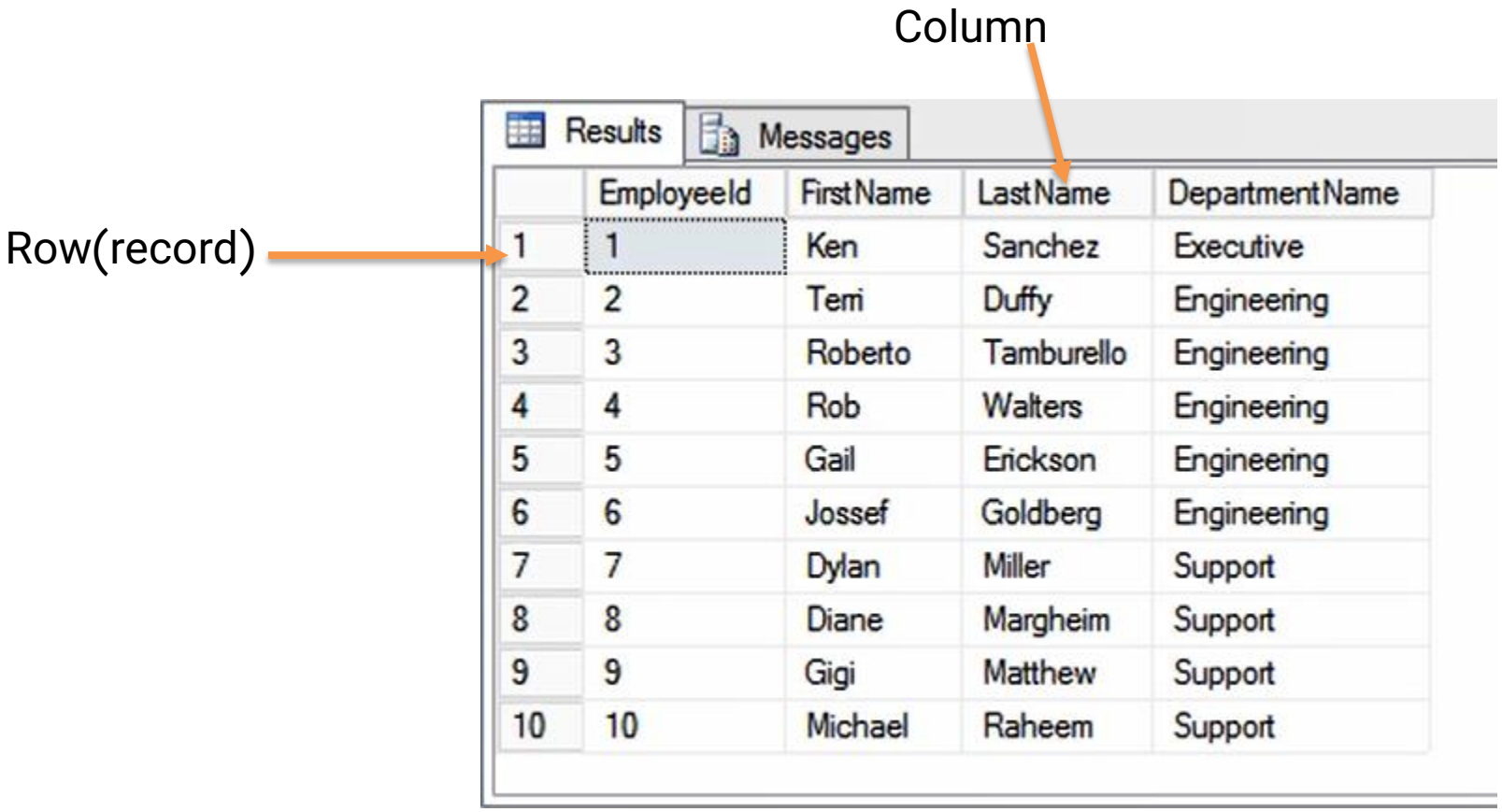
یادگیری ماشین

(پیش پردازش ۱)

محمد دهقانی

ویژگی(Attribute)

ویژگی، توصیفی در یک زمینه خاص از نمونه ارائه می‌کند.



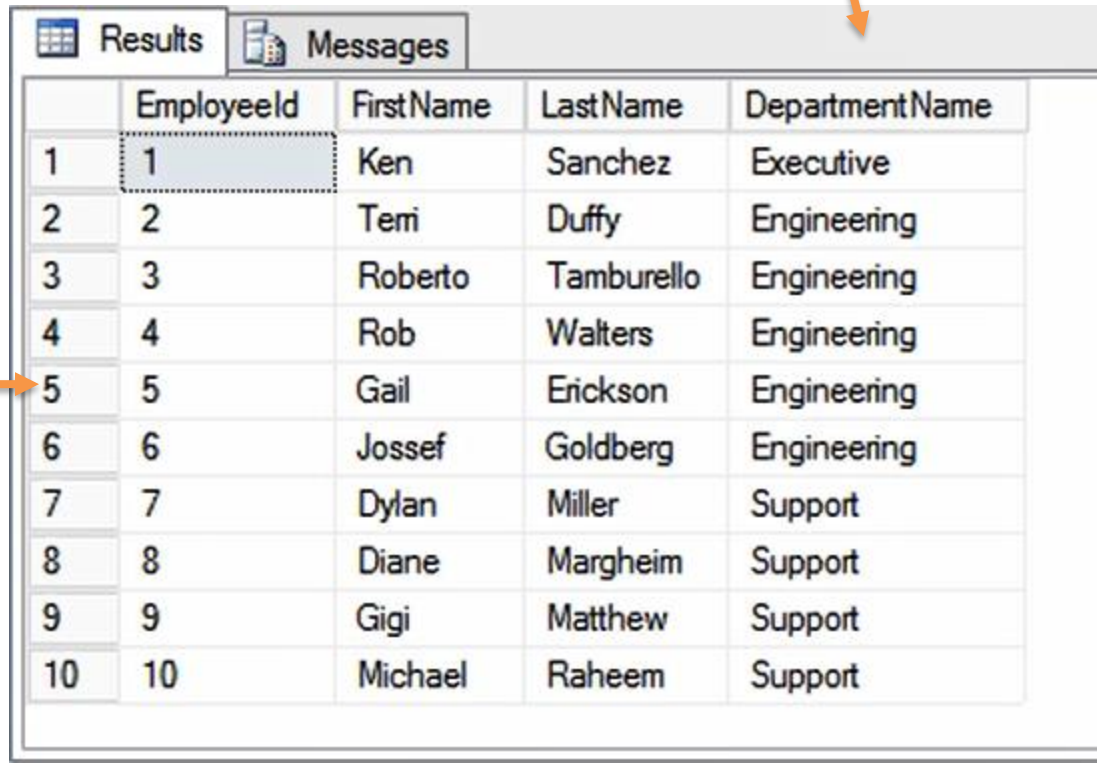
نمونه(Case)

- یک موجودیت پایه از اطلاعات می‌باشد که عملیات داده کاوی بر روی آن انجام می‌شود.
- هر نمونه شامل مجموعه ای از ویژگی‌ها می‌باشد.
- دو نوع نمونه داریم.

نمونه(ادامه)

Attribute

Case



Results | Messages

	EmployeeId	FirstName	LastName	DepartmentName
1	1	Ken	Sanchez	Executive
2	2	Terri	Duffy	Engineering
3	3	Roberto	Tamburello	Engineering
4	4	Rob	Walters	Engineering
5	5	Gail	Erickson	Engineering
6	6	Jossef	Goldberg	Engineering
7	7	Dylan	Miller	Support
8	8	Diane	Margheim	Support
9	9	Gigi	Matthew	Support
10	10	Michael	Raheem	Support

انواع ویژگی

- عددی (Numeric)
- غیرعددی (Categorical)

انواع ویژگی های عددی(کمی)

- پیوسته (Continuous) می‌توانند هر مقداری را در یک بازه از اعداد حقیقی بپذیرند.
- گسسته (Discrete) یک قلم داده که دارای مجموعه متناهی از مقادیر است

انواع ویژگی های غیرعددی(کیفی)

- عادی (nominal) (گروه خونی)
- دودویی (binary) (سالم یا بیمار)
- ترتیبی (ordinal) (درجات نظامی)

آماده سازی داده(Data Preparation)

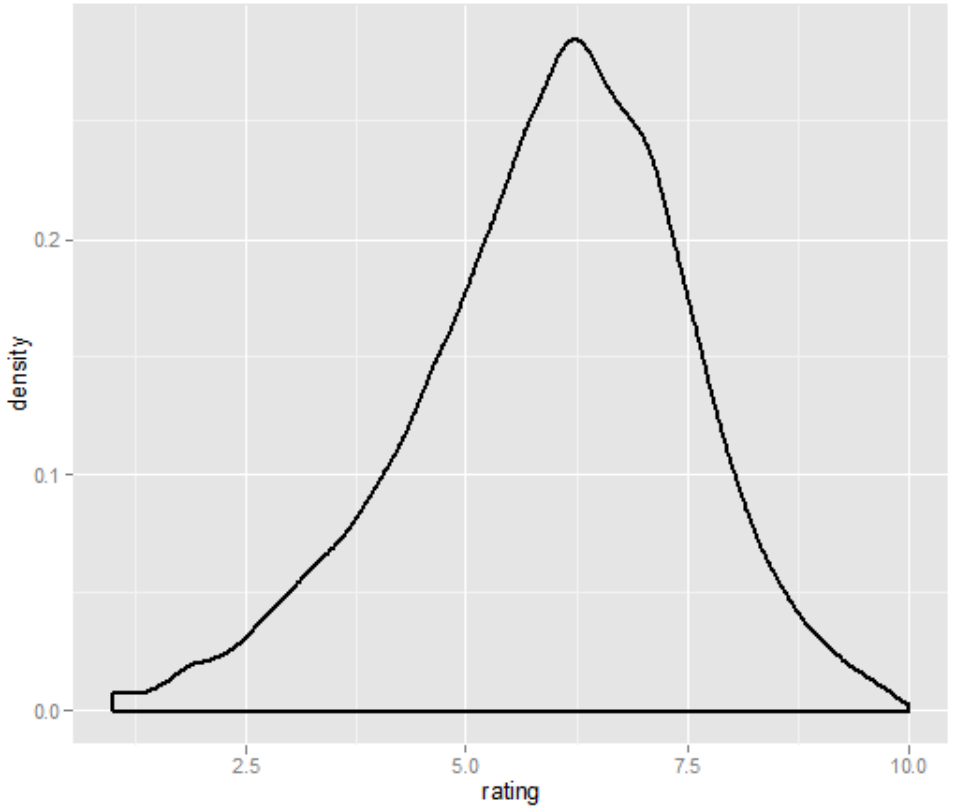
مجموعه کارهایی که منجر به پالایش داده ها شده به صورتی که قابل استفاده در الگوریتم های یادگیری ماشین باشد.

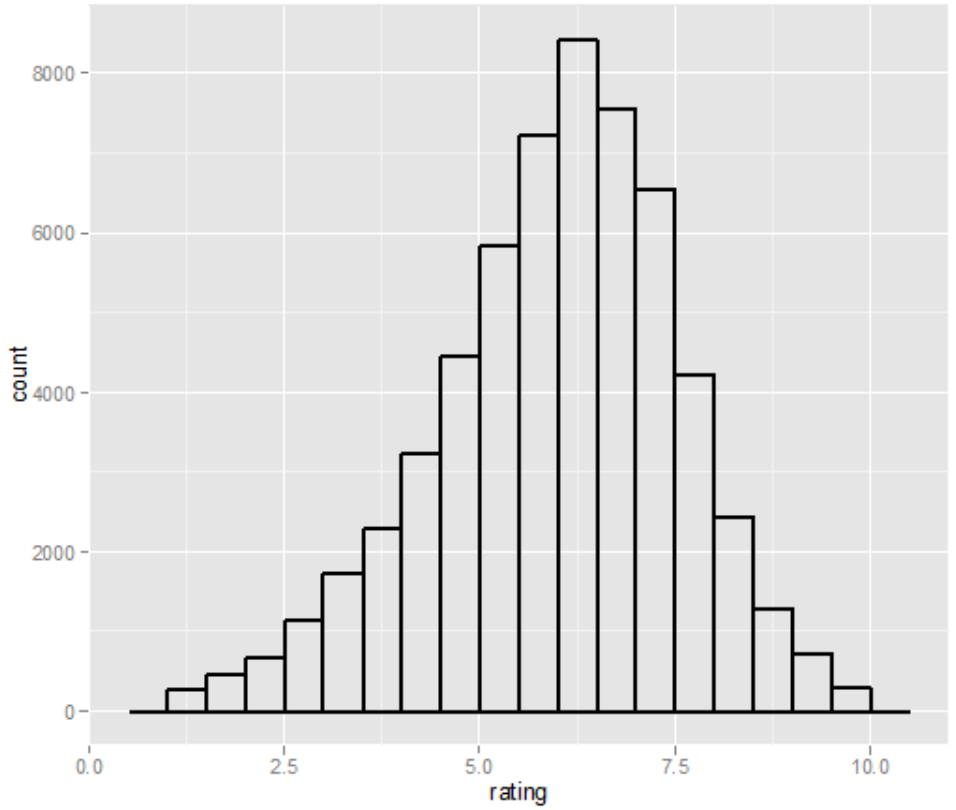
پیش پردازش داده (Data Preprocessing)

یکی از زیرمجموعه های آماده سازی داده ها

- Data Discretization
- Dimensionality Reduction
- Data Cleansing
- Feature Selection
- Data Aggregation
- Data Transformation

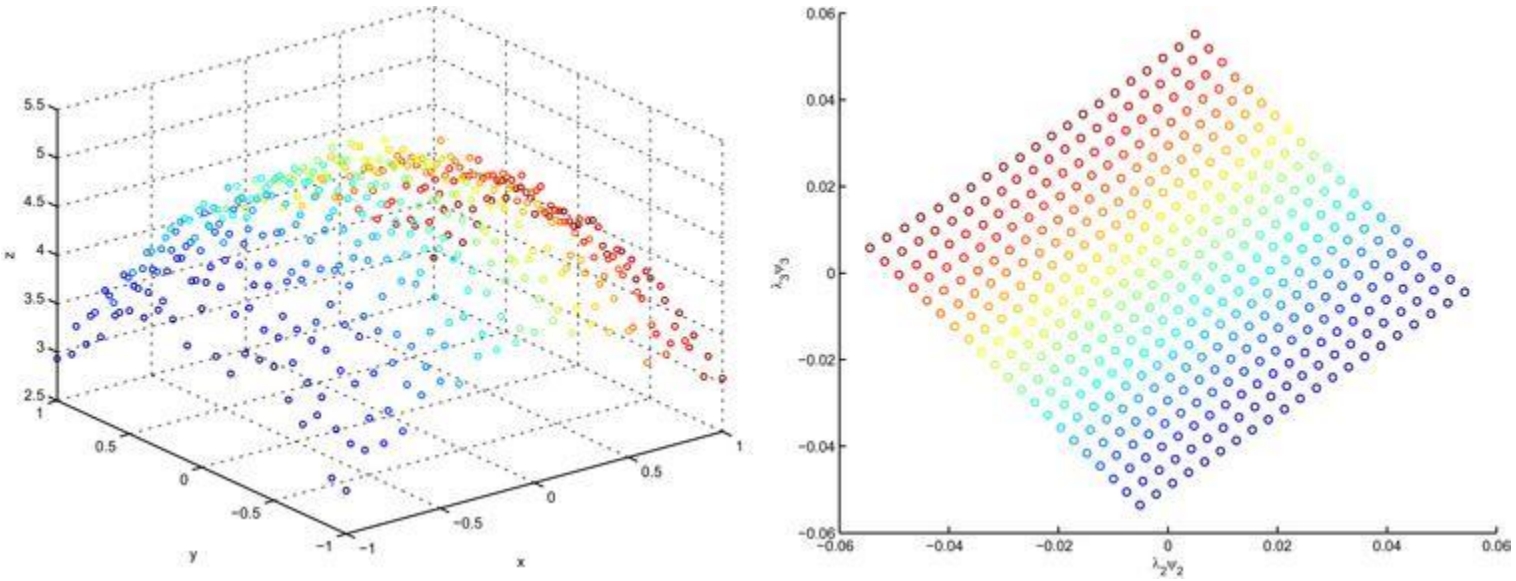
Data Discretization





Dimensionality Reduction

کاهش ویژگی ها باعث افزایش سرعت و کاهش احتمالی دقت می شود.



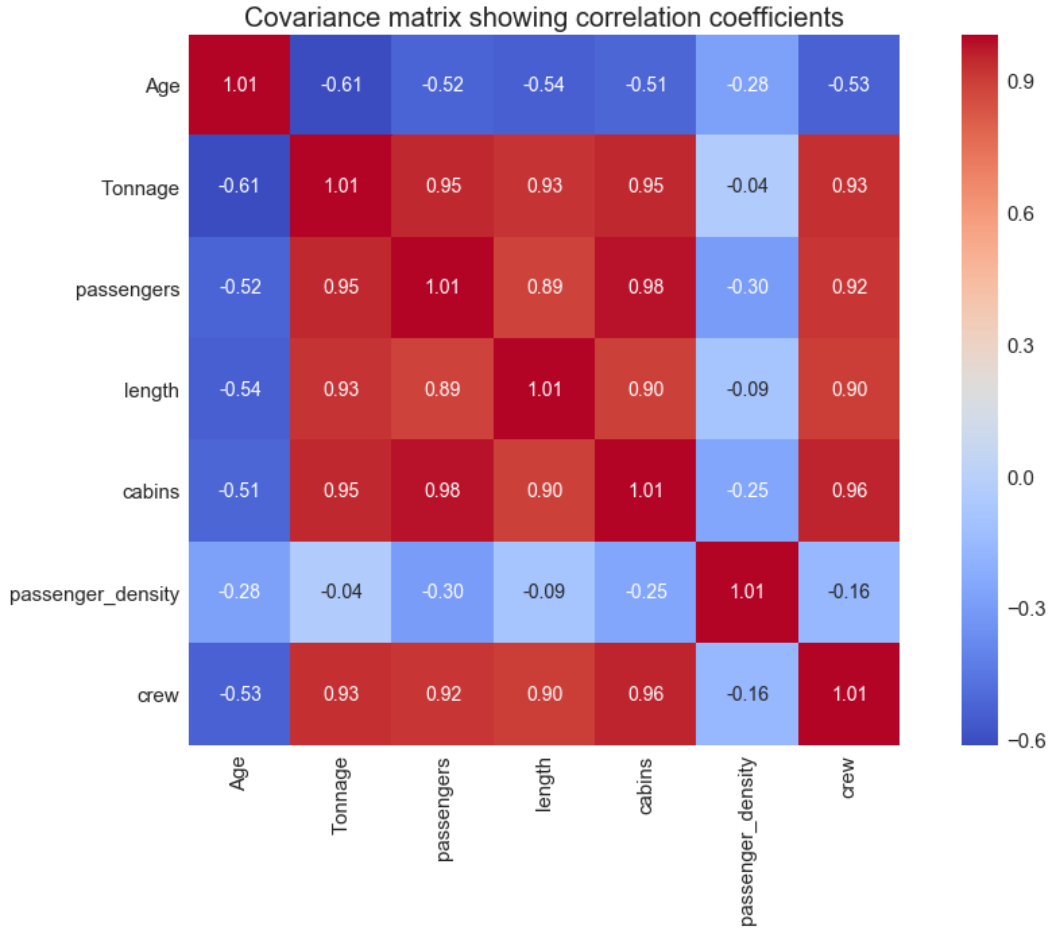
Feature Engineering

Full Feature Set

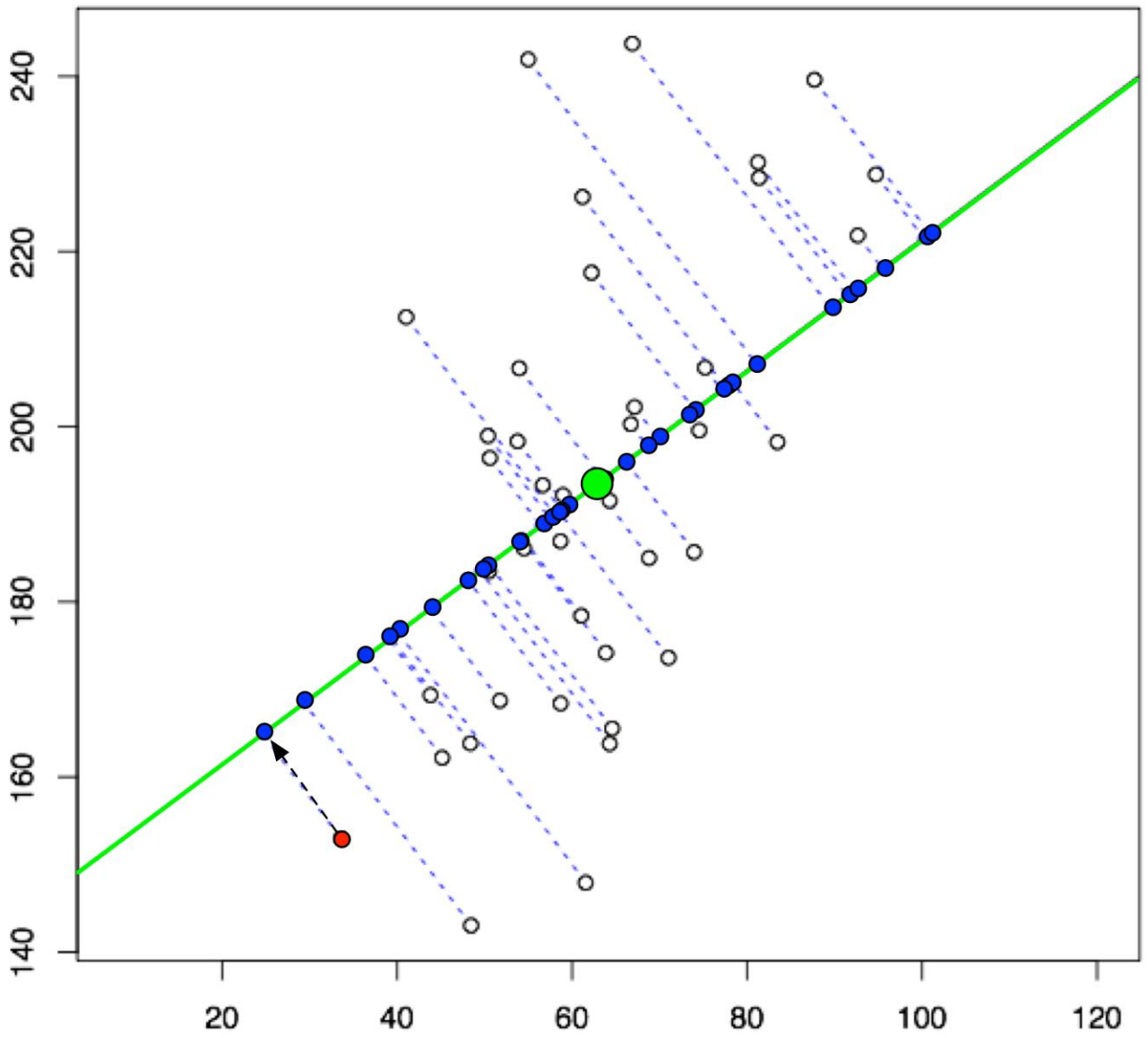
Identify Useful Features

Selected Feature Set

روش های تشخیص ویژگی های مناسب



PCA



Data Aggregation

در دو سطح ستون یا رکورد

	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6
Container 1	1	1	1	-	1	1
Container 2	0	0	0	0	0	0
Container 3	1	0	1	0	1	1

	Max	Min	Avg	Sum	Rate
Container 1	1	0	1	5	5/6
Container 2	0	0	0	0	0
Container 3	1	0	4/6	4	4/6

فشرده سازی

کاهش اطلاعات

	Protein Shake	Nike Sneakers	Adidas Boots	Fitbit	Powerade	Protein Bar	Fitness Watch	Vitamins
Buyer 1	1	1	0	1	0	5	1	0
Buyer 2	0	0	0	0	0	0	0	1
Buyer 3	3	0	1	0	5	0	0	0
Buyer 4	1	1	0	0	10	1	0	0

	Health Food	Apparel	Digital
Buyer 1	6	1	2
Buyer 2	1	0	0
Buyer 3	8	1	0
Buyer 4	12	1	0

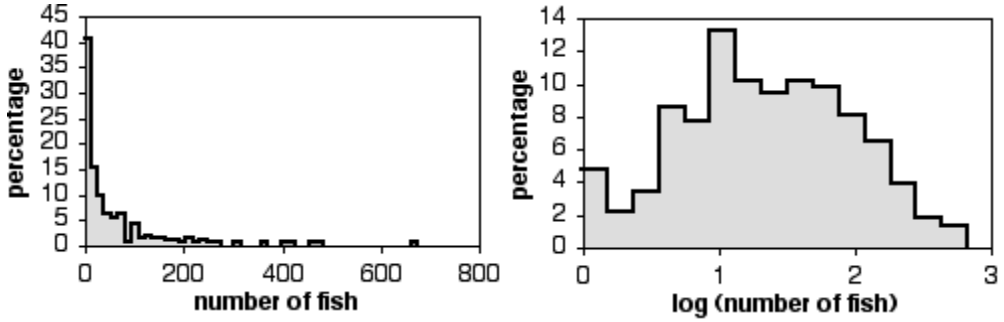
One hot coding

Name in English	Speakers	Degree of Endangerment
South Italian	7500000	Vulnerable
Sicilian	5000000	Vulnerable
Low Saxon	4800000	Vulnerable
Belarusian	4000000	Vulnerable
Lombard	3500000	Definitely endangered
Romani	3500000	Definitely endangered
Yiddish	3000000	Definitely endangered
Gondi	2713790	Vulnerable
Picard	700000	Severely endangered

Name in English	Speakers	Vulnerable	Definitely Endangered	Severely Endangered
South Italian	7500000	1	0	0
Sicilian	5000000	1	0	0
Low Saxon	4800000	1	0	0
Belarusian	4000000	1	0	0
Lombard	3500000	0	1	0
Romani	3500000	0	1	0
Yiddish	3000000	0	1	0
Gondi	2713790	1	0	0
Picard	700000	0	0	1

Data Transformation

کاهش پیچیدگی محاسبات و برطرف کردن محدودیت های الگوریتمی



Data Cleansing

هرچقدر داده ها تمیزتر شوند(همچنین تعداد ردیف ها بیشتر می شود)
پس خروجی بهتری خواهیم داشت

- Missing Value
- Duplicate Row
- Noise
- Outlier Value

Missing Value

	col1	col2	col3	col4	col5
0	2	5.0	3.0	6	NaN
1	9	NaN	9.0	0	7.0
2	19	17.0	NaN	9	NaN

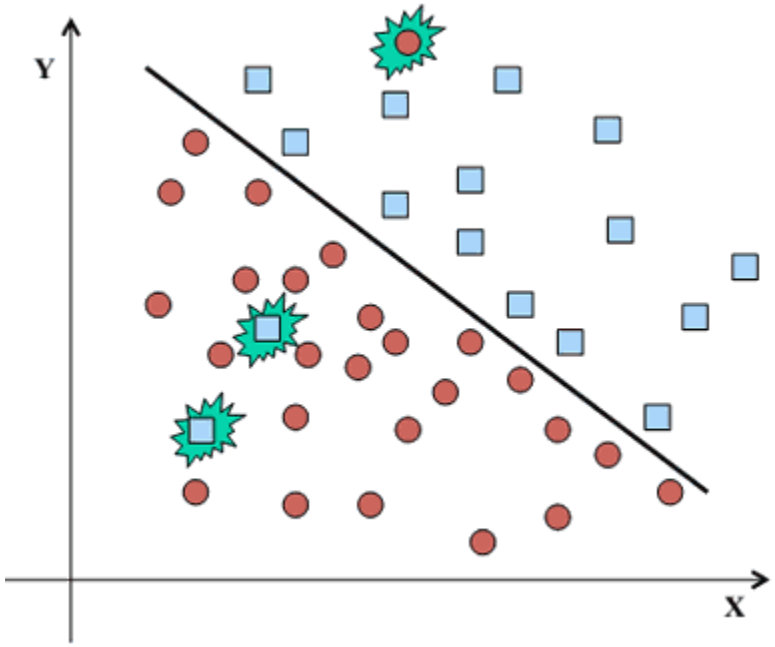
mean()

	col1	col2	col3	col4	col5
0	2.0	5.0	3.0	6.0	7.0
1	9.0	11.0	9.0	0.0	7.0
2	19.0	17.0	6.0	9.0	7.0

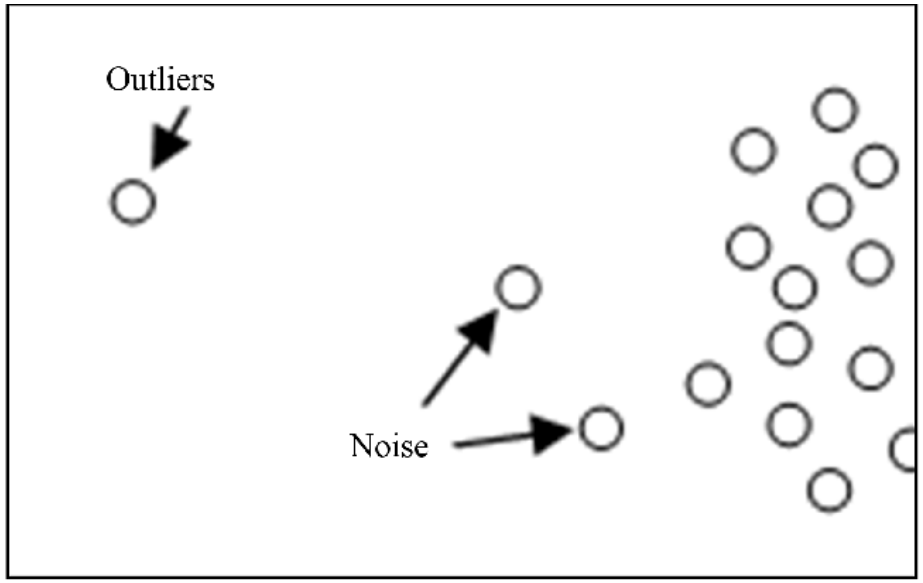
Duplicate Row

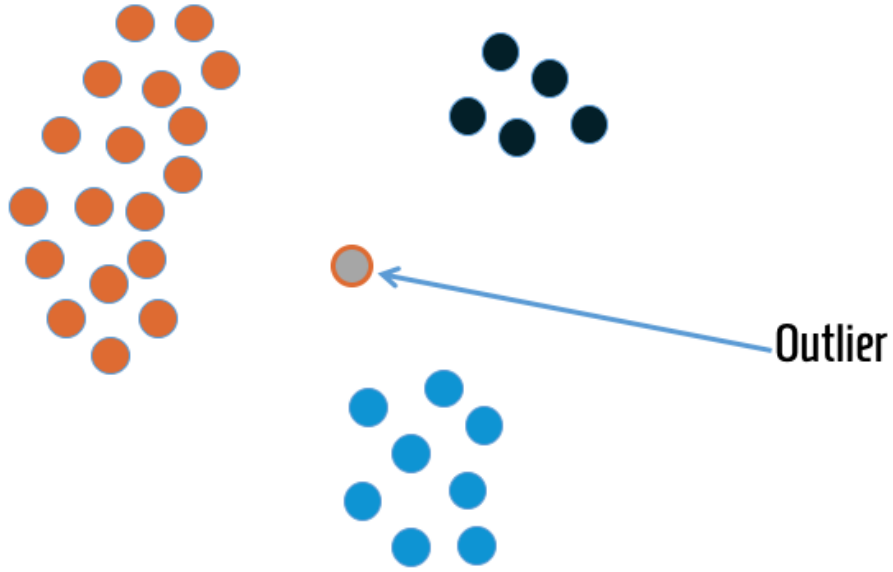
id	first_name	last_name	email
1	Carine	Schmitt	carine.schmitt@verizon.net
4	Janine	Labrune	janine.labrune@aol.com
6	Janine	Labrune	janine.labrune@aol.com
2	Jean	King	jean.king@me.com
12	Jean	King	jean.king@me.com
5	Jonas	Bergulfsen	jonas.bergulfsen@mac.com
10	Julie	Murphy	julie.murphy@yahoo.com
11	Kwai	Lee	kwai.lee@google.com
3	Peter	Ferguson	peter.ferguson@google.com
9	Roland	Keitel	roland.keitel@yahoo.com
14	Roland	Keitel	roland.keitel@yahoo.com
7	Susan	Nelson	susan.nelson@comcast.net
13	Susan	Nelson	susan.nelson@comcast.net
8	Zbyszek	Piestrzeniewicz	zbyszek.piestrzeniewicz@att.net

Noise



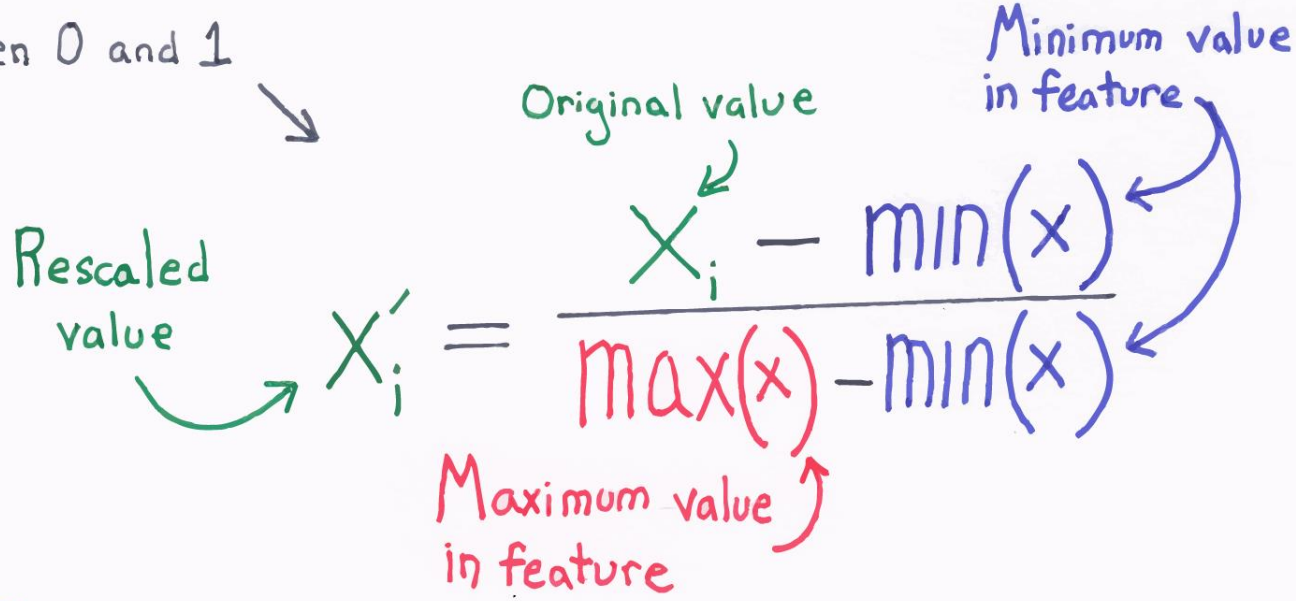
Outlier Value





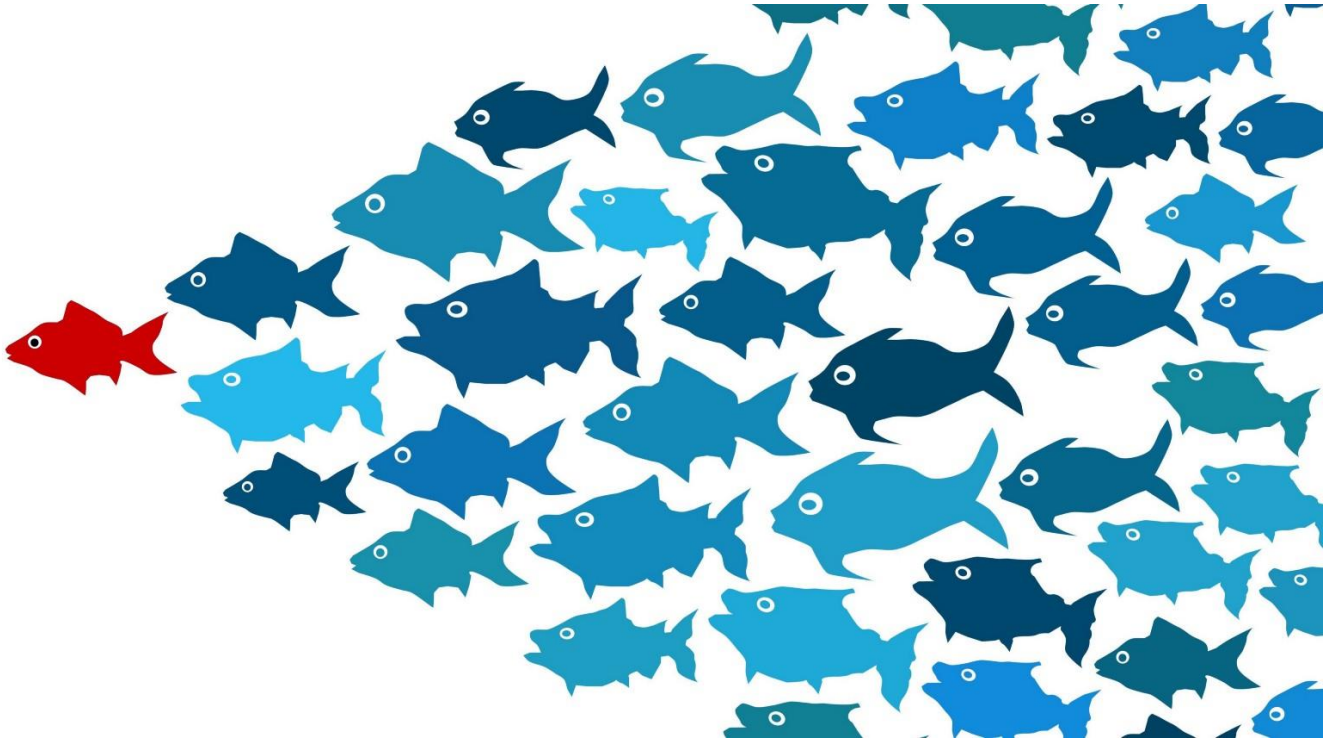
MIN MAX SCALING

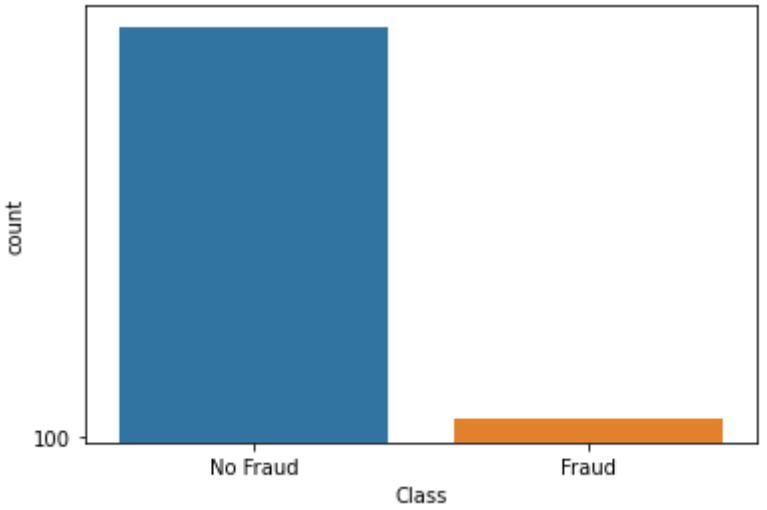
Rescales feature values to between 0 and 1



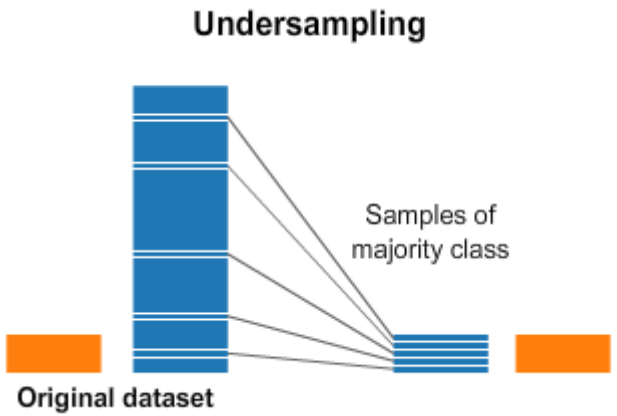
Chris Albon

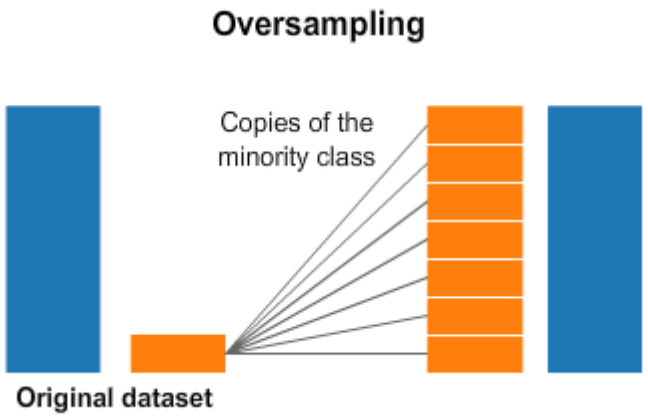
Imbalanced dataset



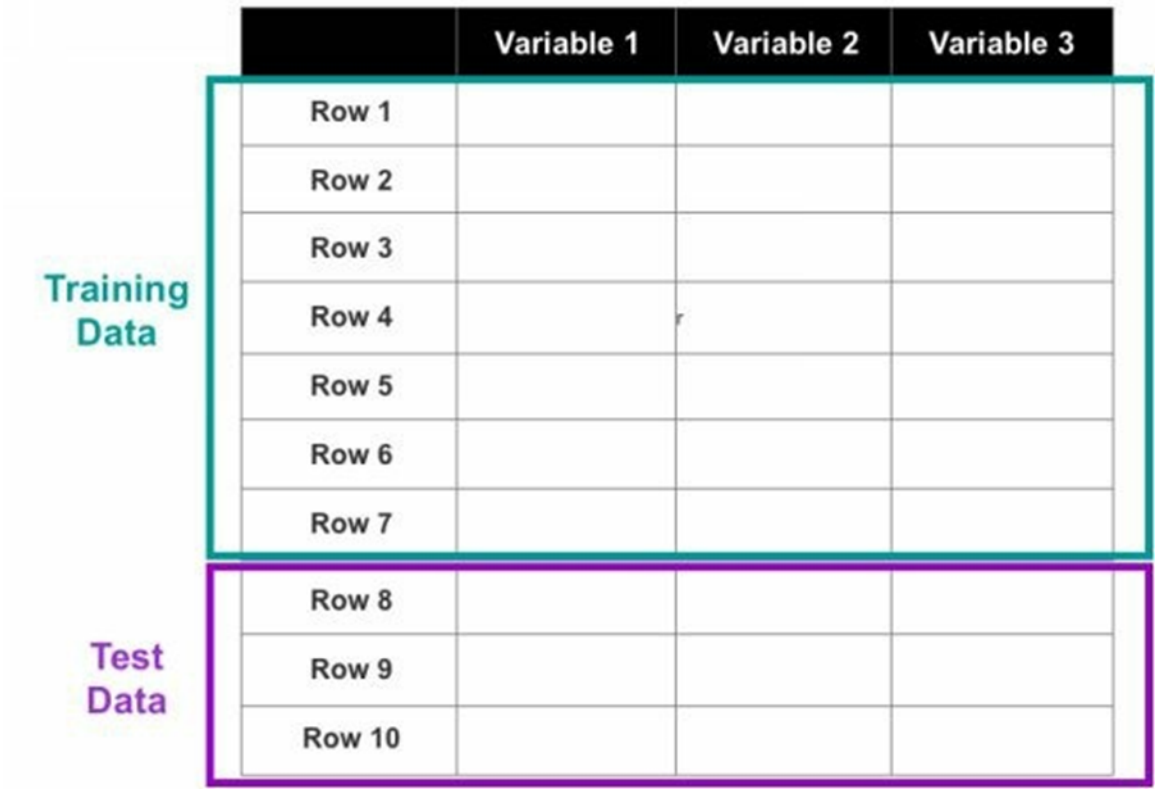


Imbalanced dataset



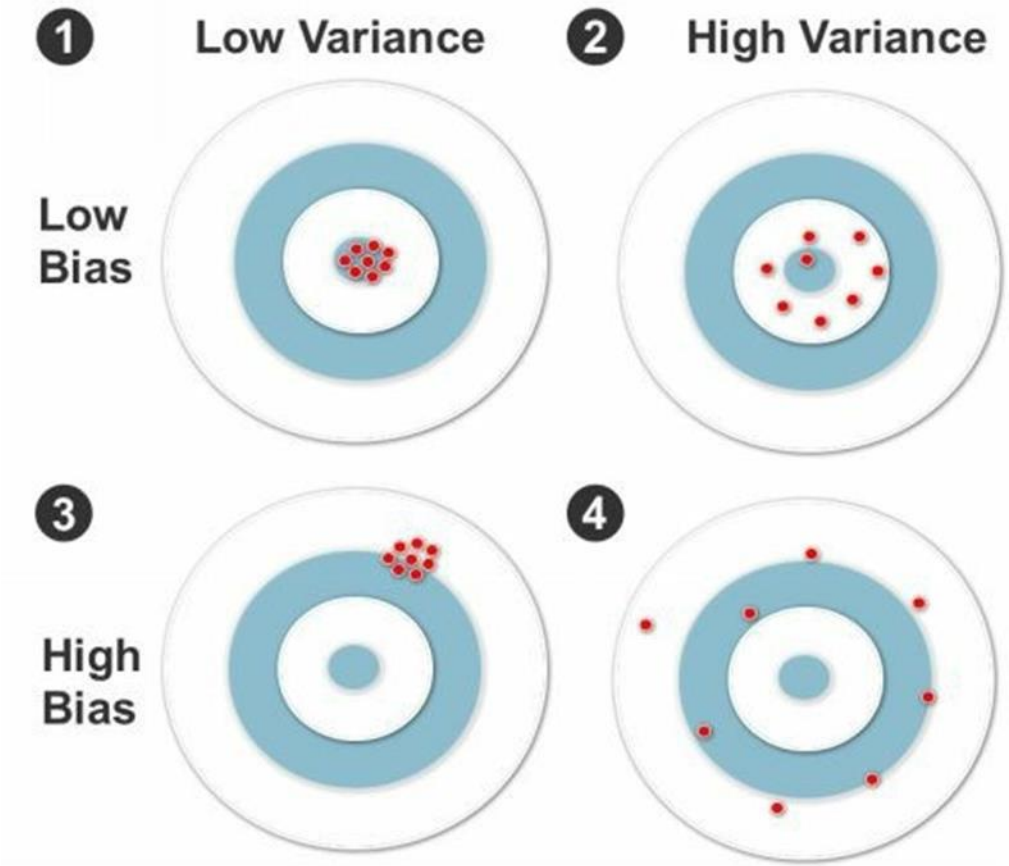


Training

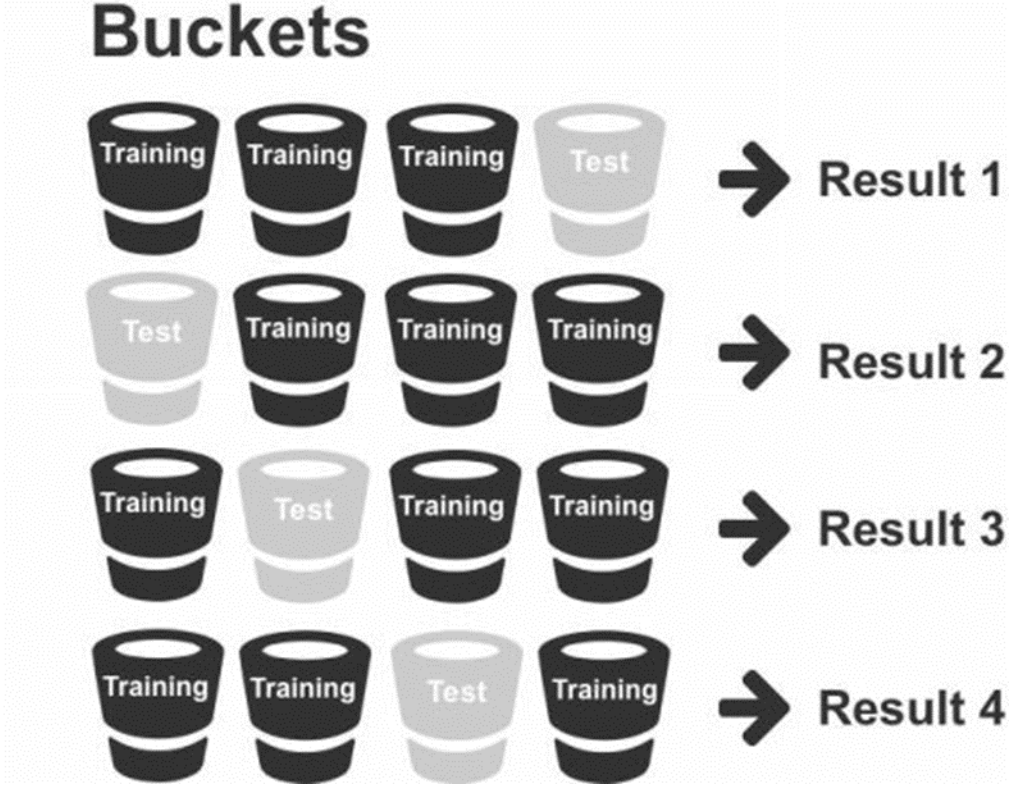


	Variable 1	Variable 2	Variable 3
Row 1			
Row 2			
Row 3			
Row 4			
Row 5			
Row 6			
Row 7			
Row 8			
Row 9			
Row 10			

Bias and Var



K-fold


Buckets
Training
Training
Training
Test
Result 1
Test
Training
Training
Training
Result 2
Training
Test
Training
Training
Result 3
Training
Training
Test
Training
Result 4

#DONTFORGETUS

آموزش های
رایگان بیشتر

www.data-hub.ir
www.t.me/data_hub_ir
www.github.com/datahub-ir
www.linkedin.com/company/data-hub-ir
www.youtube.com/channel/UCrBcbQWcD0ortWqHAlP94ug