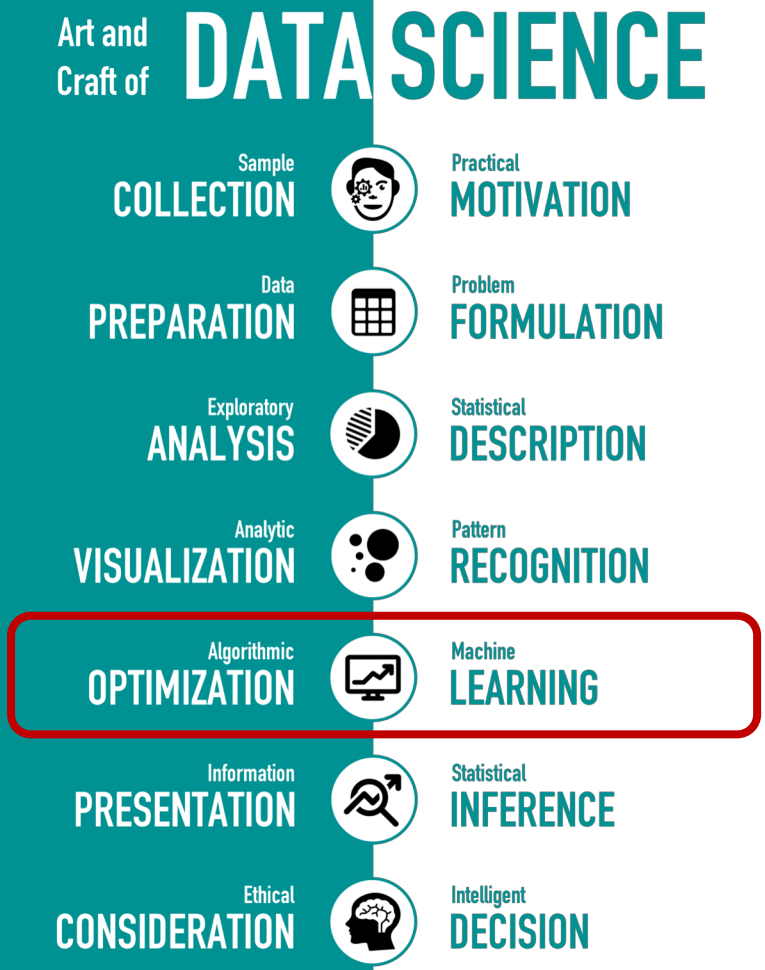
NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE

Binary Classification

Sourav SEN GUPTA
Lecturer, SCSE, NTU



Data Science

Binary Classification

Machine Learning

Are variables mutually dependent?

How to find relation between them?

How to predict one using another?

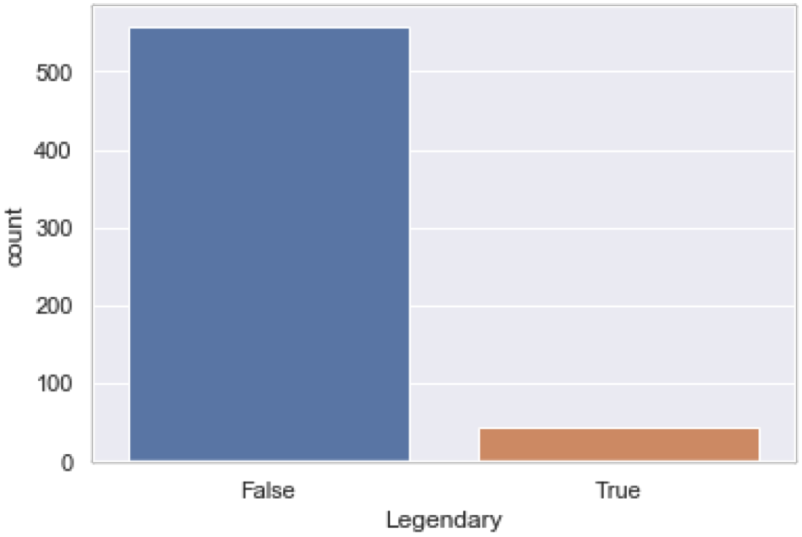
How to optimally learn from the Data?

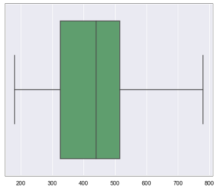
Data Science

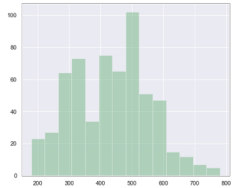
The Pokemon Dataset

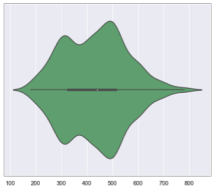
#	Name	Type 1	Type 2	Total	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed	Generation	Legendary
430	Honchkrow	Dark	Flying	505	100	125	52	105	52	71	4	False
338	Solrock	Rock	Psychic	440	70	95	85	55	65	70	3	False
32	Nidoran♂	Poison	NaN	273	46	57	40	40	40	50	1	False
442	Spiritomb	Ghost	Dark	485	50	92	108	92	108	35	4	False
480	Uxie	Psychic	NaN	580	75	75	130	75	130	95	4	True
536	Palpitoad	Water	Ground	384	75	65	55	65	55	69	5	False
360	Wynaut	Psychic	NaN	260	95	23	48	23	48	23	3	False
478	Froslass	Ice	Ghost	480	70	80	70	80	70	110	4	False
76	Golem	Rock	Ground	495	80	120	130	55	65	45	1	False
177	Natu	Psychic	Flying	320	40	50	45	70	45	70	2	False

Source : Kaggle Datasets | **Pokemon with stats** by Alberto Barradas | https://www.kaggle.com/abcsds/pokemon









Train Dataset

75% of the Data

Data Science

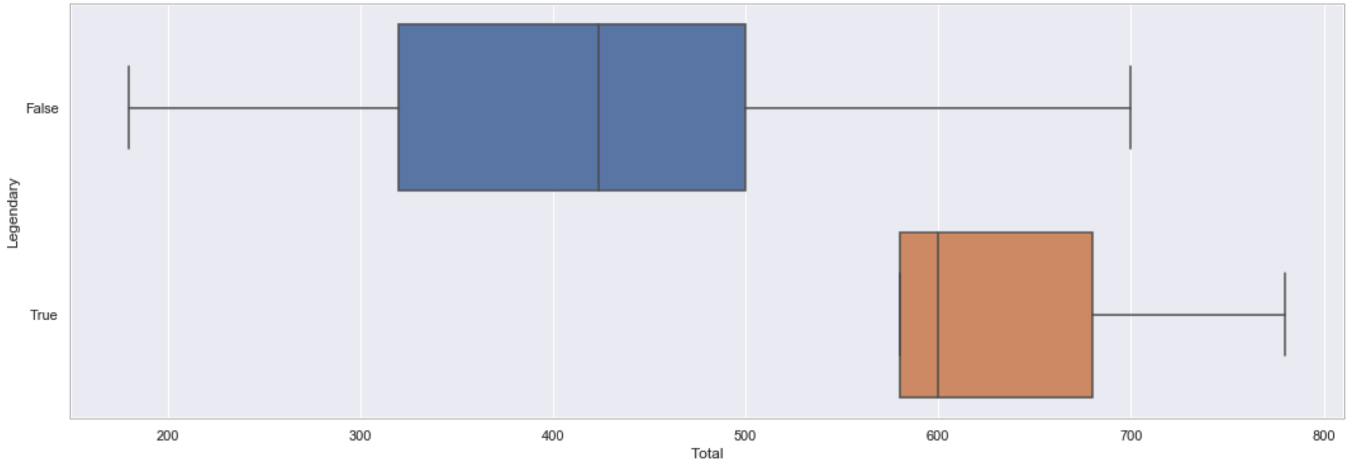
Bi-Variate Exploration

Statistical Summary

Legendary	True or False
Total	Total Points

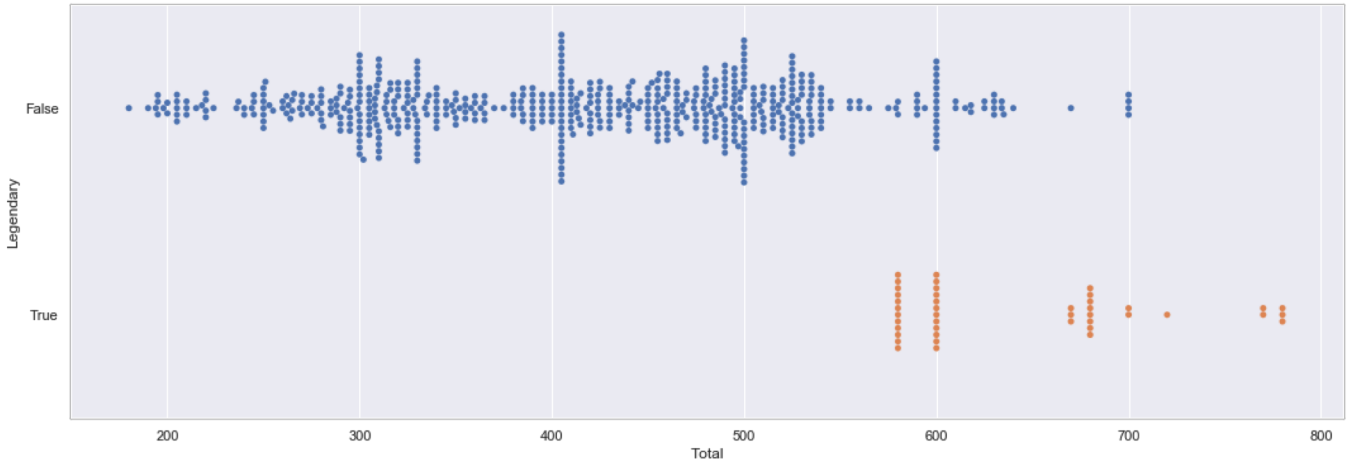
Machine Learning Questions

- What is the mutual relationship?
- Can we predict Legendary by Total?



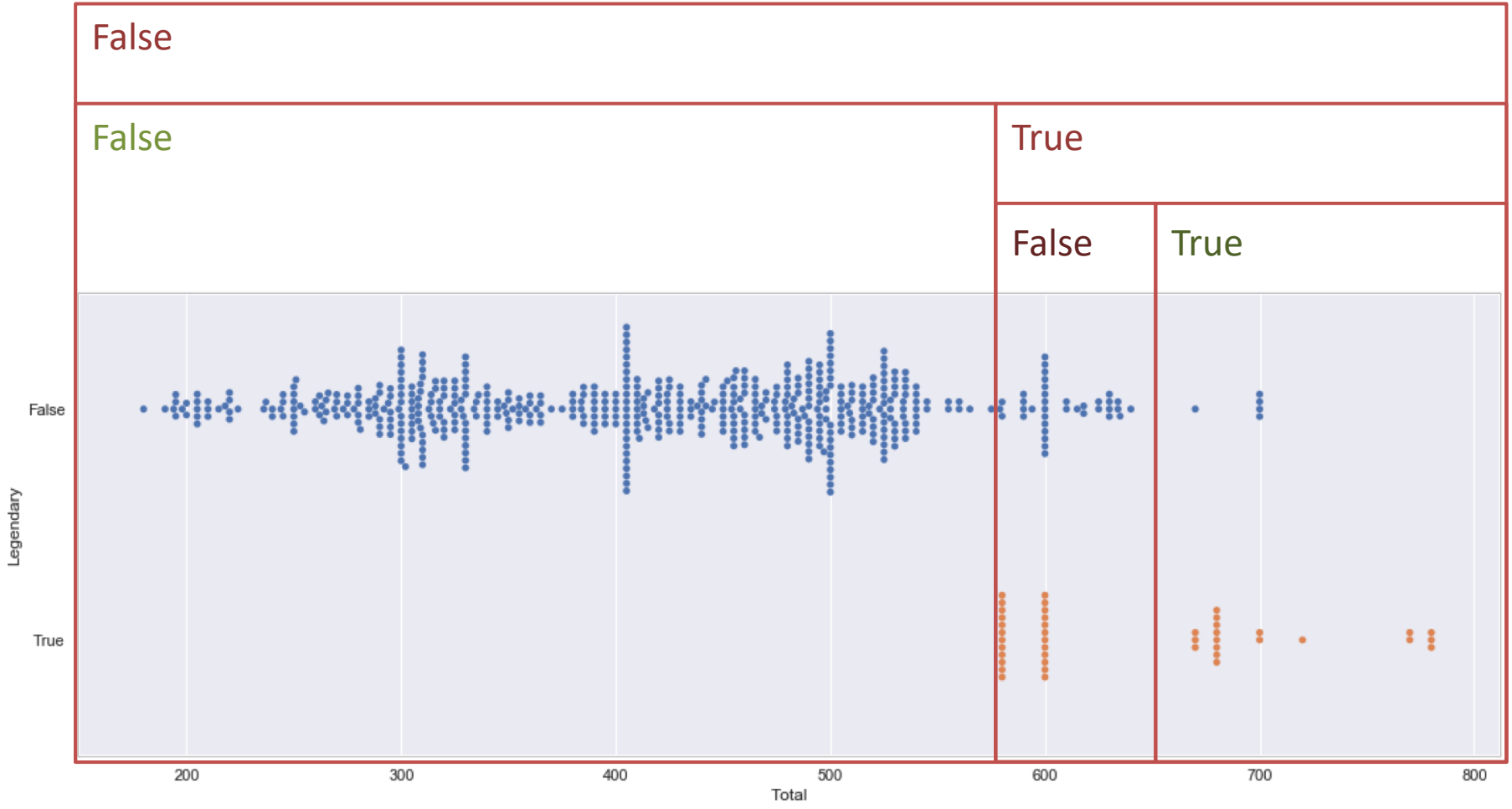
BoxPlot

Train Dataset
75% of the Data



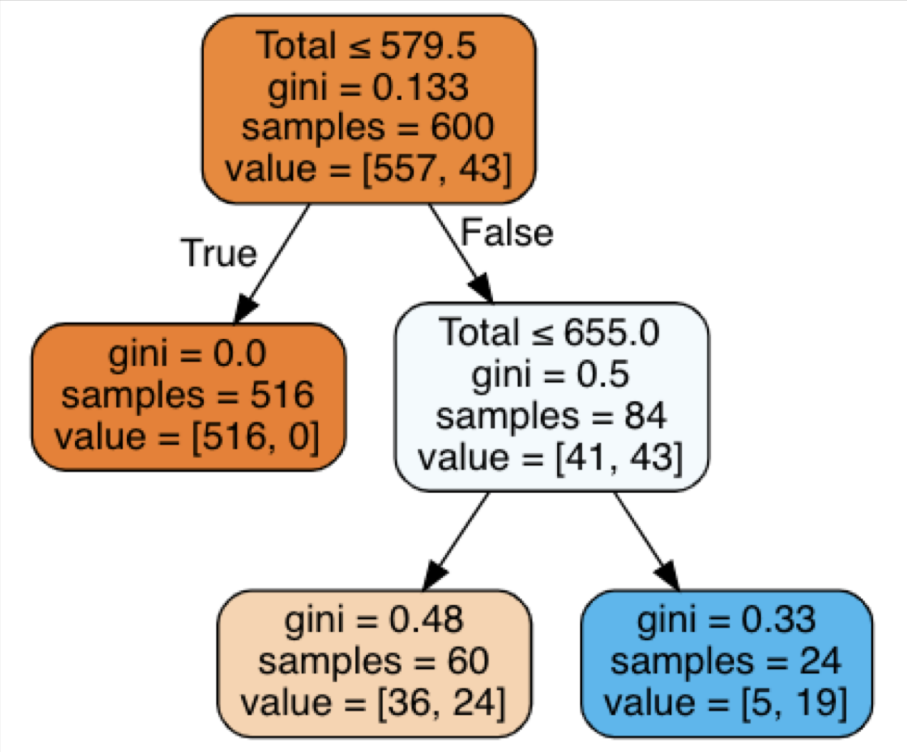
SwarmPlot

Train Dataset
75% of the Data


False
False
True
False
True
False
True
Legendary
200
300
400
500
600
700
800
Total

Binary Classification

Decision Tree



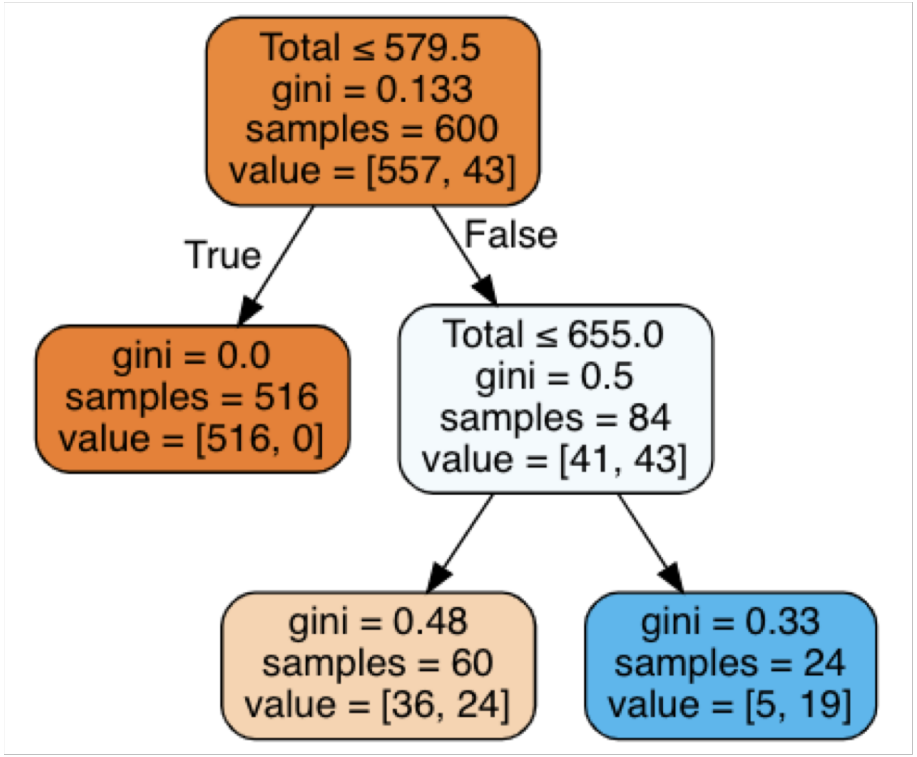
Partitions made in the Data Space methodically represented using consecutive Binary Decisions

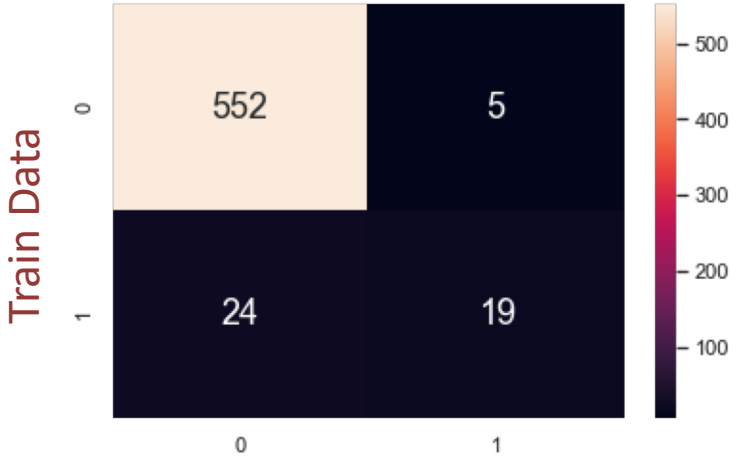
Decision of Partition depends on the Gini Index (metric of misclassification)

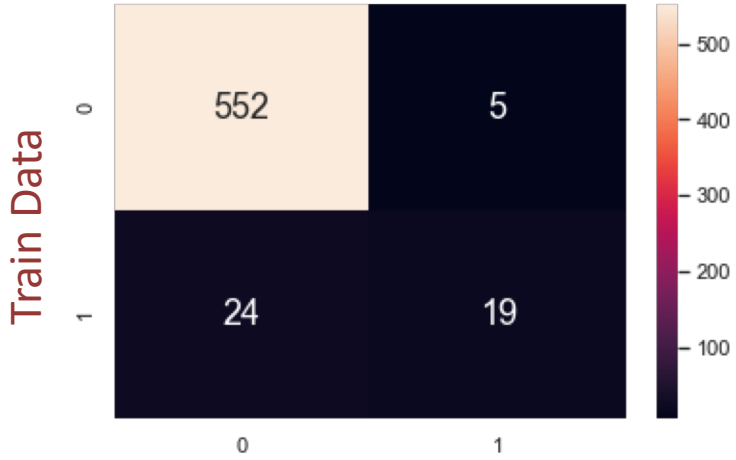
$$gini = \frac{x}{n}\left(1 - \frac{x}{n}\right) + \frac{y}{n}\left(1 - \frac{y}{n}\right)$$

Binary Classification

Prediction using Decision Tree







Data Science

Binary Classification

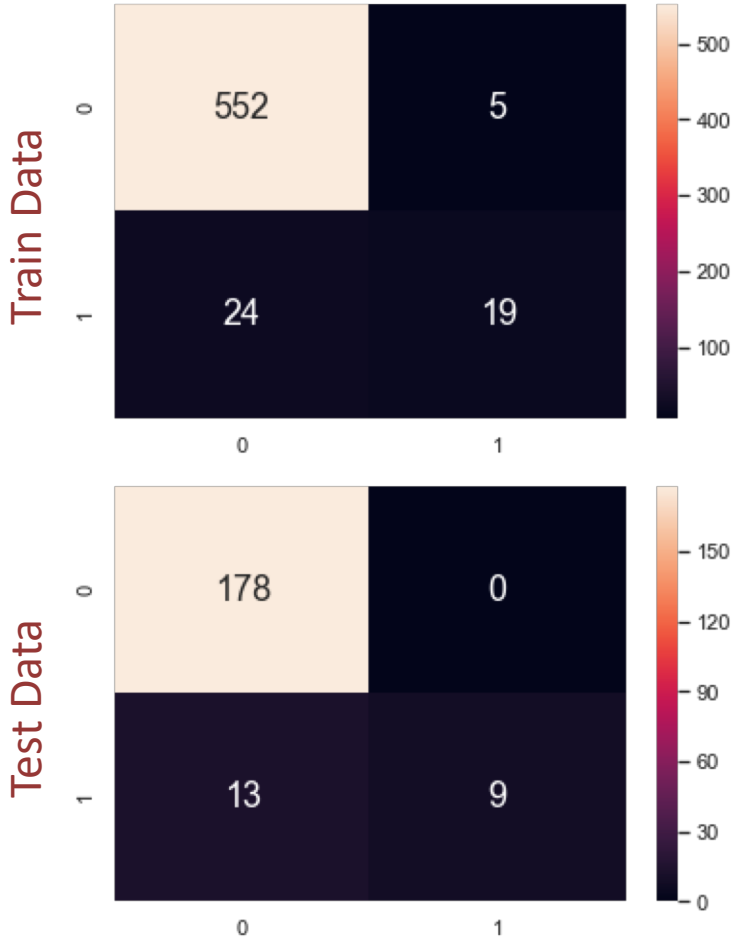
Goodness of Fit of the Model

Classification Accuracy

Fraction of Correct Predictions

TP : True predicted as True
TN : False predicted as False

Accuracy in Train Data	0.952
Accuracy in Test Data	0.935



Data Science

Binary Classification

Goodness of Fit of the Model

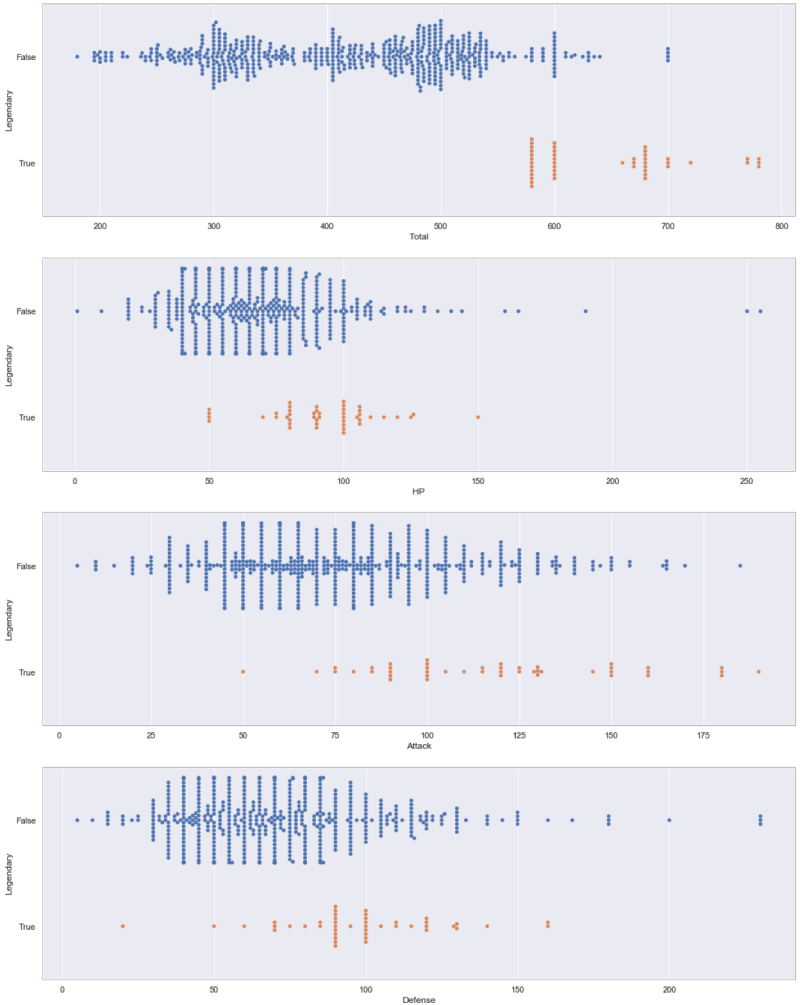
Classification Errors

FN : True predicted as False

FP : False predicted as True

Train $fpr = \frac{5}{557}$, $fnr = \frac{24}{43}$

Test $fpr = \frac{0}{178}$, $fnr = \frac{13}{22}$



Data Science

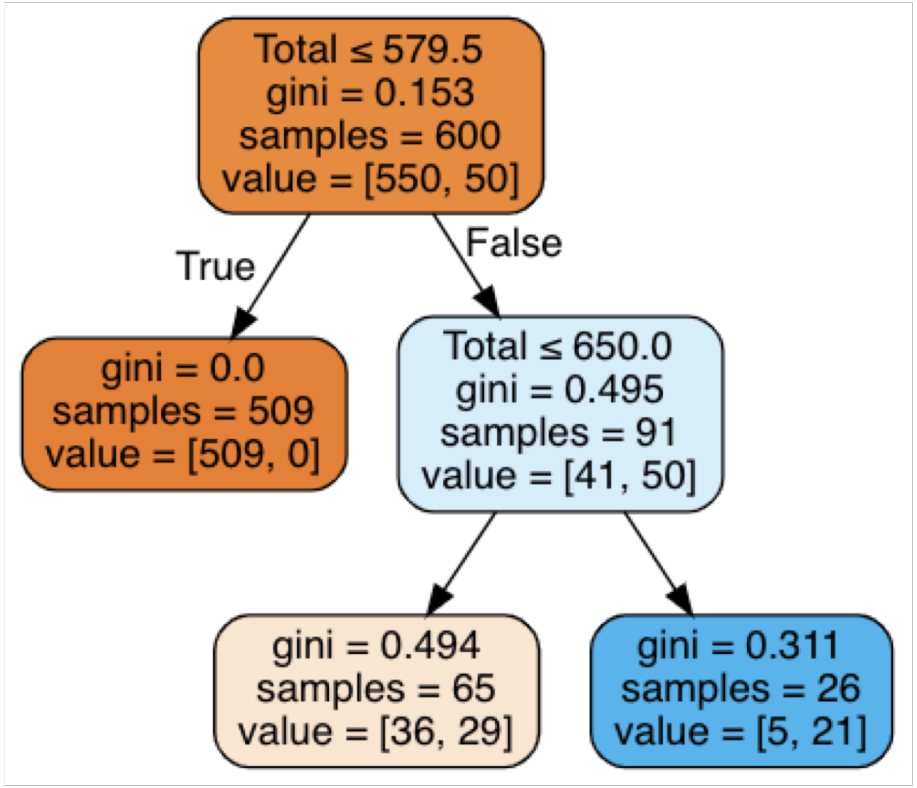
Binary Classification

Multi-Variate Decision Tree

Legendary	True or False
Total	Total Points
HP	Hit Points
Attack	Attack Points
Defense	Defense Points

Swarm Plots per Predictor

Binary Classification

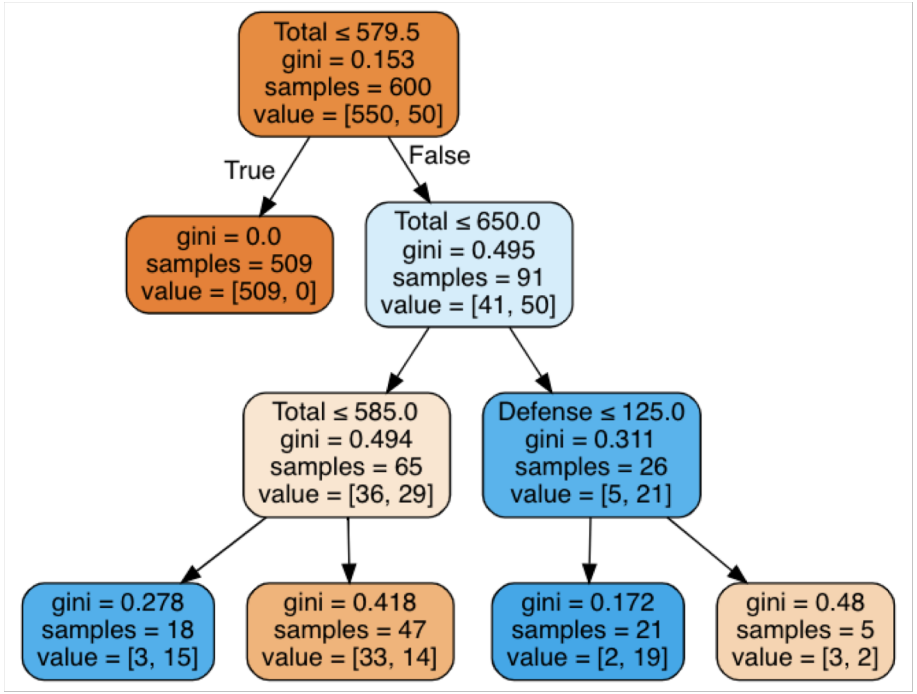


Multi-Variate Decision Tree

Legendary	True or False
Total	Total Points
HP	Hit Points
Attack	Attack Points
Defense	Defense Points

Two-Level Decision Tree

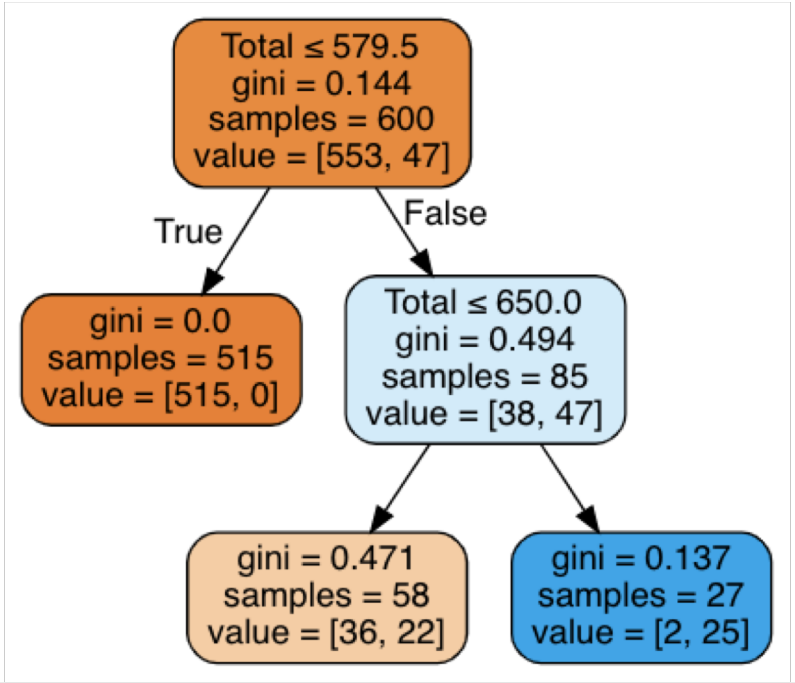
Binary Classification

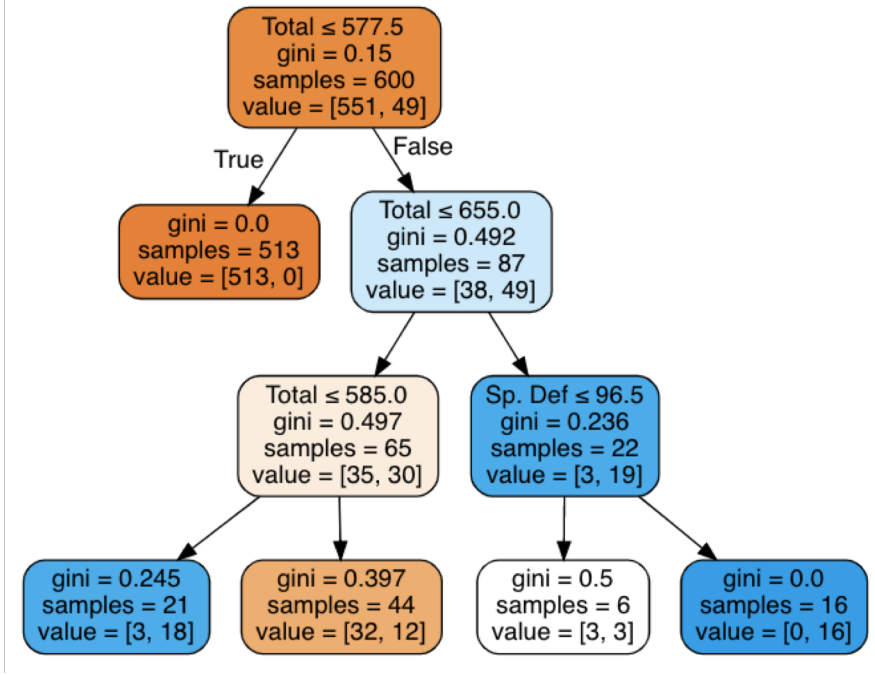


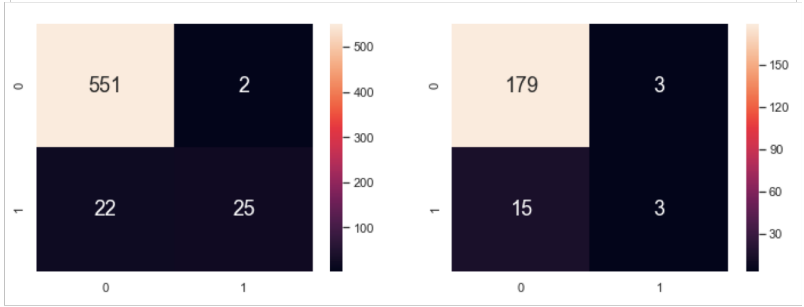
Multi-Variate Decision Tree

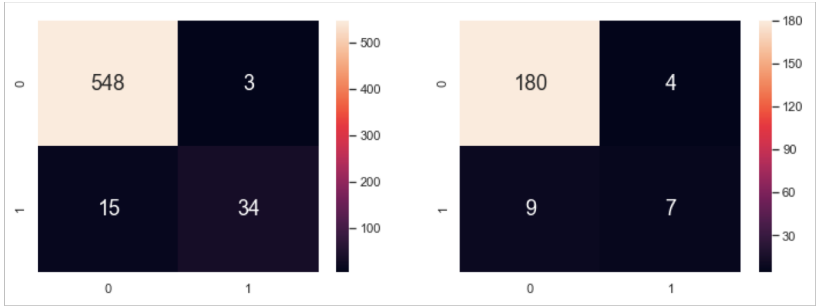
Legendary	True or False
Total	Total Points
HP	Hit Points
Attack	Attack Points
Defense	Defense Points

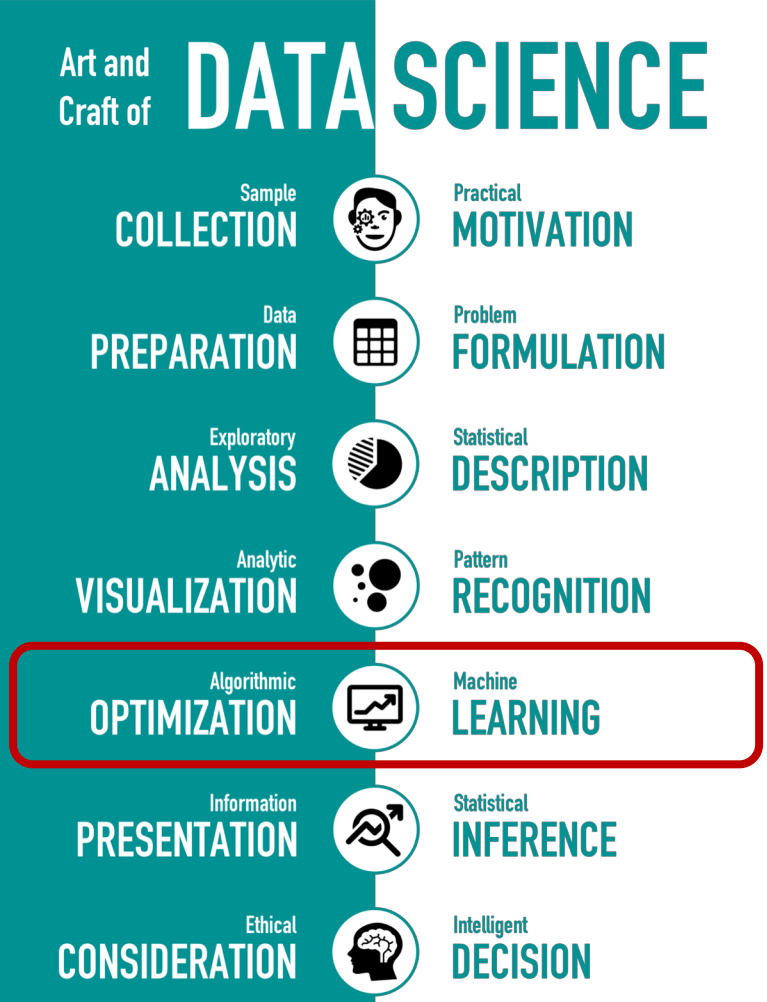
Three-Level Decision Tree


Total ≤ 579.5
gini = 0.144
samples = 600
value = [553, 47]
True
False
gini = 0.0
samples = 515
value = [515, 0]
Total ≤ 650.0
gini = 0.494
samples = 85
value = [38, 47]
gini = 0.471
samples = 58
value = [36, 22]
gini = 0.137
samples = 27
value = [2, 25]


Total ≤ 577.5
gini = 0.15
samples = 600
value = [551, 49]
True
False
gini = 0.0
samples = 513
value = [513, 0]
Total ≤ 655.0
gini = 0.492
samples = 87
value = [38, 49]
Total ≤ 585.0
gini = 0.497
samples = 65
value = [35, 30]
Sp. Def ≤ 96.5
gini = 0.236
samples = 22
value = [3, 19]
gini = 0.245
samples = 21
value = [3, 18]
gini = 0.397
samples = 44
value = [32, 12]
gini = 0.5
samples = 6
value = [3, 3]
gini = 0.0
samples = 16
value = [0, 16]


551
2
22
25
179
3
15
3


548
3
15
34
180
4
9
7



Data Science Pipeline

Machine Learning

How to learn from the acquired Data?

How to model the acquired Data?

How to predict on new Data?

How to optimally learn from the Data?