
Data Mining and Knowledge Discovery

Knowledge Discovery and Knowledge Management in e-Science

Petra Kralj Novak

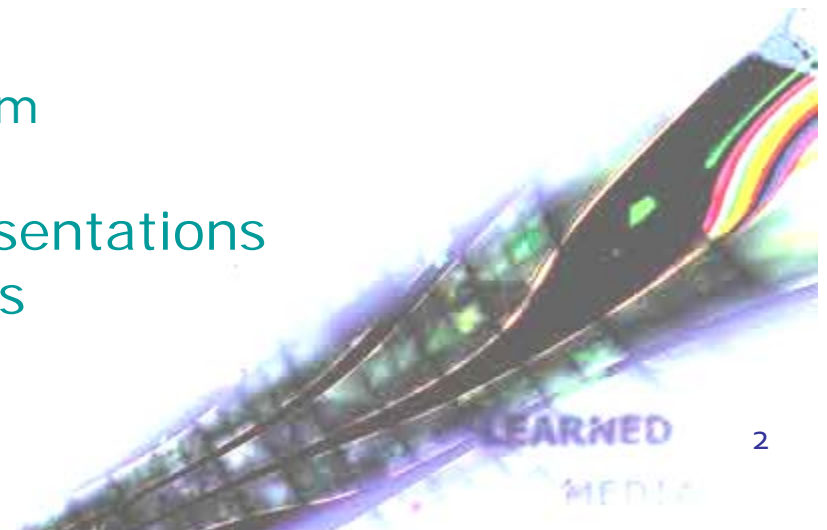
Petra.Kralj.Novak@ijs.si

Practice, 2008/11/11



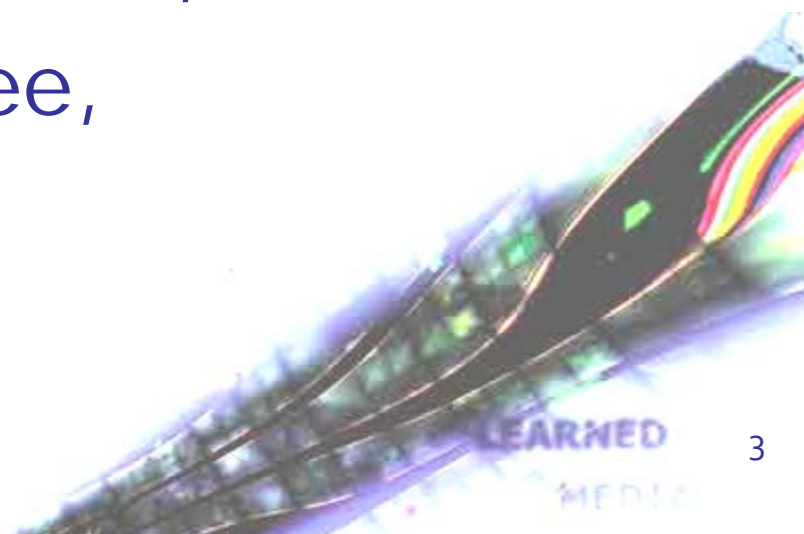
Practice plan

- 2008/10/22: Predictive data mining
 - Decision trees
 - Naïve Bayes classifier
 - Evaluating classifiers (separate test set, cross validation, confusion matrix, classification accuracy)
 - Predictive data mining in Weka
- 2008/11/11: Numeric prediction and descriptive data mining
 - Regression models
 - Regression models and evaluation in Weka
- 2008/11/12: Descriptive data mining
 - Association rules
 - Descriptive data mining in Weka
 - Discussion about seminars and exam
- 2008/12/1: Written exam
- 2008/12/8: Seminar proposals presentations
- 2009/01/14: Seminar presentations



Numeric prediction

Baseline,
Linear Regression,
Regression tree,
Model Tree,
KNN



Numeric prediction	Classification
Data: attribute-value description	
Target variable: Continuous	Target variable: Categorical (nominal)
Evaluation: cross validation, separate test set, ...	
Error: MSE, MAE, RMSE, ...	Error: 1-accuracy
Algorithms: Linear regression, regression trees, ...	Algorithms: Decision trees, Naïve Bayes, ...
Baseline predictor: Mean of the target variable	Baseline predictor: Majority class

Example

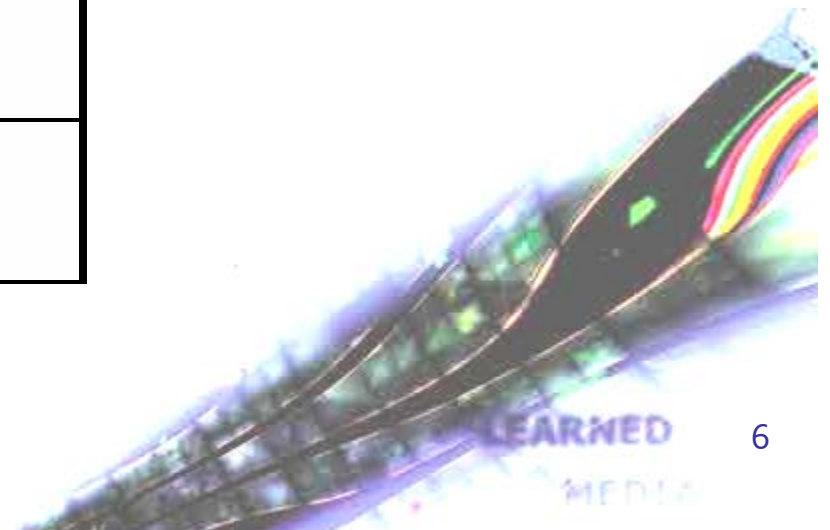
- data about 80 people:
Age and Height



Age	Height
3	1.03
5	1.19
6	1.26
9	1.39
15	1.69
19	1.67
22	1.86
25	1.85
41	1.59
48	1.60
54	1.90
71	1.82
...	...

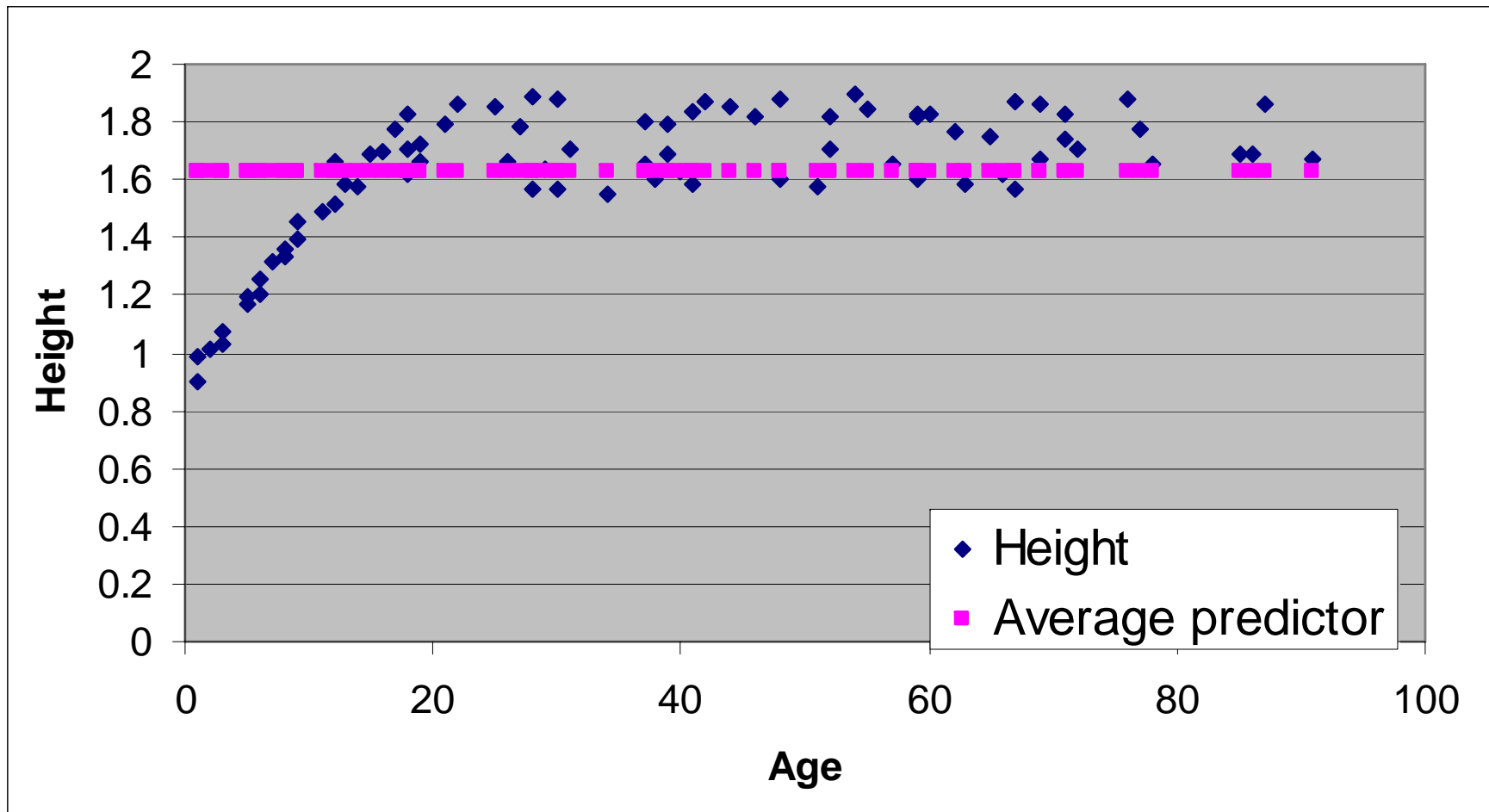
Test set

Age	Height
2	0.85
10	1.4
35	1.7
70	1.6



Baseline numeric predictor

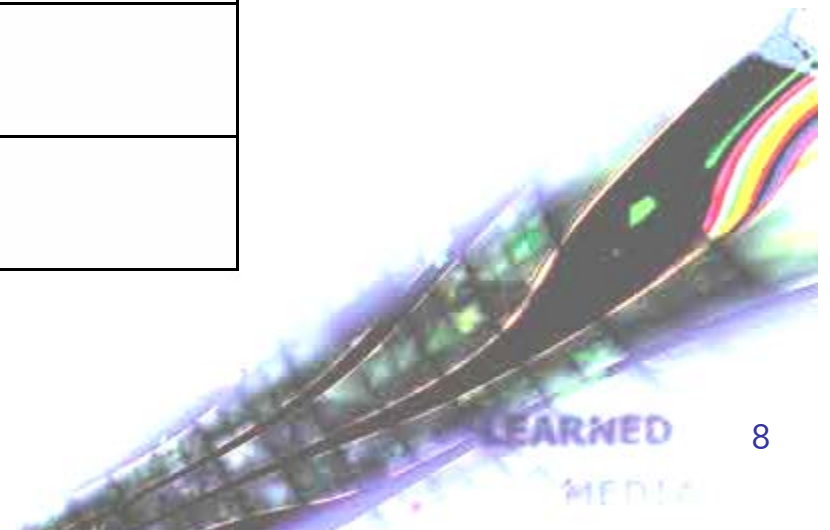
- Average of the target variable



Baseline predictor: prediction

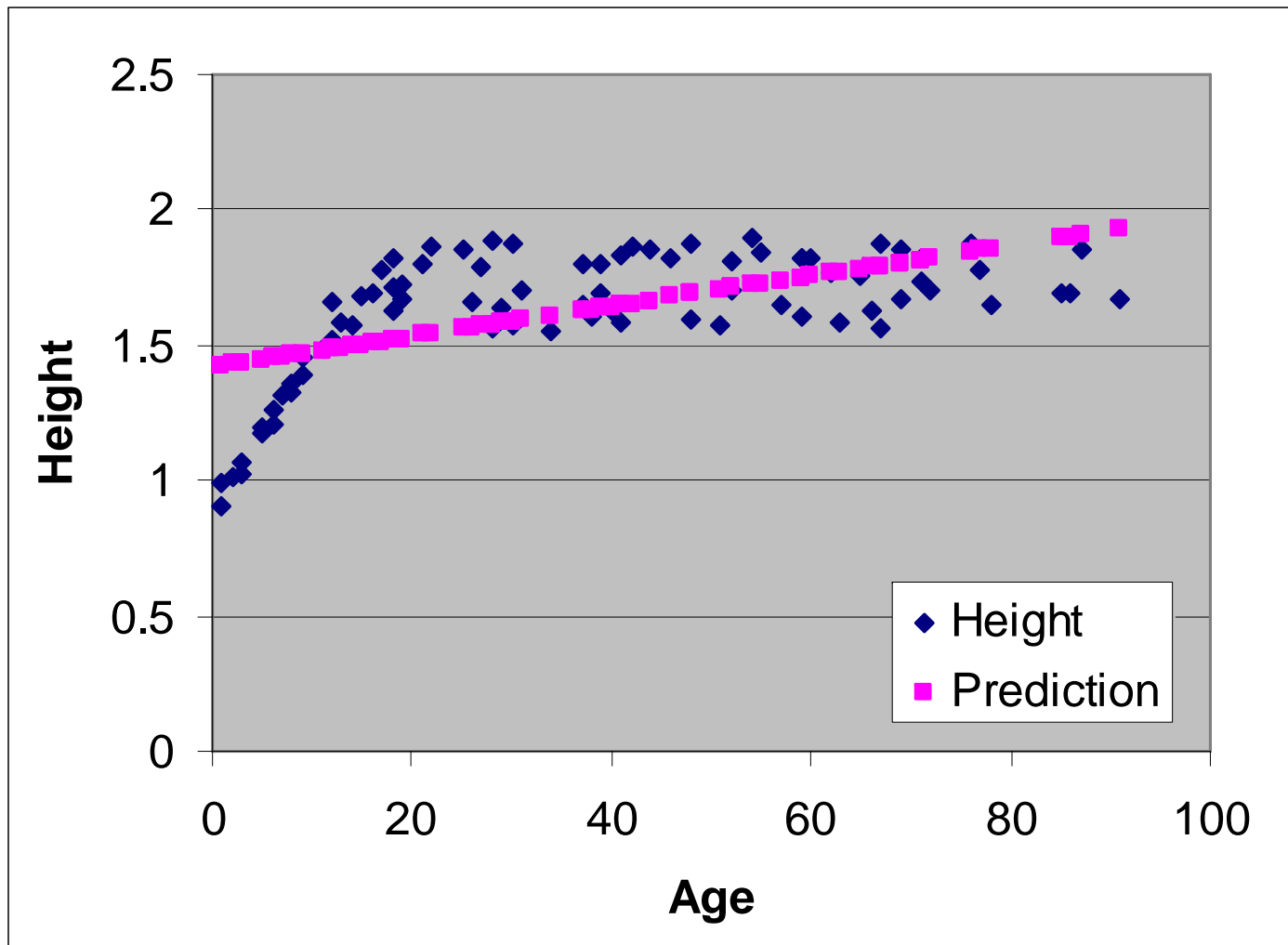
Average of the target variable is 1.63

Age	Height	Baseline
2	0.85	
10	1.4	
35	1.7	
70	1.6	



Linear Regression Model

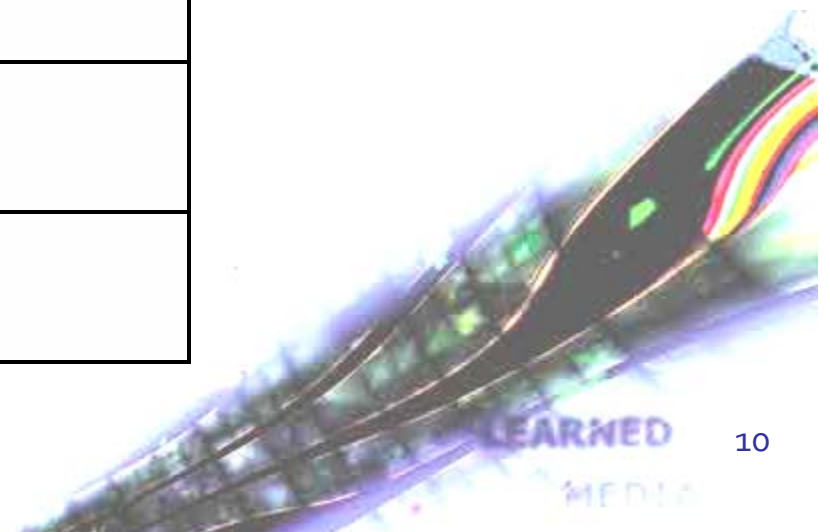
$$\text{Height} = 0.0056 * \text{Age} + 1.4181$$



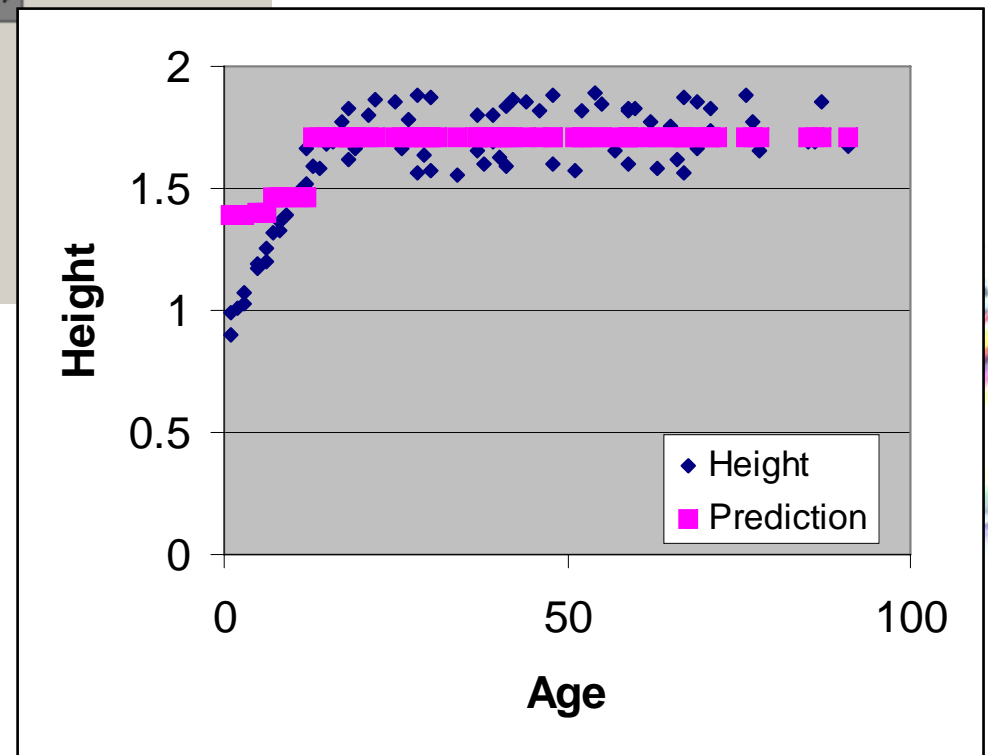
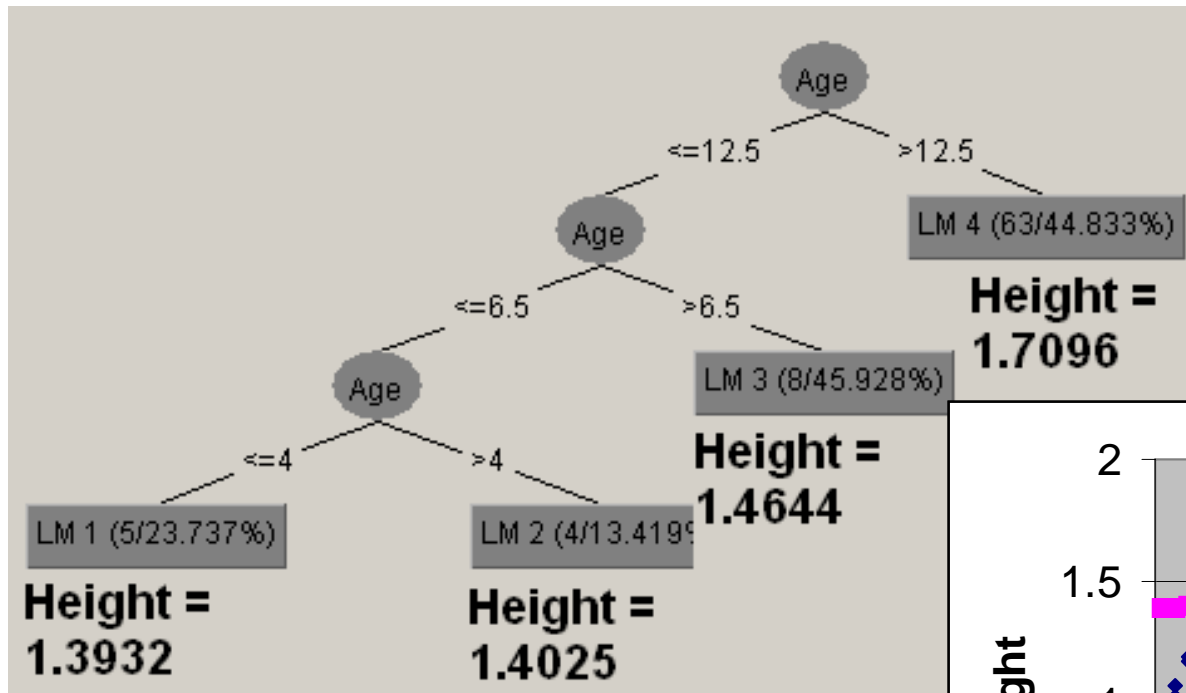
Linear Regression: prediction

$$\text{Height} = 0.0056 * \text{Age} + 1.4181$$

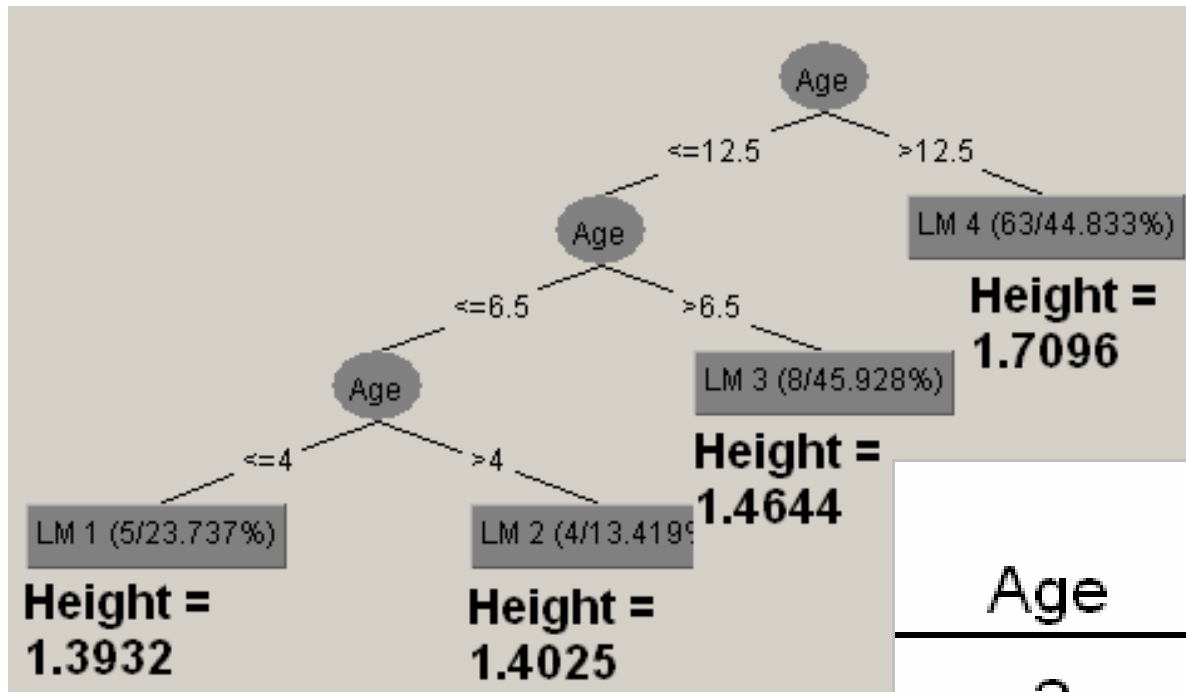
Age	Height	Linear regression
2	0.85	
10	1.4	
35	1.7	
70	1.6	



Regression tree

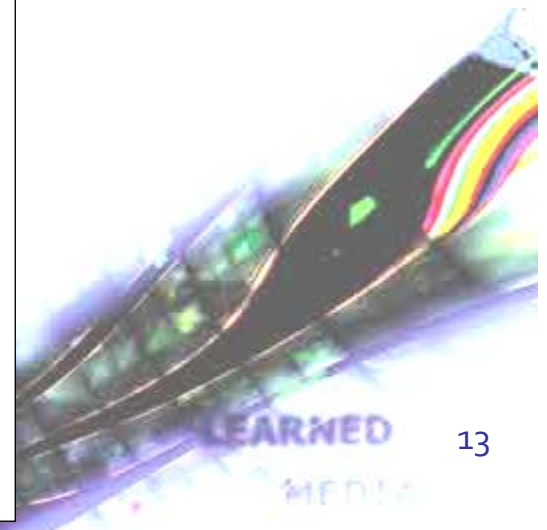
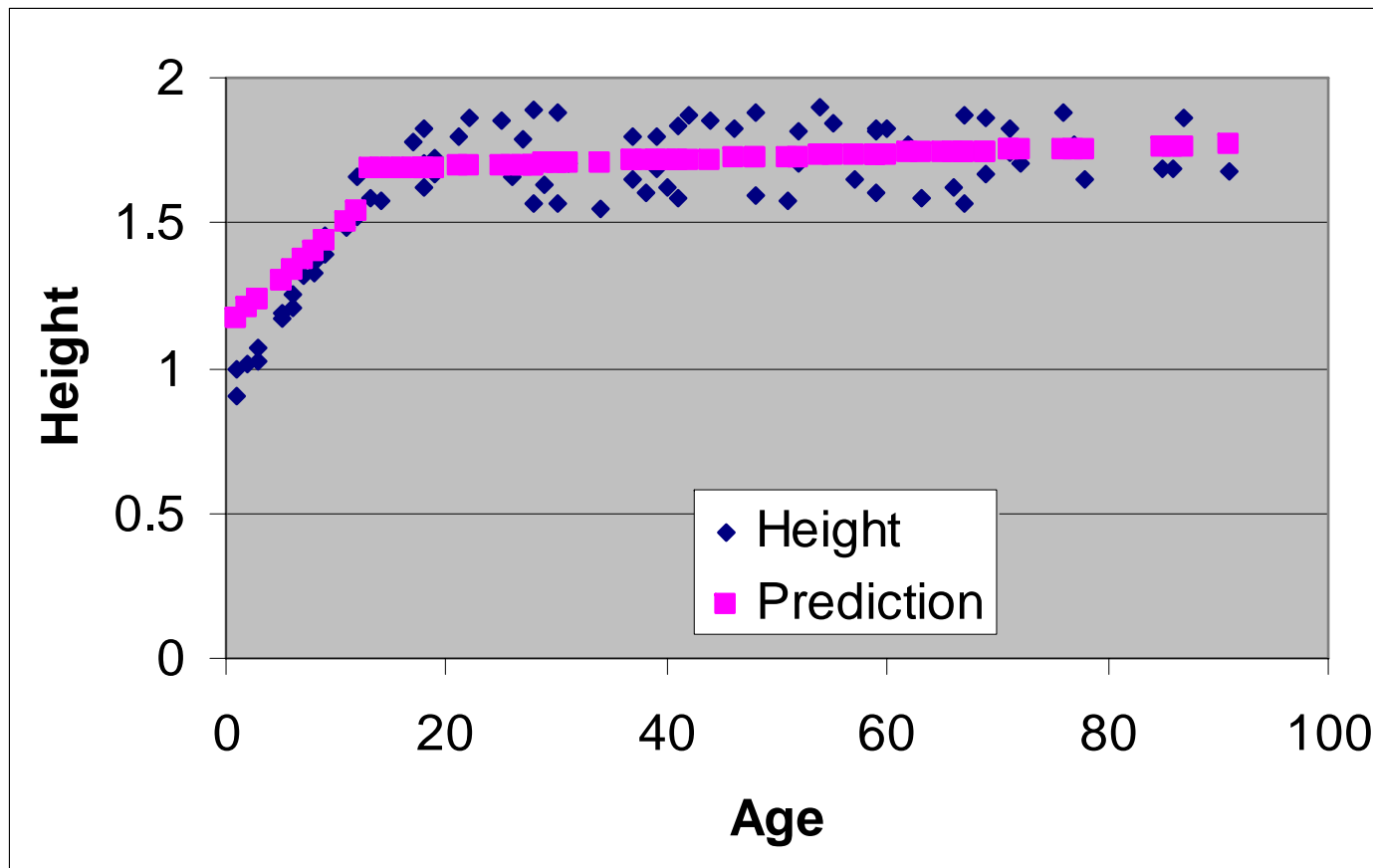
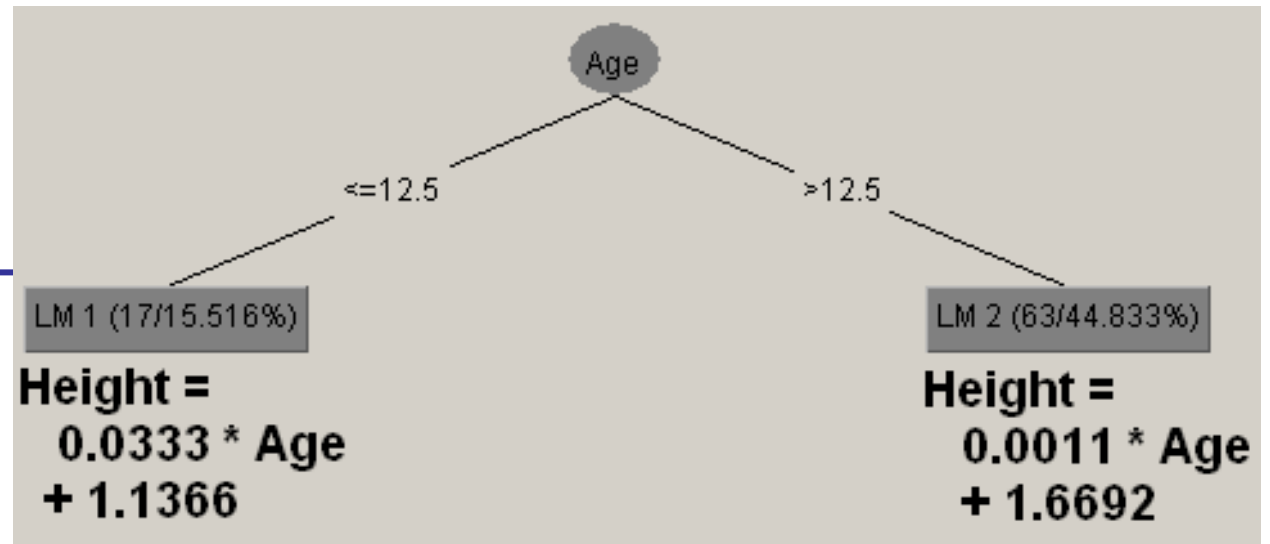


Regression tree: prediction



Age	Height	Regression tree
2	0.85	
10	1.4	
35	1.7	
70	1.6	

Model tree



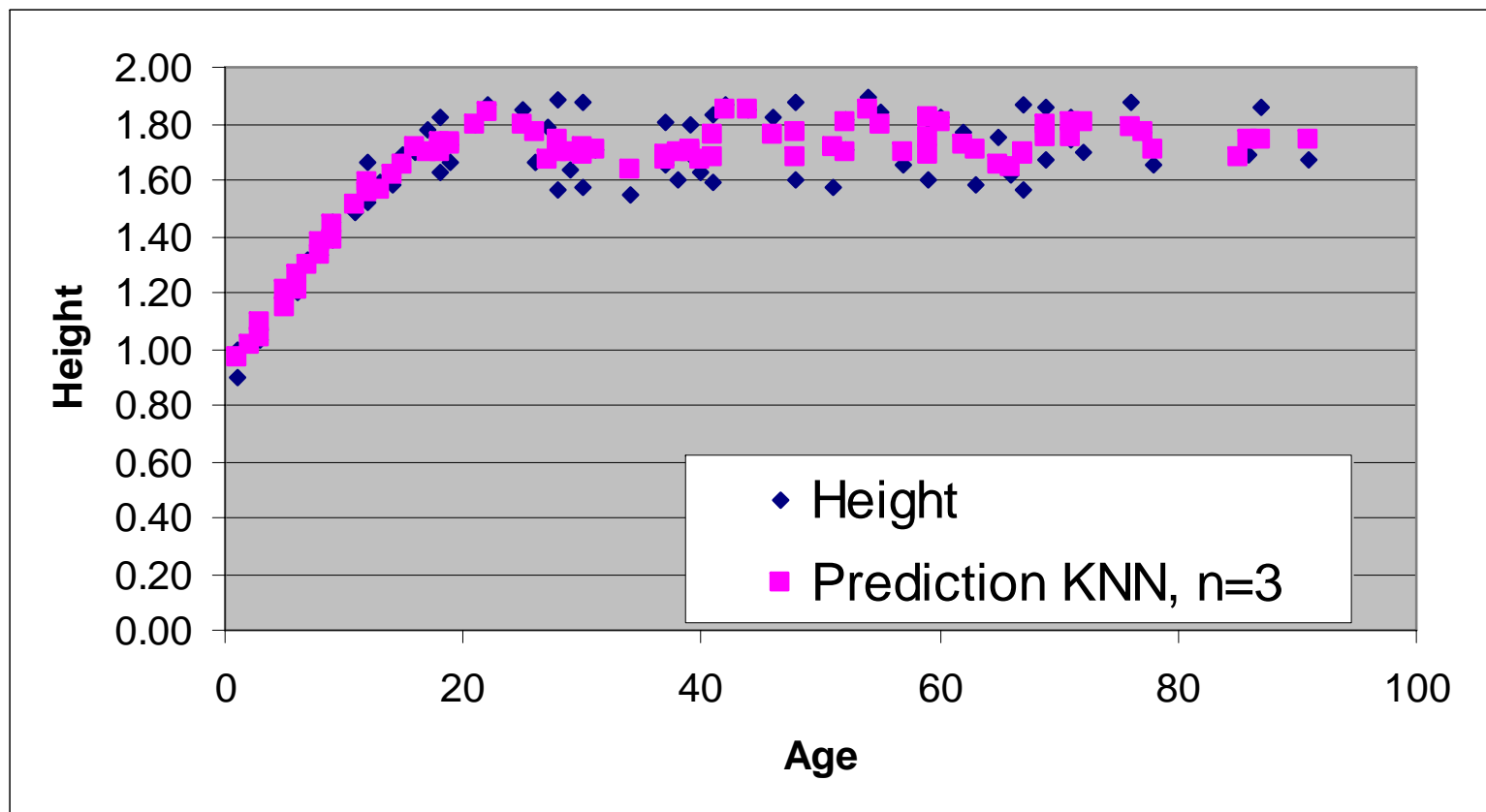
Model tree: prediction

Age	Height	Model tree
2	0.85	
10	1.4	
35	1.7	
70	1.6	



KNN – K nearest neighbors

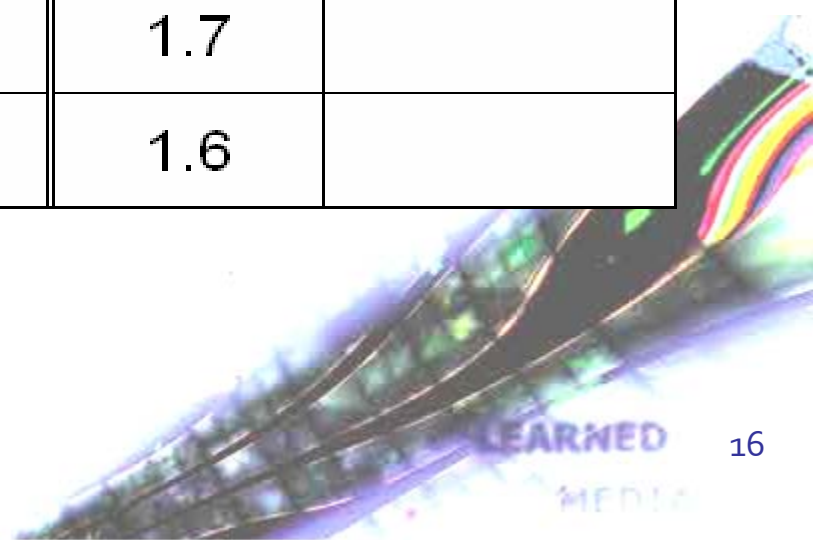
- Looks at K closest examples (by non-target attributes) and predicts the average of their target variable
- In this example, $K=3$



KNN prediction

Age	Height
1	0.90
1	0.99
2	1.01
3	1.03
3	1.07
5	1.19
5	1.17

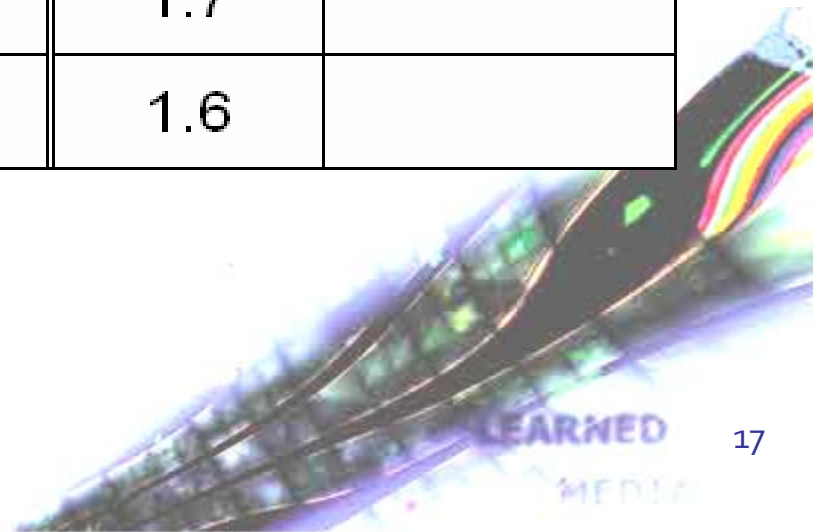
Age	Height	kNN
2	0.85	
10	1.4	
35	1.7	
70	1.6	



KNN prediction

Age	Height
8	1.36
8	1.33
9	1.45
9	1.39
11	1.49
12	1.66
12	1.52
13	1.59
14	1.58

Age	Height	kNN
2	0.85	
10	1.4	
35	1.7	
70	1.6	



KNN prediction

Age	Height
30	1.57
30	1.88
31	1.71
34	1.55
37	1.65
37	1.80
38	1.60
39	1.69
39	1.80

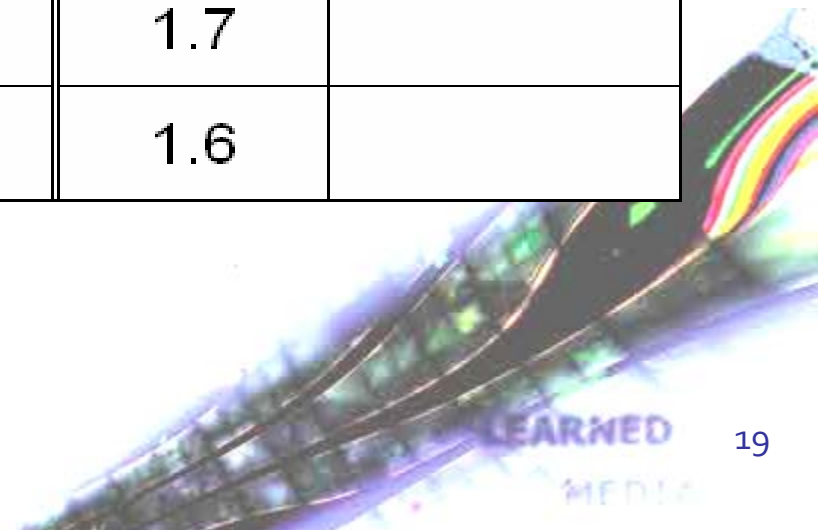
Age	Height	kNN
2	0.85	
10	1.4	
35	1.7	
70	1.6	



KNN prediction

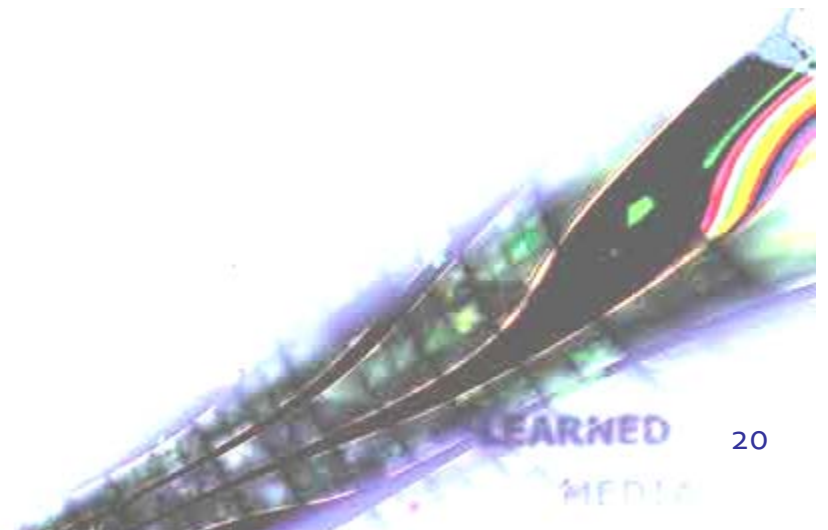
Age	Height
67	1.56
67	1.87
69	1.67
69	1.86
71	1.74
71	1.82
72	1.70
76	1.88

Age	Height	kNN
2	0.85	
10	1.4	
35	1.7	
70	1.6	



Which predictor is the best?

Age	Height	Baseline	Linear regression	Regression tree	Model tree	kNN
2	0.85	1.63	1.43	1.39	1.20	1.00
10	1.4	1.63	1.47	1.46	1.47	1.44
35	1.7	1.63	1.61	1.71	1.71	1.67
70	1.6	1.63	1.81	1.71	1.75	1.77



Evaluating numeric prediction

Performance measure	Formula
mean-squared error	$\frac{(p_1 - a_1)^2 + \dots + (p_n - a_n)^2}{n}$
root mean-squared error	$\sqrt{\frac{(p_1 - a_1)^2 + \dots + (p_n - a_n)^2}{n}}$
mean absolute error	$\frac{ p_1 - a_1 + \dots + p_n - a_n }{n}$
relative squared error	$\frac{(p_1 - a_1)^2 + \dots + (p_n - a_n)^2}{(a_1 - \bar{a})^2 + \dots + (a_n - \bar{a})^2}, \text{ where } \bar{a} = \frac{1}{n} \sum_i a_i$
root relative squared error	$\sqrt{\frac{(p_1 - a_1)^2 + \dots + (p_n - a_n)^2}{(a_1 - \bar{a})^2 + \dots + (a_n - \bar{a})^2}}$
relative absolute error	$\frac{ p_1 - a_1 + \dots + p_n - a_n }{ a_1 - \bar{a} + \dots + a_n - \bar{a} }$
correlation coefficient	$\frac{S_{PA}}{\sqrt{S_p S_A}}, \text{ where } S_{PA} = \frac{\sum_i (p_i - \bar{p})(a_i - \bar{a})}{n-1},$ $S_p = \frac{\sum_i (p_i - \bar{p})^2}{n-1}, \text{ and } S_A = \frac{\sum_i (a_i - \bar{a})^2}{n-1}$

Discussion

- Consider a dataset with a target variable with five possible values:
 1. non sufficient
 2. sufficient
 3. good
 4. very good
 5. excellent
 - Is this a classification or a numeric prediction problem?
 - What if such a variable is an attribute, is it nominal or numeric?
- Can KNN be used for classification tasks?
- Similarities between KNN and Naïve Bayes.
- Similarities and differences between decision trees and regression trees.

