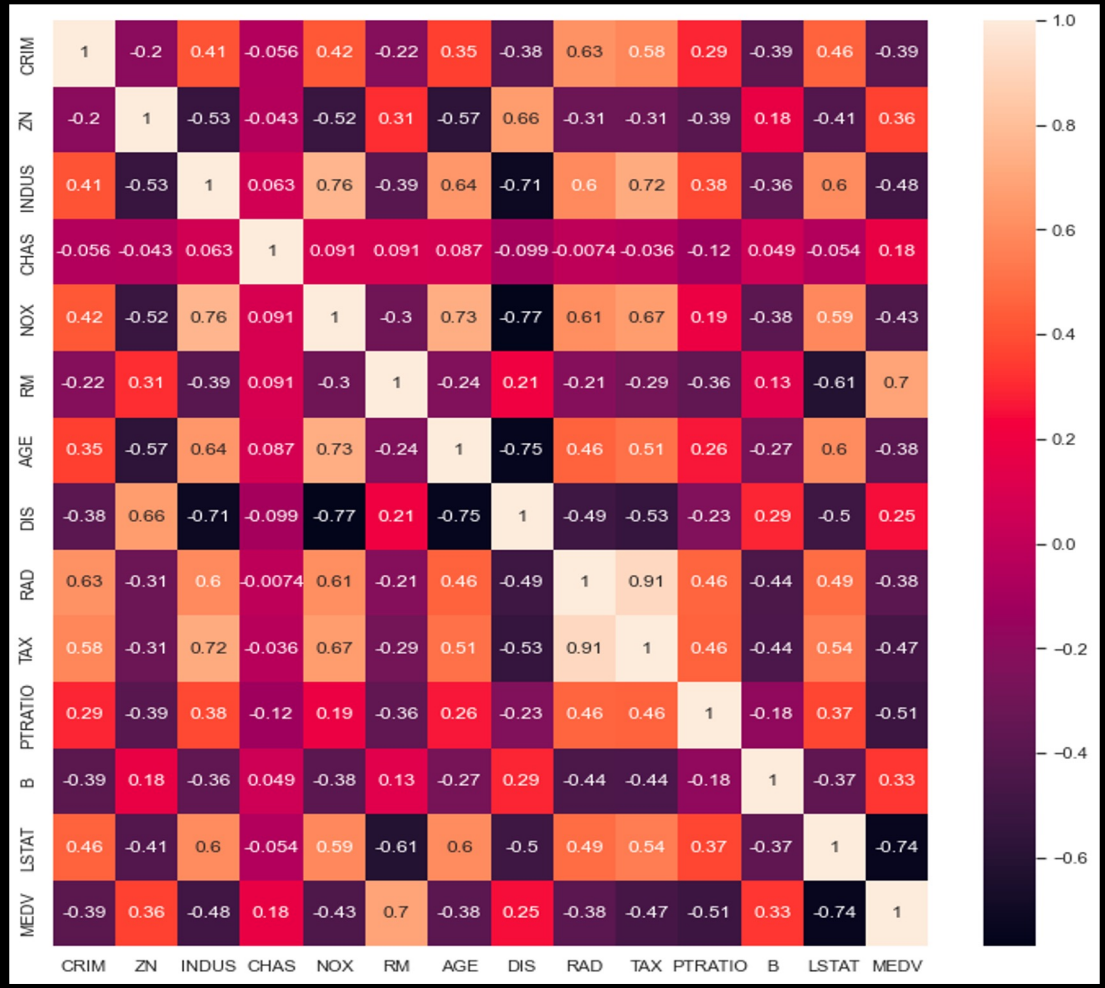
Exploratory data analysis

Heatmap

High correlation between features and target

Best:

- LSTAT: -0.74
- RM: 0.7

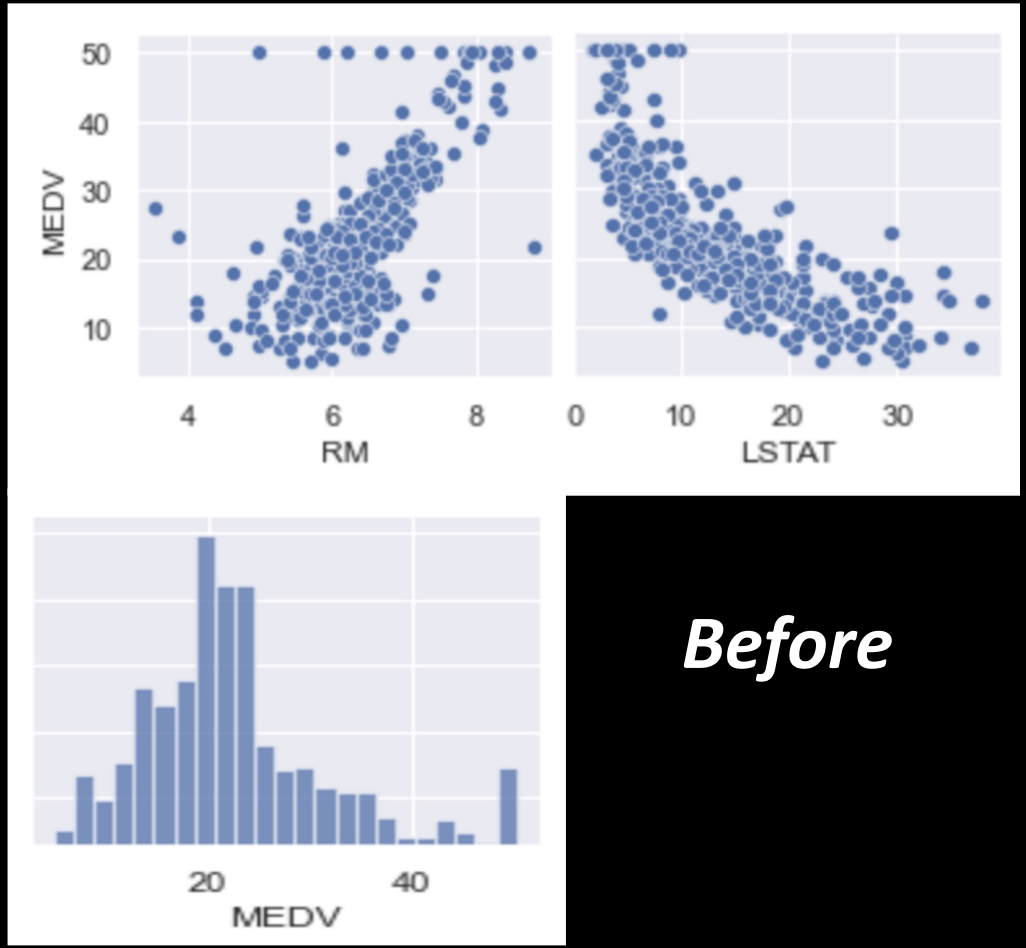


outliers

With **small datasets, outliers** become much dangerous

Drop outliers by features:

- “MEDV” == 50
- “RM” < 4
- “RM” > 8.4



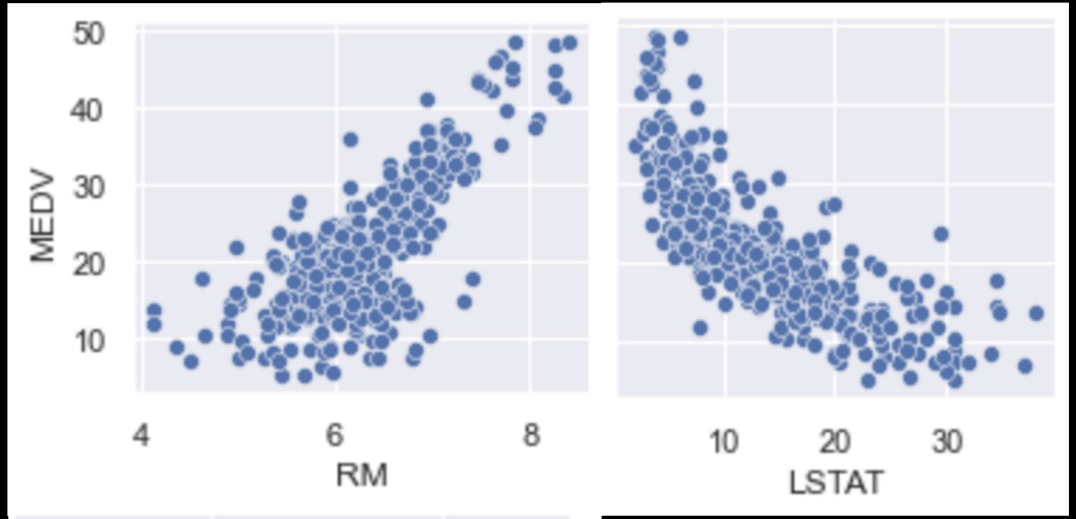
Before

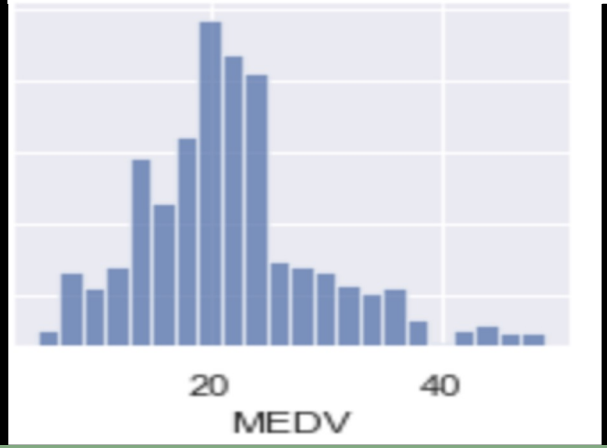
outliers

With **small datasets, outliers** become much dangerous

Drop outliers by features:

- “MEDV” == 50
- “RM” < 4
- “RM” > 8.4





After

Approaches

Metrics: R-squared

$$R^2 = 1 - \frac{\text{Unexplained Variation}}{\text{Total Variation}} = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$$

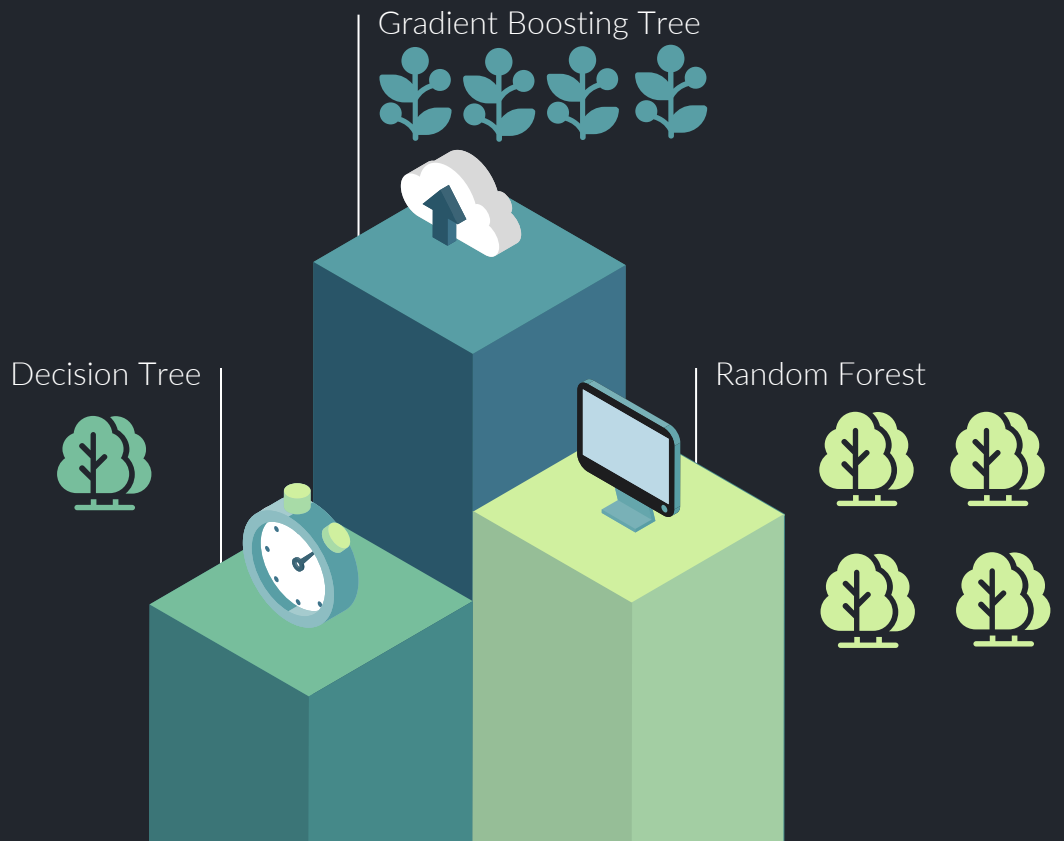
$$SS_{\text{res}} = \sum_i (y_i - f_i)^2 = \sum_i e_i^2$$

$$SS_{\text{tot}} = \sum_i (y_i - \bar{y})^2$$

R2: proportion of the variation in the dependent variable that is predictable from the independent variable(s)

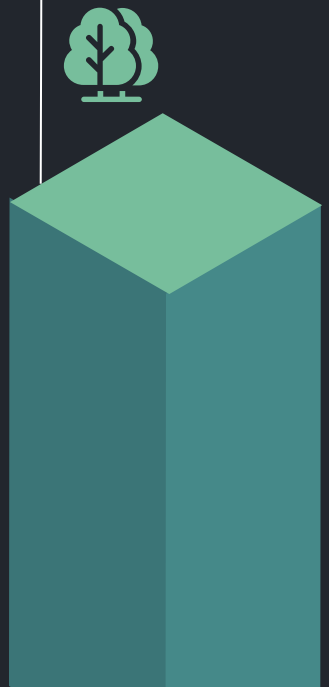
Higher R2, better model

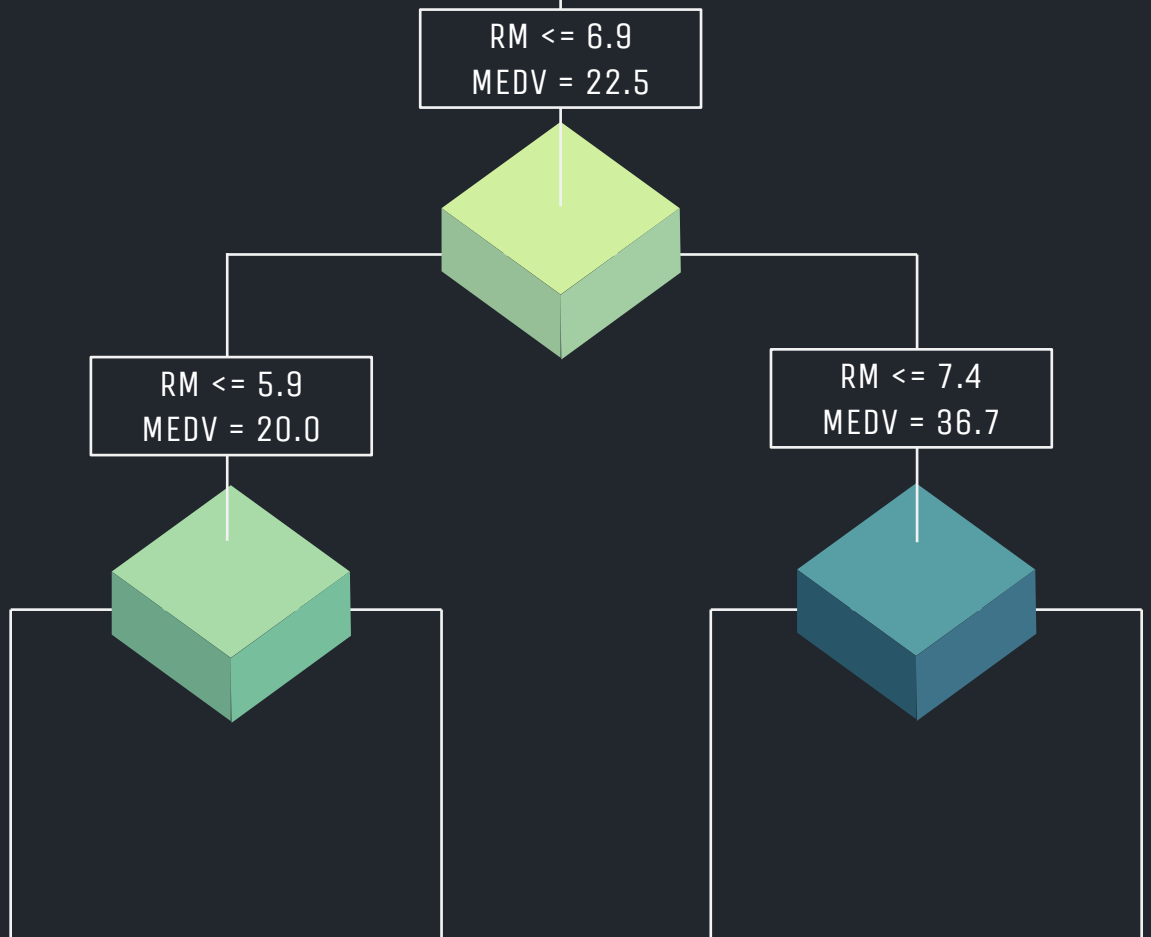
TREE-BASED ALGORITHMS



TREE-BASED ALGORITHMS

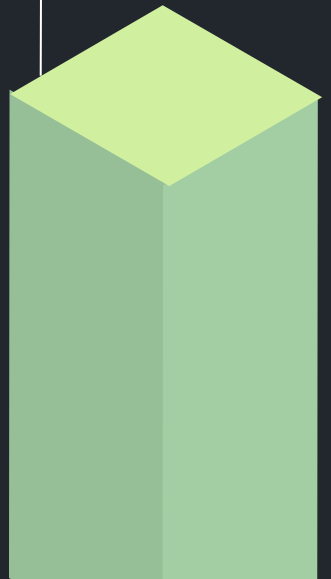
Decision Tree

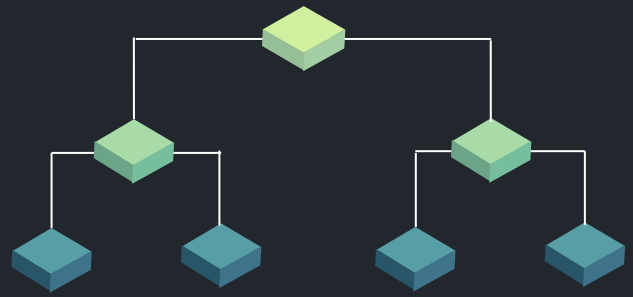


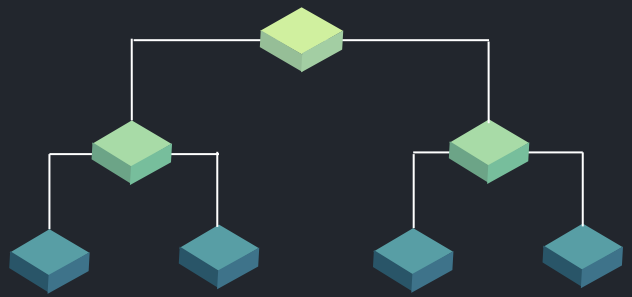


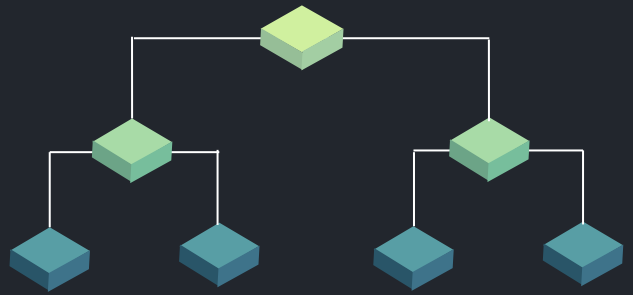
TREE-BASED ALGORITHMS

Random Forest



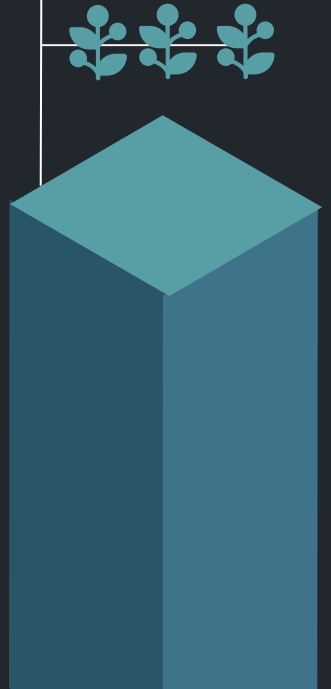


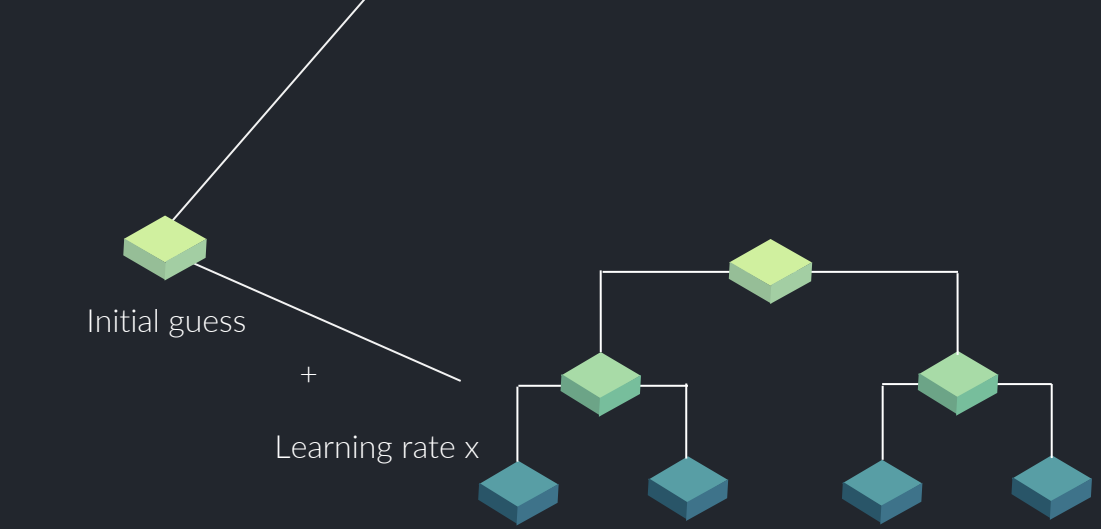


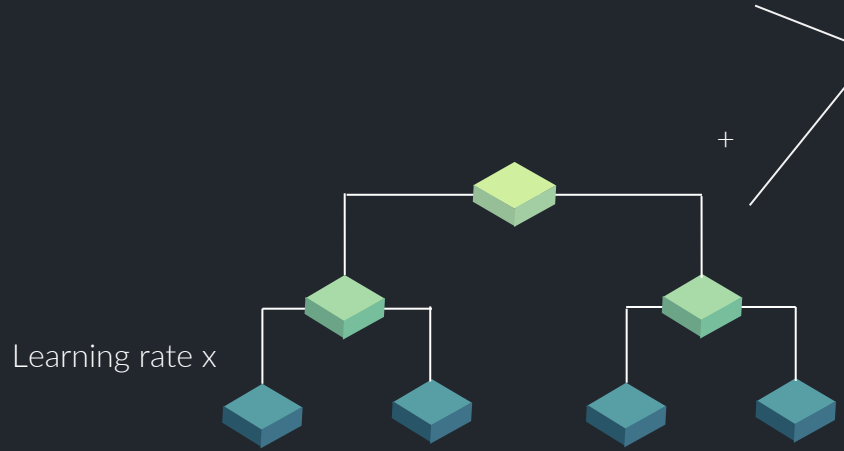


TREE-BASED ALGORITHMS

Gradient Boosting Tree

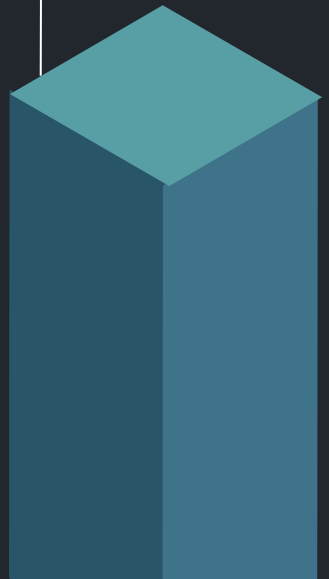






TREE-BASED ALGORITHMS

Gradient Boosting Tree



RM <= 6.95,
20.6, 21.4, 23.7, 25

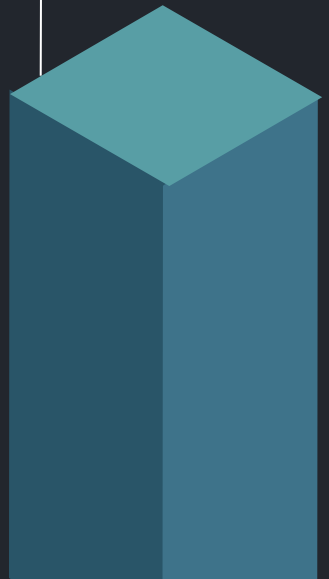
MEDV =
20.6, 21.4, 23.7

MEDV =
25

Residuals = MEDV – initial guess

TREE-BASED ALGORITHMS

Gradient Boosting Tree



RM <= 6.95,
20.6, 21.4, 23.7, 25

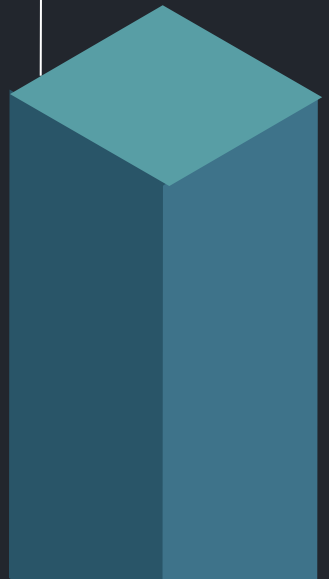
Residuals

Residuals

Residuals' = (MEDV – initial guess) + (learning rate x Residuals)

TREE-BASED ALGORITHMS

Gradient Boosting Tree



RM <= 6.95,
20.6, 21.4, 23.7, 25

Residuals'

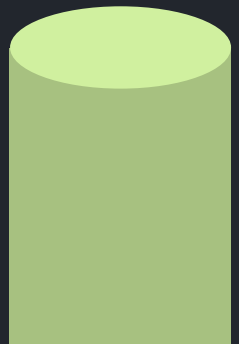
Residuals'

TRAINING

Create baseline

Decision Tree

Training score: 1.000
Testing score: 0.737



Random Forest

Training score: 0.983
Testing score: 0.861

Gradient Boosting Tree

Training score: 0.976
Testing score: 0.893

TRAINING

Hyper Tunning

Create baseline

GRADIENT BOOSTING TREE

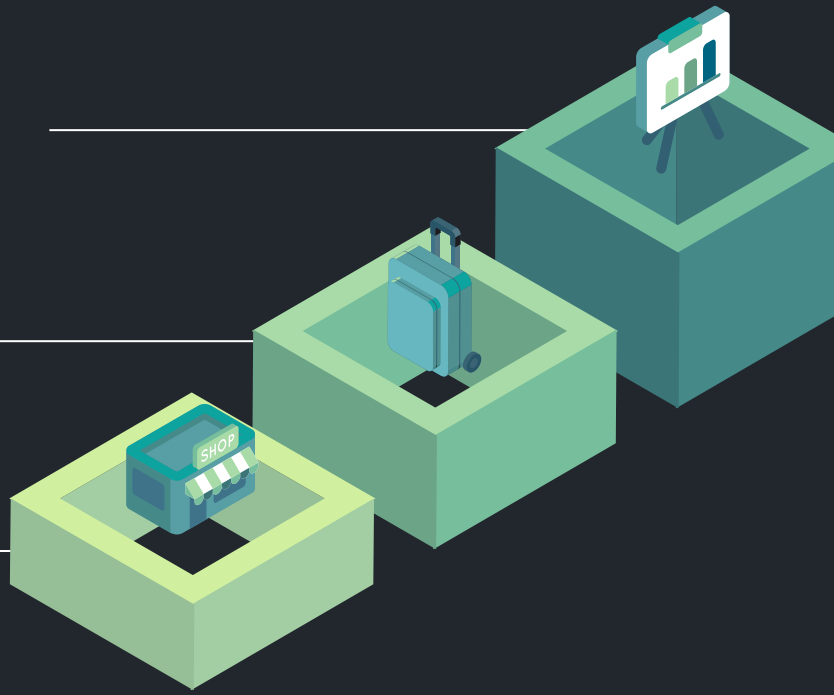
700 n_estimators
0.05 learning_rate
8 min_samples_leaf

RANDOM FOREST

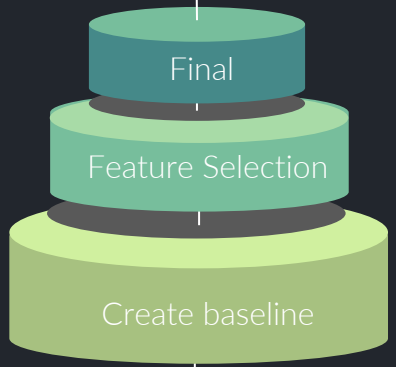
300 n_estimators
6 max_features

DECISION TREE

5 max_depth
5 max_features



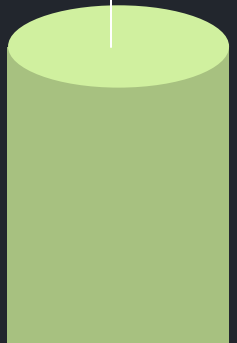
TRAINING



Training score: 1.000
Testing score: 0.737

Decision Tree

Training score: 0.910
Testing score: 0.853



Training score: 0.983
Testing score: 0.861

Random Forest

Training score: 0.985
Testing score: 0.874

Training score: 0.976
Testing score: 0.893

Gradient
Boosting Tree

Training score: 0.991
Testing score: 0.899

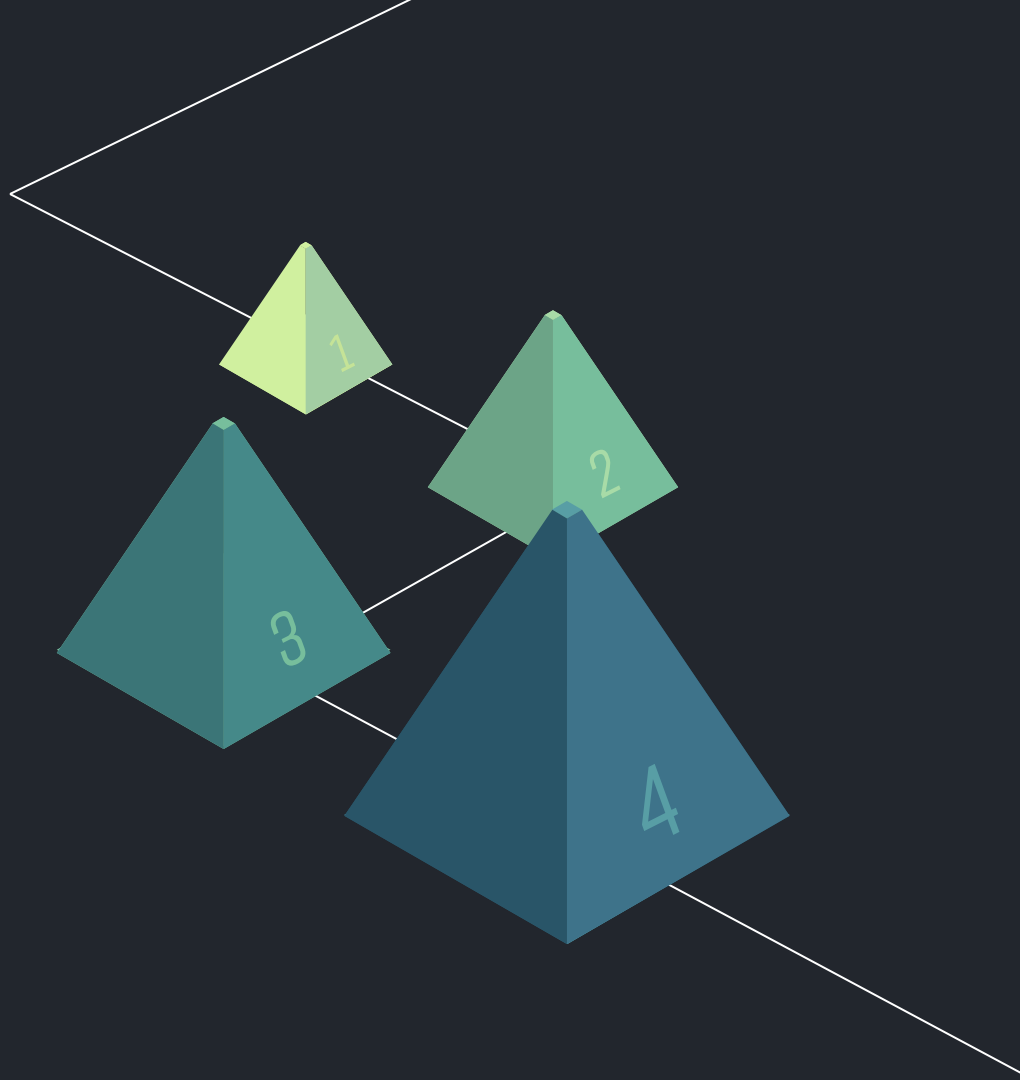
APPROACH 2

PHASE 1
EDA

PHASE 2
K-means

PHASE 3
Linear Regression

PHASE 4
Final Evaluation

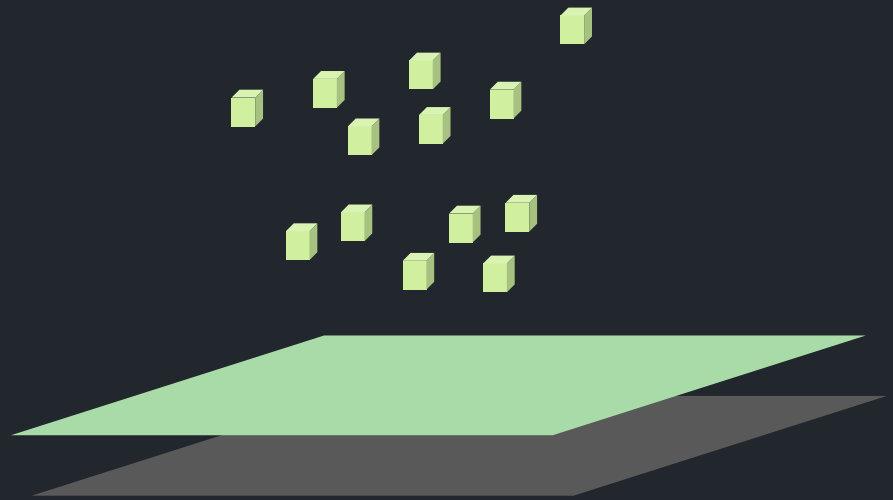


APPROACH 2

K-means

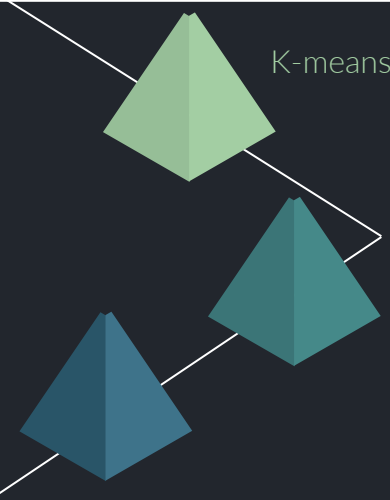
Regression

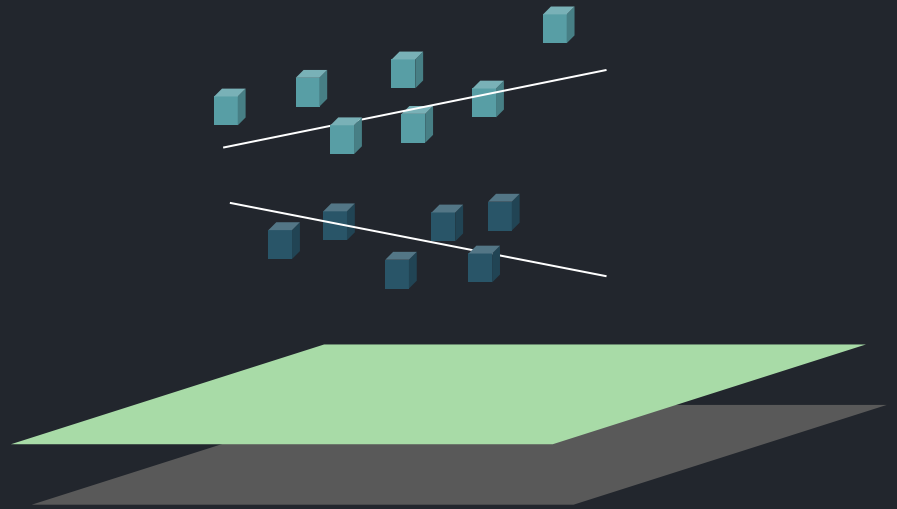
KNN



APPROACH 2

K-means

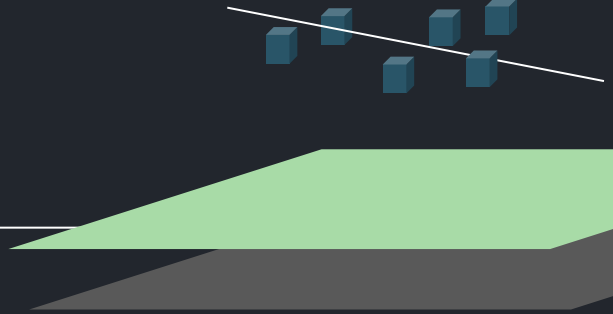




APPROACH 2

K-means

Regression

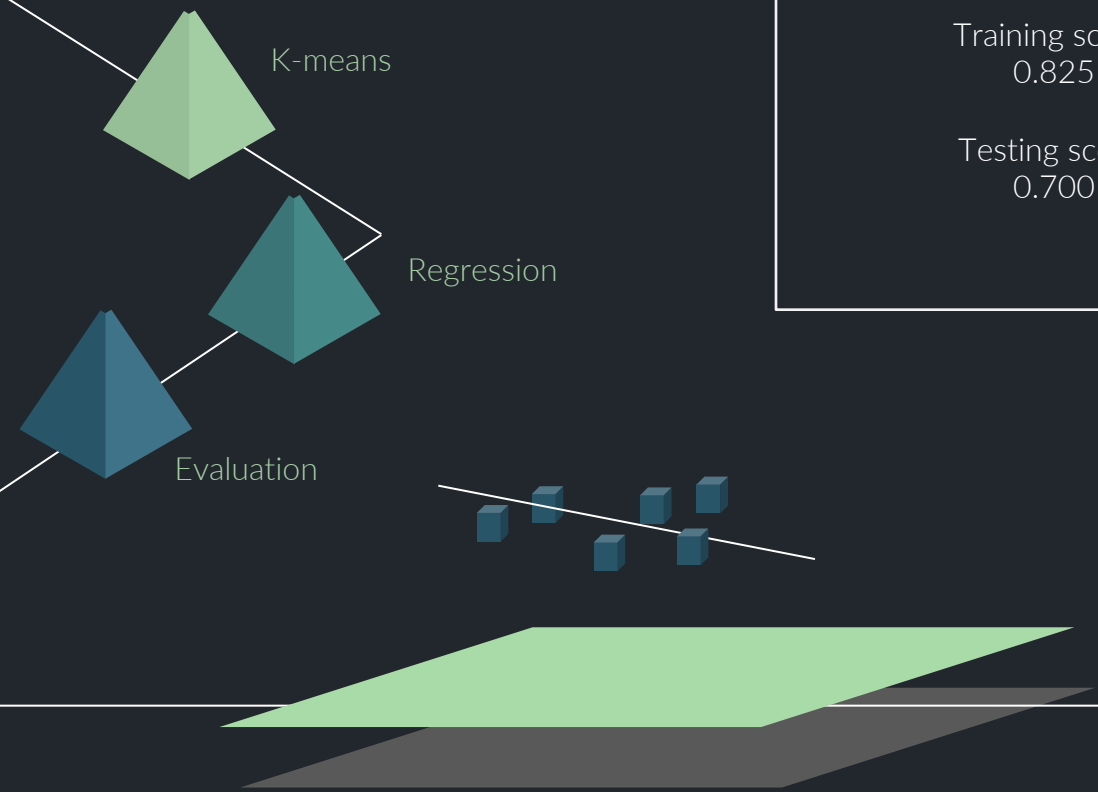


Linear Regression

Loss: $$L(y_i, x_i, \omega) = \sum_{i=1}^{N} (yi - \omega x_i)^2$$

Prediction: $$y_{i\_pred} = \omega x_i$$

APPROACH 2



Baseline	Kmeans + Linear Regression
Training score: 0.825	Training score: 0.915
Testing score: 0.700	Testing score: 0.816

THANKS FOR YOUR ATTENDING!

Nguyen Van Thanh Tung | 20190090
Nguyen Huy Hoang | 20194433
Nguyen Mau Tra | 20200624
Nguyen Thi Linh | 20200349