Data Preprocessing

Exercise 1

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

Organisational - Exercises

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

Date	Topic
28.02.2022	Data Preprocessing
07.03.2022	Ensembles
14.03.2022	Time Series
21.03.2022	Neural Networks & Deep Learning
28.03.2022	Anomaly Detection
04.04.2022	Hyperparameter Tuning
11.04.2022	- Easter Break -
18.04.2022	- Easter Break -
25.04.2022	Model Verification
02.05. – 23.05.2022	- Data Mining Cup -

Organisational - Data Mining Cup

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- Task published on **April, 12th**
- Submission due on **June, 28th** (internal submission on **21st**)
- Team Setup
  - 8-10 students per team
  - We need one volunteer to register for the DMC (data-mining-cup.com)
  - You will have to form teams on your own
  - Use this sheet to search for teammates and submit your team setup: **https://docs.google.com/spreadsheets/d/1wGPZBTPo6p9xZeoc3HomYoxyLNqkdCp-j8yRX0c3OBQ/edit?usp=sharing**

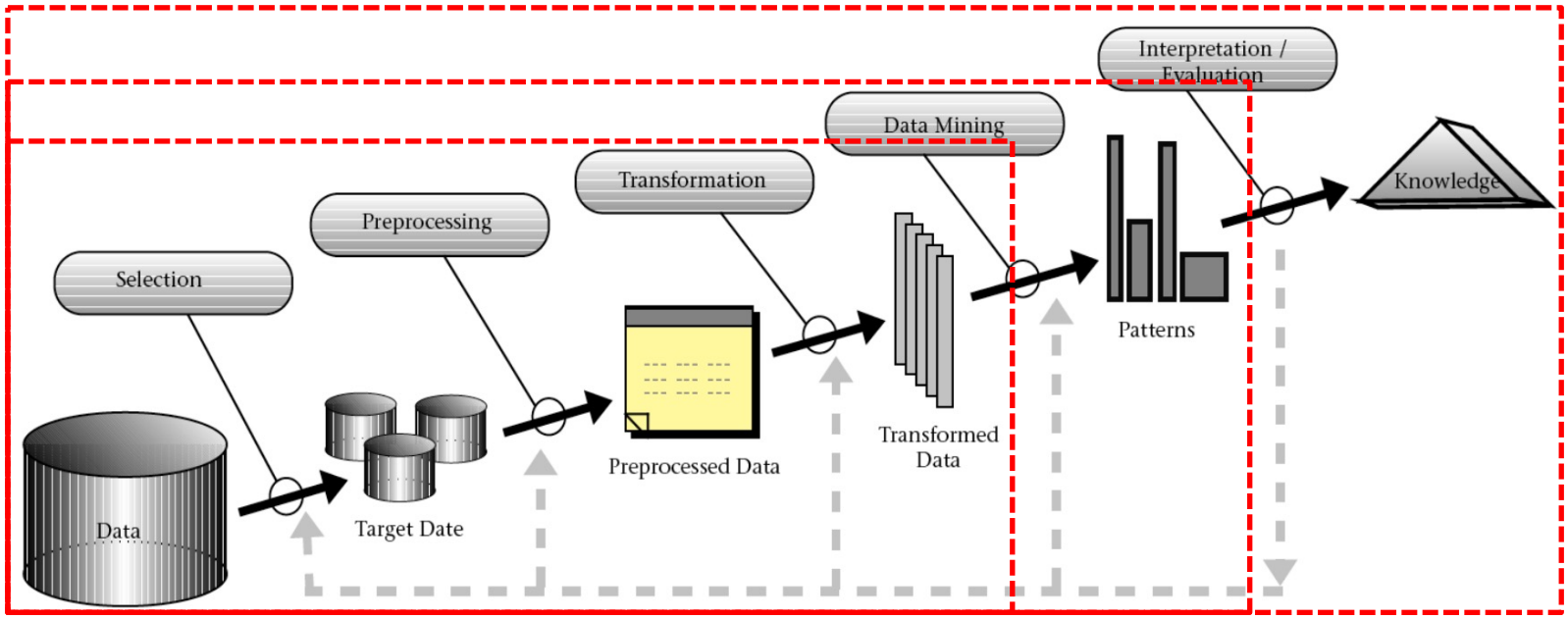
Python Exercises

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- Feel free to ask as many questions as you want (to me and to your peers!)
- Jupyter sheets for exercises are uploaded to ILIAS before the exercise
- Solutions to exercises are uploaded right after the exercise
- Python exercises are **optional and won't be discussed in detail here!**

Introduction

Data Mining Workflow

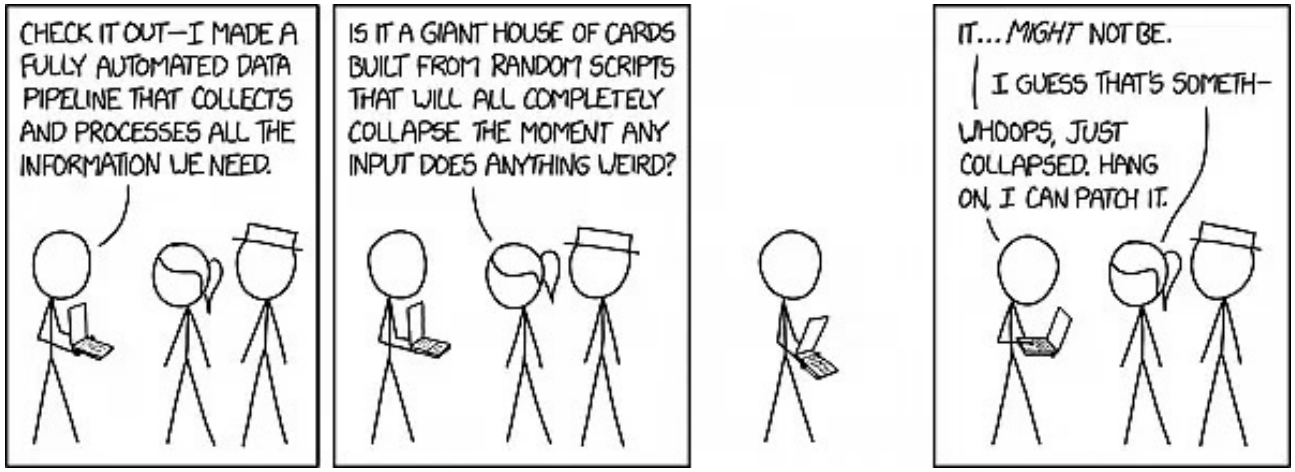


Source: Fayyad et al. (1996)

Why Preprocessing?

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

- Overall goal is **improved analysis/prediction performance**
  - Always make sure to evaluate the impact of your preprocessing methods
  - Remember Occam's razor



Source: xkcd.com

Missing Values

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

Passenger Class	Name	Sex	Age	No of Siblings ...	No of Parents ...	Ticket Nu...	Passenger Fare	Cabin	Port of Emba...	Life Boat	Survived
First	Carter, Master. William Thornton II	Male	11	1	2	113760	120	B96 B98	Southampton	4	Yes
First	Carter, Miss. Lucile Polk	Female	14	1	2	113760	120	B96 B98	Southampton	4	Yes
First	Carter, Mr. William Ernest	Male	36	1	2	113760	120	B96 B98	Southampton	C	Yes
First	Carter, Mrs. William Ernest (Lucile Polk)	Female	36	1	2	113760	120	B96 B98	Southampton	4	Yes
First	Rood, Mr. Hugh Roscoe	Male	?	0	0	113767	50	A32	Southampton	?	No
First	Marvin, Mr. Daniel Warner	Male	19	1	0	113773	53.100	D30	Southampton	?	No
First	Marvin, Mrs. Daniel Warner (Mary Graham Carmi...	Female	18	1	0	113773	53.100	D30	Southampton	10	Yes
First	Pears, Mr. Thomas Clinton	Male	29	1	0	113776	66.600	C2	Southampton	?	No
First	Pears, Mrs. Thomas (Edith Wearne)	Female	22	1	0	113776	66.600	C2	Southampton	8	Yes
First	Franklin, Mr. Thomas Parham	Male	?	0	0	113778	26.550	D34	Southampton	?	No
First	Gracie, Col. Archibald IV	Male	53	0	0	113780	28.500	C51	Cherbourg	B	Yes
First	Allison, Master. Hudson Trevor	Male	0.917	1	2	113781	151.550	C22 C26	Southampton	11	Yes
First	Allison, Miss. Helen Loraine	Female	2	1	2	113781	151.550	C22 C26	Southampton	?	No
First	Allison, Mr. Hudson Joshua Creighton	Male	30	1	2	113781	151.550	C22 C26	Southampton	?	No
First	Allison, Mrs. Hudson J C (Bessie Waldo Daniels)	Female	25	1	2	113781	151.550	C22 C26	Southampton	?	No
First	Cleaver, Miss. Alice	Female	22	0	0	113781	151.550	?	Southampton	11	Yes
First	Daniels, Miss. Sarah	Female	33	0	0	113781	151.550	?	Southampton	8	Yes
First	Bonnell, Miss. Elizabeth	Female	58	0	0	113783	26.550	C103	Southampton	8	Yes
First	Blackwell, Mr. Stephen Weart	Male	45	0	0	113784	35.500	T	Southampton	?	No
First	Peuchen, Major. Arthur Godfrey	Male	52	0	0	113786	30.500	C104	Southampton	6	Yes
First	Molson, Mr. Harry Markland	Male	55	0	0	113787	30.500	C30	Southampton	?	No
First	Sloper, Mr. William Thompson	Male	28	0	0	113788	35.500	A6	Southampton	7	Yes
First	Holverson, Mr. Alexander Oskar	Male	42	1	0	113789	52	?	Southampton	?	No
First	Holverson, Mrs. Alexander Oskar (Mary Aline To...	Female	35	1	0	113789	52	?	Southampton	8	Yes

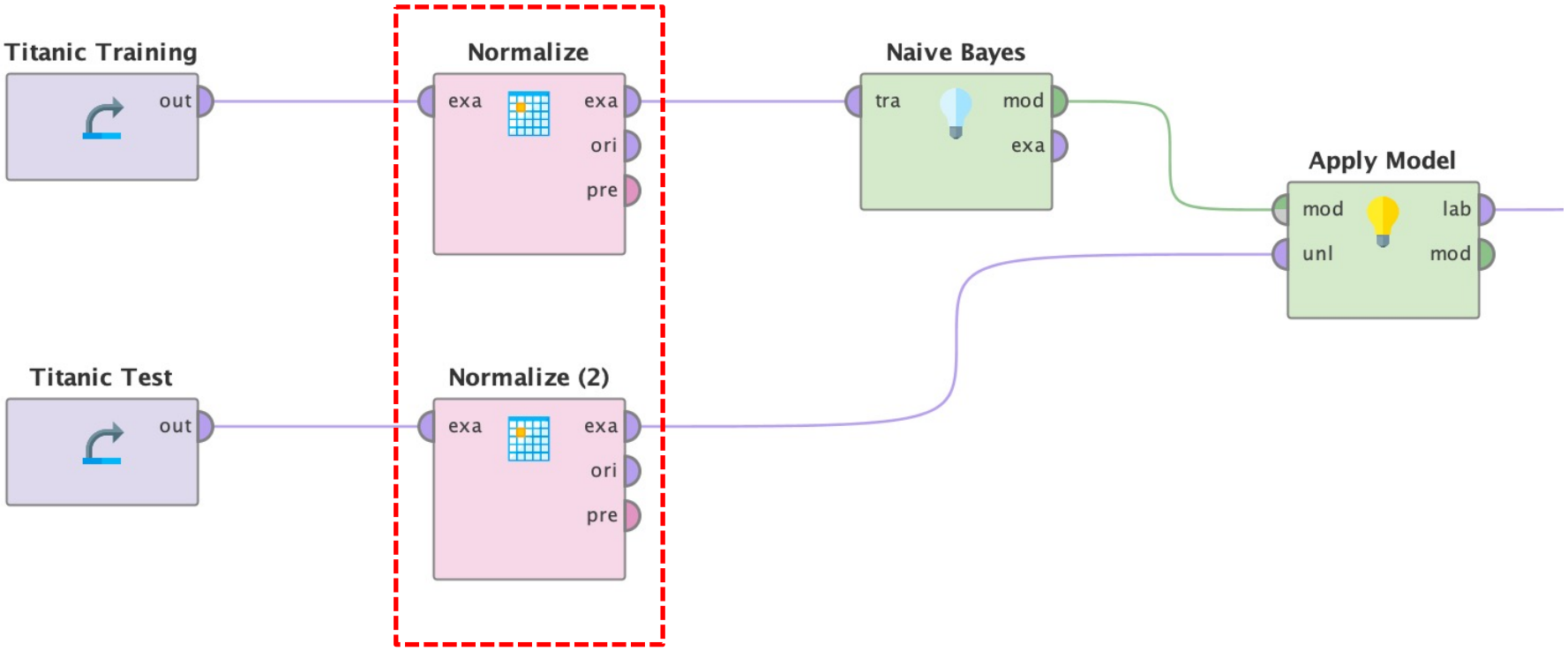
Unsupported Data Types

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

Passenger Class	Name	Sex	Age	No of Siblings ...	No of Parents ...	Ticket Nu...	Passenger Fare	Cabin	Port of Emba...	Life Boat	Survived
First	Carter, Master. William Thornton II	Male	11	1	2	113760	120	B96 B98	Southampton	4	Yes
First	Carter, Miss. Lucile Polk	Female	14	1	2	113760	120	B96 B98	Southampton	4	Yes
First	Carter, Mr. William Ernest	Male	36	1	2	113760	120	B96 B98	Southampton	C	Yes
First	Carter, Mrs. William Ernest (Lucile Polk)	Female	36	1	2	113760	120	B96 B98	Southampton	4	Yes
First	Rood, Mr. Hugh Roscoe	Male	?	0	0	113767	50	A32	Southampton	?	No
First	Marvin, Mr. Daniel Warner	Male	19	1	0	113773	53.100	D30	Southampton	?	No
First	Marvin, Mrs. Daniel Warner (Mary Graham Carmi...	Female	18	1	0	113773	53.100	D30	Southampton	10	Yes
First	Pears, Mr. Thomas Clinton	Male	29	1	0	113776	66.600	C2	Southampton	?	No
First	Pears, Mrs. Thomas (Edith Wearne)	Female	22	1	0	113776	66.600	C2	Southampton	8	Yes
First	Franklin, Mr. Thomas Parham	Male	?	0	0	113778	26.550	D34	Southampton	?	No
First	Gracie, Col. Archibald IV	Male	53	0	0	113780	28.500	C51	Cherbourg	B	Yes
First	Allison, Master. Hudson Trevor	Male	0.917	1	2	113781	151.550	C22 C26	Southampton	11	Yes
First	Allison, Miss. Helen Loraine	Female	2	1	2	113781	151.550	C22 C26	Southampton	?	No
First	Allison, Mr. Hudson Joshua Creighton	Male	30	1	2	113781	151.550	C22 C26	Southampton	?	No
First	Allison, Mrs. Hudson J C (Bessie Waldo Daniels)	Female	25	1	2	113781	151.550	C22 C26	Southampton	?	No
First	Cleaver, Miss. Alice	Female	22	0	0	113781	151.550	?	Southampton	11	Yes
First	Daniels, Miss. Sarah	Female	33	0	0	113781	151.550	?	Southampton	8	Yes
First	Bonnell, Miss. Elizabeth	Female	58	0	0	113783	26.550	C103	Southampton	8	Yes
First	Blackwell, Mr. Stephen Weart	Male	45	0	0	113784	35.500	T	Southampton	?	No
First	Peuchen, Major. Arthur Godfrey	Male	52	0	0	113786	30.500	C104	Southampton	6	Yes
First	Molson, Mr. Harry Markland	Male	55	0	0	113787	30.500	C30	Southampton	?	No
First	Sloper, Mr. William Thompson	Male	28	0	0	113788	35.500	A6	Southampton	7	Yes
First	Holverson, Mr. Alexander Oskar	Male	42	1	0	113789	52	?	Southampton	?	No
First	Holverson, Mrs. Alexander Oskar (Mary Aline To...	Female	35	1	0	113789	52	?	Southampton	8	Yes

Data Transformation

Pitfalls

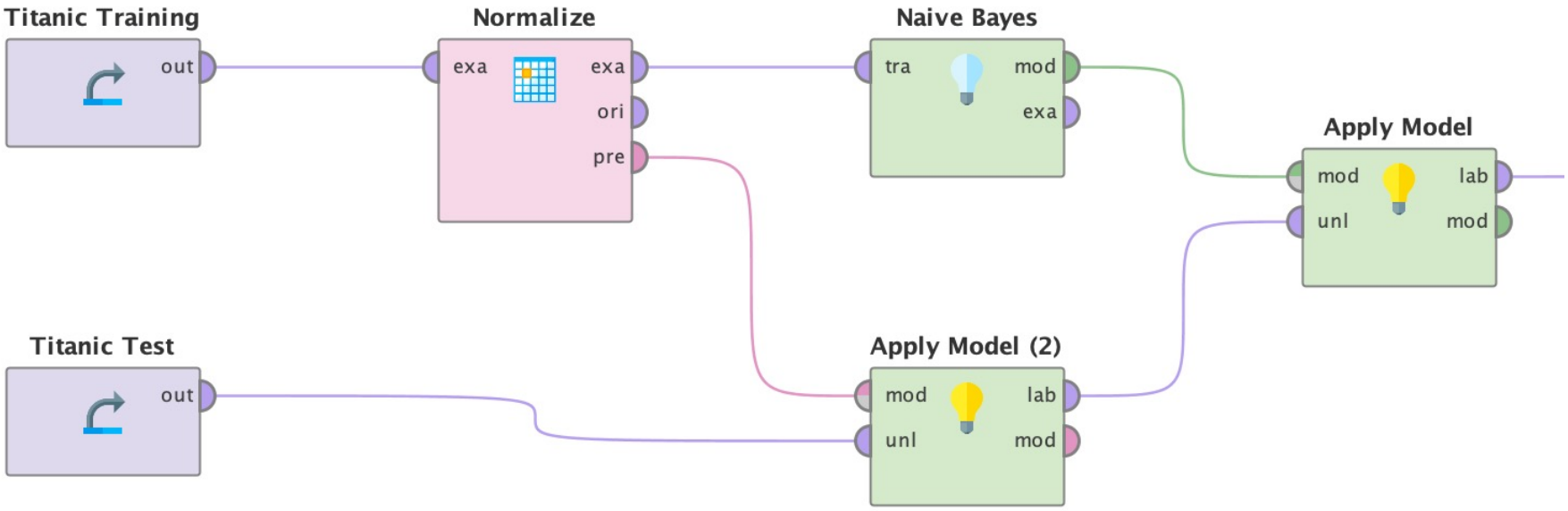


TODO: Is this setup correct?

No! Training and Test data may be normalized on different scales as they can have different minima/maxima and distributions.

Data Transformation

Pitfalls

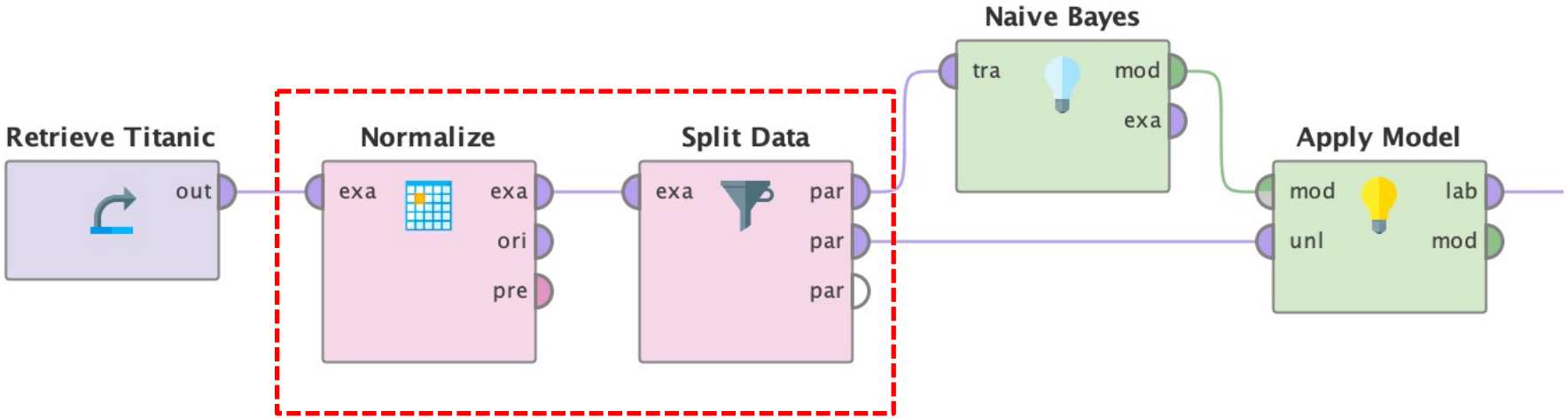


TODO: Is this setup correct?

Yes! But remember to deal with unknown values in Test data.

Data Transformation

Pitfalls



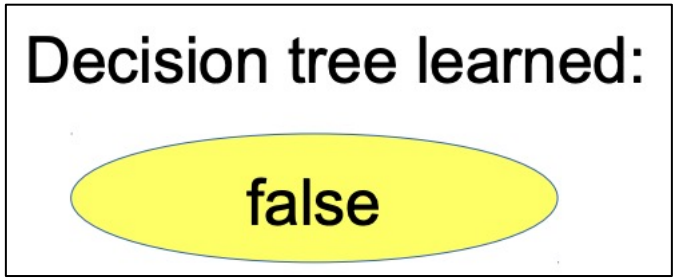
TODO: Is this setup correct?

No (in most cases)!
Potential data leakage from Test into Training.
Especially problematic for time series data.

Unbalanced Data

Problems

- Dumb but effective models



  - Sampling (Manual, SMOTE,..)
  - Use model parameters
    - E.g. XGBoost:

`scale_pos_weight` [default=1]

- Control the balance of positive and negative weights, useful for unbalanced classes. A typical value to consider: `sum(negative instances) / sum(positive instances)` . See Parameters

- Difficult to evaluate

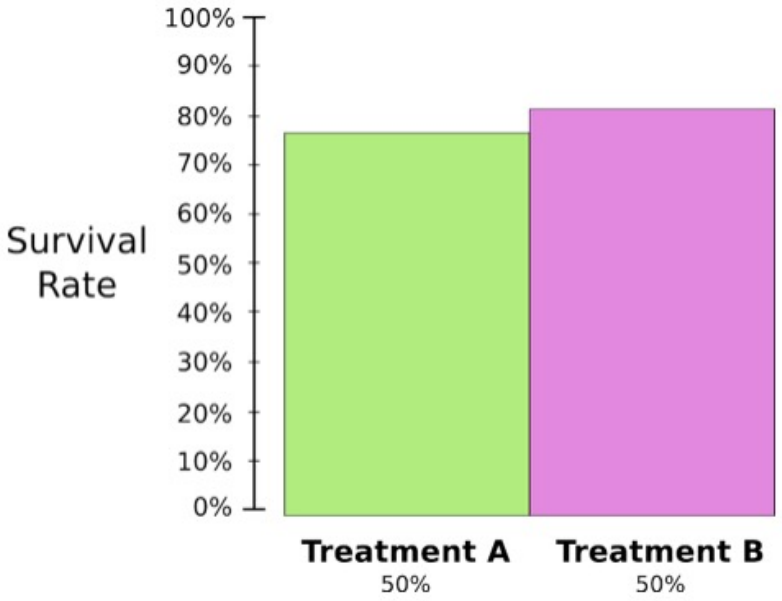
  - ~~Accuracy?~~
  - Precision?
  - Recall?
  - F1?
  - … what about Regression?

Unbalanced Data

Simpson's Paradox

Two treatments for kidney stones are tested. Half the patients are given treatment A while the other half are given treatment B. The patients who received treatment B were more likely to survive than those who received treatment A.

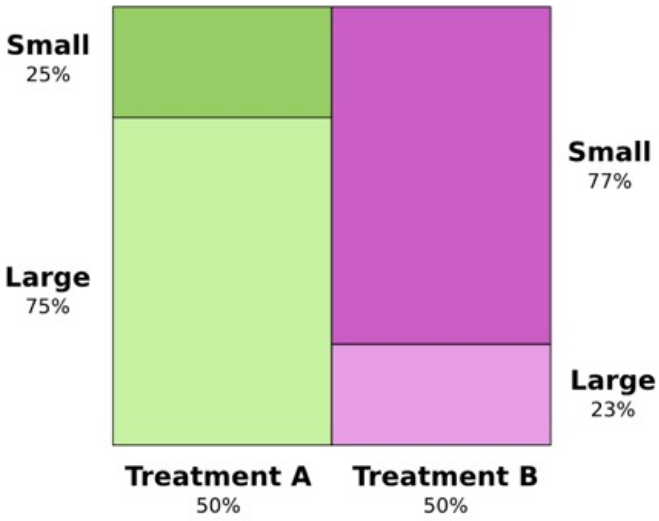
However, patients with small kidney stones were more likely to survive if they took treatment A. Patients with large kidney stones were also more likely to survive if they took treatment A!
How can this be?



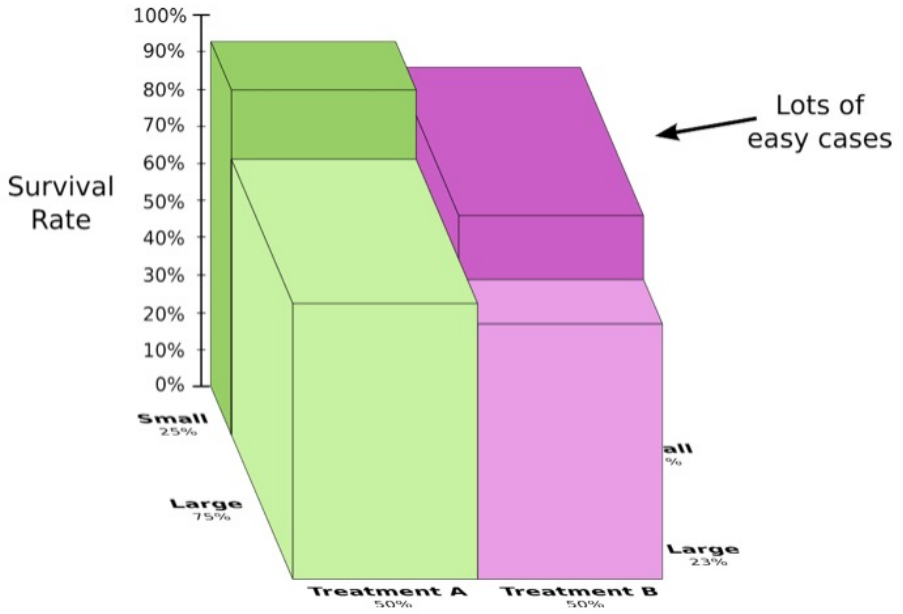
Unbalanced Data

Simpson's Paradox

The core of the issue is that the study wasn't properly randomized. The patients who received treatment A were likely to have large kidney stones, while the patients who received treatment B were more likely to have small kidney stones.



As it turns out, patients with small kidney stones are much more likely to survive in general.

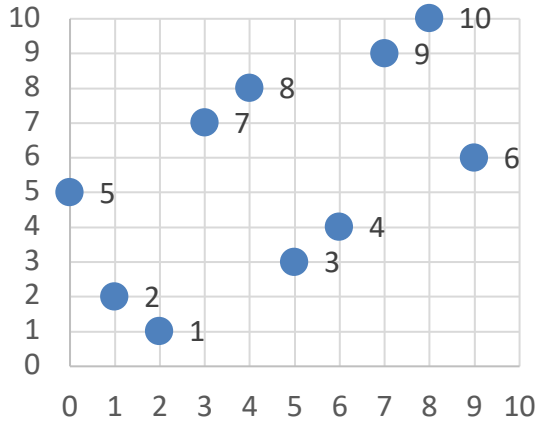


http://colah.github.io/posts/2015-09-Visual-Information/

Sampling

Kennard-Stone Sampling

1) Compute pairwise distances of points
2) Add points with largest distance from one another
3) While target sample size not reached
    1) For each candidate, find smallest distance to any point in the sample
    2) Add candidate with largest smallest distance



TODO:
Draw a 30% sample with Kennard-Stone sampling.
Use Manhattan distance as distance measure.

Sampling

Kennard-Stone Sampling

UNIVERSITY OF MANNHEIM
School of Business Informatics and Mathematics

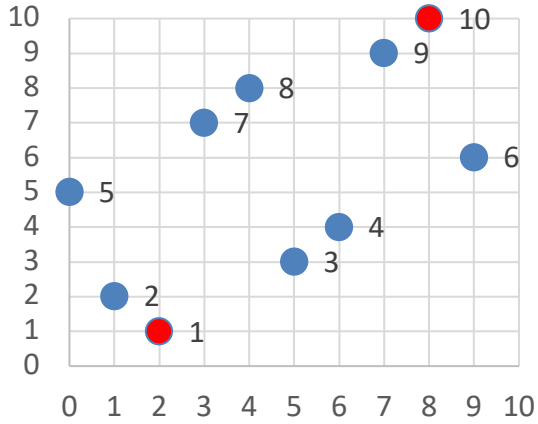
TODO:
Draw a 30% sample with Kennard-Stone sampling.
Use Manhattan distance as distance measure.

	1	2	3	4	5	6	7	8	9	10
1	-	2	5	7	6	12	7	9	13	15
2		-	5	7	4	12	7	9	13	15
3			-	2	7	7	6	6	8	10
4				-	7	5	6	6	6	8
5					-	10	5	7	11	13
6						-	7	7	5	5
7							-	2	6	8
8								-	4	6
9									-	2
10										-

Select point with highest minimum for STEP 3

STEP 1

STEP 2



STEP 3