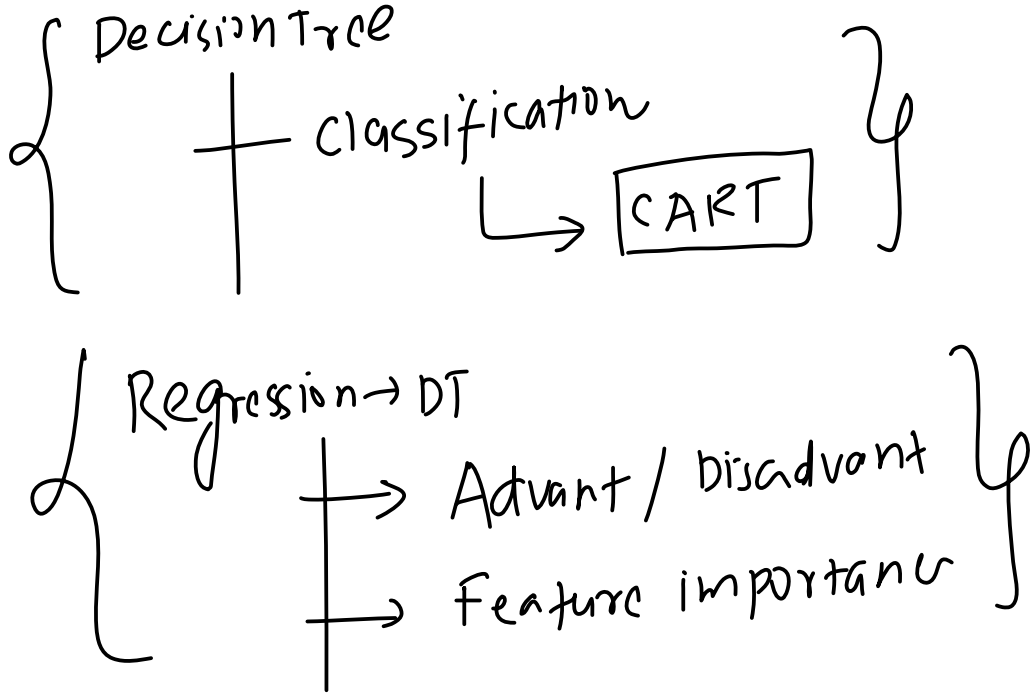
Recap

24 July 2023 13:31



CART for Regression

24 July 2023 13:31

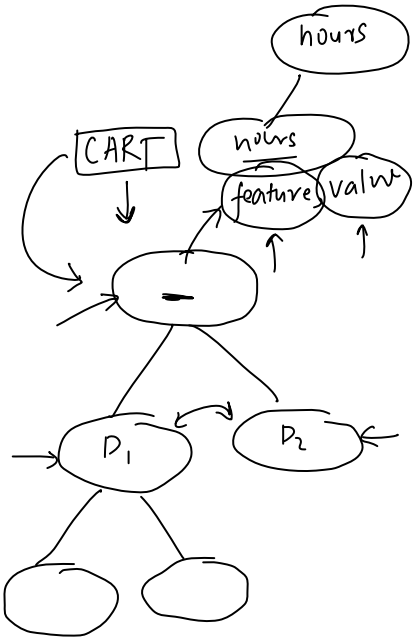
	Subject	Grade_Level	Hours_Studied	Test_Score
0	Math	Freshman	4	59
1	Physics	Freshman	1	82
2	Physics	Freshman	4	81
3	Math	Junior	6	60
4	Physics	Sophomore	1	73
5	Physics	Junior	3	85
6	Physics	Junior	4	61
7	Physics	Freshman	9	78

variance, dataset

binary, multi class, numerical

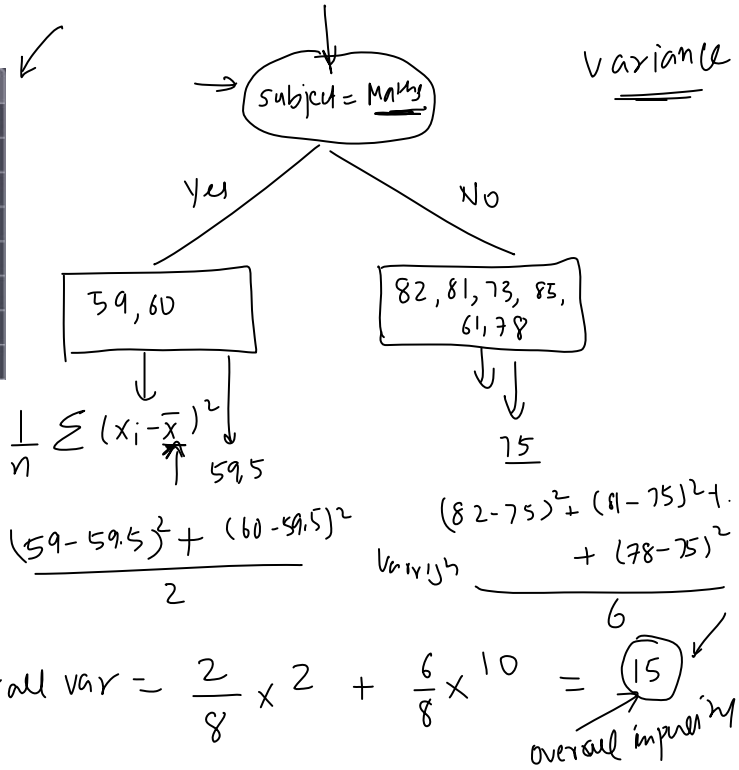
mse, variance → max reduction

impurity function → gini

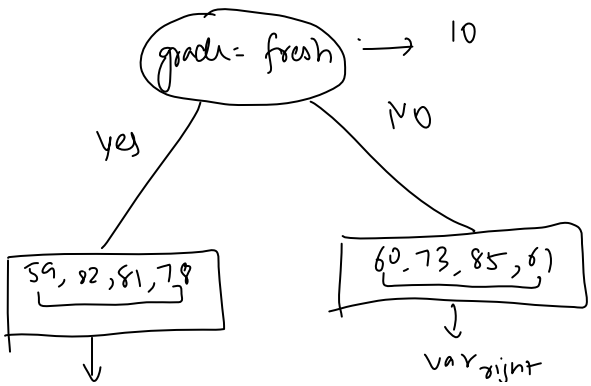


	Subject	Grade_Level	Hours_Studied	Test_Score
0	Math	Freshman	4	59
1	Physics	Freshman	1	82
2	Physics	Freshman	4	81
3	Math	Junior	6	60
4	Physics	Sophomore	1	73
5	Physics	Junior	3	85
6	Physics	Junior	4	61
7	Physics	Freshman	9	78

15, freshman = 7, hours > 2 → 6

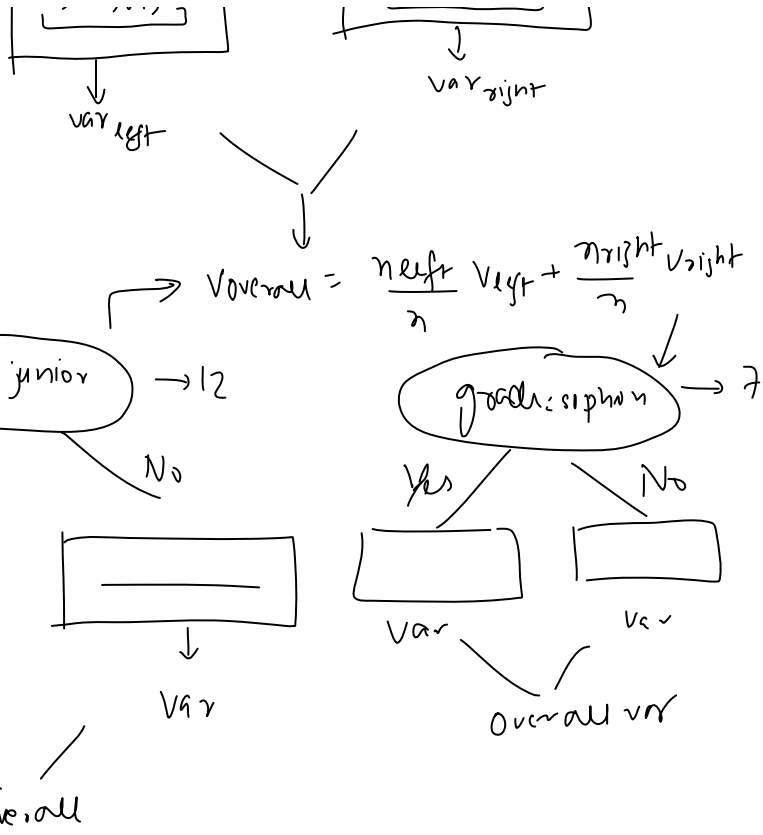


Grade_Level	Hours_Studied	Test_Score
Freshman	4	59
Freshman	1	82
Freshman	4	81
Junior	6	60
Sophomore	1	73
Junior	3	85
Junior	4	61
Freshman	9	78

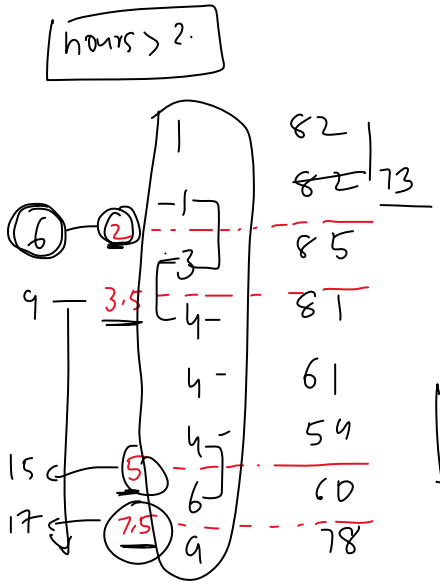


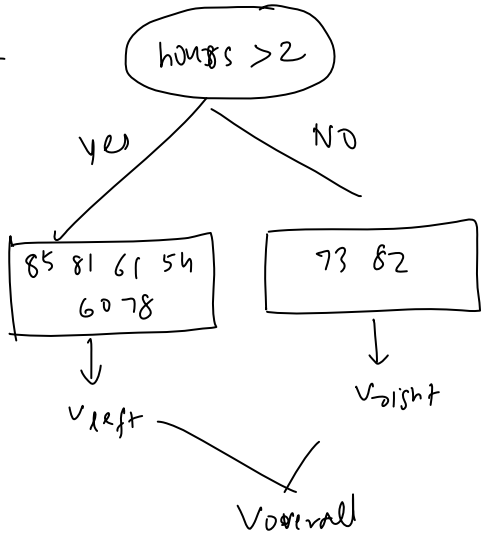
Freshman	9	78

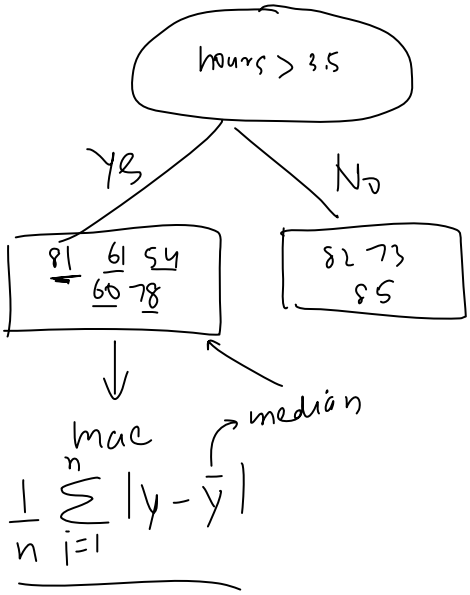
3 cate/
{ freshman, (junior, so phomore) }
{ junior, (freshman, sophomore) }
{ sophomore, (freshman, junior) }



Hours_Studied	Test_Score
4	59
1	82
4	81
6	60
1	73
3	85
4	61
9	78

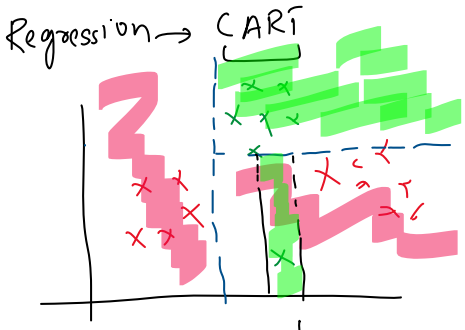




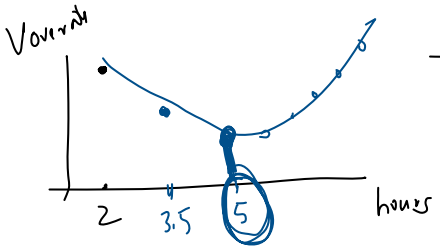


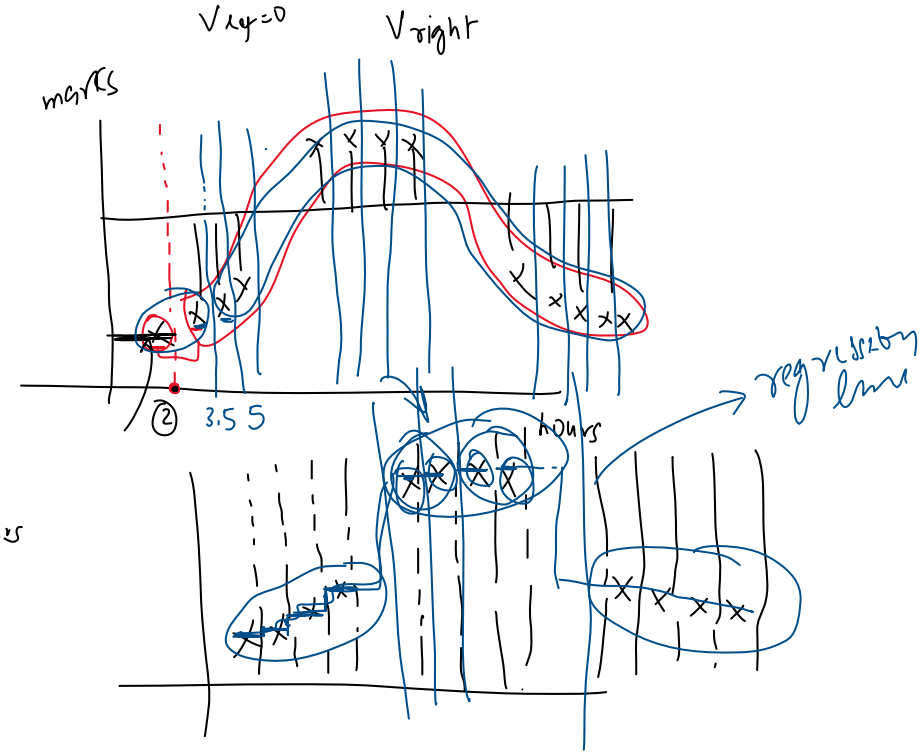
Geometric Intuition

24 July 2023 13:31

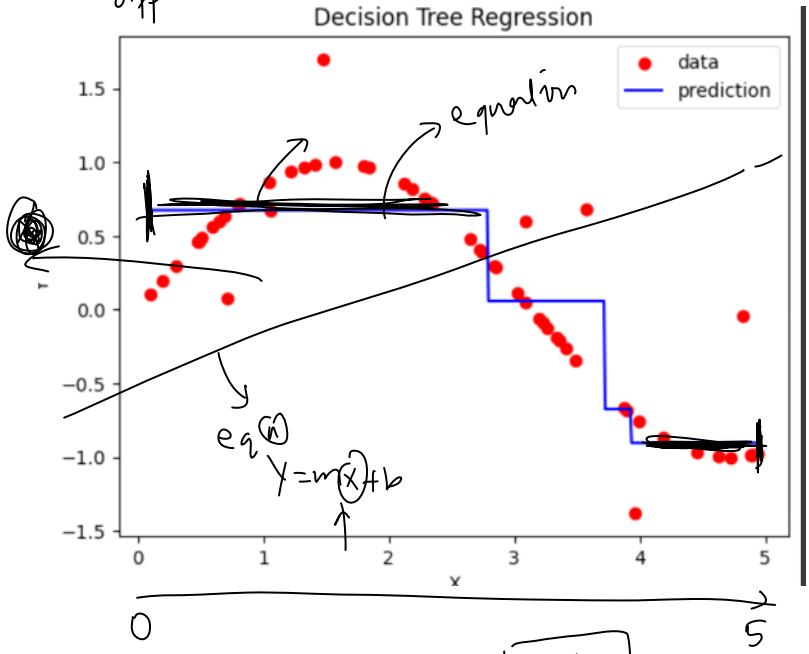


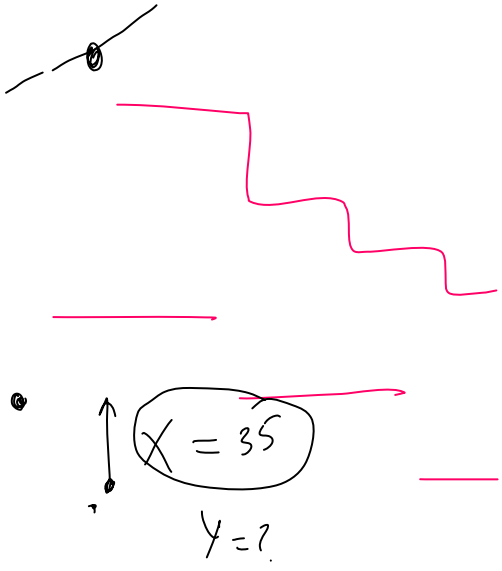
hours	exam result
5	50
6	60
10	30





piecewise constant approximation





marks (0-100)

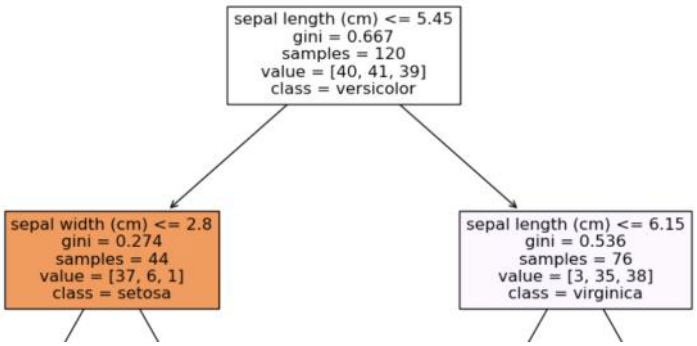
profit dis-

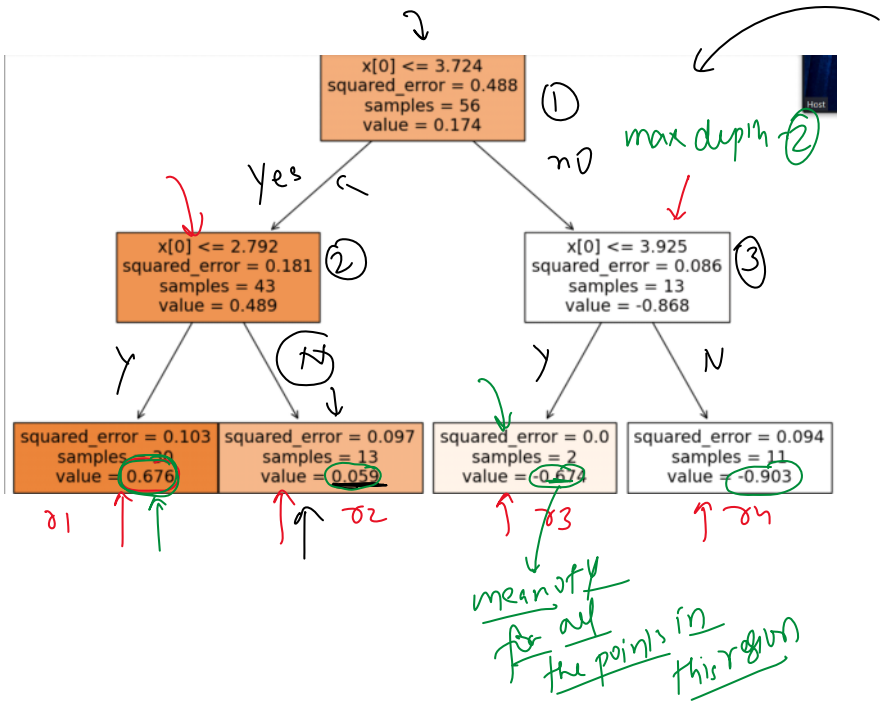
How Prediction is Done

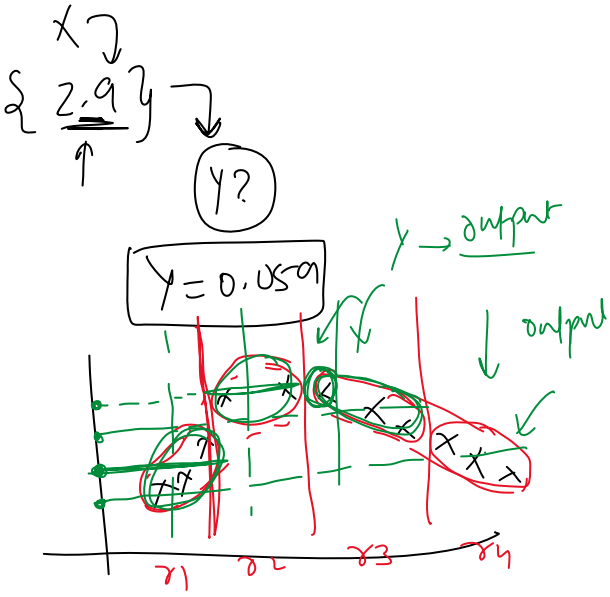
24 July 2023 13:36

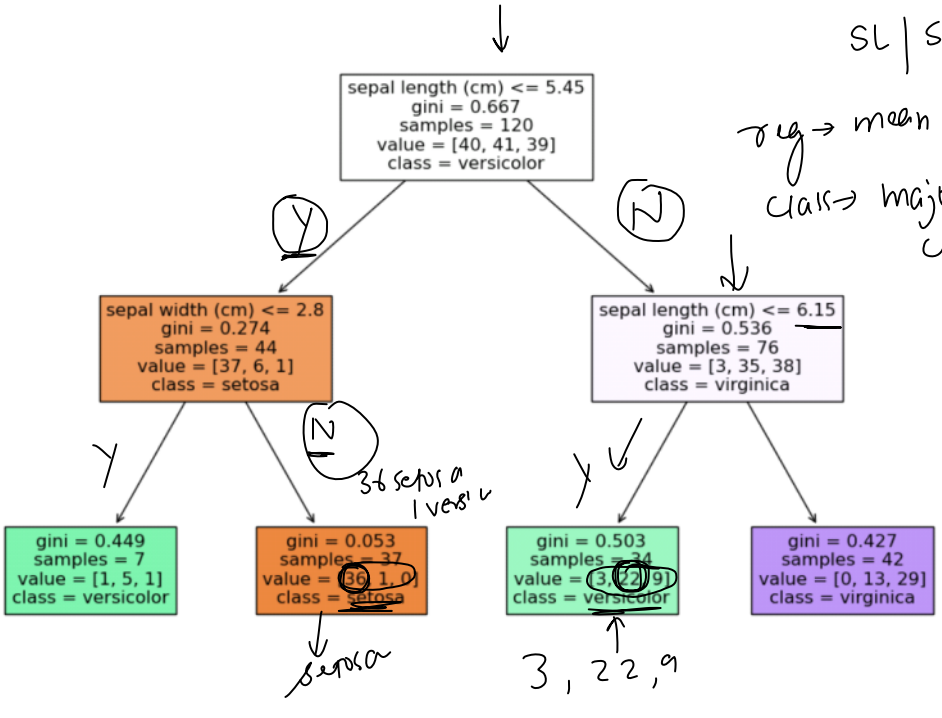
→ Classification
→ Regression

On Leaf node
On Decision Node
Probability output









SL | SW

reg → mean

class → majority count

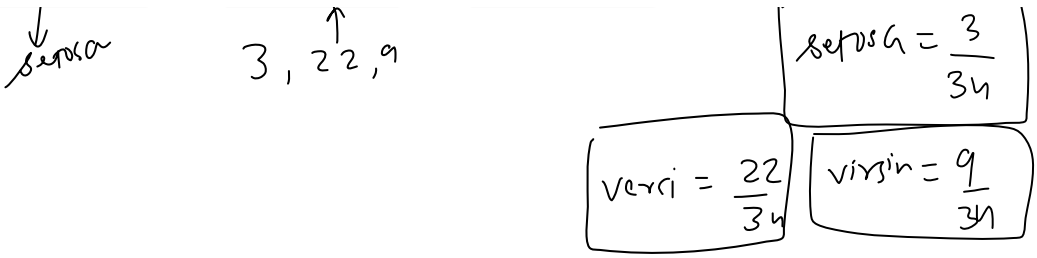
setosa
virginica
versicol

{4.9, 3}
SL SW

{5.2, 2}

{6, 2}

setosa = 3/34


setosa
3, 22, 9
setosa = 3/34
versi = 22/34
virsin = 9/34

Code

24 July 2023 13:32

Advantages & Disadvantages

24 July 2023 13:32

KNN

Advantages

- Simple to understand and to interpret. Trees can be visualized.
- Requires little data preparation. Other techniques often require data normalization, dummy variables need to be created and blank values to be removed. Note however that this module does not support missing values.
- The cost of using the tree (i.e., predicting data) is logarithmic in the number of data points used to train the tree.
- Able to handle both numerical and categorical data.
- Can work on non-linear datasets
- Can give you feature importance.

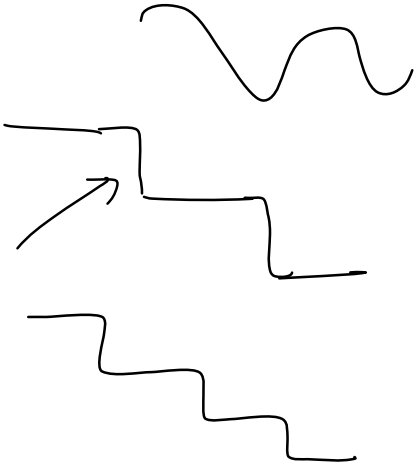
distance (log n)

500 → log(500)

Disadvantages

- Decision-tree learners can create over-complex trees that do not generalize the data well. This is called overfitting. Mechanisms such as pruning, setting the minimum number of samples required at a leaf node or setting the maximum depth of the tree are necessary to avoid this problem.
- Decision trees can be unstable because small variations in the data might result in a completely different tree being generated. This problem is mitigated by using decision trees within an ensemble.
- Predictions of decision trees are neither smooth nor continuous, but piecewise constant approximations as seen in the above figure. Therefore, they are not good at extrapolation.
- This limitation is inherent to the structure of decision tree models. They are very useful for interpretability and for handling non-linear relationships within the range of the training data, but they aren't designed for extrapolation. If extrapolation is important for your task, you might need to consider other types of models.

pruning

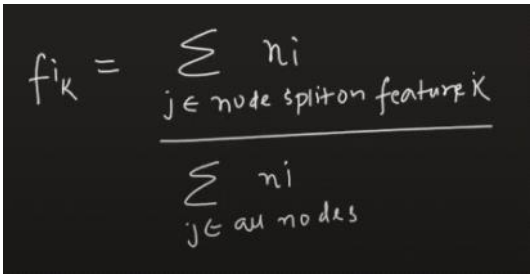


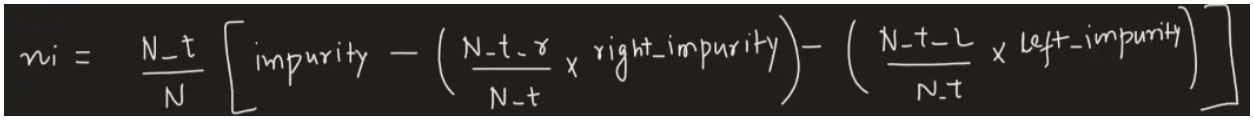
linear

Feature Importance

25 July 2023 17:40

The importance of a feature is computed as the (normalized) total reduction of the criterion brought by that feature. It is also known as the Gini importance.





Regression
-> flat tentative price system(prediction)

Recommender System
-> Suggest more flats like this
-> Society suggestion

Analysis
-> City Level
-> Sector Level
-> Insight System(Factors) -> inference(ml model)

Deploy on AWS

CI/CD pipelines
efficient