Data Mining Process

using CRISP-DM Methodology

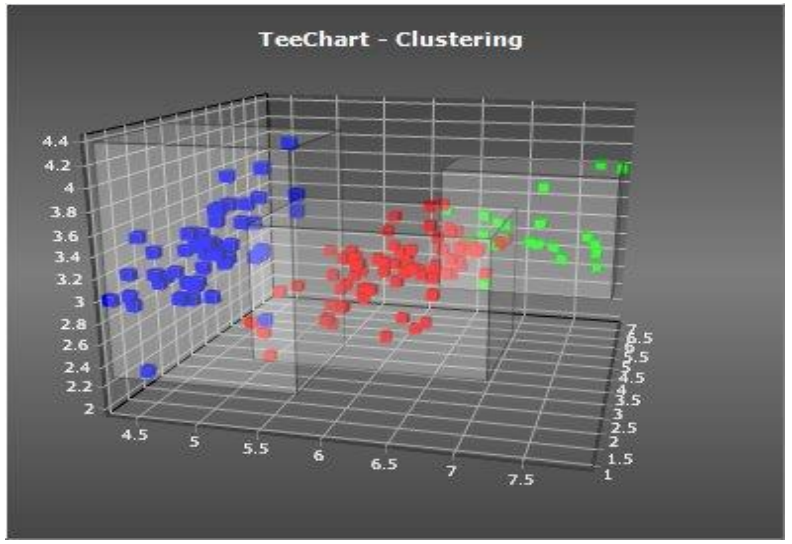
dayche.com . گروه دایکه

دای‌کِه DAYCHE

فرآیند داده کاوی

مقدمه

❑ شناسایی الگو در داده ها



If *Income = High* **then** *Credit Rate = Good*

If *Income = Low* **then** *Credit Rate = Bad*

If *Income = Middle & Car = New*
then *Credit Rate = Good*

If *Income = Middle & Car = Old*
then *Credit Rate = Bad*

X1	X2	X3	X4	Target
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.

❑ چالش های شناسایی الگوها

- انتخاب داده های مناسب
- ساخت شاخص های مناسب
- انجام تبدیلات مناسب
- انتخاب الگوریتم مناسب
- انتخاب طرح ارزیابی مناسب
- انتخاب الگوی مناسب و ...

داده کاوی
Data Mining

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

مقدمه

❑ تعریف داده کاوی

- فرآیند کشف دانش از پایگاه های داده به منظور شناسایی الگوهای موجود در داده ها که شرایط زیر را داشته باشند:
  - معتبر باشند
  - کارا و قابل استفاده باشند
  - بدیع باشند
  - توجیه پذیر باشند.

(اسامه فیاض – 1996)

One of the most used definition (Fayyad et al 1996):

Knowledge Discovery in Databases (KDD) is a **process** that aims at finding :

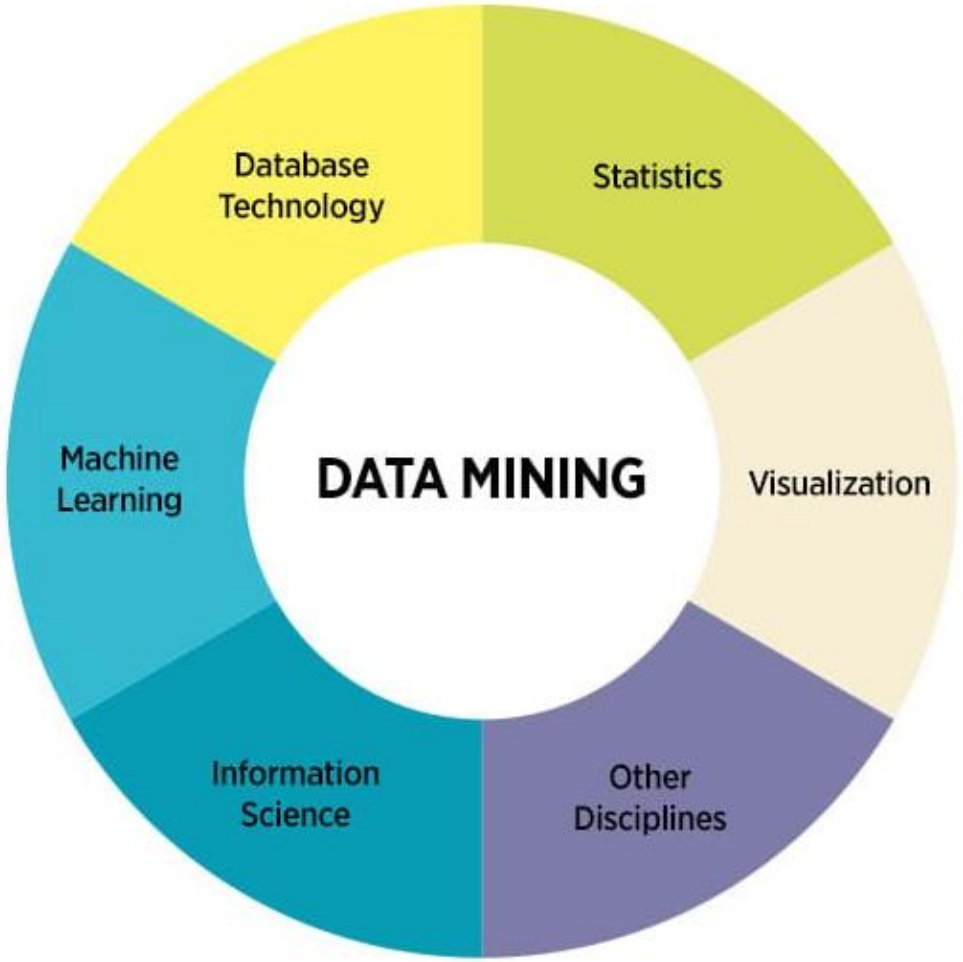
- **valid**
- **useful**
- **novel**
- **understandable**

patterns in data

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

مقدمه

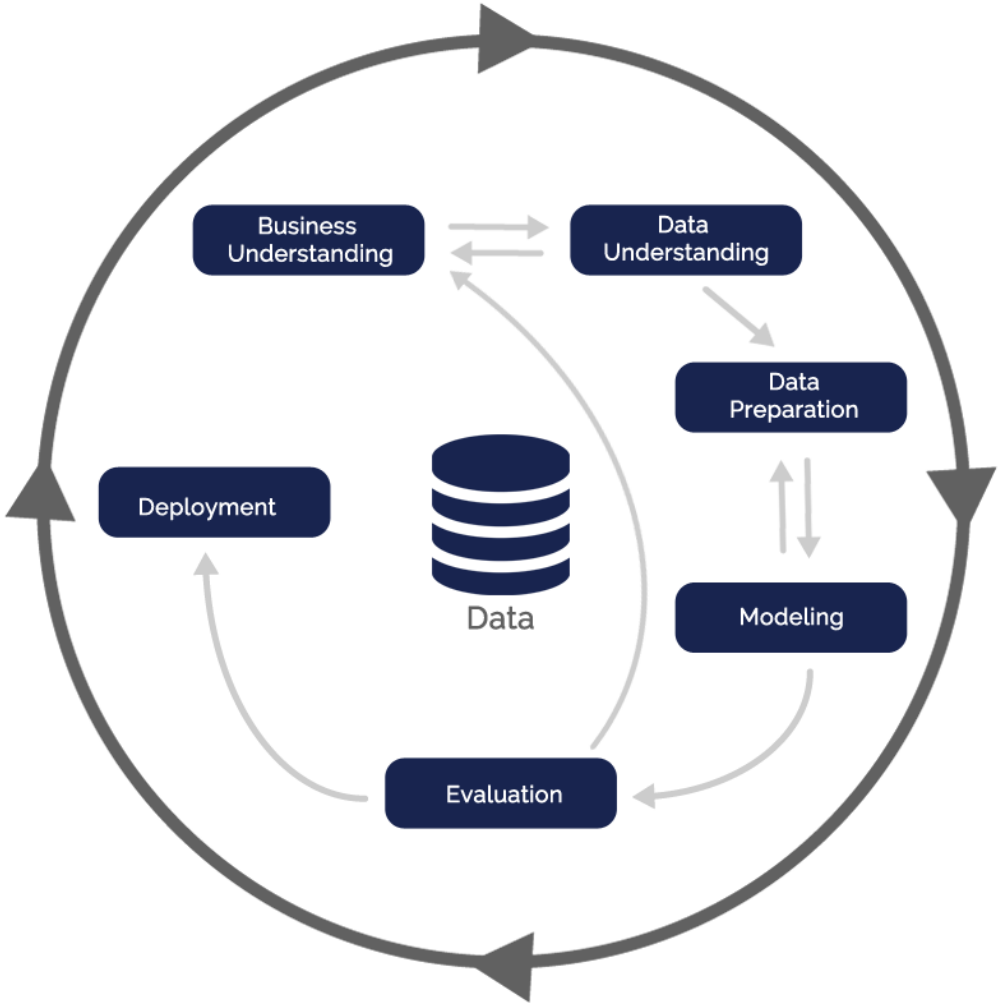


- **جعبه ابزار داده کاوی**
  - روش های آماری
  - یادگیری ماشین
  - پایگاه داده
  - مصورسازی و ...

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

مقدمه

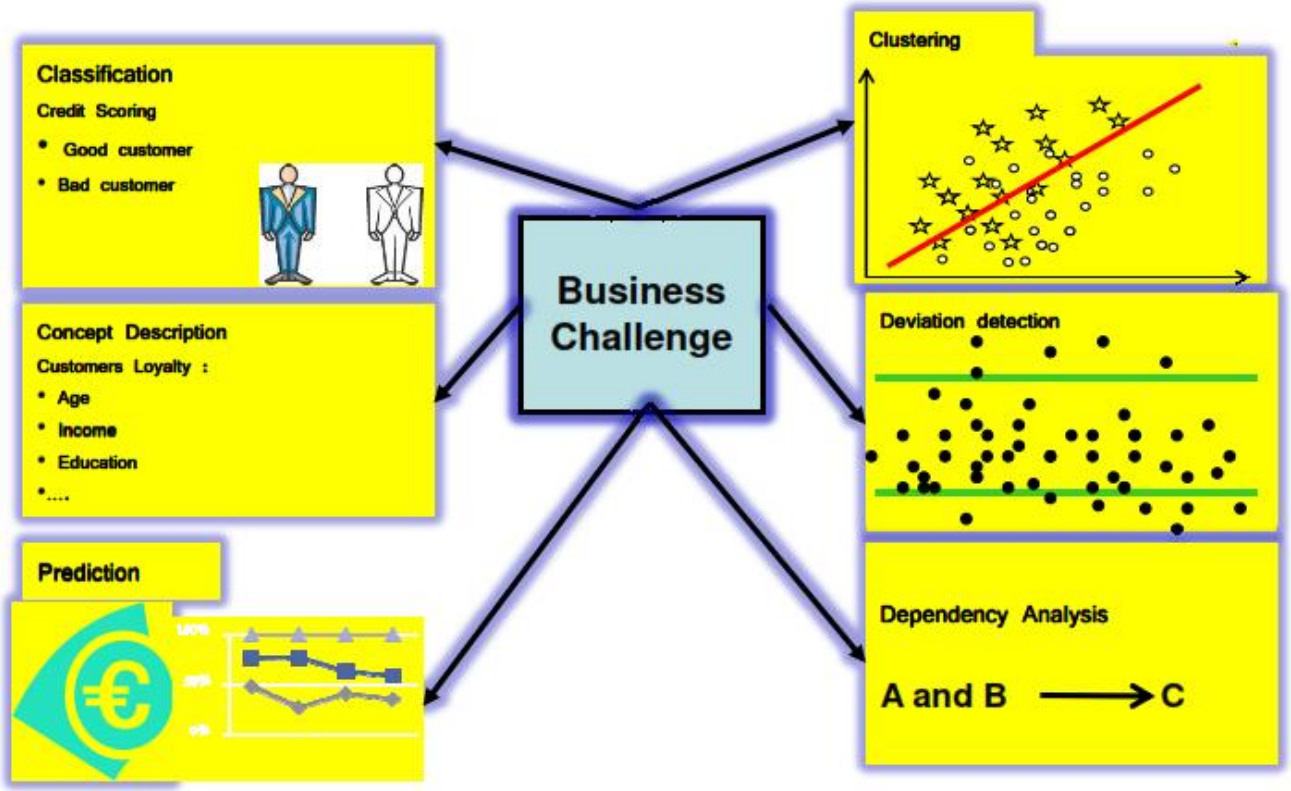


- متدولوژی CRISP-DM
  - فازهای فرآیند داده کاوی
    - شناسایی و درک مسئله
    - شناسایی و درک داده ها
    - آماده سازی داده ها
    - مدلسازی
    - ارزیابی
    - گسترش و توسعه

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

مقدمه



- **انواع مسائل (وظایف) داده کاوی**
  - پیش بینی (یا رگرسیون Regression)
  - رده بندی (Classification)
  - خوشه بندی (Clustering)
  - قوانین انجمنی (Association Rules)

.....

  - توصیف مفهوم (Concept Desc.)
  - تشخیص انحراف (Deviation Detection)

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

مقدمه

❑ انواع مسائل (وظایف) داده کاوی

- پیش بینی (یا رگرسیون Regression)
- رده بندی (Classification)

مدل های پیش بینانه
Predictive Models

بر مبنای رویکرد یادگیری
Supervised Learning

- خوشه بندی (Clustering)
- قوانین انجمنی (Association Rules)

مدل های اکتشافی
Explorative Models

بر مبنای رویکرد یادگیری
Unsupervised Learning

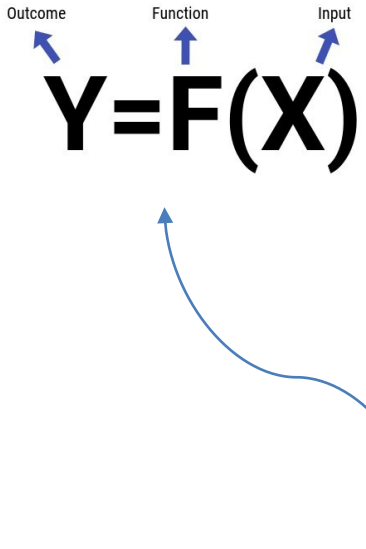
- توصیف مفهوم (Concept Desc.)
- تشخیص انحراف (Deviation Detection)

انتخاب یک یا ترکیبی از
موارد بالا بر اساس هدف

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

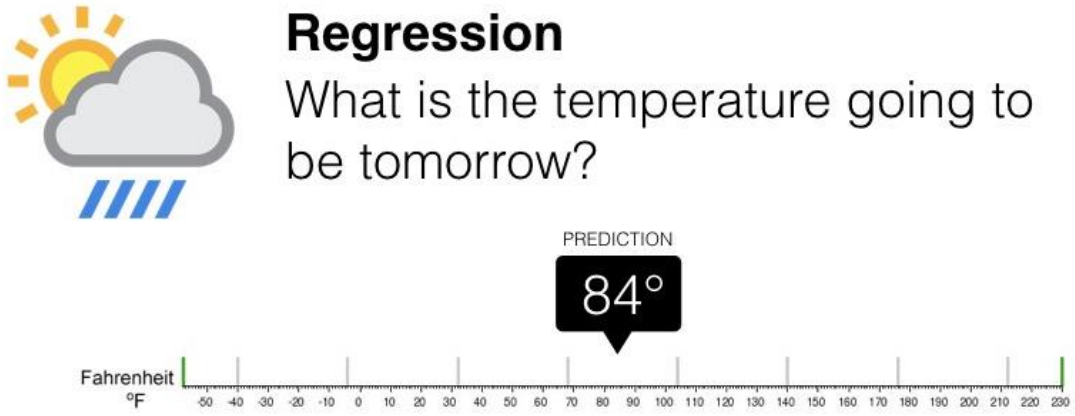
مقدمه



- **پیش بینی (یا رگرسیون Regression)**
  - یادگیری با نظارت (Supervised Learning)

پیش بینی (یا رگرسیون): پیش بینی مقدار هدف با توزیع آماری پیوسته

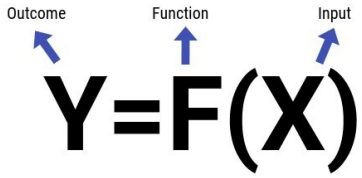
X1	X2	X3	X4	Target
.	.	.	.	54
.	.	.	.	32
.	.	.	.	65
.	.	.	.	74
.	.	.	.	45
.	.	.	.	23
.	.	.	.	18
.	.	.	.	77



تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

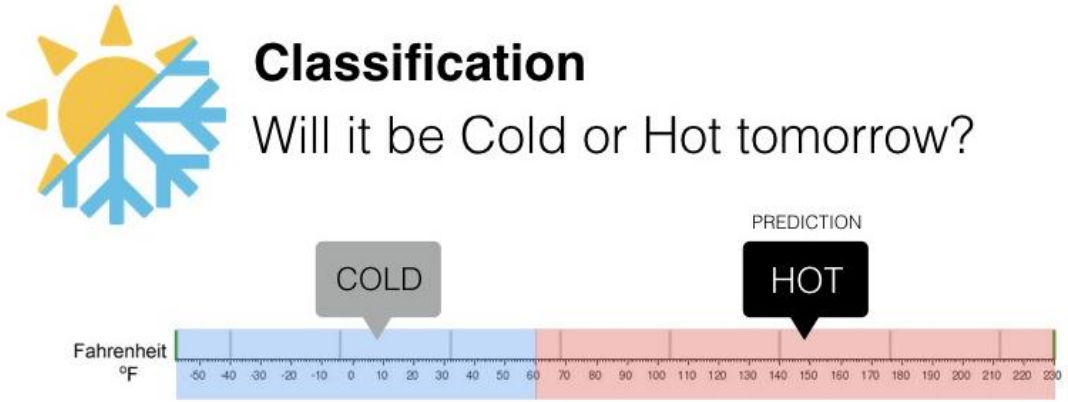
مقدمه



- **رده بندی (Classification)**
  - یادگیری با نظارت (Supervised Learning)

رده بندی: پیش بینی برچسب هدف با توزیع آماری گسسته

X1	X2	X3	X4	Lable
.	.	.	.	COLD
.	.	.	.	COLD
.	.	.	.	HOT
.	.	.	.	HOT
.	.	.	.	COLD
.	.	.	.	COLD
.	.	.	.	COLD
.	.	.	.	HOT



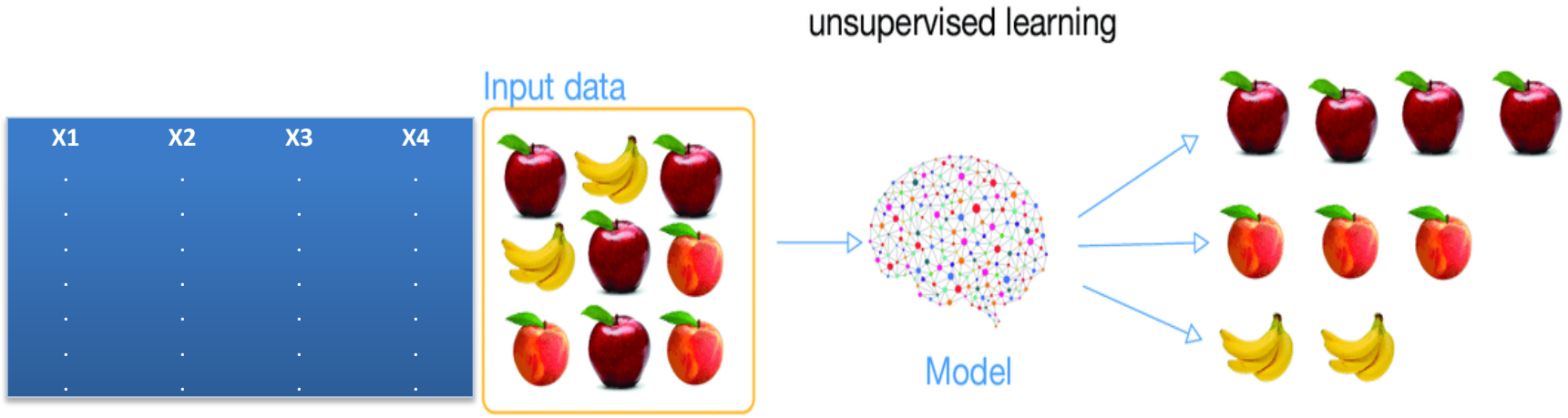
تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

مقدمه

- **خوشه بندی (Clustering)**
  - یادگیری بدون نظارت (Unsupervised Learning)

خوشه بندی: دسته بندی اتوماتیک مشاهدات بر مبنای معیار شباهت



تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

مقدمه

❑ قوانین انجمنی (Association Rules)

- یادگیری بدون نظارت (Unsupervised Learning)

قوانین انجمنی: تحلیل وابستگی بین وقوع رخدادها از نظر باهم آیی و تقدم - تاخر

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

مقدمه

❑ توصیف مفهوم (Concept Desc.)

- در بسیاری از مسئله های داده کاوی، توصیف چگونگی وقوع یک پدیده مهمتر از پیش بینی دقیق آنهاست.

در این نوع مسائل به دنبال توصیف یا تعریفی از الگوهای بدست آمده هستیم.

بطور مثال:

- مشتریان وفادار چه ویژگی هایی دارند؟
- پذیرفته شدگان در کنکور دارای چه مشخصاتی می باشند؟

- **Characteristic 1 (rule 1) :**
  If Age > 45 and annual income > 50 k $ and
  Education = university degree →Customer=royal
- **Characteristic 2 (rule 2) :....**
- **.....**

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

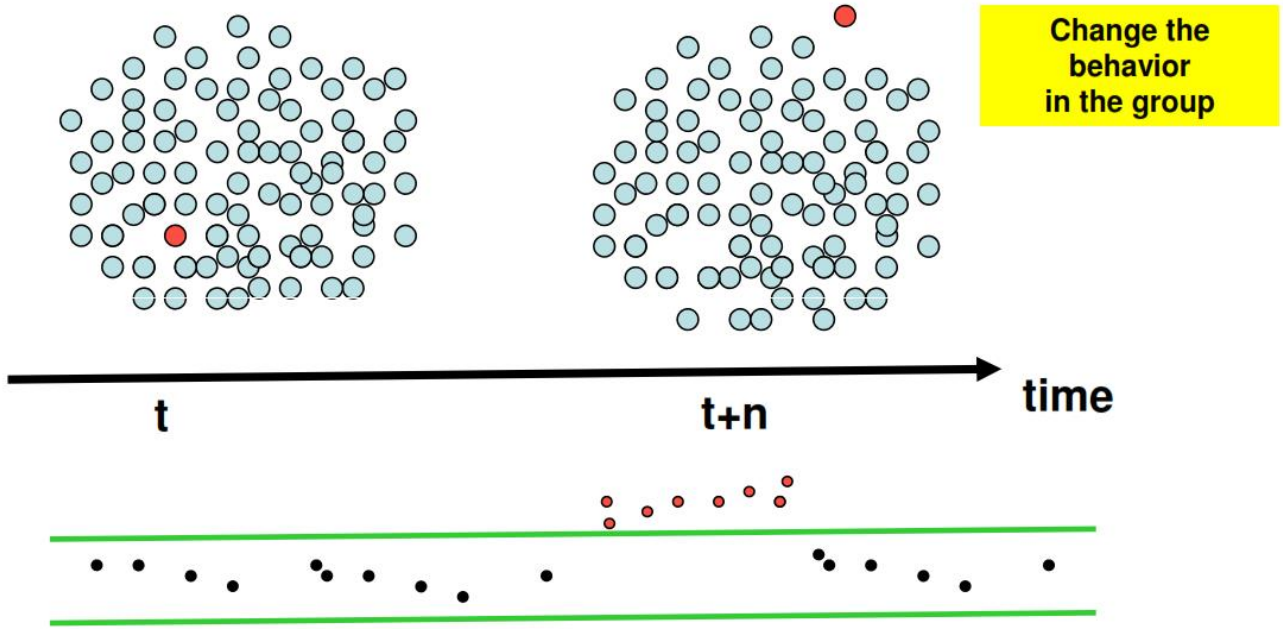
فرآیند داده کاوی

مقدمه

❑ تشخیص انحراف (Deviation Detection)

- در بسیاری از مسئله های داده کاوی، الگوی موارد نادر و خاص که به نوعی انحراف از وضعیت نرمال می باشند مد نظر می باشد.

در این نوع مسائل به دنبال شناسایی و کشف الگوهای آنومالی و پیش بینی آنها هستیم.



بطور مثال:

- شناسایی الگوی تقلب در صنعت بیمه
- پیش بینی خرابی در یک توربین گازی

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

مقدمه

❑ انواع مسائل (وظایف) داده کاوی

- پیش بینی (یا رگرسیون Regression)
- رده بندی (Classification)

مدل های پیش بینانه
Predictive Models

بر مبنای رویکرد یادگیری
Supervised Learning

- خوشه بندی (Clustering)
- قوانین انجمنی (Association Rules)

مدل های اکتشافی
Explorative Models

بر مبنای رویکرد یادگیری
Unsupervised Learning

- توصیف مفهوم (Concept Desc.)
- تشخیص انحراف (Deviation Detection)

انتخاب یک یا ترکیبی از
موارد بالا بر اساس هدف

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

شناخت و آماده سازی داده ها

مدلسازی و ارزیابی

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

تولید محتوا: زهرا ذوالقدر

daychegroup

daychegroup

dayche.com | گروه دایکه

فرآیند داده کاوی

مقدمه

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

کاهش ابعاد و انتخاب نمونه

تولید محتوا: زهرا ذوالقدر

daychegroup

daychegroup

dayche.com | گروه دایکه

فرآیند داده کاوی

مقدمه

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

مقدمه

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

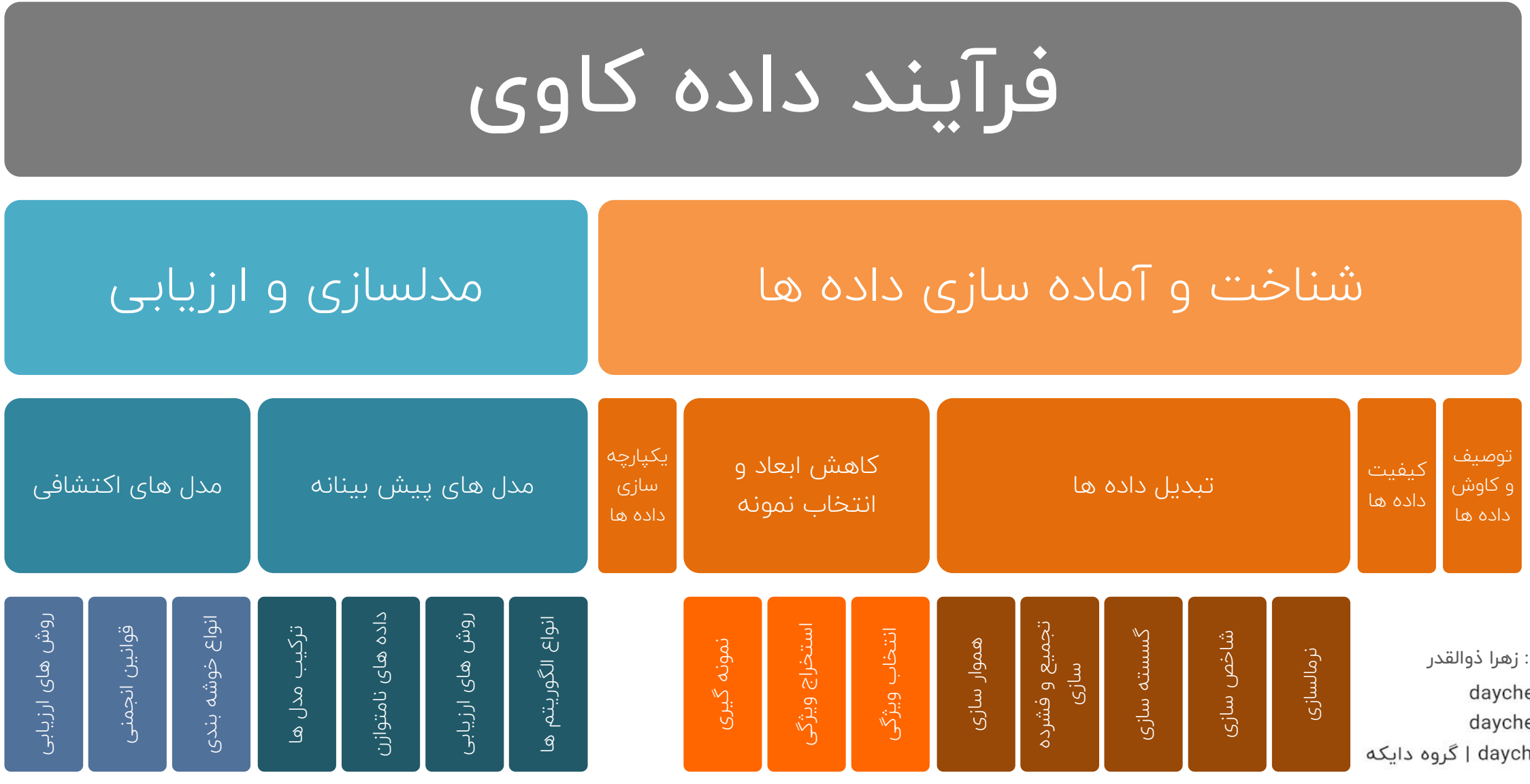
فرآیند داده کاوی

مقدمه

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

مقدمه



تولید محتوا: زهرا ذوالقدر

daychegroup

daychegroup

dayche.com | گروه دایکه

Data Description & Exploration

توصیف و کاوش داده ها

dayche.com . گروه دایکه

فرآیند داده کاوی

توصیف و کاوش داده ها

گردآوری داده ها

در اغلب موارد داده های مورد نیاز در حل مسئله در جاهای مختلف و با فرمت های متنوع پراکنده هستند.

دسترس پذیری به لیست منابع داده تعیین شده در فاز شناخت کسب و کار

انتخاب داده های متناسب با اهداف داده کاوی بر اساس معیارهای کسب و کار

برنامه ریزی جهت یکپارچه سازی و الحاق منابع داده های منتخب

انتخاب تعداد رکوردهای لازم و بازه زمانی مناسب برای انتخاب داده ها

تولید محتوا: زهرا ذوالقدر

فرآیند داده کاوی

توصیف و کاوش داده ها

مجموعه ای از ابزارها و روش های آماری به منظور شناخت دقیق از وضعیت داده ها و روابط بین آنها

توصیف داده ها

استفاده از روش های آماری توصیفی و مستند سازی نحوه دسترسی به داده ها و چارچوب آنها تاثیر زیادی در شناخت اولیه و دید شهودی نسب به داده ها می گردد.

کاوش در داده ها

بررسی آماری دقیق از همبستگی های موجود و روابط معنادار بین آنها قدم اول در انتخاب، ایجاد و تغییر مجموعه داده ها است.

نکته مهم: در اغلب پروژه های واقعی نیاز داریم ابتدا از الحاق و یکپارچگی منابع داده و همچنین کیفیت داده ها اطمینان حاصل کنیم.

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

توصیف و کاوش داده ها

ابزارها و روش های شناخت داده

1. توصیف کاملی از فرمت داده ها، ابعاد (تعداد نمونه ها و ویژگی ها) و شرح هر جدول و ارتباط بین آنها
2. بررسی دقیق ویژگی ها شامل تعیین نوع آن، دامنه، شاخص های مرکزی و پراکندگی و توزیع آماری آنها
3. استفاده از روش های مصورسازی داده ها برای درک بهتر از ساختار داده های مورد بررسی
4. بررسی و تحلیل روابط دو یا چند متغیره بین ویژگی ها از طریق آزمون های فرض و سایر روشهای اکتشافی

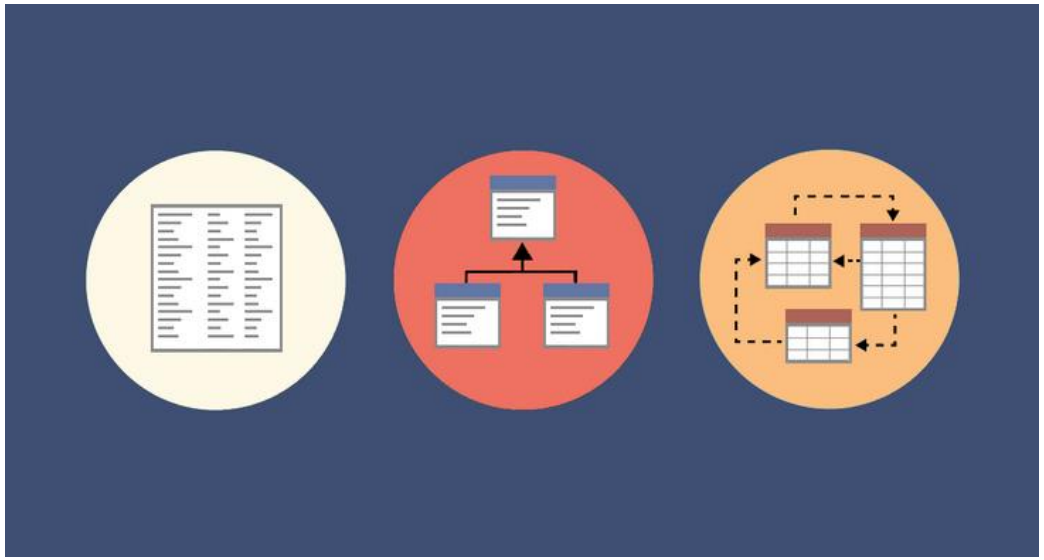
تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

توصیف و کاوش داده ها

❑ ابزارها و روش های شناخت داده

1. توصیف کاملی از فرمت داده ها، ابعاد (تعداد نمونه ها و ویژگی ها) و شرح هر جدول و ارتباط بین آنها



اولین مرحله شناخت از داده ها، شناخت کلی از ساختار جداول داده و ارتباط بین آنها می باشد. لزوم برخی از اقدامات اولیه همچون انتخاب ابزارهای مناسب برای بارگذاری داده ها، نیاز به روش های نمونه گیری، الحاق و یکپارچه سازی داده ها بر اساس شناخت اولیه از ساختار و حجم داده ها مشخص می گردد.

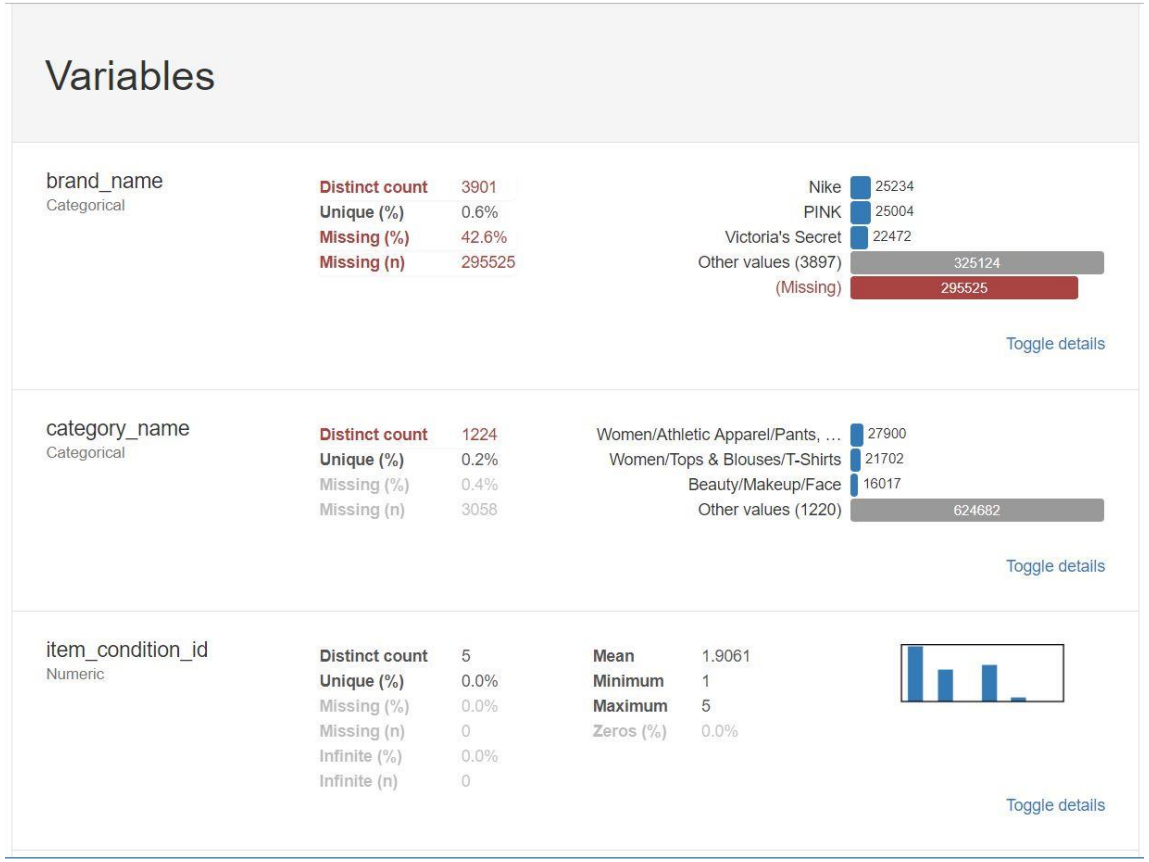
تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

توصیف و کاوش داده ها

❑ ابزارها و روش های شناخت داده

2. بررسی دقیق ویژگی ها شامل تعیین نوع آن، دامنه، شاخص های مرکزی و پراکندگی و توزیع آماری آنها



استفاده از خلاصه های آماری جهت توصیف ویژگی ها علاوه بر شناخت ماهیت داده ها، ایده های اولیه برای کیفیت داده ها و تبدیل های مورد نیاز را ارائه می دهد.

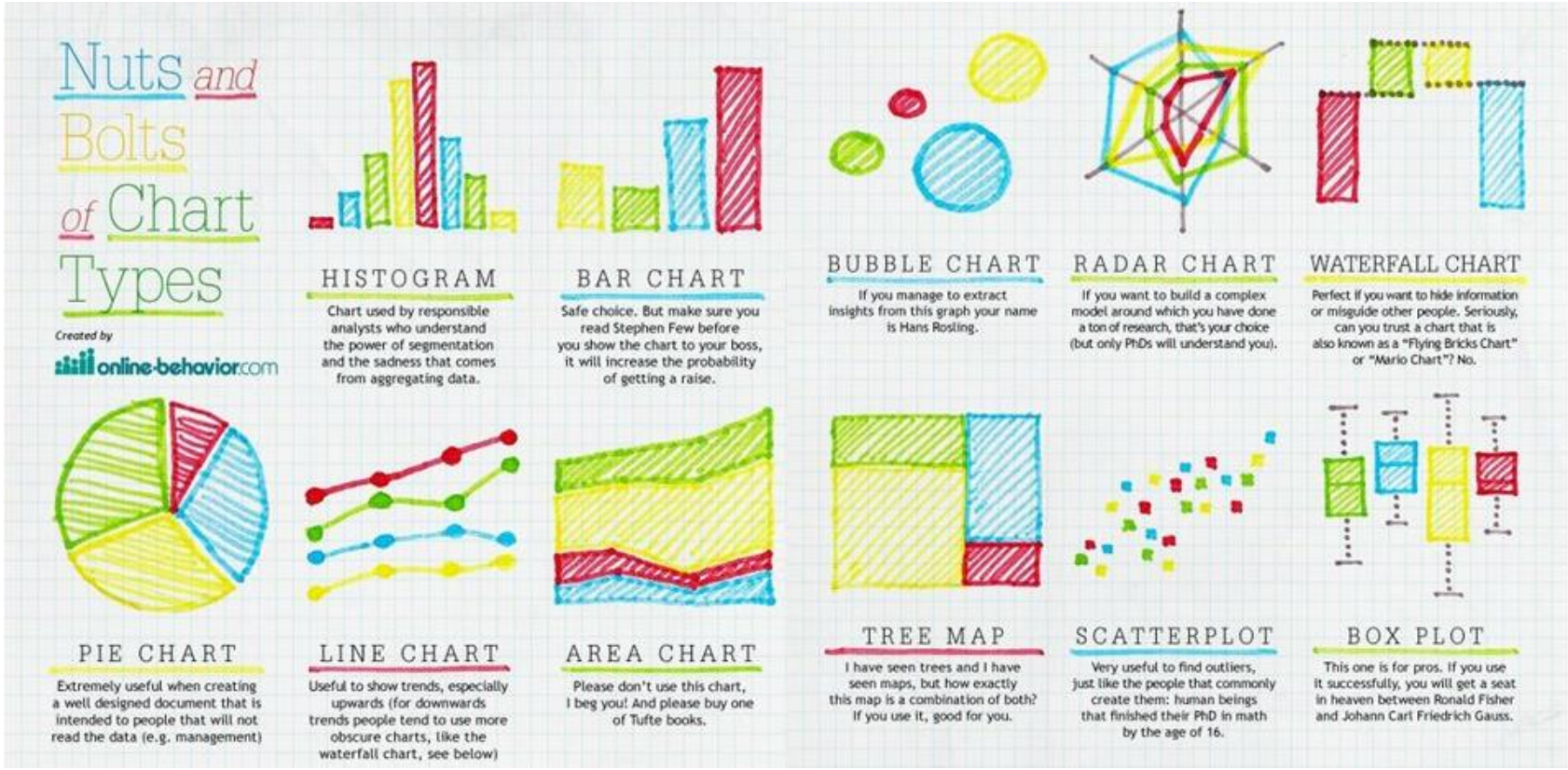
تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
گروه دایکه | dayche.com

فرآیند داده کاوی

توصیف و کاوش داده ها

❑ ابزارها و روش های شناخت داده

3. استفاده از روش های مصورسازی داده ها برای درک بهتر از ساختار داده های مورد بررسی



شناسایی الگوها بصورت بصری علاوه بر ایجاد درک شهودی از روابط بین داده ها منجر به طراحی فرضیه های تحقیق و بررسی روابط بین داده ها می شود.

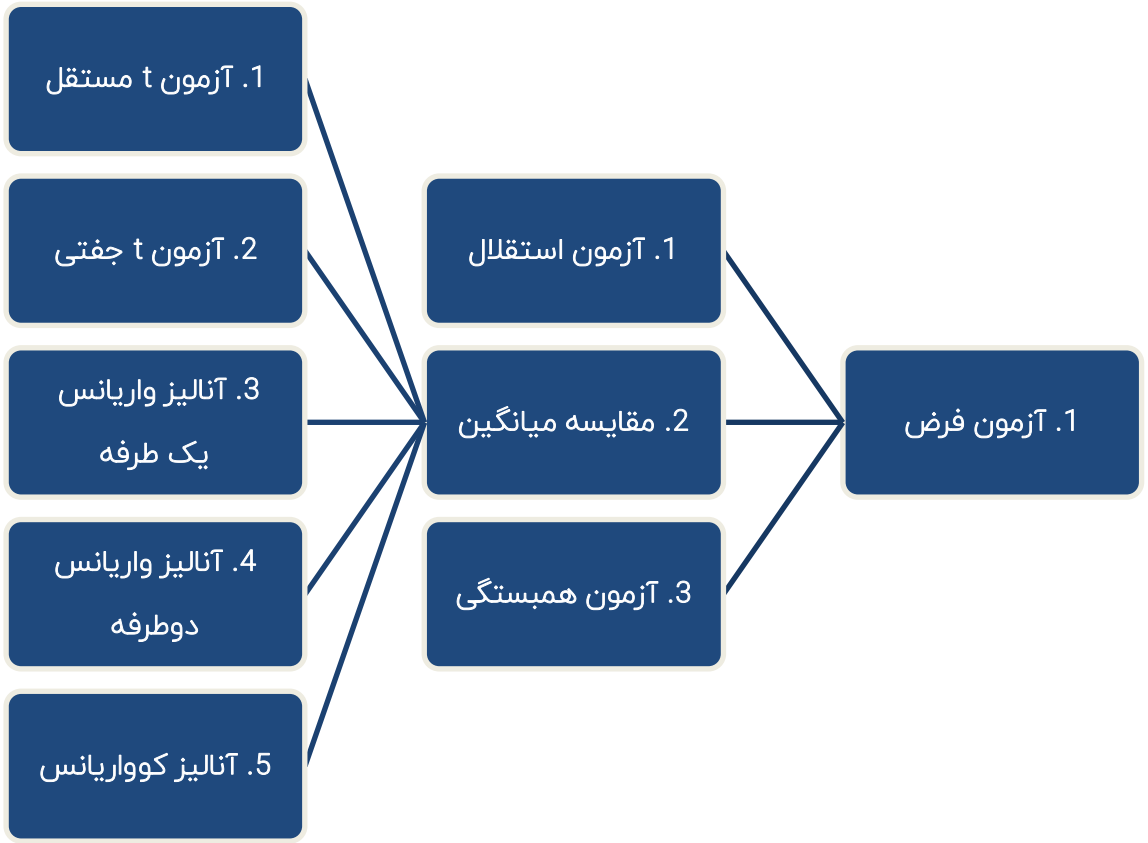
تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

توصیف و کاوش داده ها

ابزارها و روش های شناخت داده

4. بررسی و تحلیل روابط دو یا چند متغیره بین ویژگی ها از طریق آزمون های فرض و سایر روشهای اکتشافی



بررسی روابط آماری بین ویژگی ها از طریق آزمون های فرض، علاوه بر رد یا تایید فرضیات در انتخاب ویژگی ها، تبدیل داده ها، ارزیابی و تفسیر نتایج مدل ها نقش پررنگی دارد.

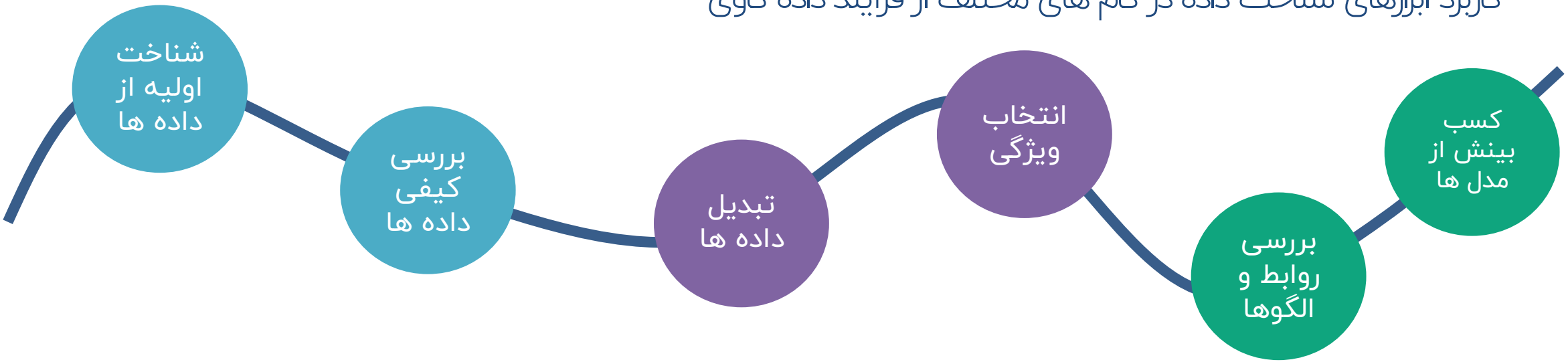
تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

توصیف و کاوش داده ها

❑ شناخت داده ها در فرآیند داده کاوی

کاربرد ابزارهای شناخت داده در گام های مختلف از فرآیند داده کاوی



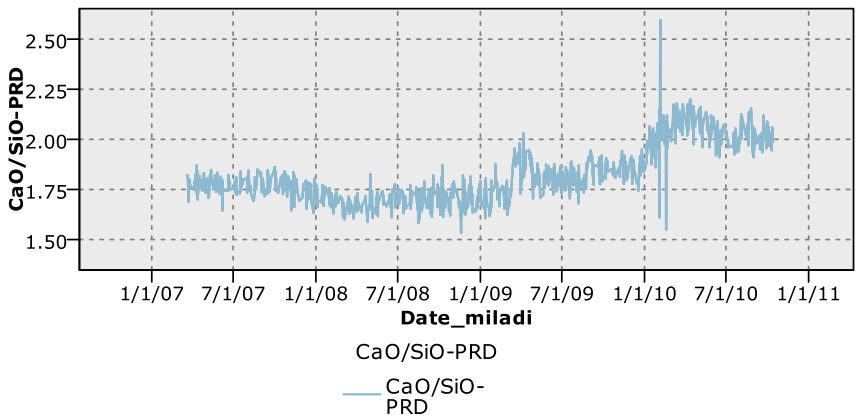
تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

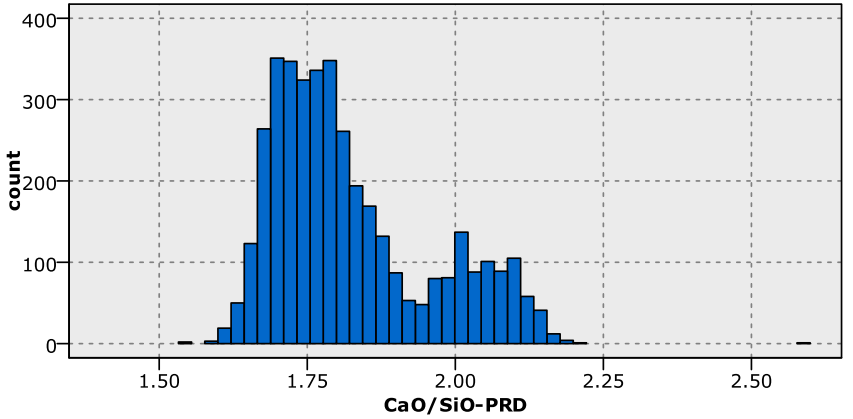
فرآیند داده کاوی

توصیف و کاوش داده ها

❑ شناخت داده ها در فرآیند داده کاوی

کاربرد 1: استفاده از ابزارهای شناخت داده در نقطه شروع فرآیند داده کاوی





تولید محتوا: زهرا ذوالقدر

daychegroup

daychegroup

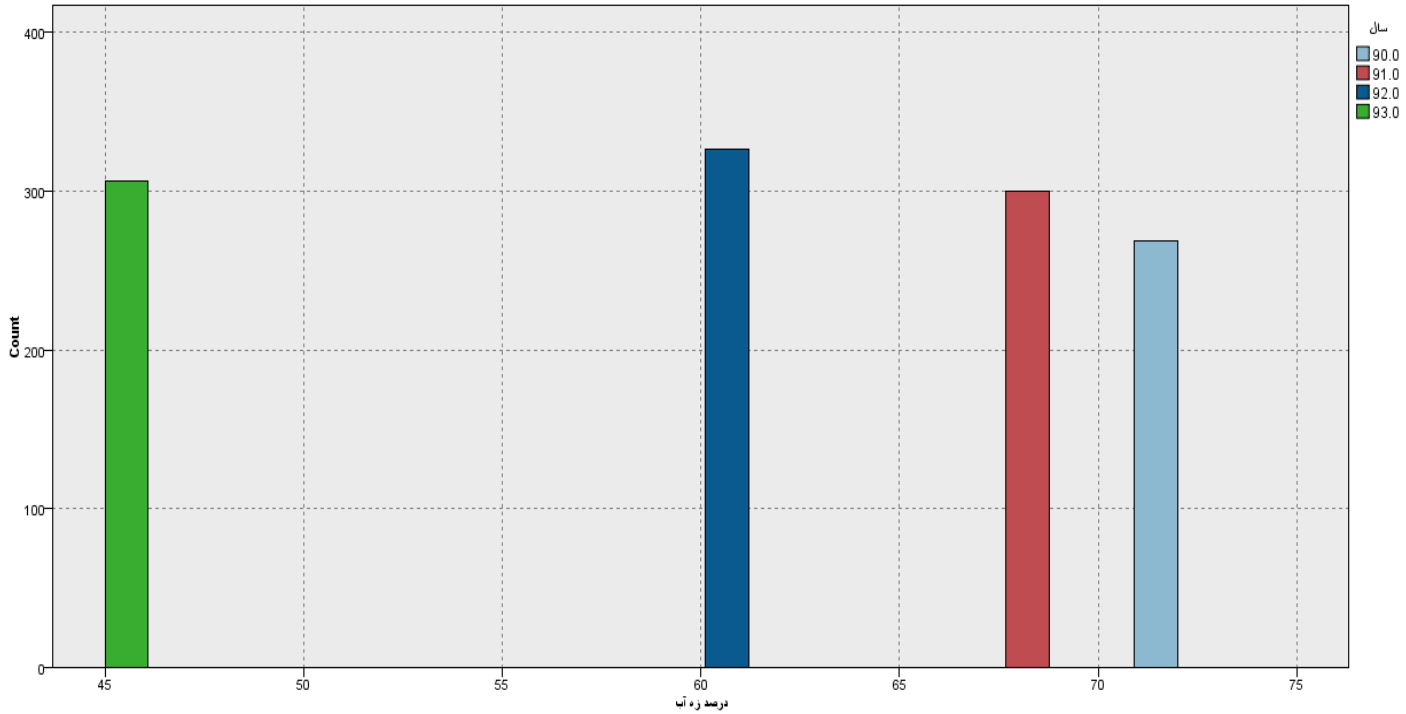
dayche.com | گروه دایکه

فرآیند داده کاوی

توصیف و کاوش داده ها

❑ شناخت داده ها در فرآیند داده کاوی

کاربرد 1: استفاده از ابزارهای شناخت داده در نقطه شروع فرآیند داده کاوی



تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

❑ شناخت داده ها در فرآیند داده کاوی

کاربرد 2: استفاده از ابزارهای شناخت داده در بررسی کیفی داده ها

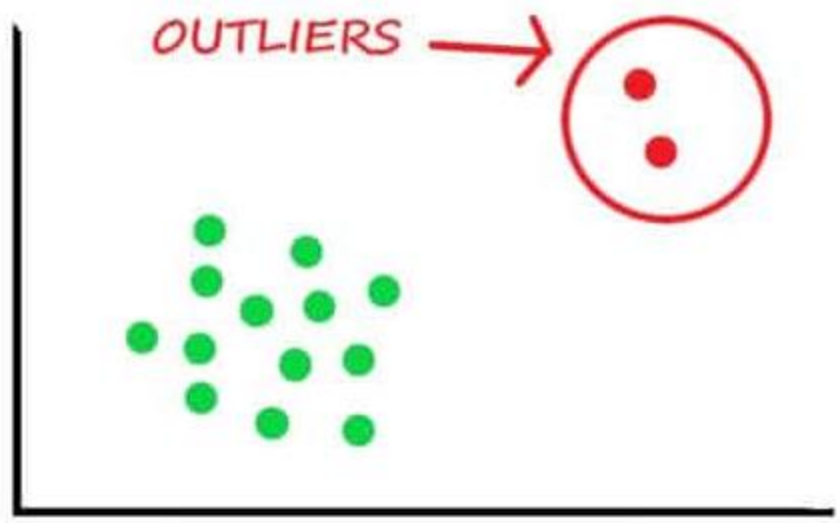
مثال:

- استفاده از جداول و آماره های توصیفی به منظور شناسایی ناسازگاری در داده ها

Value	Proportion	%	Count
		0.01	9
AUTO		96.46	70398
Manual		0.0	1
MANUAL		3.53	2575

❑ شناخت داده ها در فرآیند داده کاوی

کاربرد 2: استفاده از ابزارهای شناخت داده در بررسی کیفی داده ها



مثال:

- استفاده از نمودارهای Box Plot و Scatter Plot جهت شناسایی نقاط پرت

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

توصیف و کاوش داده ها

❑ شناخت داده ها در فرآیند داده کاوی

کاربرد 2: استفاده از ابزارهای شناخت داده در بررسی کیفی داده ها

مثال:

- استفاده از آزمون فرض جهت تعیین استراتژی مناسب برای برخورد با مقادیر گمشده

Sig<0.05

Field	Measurement	% Complete
PRIMEUNIT	Nominal	4.685
AUCGUART	Nominal	4.685

Sig>0.05

Field	Measurement	Values
IsBadBuy	Flag	1/0

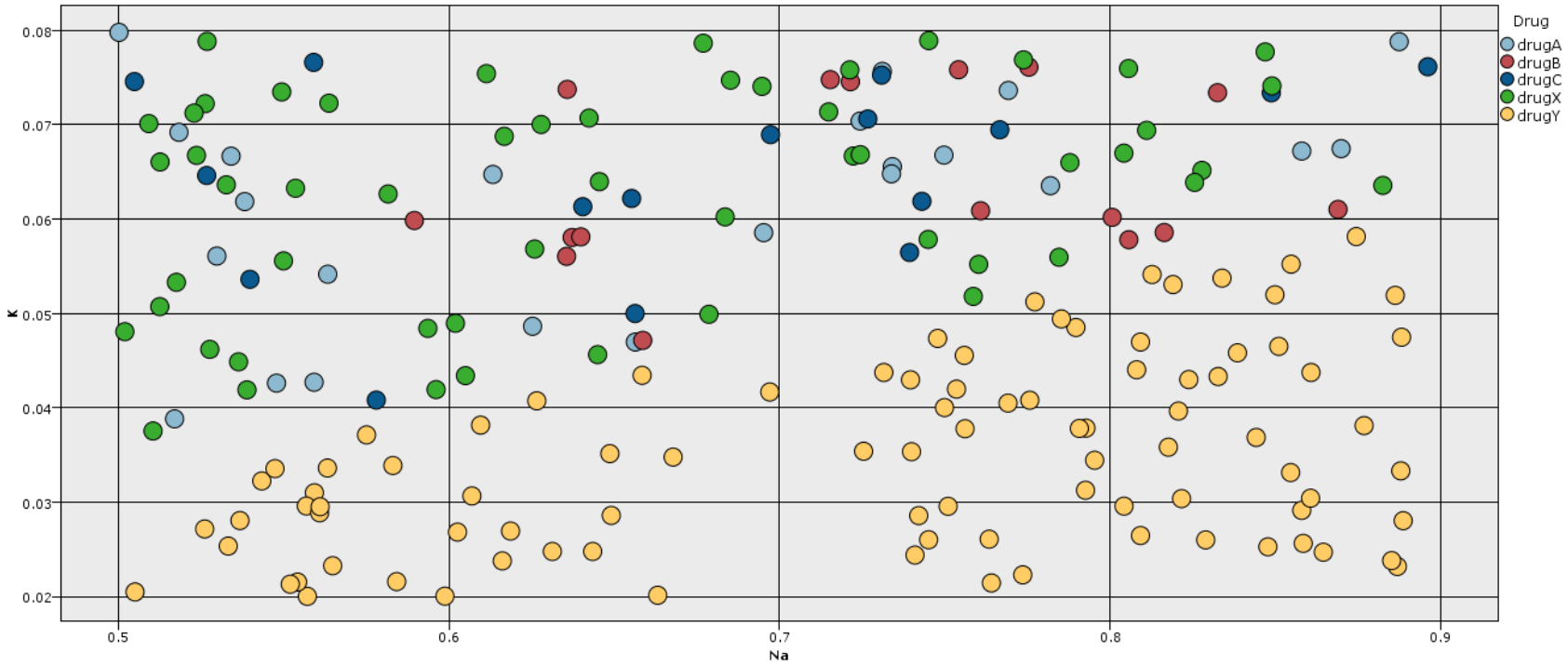
تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

توصیف و کاوش داده ها

❑ شناخت داده ها در فرآیند داده کاوی

کاربرد 3: استفاده از ابزارهای شناخت داده در تبدیل و شاخص سازی داده ها



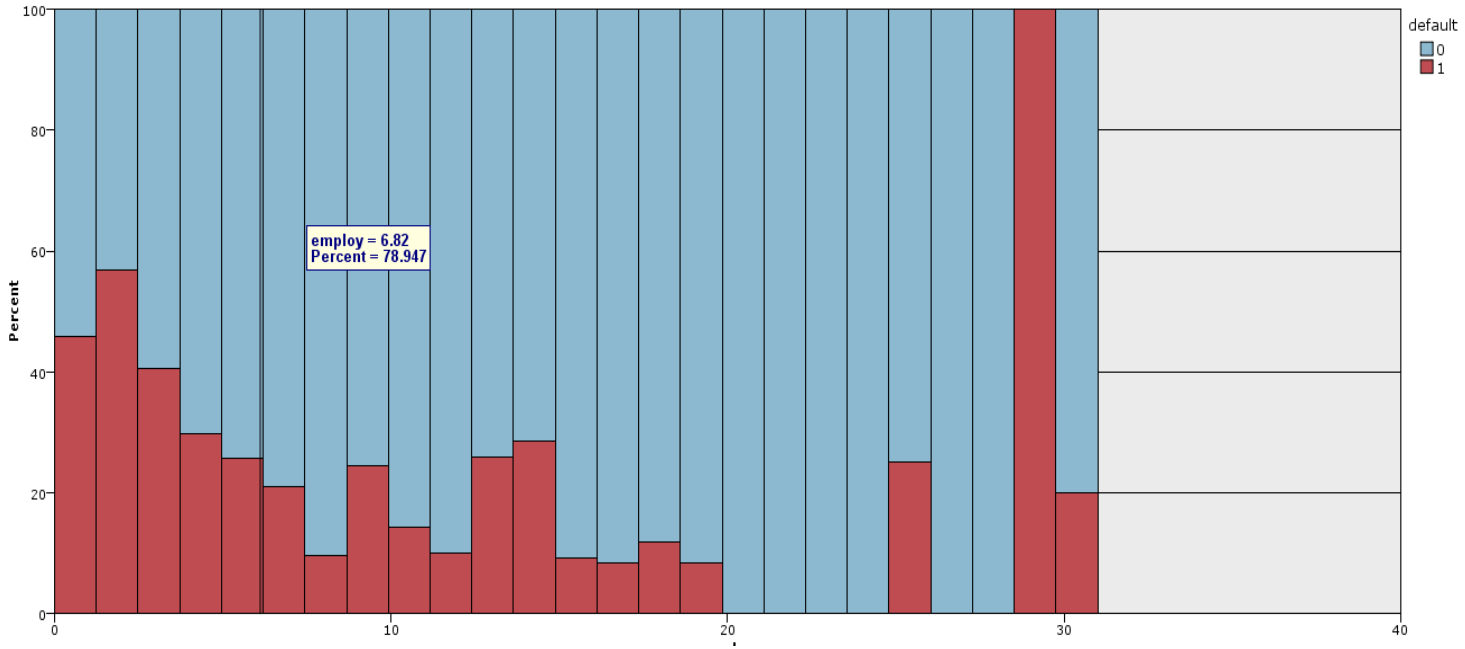
تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

توصیف و کاوش داده ها

❑ شناخت داده ها در فرآیند داده کاوی

کاربرد 4: استفاده از ابزارهای شناخت داده در انتخاب ویژگی های موثر



employ_OPTIMAL	1	Count	162	118	280
		Expected Count	206.8	73.2	280.0
		% within employ_OPTIMAL	57.9%	42.1%	100.0%
		Residual	-44.8	44.8	
	2	Count	355	65	420
		Expected Count	310.2	109.8	420.0
		% within employ_OPTIMAL	84.5%	15.5%	100.0%
		Residual	44.8	-44.8	
Total		Count	517	183	700
		Expected Count	517.0	183.0	700.0
		% within employ_OPTIMAL	73.9%	26.1%	100.0%

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	61.873[a]	1	.000		
Continuity Correction[b]	60.500	1	.000		
Likelihood Ratio	61.205	1	.000		
Fisher's Exact Test				.000	.000
Linear-by-Linear Association	61.785	1	.000		
N of Valid Cases	700				

تولید محتوا: زهرا ذوالقدر

daychegroup

daychegroup

dayche.com | گروه دایکه

فرآیند داده کاوی

توصیف و کاوش داده ها

t-test

Independency test

❑ شناخت داده ها در فرآیند داده کاوی

کاربرد 5: استفاده از ابزارهای شناخت داده در بررسی روابط و الگوها

IsBadBuy	Auction	VehicleAge	Transmission	WheelType	VehOdo	TopThreeAmericanName	VehBCost	IsOnlineSale	WarrantyCost
0	ADESA	3	AUTO	1	89046	OTHER	7100	0	1113
0	ADESA	5	AUTO	1	93593	CHRYSLER	7600	0	1053
0	ADESA	4	AUTO	2	73807	CHRYSLER	4900	0	1389
0	ADESA	5	AUTO	1	65617	CHRYSLER	4100	0	630
0	ADESA	4	MANUAL	2	69367	FORD	4000	0	1020
0	ADESA	5	AUTO	2	81054	OTHER	5600	0	594
0	ADESA	5	AUTO	2	65328	OTHER	4200	0	533
0	ADESA	4	AUTO	2	65805	FORD	4500	0	825
0	ADESA	2	AUTO	2	49921	OTHER	5600	0	482
0	ADESA	2	AUTO	1	84872	FORD	7700	0	1633
0	ADESA	4	AUTO	1	80080	GM	5500	0	1373
0	ADESA	8	MANUAL	1	75419	FORD	5300	0	869
1	ADESA	4	AUTO	1	79315	CHRYSLER	5400	0	1623
0	ADESA	4	AUTO	2	71254	OTHER	7800	0	686
0	ADESA	3	AUTO	1	74722	CHRYSLER	6900	0	1623
0	ADESA	6	MANUAL	2	72132	GM	3300	0	1455
0	ADESA	4	AUTO	1	80736	GM	6800	0	1243
0	ADESA	6	AUTO	2	75156	GM	4900	0	1923
0	ADESA	4	AUTO	2	65925	GM	5700	0	1703
0	ADESA	5	AUTO	1	84498	GM	7100	0	1243
0	ADESA	7	AUTO	2	54586	CHRYSLER	4800	0	1551
0	ADESA	7	AUTO	1	66536	GM	5800	0	2003
0	ADESA	5	AUTO	2	98130	GM	4100	0	5613
0	ADESA	3	AUTO	2	59789	GM	7700	0	671

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

توصیف و کاوش داده ها

❑ شناخت داده ها در فرآیند داده کاوی

کاربرد 6: استفاده از ابزارهای شناخت داده در کسب بینش از مدل ها

Clusters

Input (Predictor) Importance

1.0 0.8 0.6 0.4 0.2 0.0

Cluster	cluster-3	cluster-6	cluster-1	cluster-4	cluster-2	cluster-5
Label						
Size	27.3% (571)	21.0% (439)	19.1% (399)	13.3% (278)	12.4% (258)	6.9% (144)
Inputs	Frequency Score 4.50	Frequency Score 1.31	Frequency Score 1.28	Frequency Score 1.56	Frequency Score 3.56	Frequency Score 4.33
	Monetary Score 3.61	Monetary Score 1.62	Monetary Score 4.64	Monetary Score 1.66	Monetary Score 2.07	Monetary Score 4.21
	Recency Score 3.86	Recency Score 1.49	Recency Score 2.26	Recency Score 3.63	Recency Score 2.10	Recency Score 1.94

Gender * Clusters Crosstabulation

			Clusters						
			cluster-1	cluster-2	cluster-3	cluster-4	cluster-5	cluster-6	Total
Gender	F	Count	8	9	10	7	1	11	46
		Expected Count	8.8	5.7	12.6	6.1	3.2	9.7	46.0
		% within Gender	17.4%	19.6%	21.7%	15.2%	2.2%	23.9%	100.0%
		Residual	-.8	3.3	-2.6	.9	-2.2	1.3	
	M	Count	391	249	561	271	143	428	2043
		Expected Count	390.2	252.3	558.4	271.9	140.8	429.3	2043.0
		% within Gender	19.1%	12.2%	27.5%	13.3%	7.0%	20.9%	100.0%
		Residual	.8	-3.3	2.6	-.9	2.2	-1.3	
Total		Count	399	258	571	278	144	439	2089
		Expected Count	399.0	258.0	571.0	278.0	144.0	439.0	2089.0
		% within Gender	19.1%	12.4%	27.3%	13.3%	6.9%	21.0%	100.0%

Chi-Square Tests

	Value	df	Asymptotic Significance (2-sided)
Pearson Chi-Square	4.430[a]	5	.489
Likelihood Ratio	4.711	5	.452
N of Valid Cases	2089		

a. 1 cells (8.3%) have expected count less than 5. The minimum expected count is 3.17.

تولید محتوا: زهرا ذوالقدر

Data Quality

کیفیت داده ها

dayche.com . گروه دایکه

فرآیند داده کاوی

کیفیت داده ها

❑ انواع شاخص کیفیت داده ها

داده های واقعی عموما دارای انواع مشکلات کیفی هستند که نیاز به پاکسازی را ضروری می نماید. استفاده از داده های خام و دارای مشکلات کیفی، منجر به کاهش عملکرد الگوریتم ها، اختلال در شناسایی الگوها و نتایج گمراه کننده می شود.

کد مشتری	کد پستی	سن	جنسیت	درآمد (تومان)	شغل
c-25687	14546-32148	32	مرد	15.000.000	مدیر پروژه
c-62578	65489-32698	698	مر د	7.000.000	کارشناس
c-32689	32478-89543	37	زن	0	خانه دار
c-11457	2564898562		مرد	-14	آزاد
11111111	98742-25691	17			کارمند
c-19871	32659	56	زن	60.000.000	پزشک
c-36958	21698-52149	1368	زن	12.000.000	مدیرپروژه

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

کیفیت داده ها

❑ انواع شاخص کیفیت داده ها

داده های واقعی عموما دارای انواع مشکلات کیفی هستند که نیاز به پاکسازی را ضروری می نماید. استفاده از داده های خام و دارای مشکلات کیفی، منجر به کاهش عملکرد الگوریتم ها، اختلال در شناسایی الگوها و نتایج گمراه کننده می شود.

کد مشتری	کد پستی	سن	جنسیت	درآمد (تومان)	شغل
c-25687	14546-32148	32	مرد	15.000.000	مدیر پروژه
c-62578	65489-32698	698	مر د	7.000.000	کارشناس
c-32689	32478-89543	37	زن	0	خانه دار
c-11457	2564898562		مرد	-14	آزاد
11111111	98742-25691	17			کارمند
c-19871	32659	56	زن	60.000.000	پزشک
c-36958	21698-52149	1368	زن	12.000.000	مدیرپروژه

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

کیفیت داده ها

❑ انواع شاخص کیفیت داده ها

داده های واقعی عموما دارای انواع مشکلات کیفی هستند که نیاز به پاکسازی را ضروری می نماید. استفاده از داده های خام و دارای مشکلات کیفی، منجر به کاهش عملکرد الگوریتم ها، اختلال در شناسایی الگوها و نتایج گمراه کننده می شود.

خارج از بازه منطقی Out of Logical Range	ناسازگار Inconsistent	داده های پرت Outliers	داده های گمشده Missing Data

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

کیفیت داده ها

❑ مقادیر خارج از بازه منطقی

شرایطی که در آن یک یا چند مقدار با استانداردهای تعریف شده مطابقت ندارد.

بطور مثال

- مقادیر وزن منفی یا مقدار درصد بیشتر از 100
- مقدار سن کمتر از سن قانونی برای دریافت کننده تسهیلات بانکی

روش شناسایی

فیلترینگ بر اساس دانش زمینه ای داده ها (بطور مثال با داشتن دانش زمینه ای در باره قرار گرفتن منطقی مقادیر ویژگی ها در یک بازه خاص و مقایسه آن با مقادیر مینیمم و ماکسیمم داده ها می توان مقادیر خارج از بازه منطقی در آن ویژگی را فیلتر نمود.)

تولید محتوا: زهرا ذوالقدر

فرآیند داده کاوی

کیفیت داده ها

❑ مقادیر خارج از بازه منطقی

روش های پاکسازی

- در صورت دسترسی به مقدار صحیح: جایگزینی با مقدار صحیح
- در صورت عدم دسترسی به مقدار صحیح: حذف مقدار اشتباه

شناسایی مقادیر خارج از بازه منطقی اولین گام مهم در پاکسازی داده ها می باشد.

چرا که بطور مستقیم در توزیع آماری داده ها اثر منفی می گذارد.

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

کیفیت داده ها

❑ ناسازگاری در داده ها

ناسازگاری در داده ها به معنای عدم همخوانی با سایر داده ها می باشد.

بطور مثال

- تعریف کدهای متفاوت برای رنگ بدنه خودرو
- ثبت نام شهر با نوشتار های متفاوت در مجموعه داده ها

روش شناسایی

- **خلاصه سازی آماری داده** (بطور مثال استفاده از جدول فراوانی و یا نمودار توزیع داده ها، می تواند منجر به شناسایی سریع برچسب های اشتباه یا ناسازگار در مقادیر یک ویژگی شود)

فرآیند داده کاوی

کیفیت داده ها

- **ناسازگاری در داده ها**

روش های پاکسازی

- اصلاح و یکسان سازی کدها

شناسایی مقادیر ناسازگار و اشتباه اولین گام مهم در پاکسازی داده ها می باشد.

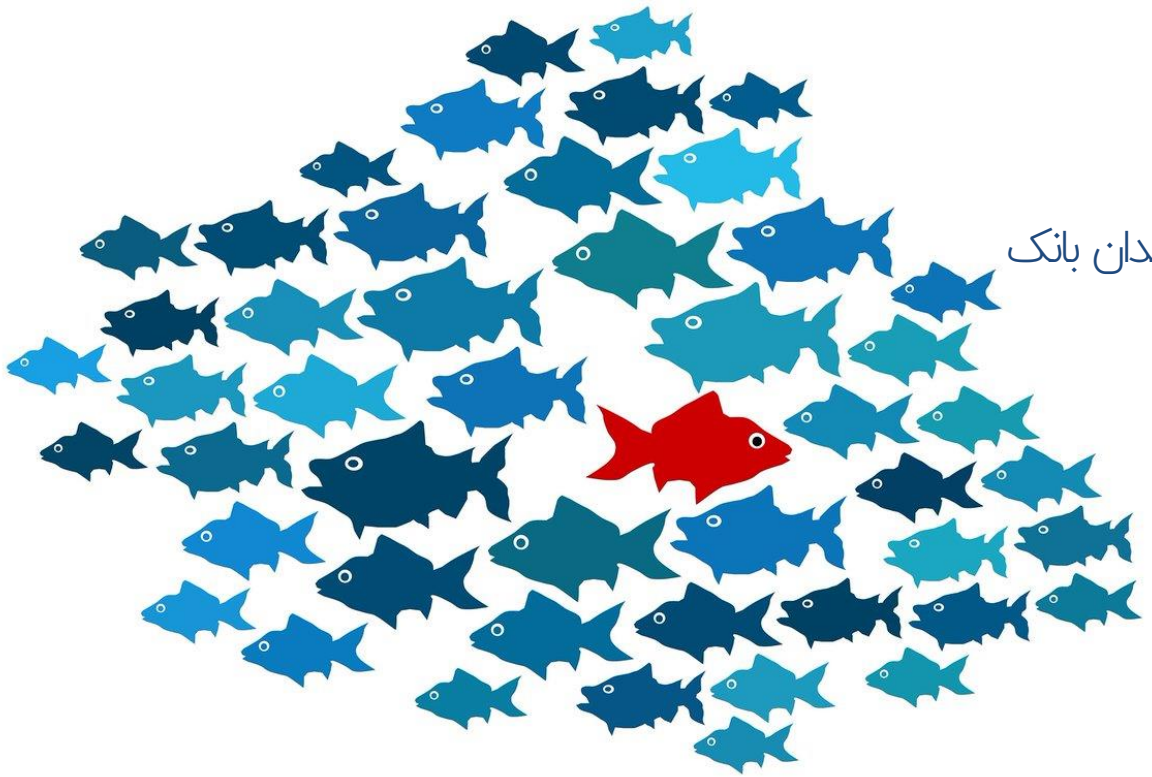
چرا که بطور مستقیم در توزیع آماری داده ها اثر منفی می گذارد.

فرآیند داده کاوی

کیفیت داده ها

❑ داده های پرت

داده های پرت الزاما داده های اشتباه نیستند بلکه از توزیع آماری بدنه اصلی داده ها پیروی نمی کنند.



بطور مثال

- مقدار درآمد ثبت شده 60 میلیونی در یک گروه تصادفی از کارمندان بانک
- وجود یک لکه متمایز در تصویر سی تی اسکن ریه
- سن 48 سال در داده های دانشجویان ترم اول کارشناسی

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

کیفیت داده ها

❑ داده های پرت

داده های پرت به علت فاصله زیاد با سایر داده ها، در اغلب الگوریتم ها منجر به ایجاد اختلال در شناسایی الگوها می شوند. بنابراین شناسایی و برخود مناسب با آنها از اهمیت بالایی برخوردار می باشد.

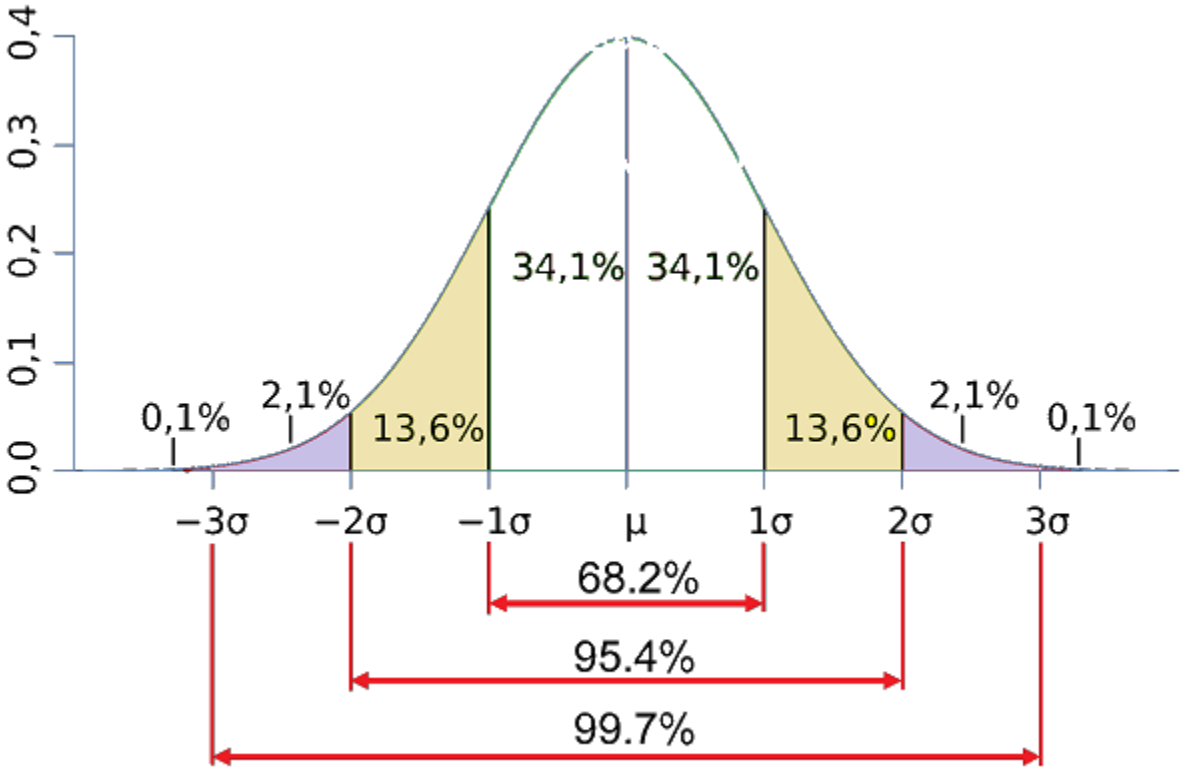
انواع داده های پرت

- **داده های پرت تک متغیره** (عمدتا با استفاده از روش های توصیفی آماری و مفاهیم توزیع آماری روی یک ویژگی **مقادیر پرت** شناسایی می شوند)
- **داده های پرت چند متغیره** (عمدتا با استفاده از روش های تحلیل چندمتغیره و در فضای n-بعدی **رکوردهای پرت** شناسایی می شوند)

فرآیند داده کاوی

کیفیت داده ها

❑ داده های پرت



شناسایی مقادیر پرت

- بر اساس توزیع نرمال (روش تک متغیره - پارامتری)

نقاط ضعف و قوت

- مناسب برای مجموعه داده های با تعداد کم ویژگی های کمی
- روش سریع و با هزینه محاسباتی کم در تعداد رکوردهای زیاد
- نیازمند فرض توزیع نرمال برای ویژگی های مورد بررسی

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

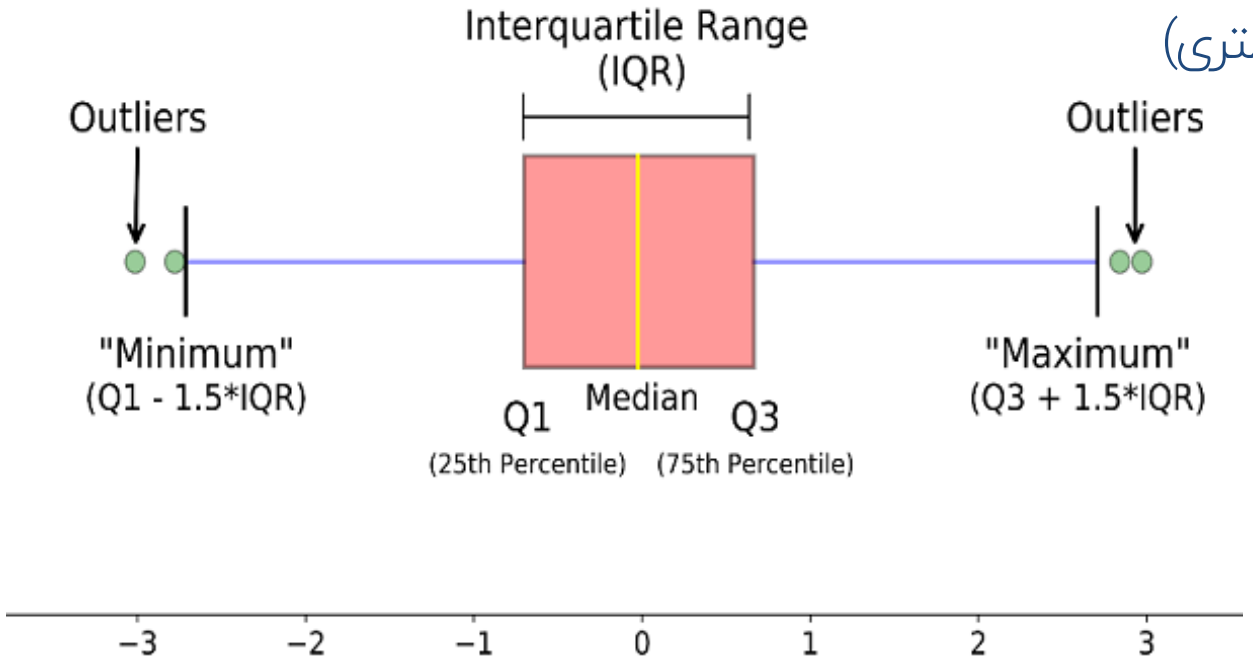
فرآیند داده کاوی

کیفیت داده ها

❑ داده های پرت

شناسایی مقادیر پرت

- بر اساس دامنه میان چارکی (روش تک متغیره – ناپارامتری)



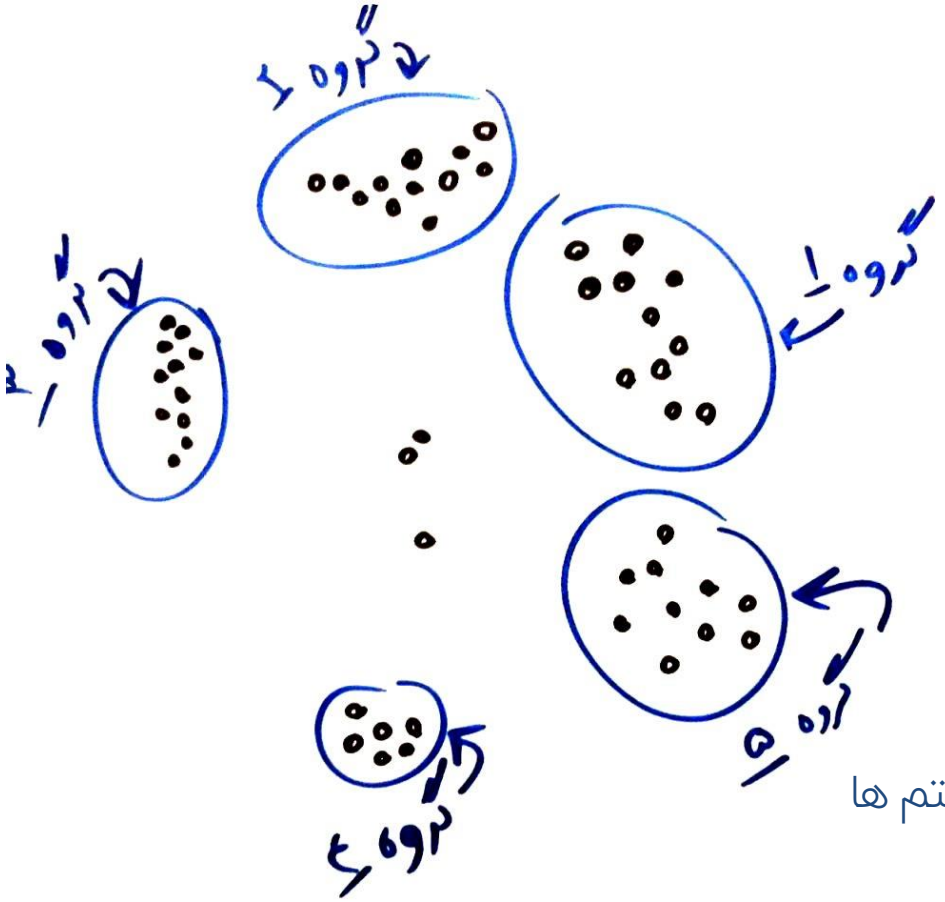
نقاط ضعف و قوت

- مناسب برای مجموعه داده های با تعداد کم ویژگی های کمی
- هزینه محاسباتی بالا در داده های با تعداد رکوردهای زیاد
- مستقل از توزیع آماری داده ها

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

کیفیت داده ها



❑ داده های پرت

شناسایی رکوردهای پرت

- بر اساس خوشه بندی (روش چند متغیره)

نقاط ضعف و قوت

- قابل استفاده در داده های با تعداد ویژگی های زیاد
- استفاده از الگو ها و روابط بین ویژگی ها در تشخیص رکوردهای پرت
- پیچیدگی اجرایی و نیاز به دانش در خصوص بهینه سازی پارمترها و اجرای الگوریتم ها

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

کیفیت داده ها

❑ داده های پرت

برخورد با داده های پرت

برخورد با داده های پرت به هدف مسئله و موضوع زمینه آن وابسته هست.

در برخی از مسائل، هدف اصلی کشف الگوهای نادر و آنومالی ها می باشد.

بطور مثال

مدل تشخیص غده بدخیم در تصاویر پزشکی

شناسایی الگوهای تخلف در پرونده های خسارتی یک شرکت بیمه

ولی در اغلب مسائل به دنبال الگوی بدنه اصلی داده ها هستیم و وجود داده های پرت منجر به کاهش کارایی مدلها می شوند.

فرآیند داده کاوی

کیفیت داده ها

❑ داده های پرت

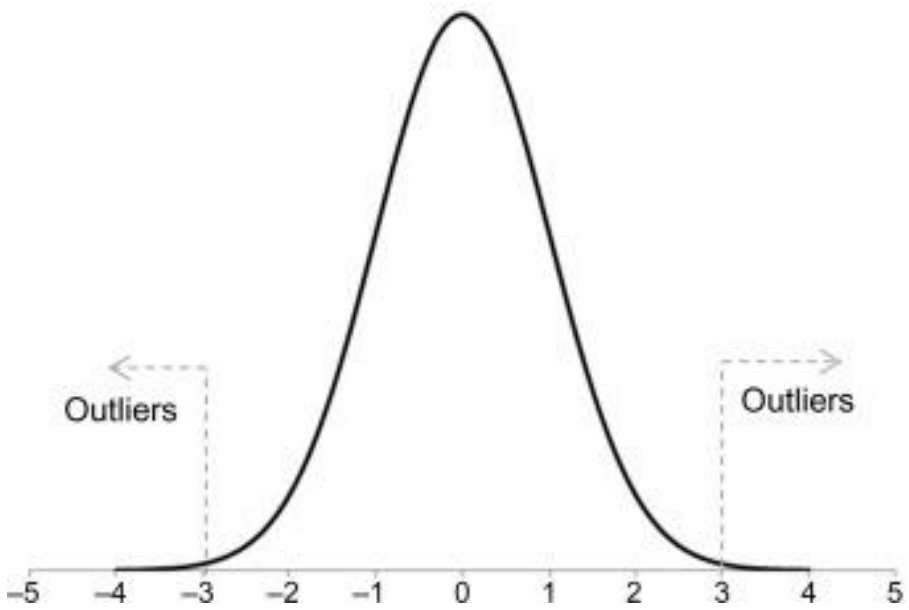
روش های برخورد با مقادیر پرت

- روش جایگزینی

یکی از رایج ترین روش های اصلاح در مقادیر پرت، جایگزینی مقادیر پرت با مقادیر آستانه ای قابل قبول در توزیع داده ها می باشد.

بطور مثال در تصویر روبرو، با استفاده از روش مبتنی بر توزیع نرمال، کلیه مقادیر خارج از محدوده سه انحراف معیار، با نقاط آستانه ای جایگزین می شوند. این تغییر منجر به تغییر در توزیع داده می شود.

بنابراین با تغییر پارامترهای توزیع، تکرار این روش می تواند مجددا داده های دیگری را به عنوان نقاط پرت شناسایی کند.



تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

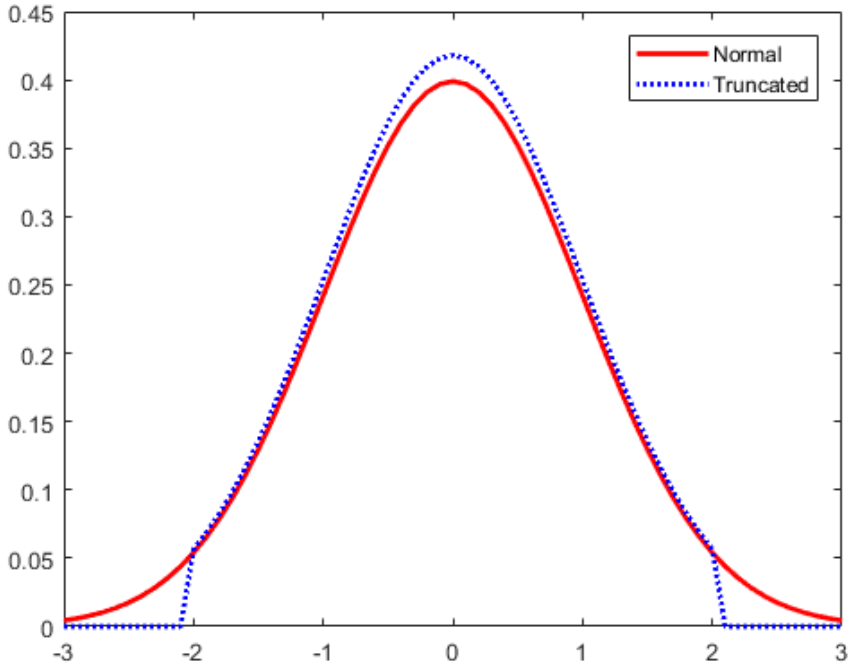
کیفیت داده ها

❑ داده های پرت

روش های برخورد با مقادیر پرت

- روش حذفی

روش حذفی در برخورد با مقادیر پرت عموما به معنای پاک کردن مقدار پرت یک ویژگی می باشد.



بطور مثال در تصویر روبرو، با استفاده از روش مبتنی بر توزیع نرمال، کلیه مقادیر خارج از محدوده دو انحراف معیار، حذف می شوند. این تغییر منجر به تغییر در توزیع داده می شود. بنابراین با تغییر پارامترهای توزیع، تکرار این روش می تواند مجددا داده های دیگری را به عنوان نقاط پرت شناسایی کند.

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

کیفیت داده ها

❑ داده های پرت

روش های برخورد با رکورد های پرت

- **حفظ رکوردهای پرت**

این استراتژی در مسائلی از نوع شناسایی آنومالی ها که هدف اصلی می باشد بکار می رود. در واقع هدف کشف الگوی رکوردهای پرت هست.

- **حذف رکوردهای پرت**

در اغلب مسائل رایج ترین استراتژی در مواجهه با رکوردهای پرت، حذف آنها پس از بررسی علل و چرایی آنهاست.

- **جداسازی رکوردهای پرت و مدلسازی جداگانه**

در مواقعی که تعداد رکوردهای پرت شناسایی شده به اندازه کافی بزرگ باشد، در صورت اهمیت موضوع می توان بصورت جداگانه اقدام به مدلسازی بر روی آنها کرد.

فرآیند داده کاوی

کیفیت داده ها

❑ مقادیر گمشده

مقادیر گمشده به داده هایی گفته می شود که به دلایل مختلفی همچون عدم ثبت (یا پاسخ دهی) عمدی یا سهوی در مجموعه داده ها ایجاد شده است. همچنین داده های گمشده می تواند به علت استفاده از استراتژی های پاکسازی مقادیر خارج از بازه منطقی یا داده های پرت بوجود آمده باشد.

بررسی علت شکل گیری داده های گمشده، جهت انتخاب روش برخورد از اهمیت بالایی برخوردار می باشد.

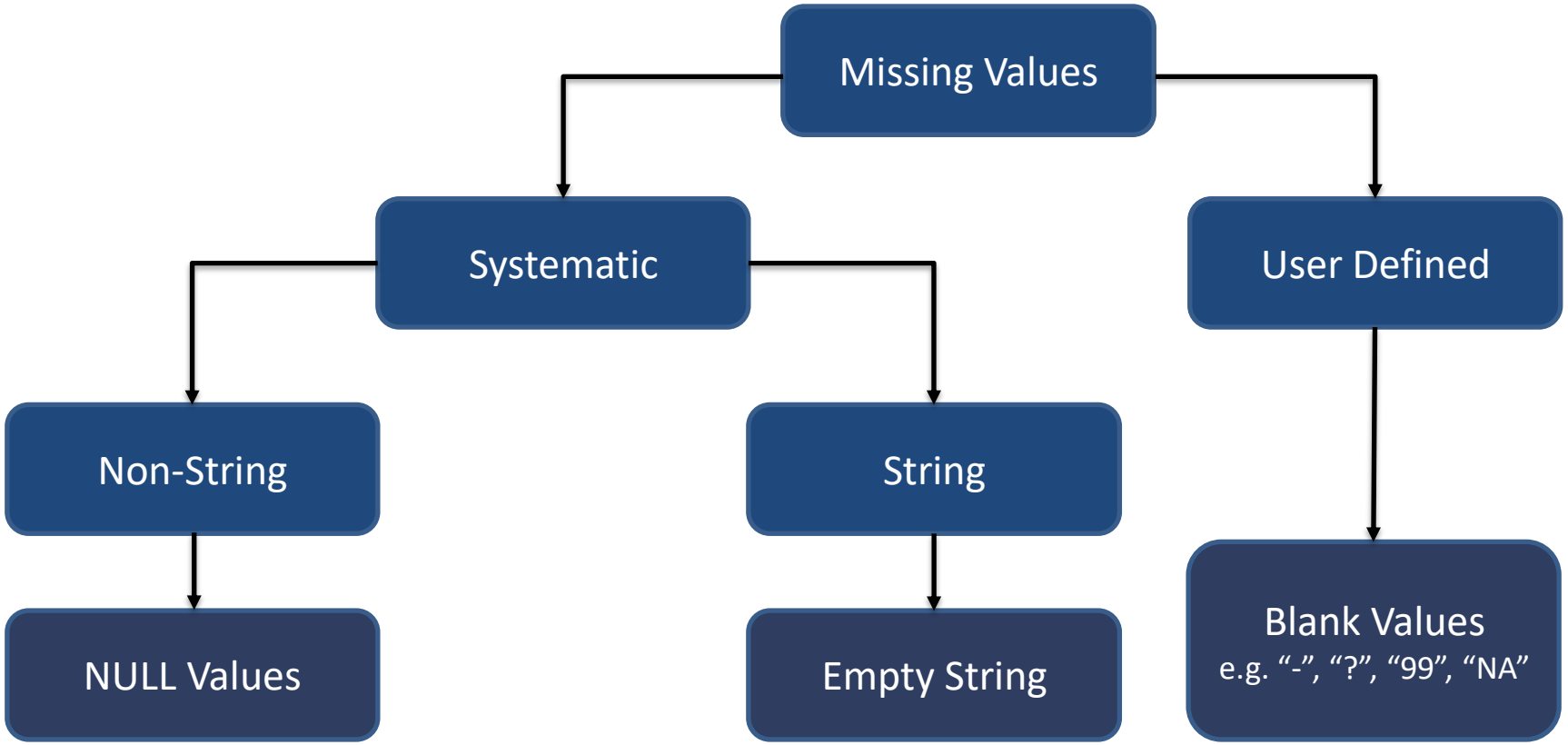
تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

کیفیت داده ها

❑ داده های گمشده

داده های گمشده به شکل های مختلفی در مجموعه داده ها ثبت و شناسایی می شوند:



تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
گروه دایکه | dayche.com

❑ مقادیر گمشده

استراتژی های برخورد با مقادیر گمشده

انتخاب روش برخورد با مقادیر گمشده به پارامترهای مختلفی وابسته هست؛ همچون:

- ماهیت و موضوع زمینه ای داده های گمشده
- پراکندگی وقوع مقادیر گمشده
- تعداد ویژگی های دارای مقادیر گمشده در هر رکورد
- تعداد رکوردهای دارای مقادیر گمشده در هر ویژگی
- اهمیت ویژگی های دارای مقادیر گمشده و ارتباط آنها با سایر ویژگی ها

فرآیند داده کاوی

کیفیت داده ها

❑ مقادیر گمشده

استراتژی های برخورد با مقادیر گمشده

بررسی دقیق از ماهیت و ارتباط مقادیر یک ویژگی با سایر ویژگی ها گام اول جهت شناسایی و برخورد با داده های گمشده می باشد.

شناسه	وضعیت تأهل	جنسیت	تعداد فرزند
1235	مجرد	مرد	
1785	متأهل	مرد	2
1359	متأهل	زن	
1365	مجرد	مرد	
1248	مجرد	زن	
1456	متأهل	زن	0
1532	متأهل	زن	1

بطور مثال: در جدول داده روبرو تعداد 4 مقدار گمشده مشاهده می شود. با دقت در ماهیت داده ها خواهیم دید 3 مقدار گمشده مربوط به رکوردهایی می باشد که وضعیت تأهل آنها "مجرد" هست. بنابراین بصورت ذاتی آنها بدون مقدار می باشند.

نحوه برخورد: برخورد با مقادیر گمشده که بصورت ذاتی دارای مقدار نمی باشند اغلب با تعریف کلاس جدید مانند "Unknown" یا مقداردهی عدد ثابت بر اساس موضوع داده مورد بررسی و یا ادغام ویژگی های مرتبط انجام می شود.

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

کیفیت داده ها

❑ مقادیر گمشده

استراتژی های برخورد با مقادیر گمشده

روش های عمده در برخورد با مقادیر گمشده شامل موارد زیر است:

روش حذفی

Dropping

روش جانشینی

Imputation

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

کیفیت داده ها

❑ مقادیر گمشده

استراتژی های برخورد با مقادیر گمشده

- روش حذفی

حذف رکوردها یا ویژگی هایی که دارای مقادیر گمشده می باشند یکی از سریع ترین و راحت ترین استراتژی های برخورد می باشد.

حذف ویژگی ها: در اغلب موارد ویژگی هایی که دارای تعداد زیادی مقادیر گمشده باشد (بیش از 50%) و امکان دسترسی به داده های صحیح نباشد به علت جلوگیری از بایاس (اریبی) از تحلیل کنار گذاشته می شوند.

در مواردی که ویژگی مورد نظر اهمیت بسیار زیادی در تحلیل داشته باشد می توان از کلاس " unknown" برای آنها استفاده کرد.

حذف رکوردها: حذف کلیه رکوردهای دارای مقادیر گمشده، امکان از دست دادن تعداد قابل توجهی از مجموعه داده ها را در پی دارد. بنابراین این استراتژی عموما برای رکوردهای بی کیفیت که تعداد ویژگی های مفقود شده آن بیش از 50% کل ویژگی ها باشد استفاده می شود و یا در مواردی که داده های بسیار بزرگ در اختیار داریم قابل استفاده است.

فرآیند داده کاوی

کیفیت داده ها

❑ مقادیر گمشده

استراتژی های برخورد با مقادیر گمشده

- روش جانشینی با اعداد ثابت یا تصادفی

روش های جانشینی عمدتا بر اساس پارامترهای توزیع هر ویژگی مانند میانگین، میانه، مد و ... و یا شبیه سازی توزیع آماری ویژگی ها انجام می شود. استفاده از این روش در تعداد زیاد مقادیر گمشده، مستعد بایاس و خطا در تحلیل هست و بایستی با توجه به شناخت از ماهیت و کمیت داده ها انجام شود.

نکته 1: در توزیع های دارای چولگی استفاده از میانه جایگزین بهتری نسبت به میانگین می باشد.

نکته 2: در مجموعه داده هایی که رکوردهای آن دارای توالی زمانی هستند، استفاده از درونیابی و یا پارامترهای متحرک (مانند میانگین متحرک) مفید می باشد.

نکته 3: استفاده از شبیه سازی توزیع آماری (بطور مثال تولید اعداد تصادفی از توزیع نرمال، یکنواخت و یا توزیع تجربی) و جانشینی آن با مقادیر گمشده، منجر به جلوگیری از بایاس بودن نتایج می شود، اما همچنان ریسک ایجاد خطا در داده ها وجود دارد.

فرآیند داده کاوی

کیفیت داده ها

❑ مقادیر گمشده

استراتژی های برخورد با مقادیر گمشده

- روش جانشینی با الگوریتم

روش های الگوریتمی از نظر جلوگیری از بایاس و ناسازگاری در داده ها بهترین روش می باشند اما دارای پیچیدگی های محاسباتی و اجرایی هستند.

- CART
- KNN
- Random Forest
- Regression , ...

استفاده از روش های الگوریتمی دارای فرض ارتباط منطقی بین ویژگی دارای مقادیر گمشده با سایر ویژگی ها هست. بنابر این فرض، سایر ویژگی های مرتبط به عنوان ورودی الگوریتم برای حل مسئله رگرسیون (پیش بینی) یا رده بندی در نظر گرفته می شود.

مقادیر پیش بینی شده جایگزین مقادیر گمشده در ویژگی مورد بررسی می شود.

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

Data Transformation

تبدیل داده ها

dayche.com . گروه دایکه

دای‌کِه DAYCHE

فرآیند داده کاوی

تبدیل داده ها

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

تبدیل داده ها

❑ نرمال سازی داده ها (Normalization)

نرمال سازی داده ها با دو هدف عمده انجام می شود:

هم مقیاس سازی مقادیر ویژگی های ورودی

تغییر توزیع آماری مقادیر ویژگی ها و کاهش چولگی داده ها

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

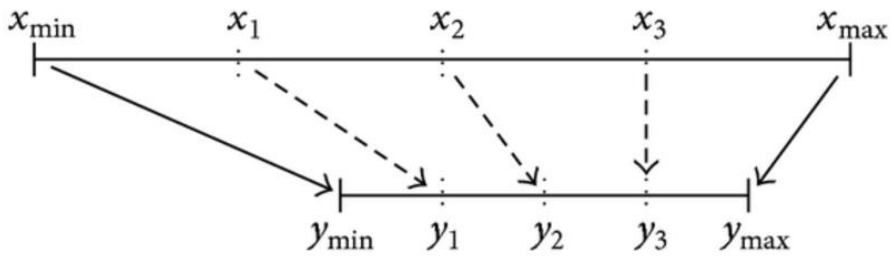
فرآیند داده کاوی

تبدیل داده ها

❑ نرمال سازی داده ها (Normalization) – هم مقیاس سازی

مقیاس و واحد اندازه گیری داده ها نباید در مدلسازی و تحلیل داده ها اثرگذار باشد. از این رو از روش هایی برای نرمالسازی (یا استانداردسازی) داده ها استفاده می شود.

بطور مثال: حضور همزمان دو ویژگی سن با دامنه 65 – 20 سال و درآمد با دامنه 25.000.000 – 2.500.000 تومان در یک مدل رده بندی یا خوشه بندی می تواند منجر به کاهش اثر ویژگی سن (بطور کاذب) در مدل شود.



روش های نرمالسازی عمدتا دامنه داده های خام را به دامنه کنترل شده ای مانند [0, 1] ، یا [1, -1] انتقال می دهند تا اثر واحد اندازه گیری را خنثی کنند.

روش Min-Max | روش Z-Score | روش Robust Scaling | روش Decimal Scaling

روش های نرمالسازی جهت هم مقیاس سازی به معنای نرمال کردن توزیع آماری نیست.

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

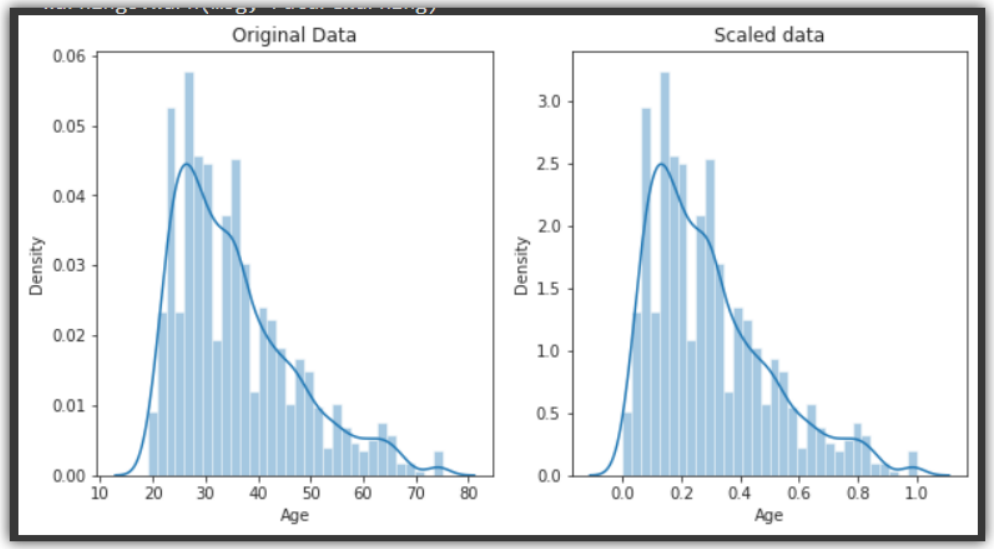
فرآیند داده کاوی

تبدیل داده ها

❑ نرمال سازی داده ها (Normalization) – هم مقیاس سازی

- روش Min – Max

با تبدیل خطی بروی داده ها، دامنه آنها را به بازه 0 تا 1 نگاشت می دهد.



$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

سوال: با تغییر در رابطه فوق دامنه تبدیل Min – Max را به بازه a تا b نگاشت دهید.

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

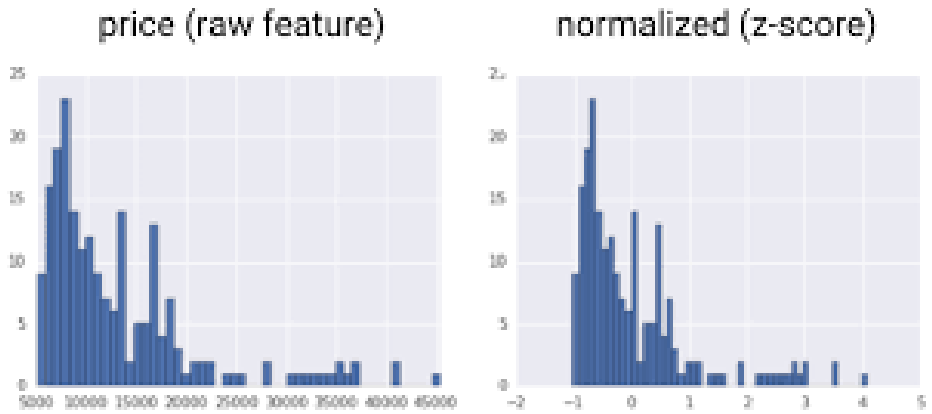
فرآیند داده کاوی

تبدیل داده ها

❑ نرمال سازی داده ها (Normalization) – هم مقیاس سازی

- روش Z-Score

با تبدیل خطی بر اساس میانگین و انحراف معیار توزیع داده ها، میانگین و واریانس ویژگی به مقادیر 0 و 1 تبدیل می شوند.



$$Z = \frac{X - \bar{X}}{s}$$

نکته: تبدیل Z-Score توزیع داده ها را نرمال نمی کند و منجر به کاهش چولگی نمی شود.

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

تبدیل داده ها

❑ نرمال سازی داده ها (Normalization) – هم مقیاس سازی

- روش Robust Scaling

با تبدیل خطی بر اساس میانه و دامنه میان چارکی داده ها، نسبت به اثر داده های پرت مقاوم است.

نرمال سازی مقاوم مرکزیت داده ها را به صفر نگاشت می دهد

ولی مقادیر مینیمم و ماکسیمم آن متغیر است.

$$\bar{x} = \frac{X - median(X)}{75\text{th quantile(x)} - 25\text{th quantile(x)}}$$

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

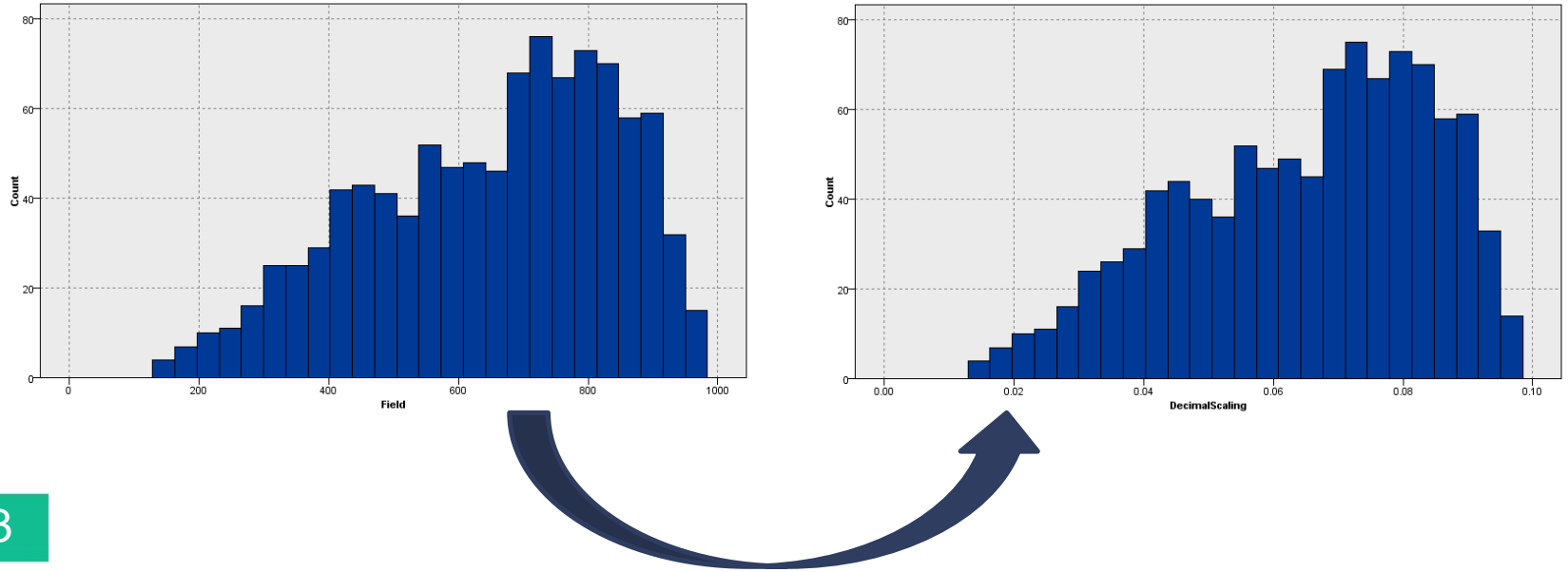
فرآیند داده کاوی

تبدیل داده ها

❑ نرمال سازی داده ها (Normalization) – هم مقیاس سازی

- روش Decimal Scaling

با تبدیل خطی، با تغییر مکان اعشار به تعداد ارقام مقدار حداکثر مطلق انجام می شود. در رابطه زیر مقدار j کوچکترین عدد صحیحی است که مقدار قدر مطلق دادهای نگاشت داده شده را کوچکتر از مقدار 1 باشد.



$$v' = \frac{v}{10^j}$$

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

تبدیل داده ها

❑ نرمال سازی داده ها (Normalization) – تغییر توزیع آماری

در برخی از مدل ها به خصوص مدل های آماری، نیاز به وجود فرض نرمال بودن توزیع آماری یکی از فرضیات مدلسازی است. برخی تبدیل های رایج همچون تبدیل لگاریتمی، نمایی، توانی و ... با کاهش چولگی توزیع داده ها می تواند به این هدف کمک کند.

e^y $y^{\lambda>1}$

این توابع با متورم کردن اعداد بزرگ در فضای جدید، تاثیر مقادیر کوچک در توزیع داده ها را کاهش می دهند.

$\sqrt{y}$ $1/y$ $\ln y$

این توابع با کوچک کردن اعداد بزرگ در فضای جدید، چولگی راست توزیع داده ها را کاهش می دهند.

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

تبدیل داده ها

❑ نرمال سازی داده ها (Normalization) – تغییر توزیع آماری

- تبدیل Box-Cox

یکی از رایج ترین تبدیل های توانی هست که جهت نرمالسازی توزیع آماری ویژگی های دارای توزیع چوله استفاده می شود.

$$y(\lambda) = \begin{cases} \frac{y^{\lambda}-1}{\lambda}, & \text{if } \lambda \neq 0; \\ \log y, & \text{if } \lambda = 0. \end{cases}$$

در صورتی که کلیه مقادیر ویژگی y بزرگتر از صفر باشد

$$y(\boldsymbol{\lambda}) = \begin{cases} \frac{(y+\lambda_2)^{\lambda_1}-1}{\lambda_1}, & \text{if } \lambda_1 \neq 0; \\ \log(y+\lambda_2), & \text{if } \lambda_1 = 0. \end{cases}$$

در صورتی که ویژگی y دارای مقادیر کوچکتر از صفر باشد

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

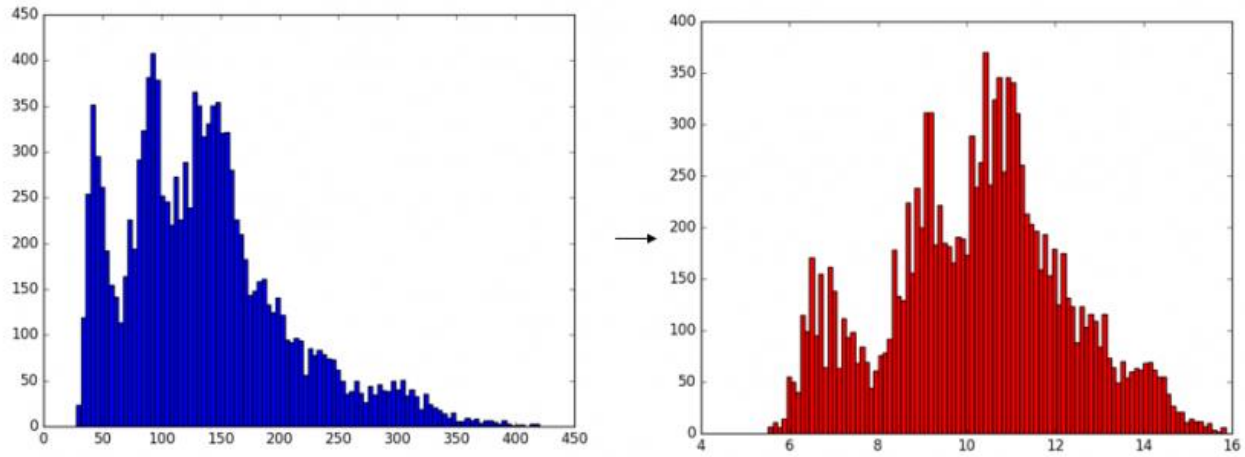
تبدیل داده ها

❑ نرمال سازی داده ها (Normalization) – تغییر توزیع آماری

- تبدیل Box-Cox

برای تعیین بهترین مقدار پارامتر λ معمولا در دامنه 5- تا 5 الگوریتم باکس کاکس اجرا می شود تا بهترین مقدار برای نرمالسازی بر اساس آزمون آماری نیکویی برازش بدست آید.

Lambda value (λ)	Transformed data (Y')
-3	1/Y^3
-2	1/Y^2
-1	1/Y
-0.5	1/(√(Y))
0	log(Y)
0.5	√(Y)
1	Y
2	Y^2
3	Y^3



تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

تبدیل داده ها

❑ ساخت ویژگی (Construction)

یکی از رایج ترین مراحل آماده سازی داده ها، ساخت ویژگی های جدید و اثربخش برای ورود به مدل ها می باشد. این مرحله به عنوان فرآیند شاخص سازی با اهداف مختلفی همچون تفسیرپذیری، تعمیم پذیری، افزایش کارایی مدل ها و ... در داده های خام انجام می شود که شامل رویکردهای زیر می باشد:

- استفاده از دانش زمینه ای در خصوص داده ها و روابط بین آنها
- استفاده از روابط آماری بین داده ها بر اساس روش های توصیف و کاوش
- کدگذاری عددی مقادیر ویژگی های اسمی

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

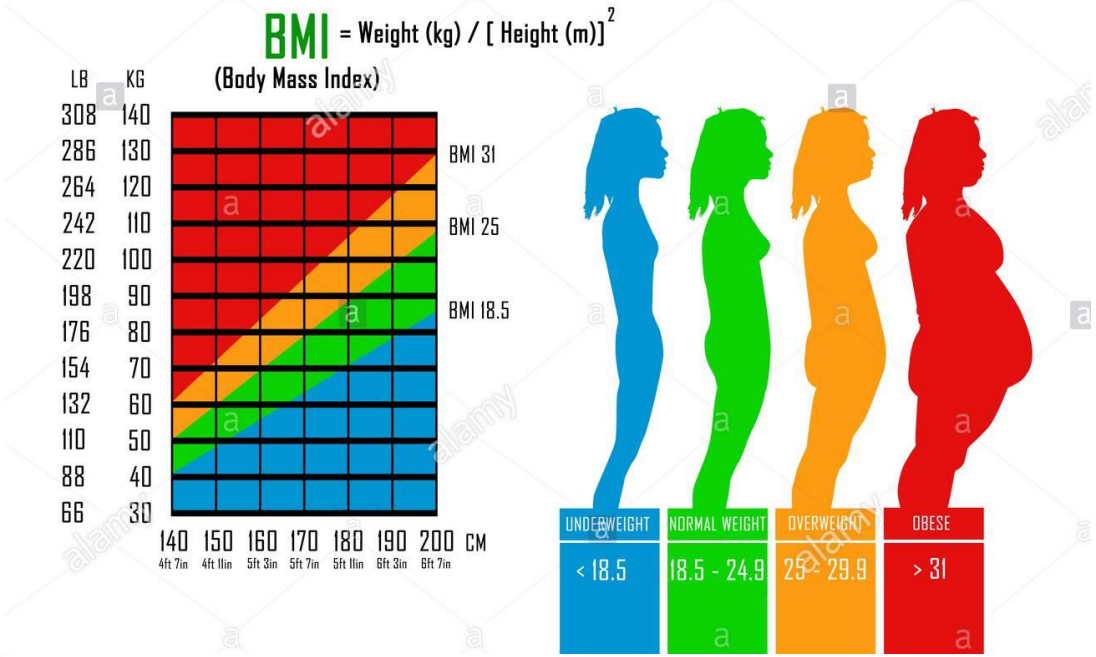
فرآیند داده کاوی

تبدیل داده ها

❑ ساخت ویژگی (Construction)

- **استفاده از دانش زمینه ای**

استفاده از دانش زمینه ای در خصوص رابطه متقابل بین ویژگی ها و یا ارتباط آنها با فیلد هدف می تواند ایده هایی مبنی بر ترکیب ویژگی های اولیه و ساخت ویژگی های جدید بیان کند.



مثال 1: با داشتن ویژگی های وزن و قد در بررسی عوامل موثر بر کنترل دیابت، می توان شاخص توده بدنی (BMI) را جایگزین ویژگی های اولیه نمود.

مثال 2: ساخت شاخص درآمد با تقسیم درآمد بر میزان حداقل دستمزد سالانه، جهت تفسیرپذیری و تعمیم پذیری الگوهای بدست آمده.

مثال 3: ساخت سلسله مراتب مفهومی از ویژگی های اسمی مانند: تبدیل شهر به استان، تبدیل شغل به گروه های شغلی و ... .

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

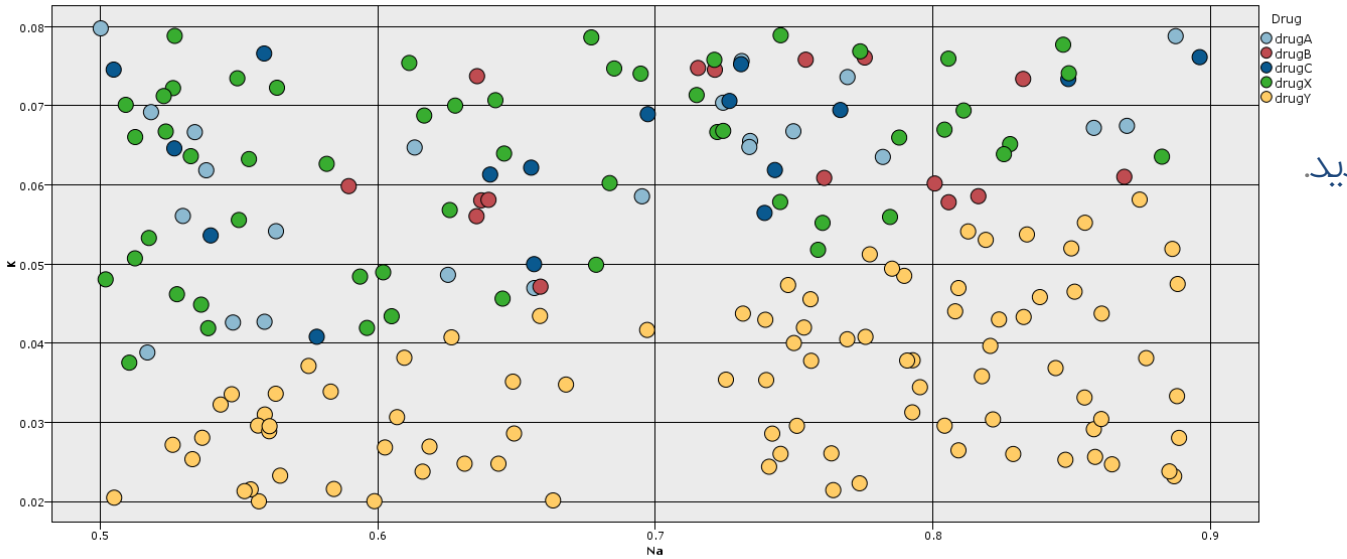
فرآیند داده کاوی

تبدیل داده ها

❑ ساخت ویژگی (Construction)

- **استفاده از روابط آماری**

استفاده از روابط کشف شده در توصیف و کاوش داده ها، مبنای ساخت ویژگی های جدید قرار می گیرد. در این حالت به دو شکل اقدام به ساخت ویژگی های جدید می شود:



- **حالت اول:** استفاده از دانش بدست آمده برای ساخت ویژگی جدید.

بطور مثال، تقسیم سدیم به پتاسیم برای پیش بینی نوع دارو بر اساس رابطه شناسایی شده در نمودار پراکنش

تولید محتوا: زهرا ذوالقدر

daychegroup

daychegroup

dayche.com | گروه دایکه

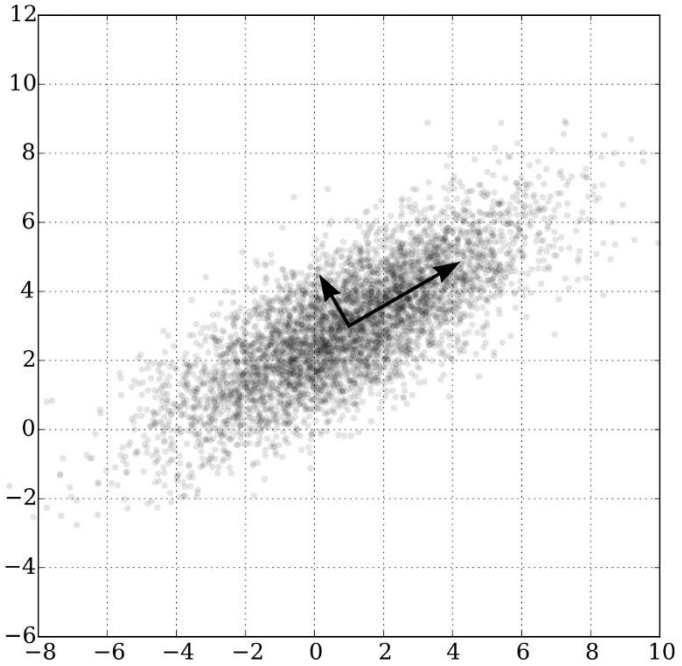
فرآیند داده کاوی

تبدیل داده ها

❑ ساخت ویژگی (Construction)

- **استفاده از روابط آماری**

استفاده از روابط کشف شده در توصیف و کاوش داده ها، مبنای ساخت ویژگی های جدید قرار می گیرد. در این حالت به دو شکل اقدام به ساخت ویژگی های جدید می شود:



- **حالت دوم:** استفاده از الگوریتم های آماری مانند PCA یا SVD.

بطور مثال، ساخت یک مولفه از ترکیب خطی ویژگی ها بر اساس ساختار واریانس کوواریانس بین داده ها.

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

تبدیل داده ها

❑ ساخت ویژگی (Construction)

- کدگذاری مقادیر ویژگی های اسمی

بسیاری از الگوریتم های یادگیری ماشین و روش های آماری نیاز به داده های عددی دارند. بنابراین یکی از اقدامات لازم در آماده سازی داده ها کدگذاری مقادیر ویژگی های اسمی می باشد. برخی از روش های رایج کدگذاری موارد زیر است:

- One-hot encoding
- Ordinal or label encoding
- Ordered label encoding
- Probability ratio encoding
- Rare labels encoding

نوع الگوریتم های مورد استفاده در مدلسازی در انتخاب نوع روش کدگذاری موثر می باشد.

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

تبدیل داده ها

❑ ساخت ویژگی (Construction)

- کدگذاری مقادیر ویژگی های اسمی
  - One-hot encoding
  - Ordinal or label encoding
  - Ordered label encoding
  - Probability ratio encoding
  - Rare labels encoding

Color	Target
Yellow	0
Yellow	1
Blue	1
Yellow	1
Red	1
Yellow	0
Red	1
Red	0
Yellow	1
Blue	0

Color_Y	Color_B	Color_R	Target
1	0	0	0
1	0	0	1
0	1	0	1
1	0	0	1
0	0	1	1
1	0	0	0
0	0	1	1
0	0	1	0
1	0	0	1
0	1	0	0

One Hot Encoding

One-hot encoding into k-1 or k variables?

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

تبدیل داده ها

❑ ساخت ویژگی (Construction)

- کدگذاری مقادیر ویژگی های اسمی
  - One-hot encoding
  - Ordinal or label encoding
  - Ordered label encoding
  - Probability ratio encoding
  - Rare labels encoding

Color	Target
Yellow	0
Yellow	1
Blue	1
Yellow	1
Red	1
Yellow	0
Red	1
Red	0
Yellow	1
Blue	0

Color	Target
1	0
1	1
2	1
1	1
3	1
1	0
3	1
3	0
1	1
2	0

Label Encoding

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

تبدیل داده ها

❑ ساخت ویژگی (Construction)

- کدگذاری مقادیر ویژگی های اسمی
  - One-hot encoding
  - Ordinal or label encoding
  - Ordered label encoding
  - Probability ratio encoding
  - Rare labels encoding

Color	Target
Yellow	0
Yellow	1
Blue	1
Yellow	1
Red	1
Yellow	0
Red	1
Red	0
Yellow	1
Blue	0

Color	Target Mean	Order
Yellow	0.6	2
Blue	0.5	3
Red	0.66	1

Color	Target
2	0
2	1
3	1
2	1
1	1
2	0
1	1
1	0
2	1
3	0

Ordered Label Encoding

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
گروه دایکه | dayche.com

فرآیند داده کاوی

تبدیل داده ها

❑ ساخت ویژگی (Construction)

- کدگذاری مقادیر ویژگی های اسمی
  - One-hot encoding
  - Ordinal or label encoding
  - Ordered label encoding
  - Probability ratio encoding
  - Rare labels encoding

Color	Target
Yellow	0
Yellow	1
Blue	1
Yellow	1
Red	1
Yellow	0
Red	1
Red	0
Yellow	1
Blue	0

Color	P (1)	P (0)	Ratio
Yellow	0.6	0.4	1.5
Blue	0.5	0.5	1
Red	0.66	0.34	1.94

Color	Target
1.5	0
1.5	1
1	1
1.5	1
1.94	1
1.5	0
1.94	1
1.94	0
1.5	1
1	0

Probability Ratio Encoding

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

تبدیل داده ها

❑ ساخت ویژگی (Construction)

- کدگذاری مقادیر ویژگی های اسمی
  - One-hot encoding
  - Ordinal or label encoding
  - Ordered label encoding
  - Probability ratio encoding
  - Rare labels encoding

Color	Target
Yellow	0
Yellow	1
Blue	1
Purple	1
Red	1
Yellow	0
Red	1
Orange	0
Yellow	1
Blue	0
Blue	1
Yellow	1
Green	1
Yellow	0
Red	1
Gray	0
Orange	1

Rare Label Encoding

Color	Target
Yellow	0
Yellow	1
Blue	1
Others	1
Red	1
Yellow	0
Red	1
Others	0
Yellow	1
Blue	0
Blue	1
Yellow	1
Others	1
Yellow	0
Red	1
Others	0
Others	1

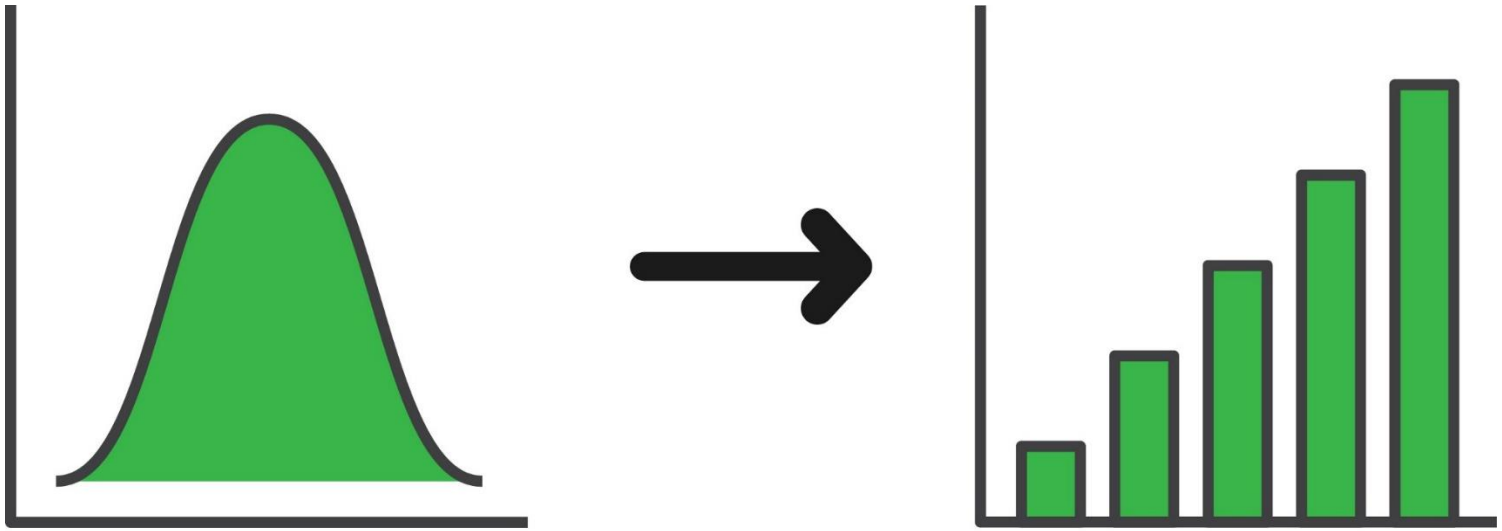
تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

تبدیل داده ها

❑ گسسته سازی (Discretization / Binning)

فرآیند تبدیل داده های پیوسته به مقادیر گسسته در قالب فواصل مجاوری که داده های پیوسته را درون خود قرار داده است.



Discretization Process

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

تبدیل داده ها

❑ گسسته سازی (Discretization / Binning)

- اهمیت گسسته سازی

درک ساده تر از داده های کیفی در صورت تناسب با نوع مسئله؛ بطور مثال سطح درآمد کم، متوسط و زیاد

تفسیر پذیری بیشتر مدل ها از طریق روابط معنادار آماری بین مقادیر گسسته و فیلد هدف

تطابق بیشتر با برخی الگوریتم های یادگیری ماشین

حذف نویز های موجود در ثبت اندازه های کمی

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

تبدیل داده ها

❑ گسسته سازی (Discretization / Binning)

- اهمیت گسسته سازی

درک ساده تر از داده های کیفی در صورت تناسب با نوع مسئله؛ بطور مثال سطح درآمد کم، متوسط و زیاد

تفسیر پذیری بیشتر مدل ها از طریق روابط معنادار آماری بین مقادیر گسسته و فیلد هدف

تطابق بیشتر با برخی الگوریتم های یادگیری ماشین

حذف نویز های موجود در ثبت اندازه های کمی

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

تبدیل داده ها

❑ گسسته سازی (Discretization / Binning)

- اهمیت گسسته سازی

درک ساده تر از داده های کیفی در صورت تناسب با نوع مسئله؛ بطور مثال سطح درآمد کم، متوسط و زیاد

تفسیر پذیری بیشتر مدل ها از طریق روابط معنادار آماری بین مقادیر گسسته و فیلد هدف

تطابق بیشتر با برخی الگوریتم های یادگیری ماشین

حذف نویز های موجود در ثبت اندازه های کمی

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

تبدیل داده ها

❑ گسسته سازی (Discretization / Binning)

- اهمیت گسسته سازی

درک ساده تر از داده های کیفی در صورت تناسب با نوع مسئله؛ بطور مثال سطح درآمد کم، متوسط و زیاد

تفسیر پذیری بیشتر مدل ها از طریق روابط معنادار آماری بین مقادیر گسسته و فیلد هدف

تطابق بیشتر با برخی الگوریتم های یادگیری ماشین

حذف نویز های موجود در ثبت اندازه های کمی

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

تبدیل داده ها

❑ گسسته سازی (Discretization / Binning)

- رویکردهای گسسته سازی

Domain Knowledge

- Custom discretization

Unsupervised Approaches

- Equal-width discretization
- Equal-frequency discretization
- K-means discretization

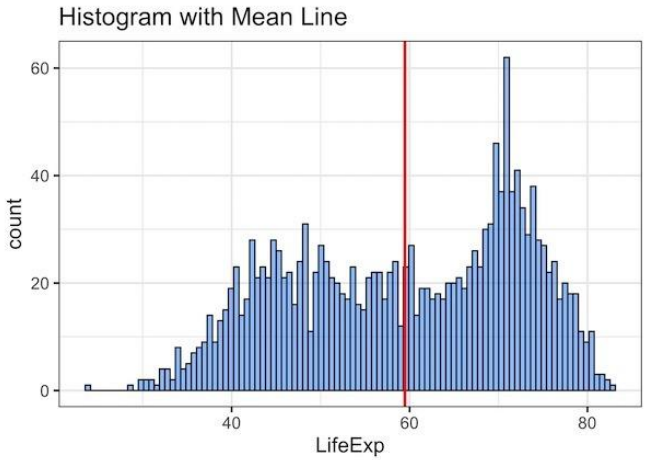
Supervised Approach

- Discretization with correlation (ChiMerge Alg.)
- Discretization with decision trees

در بسیاری از موارد اطلاع از دانش زمینه ای و قواعد کسب و کار مرتبط با داده ها، جهت گسسته سازی ویژگی های کمی، تعیین کننده است.

مثال 1: تعریف BMI < 18.5 : کمبود وزن / 18.5 < BMI < 24.9 : نرمال و ...

مثال 2: تعریف Age < 10 : کودک / 10 < Age < 18 : نوجوان و ...



استفاده از هیستوگرام توزیع داده ها، ابزار رایج و قدرتمندی در انتخاب بازه های مناسب هست.

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

❑ گسسته سازی (Discretization / Binning)

- رویکردهای گسسته سازی

Domain Knowledge

- Custom discretization

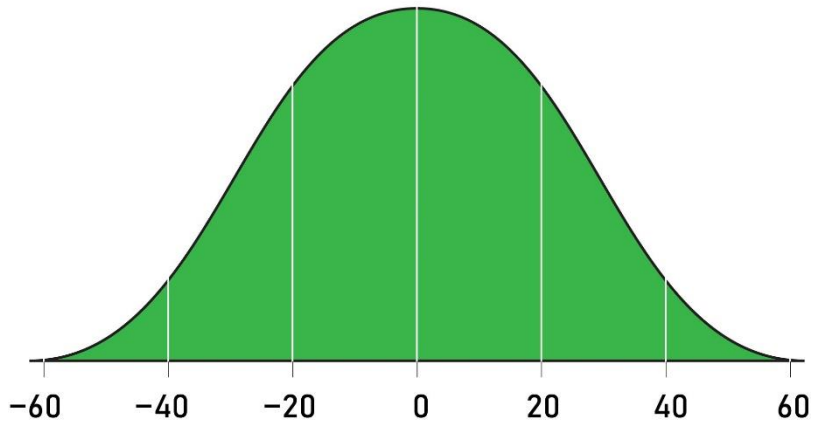
Unsupervised Approaches

- Equal-width discretization
- Equal-frequency discretization
- K-means discretization

Supervised Approach

- Discretization with correlation (ChiMerge Alg.)
- Discretization with decision trees

$$width = \frac{maxvalue - minvalue}{N}$$



Equal width Discretization

تولید محتوا: زهرا ذوالقدر

daychegroup

daychegroup

dayche.com | گروه دایکه

فرآیند داده کاوی

تبدیل داده ها

❑ گسسته سازی (Discretization / Binning)

- رویکردهای گسسته سازی

رویکرد مبتنی بر برابری فراوانی، بر اساس محاسبه چندک ها (Quantiles) بدست می آید. در این روش تا حد امکان توازن کلاسها حفظ می شود و به همین دلیل یکی از ابزارهای رایج در گسسته سازی می باشد.

Domain Knowledge

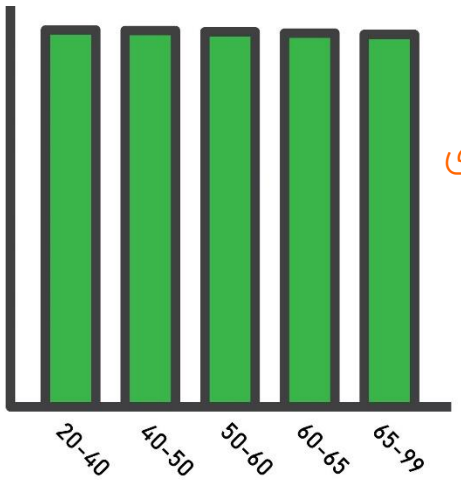
- Custom discretization

Unsupervised Approaches

- Equal-width discretization
- Equal-frequency discretization
- K-means discretization

Supervised Approach

- Discretization with correlation (ChiMerge Alg.
- Discretization with decision trees



Equal-Frequency Discretization

در حالت خاص از این رویکرد، بجای برابری فراوانی، برابری مجموع ارزش ویژگی، مبنای تعیین دسته می شود.

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

تبدیل داده ها

❑ گسسته سازی (Discretization / Binning)

- رویکردهای گسسته سازی

Domain Knowledge

- Custom discretization

Unsupervised Approaches

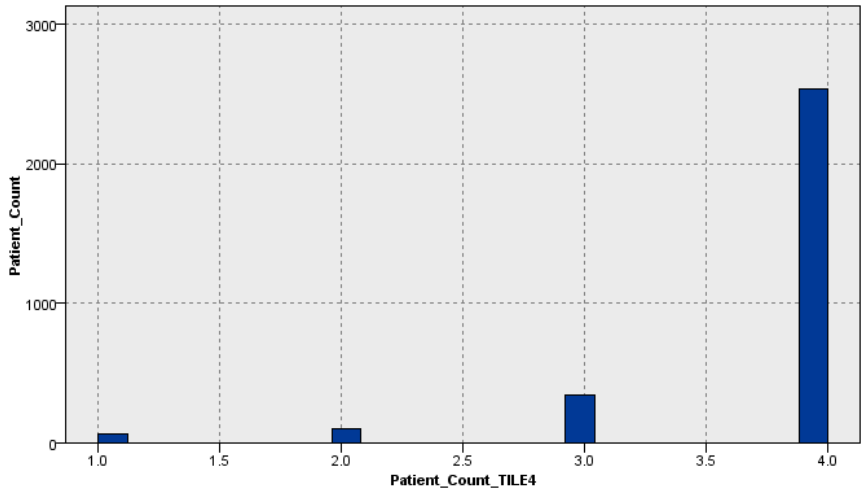
- Equal-width discretization
- Equal-frequency discretization
- K-means discretization

Supervised Approach

- Discretization with correlation (ChiMerge Alg.)
- Discretization with decision trees

مثال 1: گسسته سازی ویژگی "تعداد بیماران هر پزشک" در چهار گروه برابر از نظر فراوانی در هر گروه

Bin	Lower	Upper	Proportion	%	Count
1	>= 1	<= 1		31.08	69
2	> 1	<= 3		19.37	43
3	> 3	<= 9		26.13	58
4	> 9	<= 615		23.42	52



تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

تبدیل داده ها

❑ گسسته سازی (Discretization / Binning)

- رویکردهای گسسته سازی

Domain Knowledge

- Custom discretization

Unsupervised Approaches

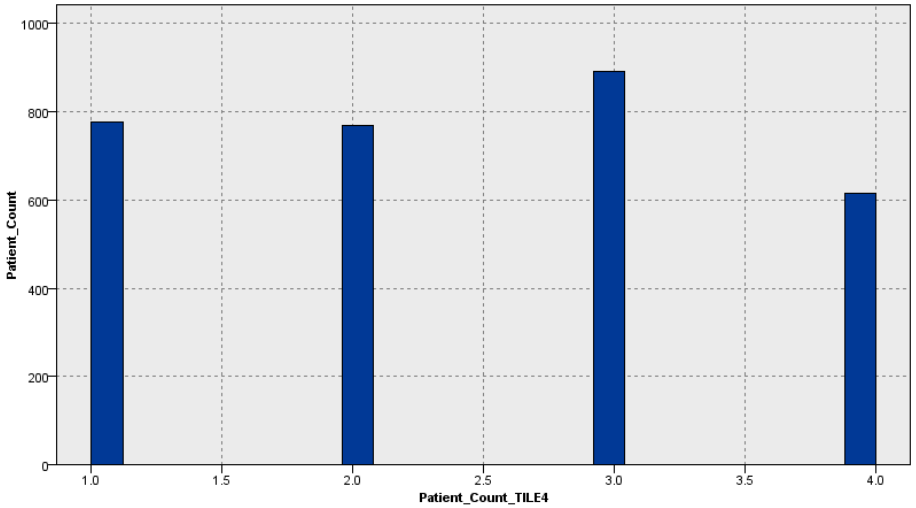
- Equal-width discretization
- Equal-frequency discretization
- K-means discretization

Supervised Approach

- Discretization with correlation (ChiMerge Alg.)
- Discretization with decision trees

مثال 2: گسسته سازی ویژگی "تعداد بیماران هر پزشک" در چهار گروه بر اساس برابری مجموع بیماران در هر گروه

Bin	Lower	Upper	Proportion	%	Count
1	>= 1	< 18		85.59	190
2	>= 18	< 80		10.81	24
3	>= 80	< 615		3.15	7
4	>= 615	<= 615		0.45	1



تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

تبدیل داده ها

❑ گسسته سازی (Discretization / Binning)

- رویکردهای گسسته سازی

Domain Knowledge

- Custom discretization

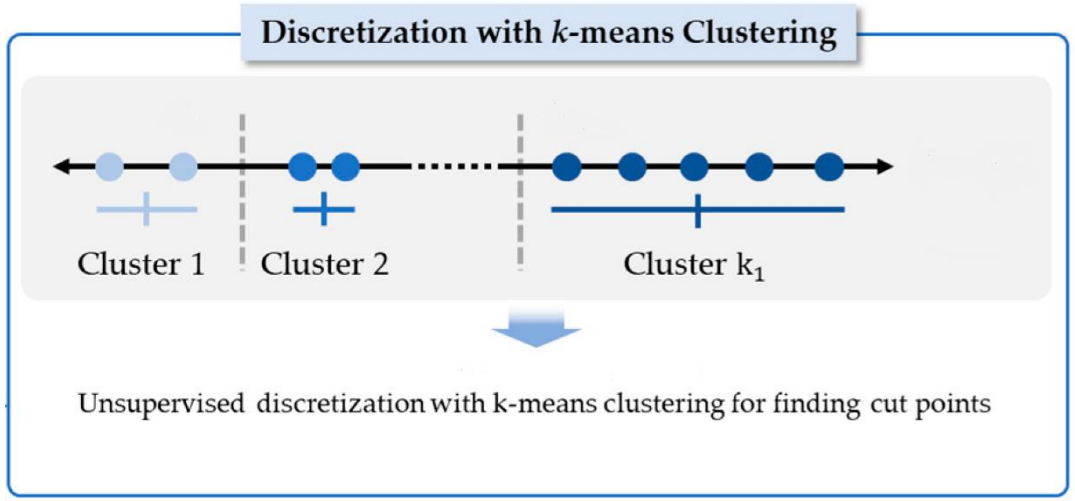
Unsupervised Approaches

- Equal-width discretization
- Equal-frequency discretization
- K-means discretization

Supervised Approach

- Discretization with correlation (ChiMerge Alg.)
- Discretization with decision trees

استفاده از الگوریتم K-Means جهت خوشه بندی تک متغیره ویژگی کمی، می تواند یکی از رویکردهای گسسته سازی به تعداد k دسته باشد.



فرآیند داده کاوی

تبدیل داده ها

❑ گسسته سازی (Discretization / Binning)

- رویکردهای گسسته سازی

Domain Knowledge

- Custom discretization

Unsupervised Approaches

- Equal-width discretization
- Equal-frequency discretization
- K-means discretization

Supervised Approach

- Discretization with correlation (ChiMerge Alg.)
- Discretization with decision trees

این رویکرد بصورت هدایت شده بر اساس برچسب کلاس هدف، با ادغام فواصل مجاوری که میزان χ^2 کوچکی دارند ادامه پیدا میکند تا در نهایت دسته هایی را با توزیع آماری متمایز از برچسب کلاس هدف ایجاد نماید.

Sample	F	K	Intervals
1	1	1	{0,2}
2	3	2	{2,5}
3	7	1	{5,7.5}
4	8	1	{7.5,8.5}
5	9	1	{8.5,10}
6	11	2	{10,17}
7	23	2	{17,30}
8	37	1	{30,38}
9	39	2	{38,42}
10	45	1	{42,45.5}
11	46	1	{45.5,52}
12	59	1	{52,60}

ChiMerge Discretization Example

•Sort and order the attributes that you want to group (in this example attribute F).

•Start with having every unique value in the attribute be in its own interval.

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
گروه دایکه | dayche.com

فرآیند داده کاوی

تبدیل داده ها

❑ گسسته سازی (Discretization / Binning)

- رویکردهای گسسته سازی

Domain Knowledge

- Custom discretization

Unsupervised Approaches

- Equal-width discretization
- Equal-frequency discretization
- K-means discretization

Supervised Approach

- Discretization with correlation (ChiMerge Alg.)
- Discretization with decision trees

Sample	F	K
1	1	1
2	3	2
3	7	1
4	8	1
5	9	1
6	11	2
7	23	2
8	37	1
9	39	2
10	45	1
11	46	1
12	59	1

ChiMerge Discretization Example

•Begin calculating the Chi Square test on every interval

Sample	K=1	K=2	
2	0	1	1
3	1	0	1
total	1	1	2

Sample	K=1	K=2	
3	1	0	1
4	1	0	1
total	2	0	2

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

❑ گسسته سازی (Discretization / Binning)

- رویکردهای گسسته سازی

Domain Knowledge

- Custom discretization

Unsupervised Approaches

- Equal-width discretization
- Equal-frequency discretization
- K-means discretization

Supervised Approach

- Discretization with correlation (ChiMerge Alg.)
- Discretization with decision trees

ChiMerge Discretization Example

Sample	K=1	K=2	
2	0	1	1
3	1	0	1
total	1	1	2

$E_{11} = (1/4)*2 = .5$
$E_{12} = (1/4)*2 = .5$
$E_{21} = (1/4)*2 = .5$
$E_{22} = (1/4)*2 = .5$

$X^2 = (0-.5)^2/.5 + (1-.5)^2/.5 + (0-.5)^2/.5 + (1-.5)^2/.5 = \mathbf{2}$

Sample	K=1	K=2	
3	1	0	1
4	1	0	1
total	2	0	2

$E_{11} = (1/2)*2 = 1$
$E_{12} = (0/2)*2 = 0$
$E_{21} = (1/2)*2 = 1$
$E_{22} = (0/2)*2 = 0$

$X^2 = (1-1)^2/1+(0-0)^2/0+ (1-1)^2/1+(0-0)^2/0 = \mathbf{0}$

Threshold .1 with df=1 from Chi square distribution chart merge if $X^2 < 2.7024$

فرآیند داده کاوی

تبدیل داده ها

❑ گسسته سازی (Discretization / Binning)

- رویکردهای گسسته سازی

Domain Knowledge

- Custom discretization

Unsupervised Approaches

- Equal-width discretization
- Equal-frequency discretization
- K-means discretization

Supervised Approach

- Discretization with correlation (ChiMerge Alg.)
- Discretization with decision trees

ChiMerge Discretization Example

Sample	F	K	Intervals	Chi²
1	1	1	{0,2}	2
2	3	2	{2,5}	2
3	7	1	{5,7.5}	0
4	8	1	{7.5,8.5}	0
5	9	1	{8.5,10}	2
6	11	2	{10,17}	0
7	23	2	{17,30}	2
8	37	1	{30,38}	2
9	39	2	{38,42}	2
10	45	1	{42,45.5}	0
11	46	1	{45.5,52}	0
12	59	1	{52,60}	

•Calculate all the Chi Square value for all intervals

•Merge the intervals with the smallest Chi values

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
گروه دایکه | dayche.com

فرآیند داده کاوی

تبدیل داده ها

❑ گسسته سازی (Discretization / Binning)

- رویکردهای گسسته سازی

Domain Knowledge

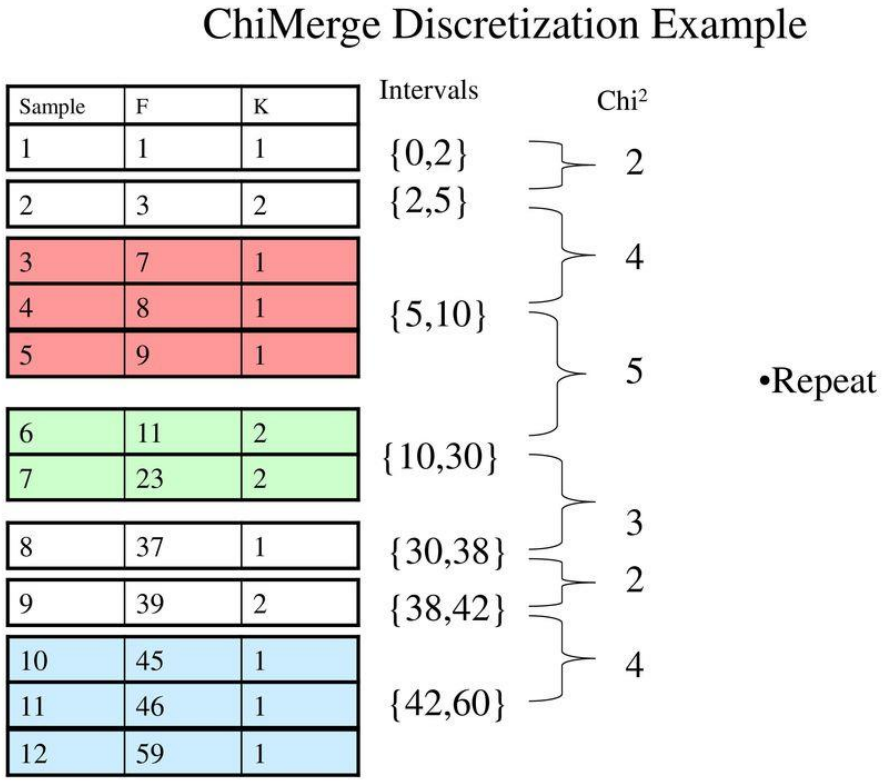
- Custom discretization

Unsupervised Approaches

- Equal-width discretization
- Equal-frequency discretization
- K-means discretization

Supervised Approach

- Discretization with correlation (ChiMerge Alg.)
- Discretization with decision trees



تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

تبدیل داده ها

❑ گسسته سازی (Discretization / Binning)

- رویکردهای گسسته سازی

Domain Knowledge

- Custom discretization

Unsupervised Approaches

- Equal-width discretization
- Equal-frequency discretization
- K-means discretization

Supervised Approach

- Discretization with correlation (ChiMerge Alg.)
- Discretization with decision trees

ChiMerge Discretization Example

Sample	F	K	Intervals	Chi²
1	1	1	{0,5}	
2	3	2		1.875
3	7	1	{5,10}	
4	8	1		5
5	9	1		
6	11	2	{10,30}	
7	23	2		1.33
8	37	1	{30,42}	
9	39	2		1.875
10	45	1	{42,60}	
11	46	1		
12	59	1		

•Again

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

تبدیل داده ها

❑ گسسته سازی (Discretization / Binning)

- رویکردهای گسسته سازی

Domain Knowledge

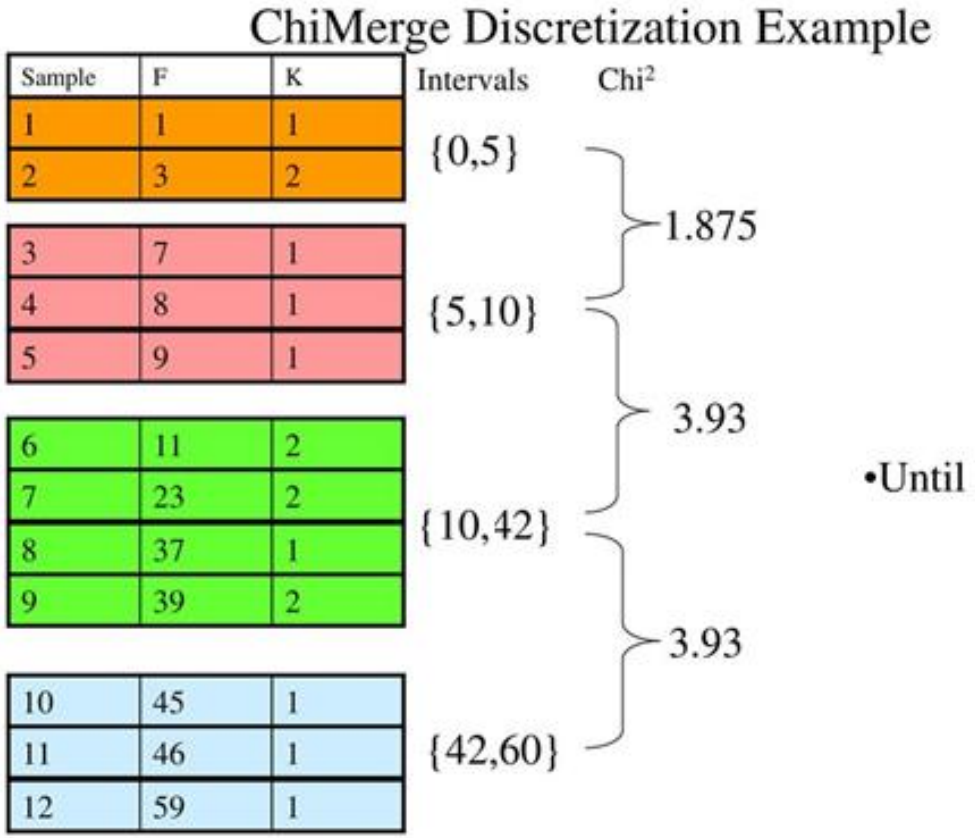
- Custom discretization

Unsupervised Approaches

- Equal-width discretization
- Equal-frequency discretization
- K-means discretization

Supervised Approach

- Discretization with correlation (ChiMerge Alg.)
- Discretization with decision trees



تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

تبدیل داده ها

❑ گسسته سازی (Discretization / Binning)

- رویکردهای گسسته سازی

Domain Knowledge

- Custom discretization

Unsupervised Approaches

- Equal-width discretization
- Equal-frequency discretization
- K-means discretization

Supervised Approach

- Discretization with correlation (ChiMerge Alg.)
- Discretization with decision trees

ChiMerge Discretization Example

Sample	F	K	Intervals	Chi²
1	1	1		
2	3	2	{0,10}	
3	7	1		
4	8	1		
5	9	1		2.72
6	11	2		
7	23	2		
8	37	1	{10,42}	
9	39	2		
10	45	1		3.93
11	46	1	{42,60}	
12	59	1		

•There are no more intervals that can satisfy the Chi Square test.

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

تبدیل داده ها

❑ گسسته سازی (Discretization / Binning)

- رویکردهای گسسته سازی

Domain Knowledge

- Custom discretization

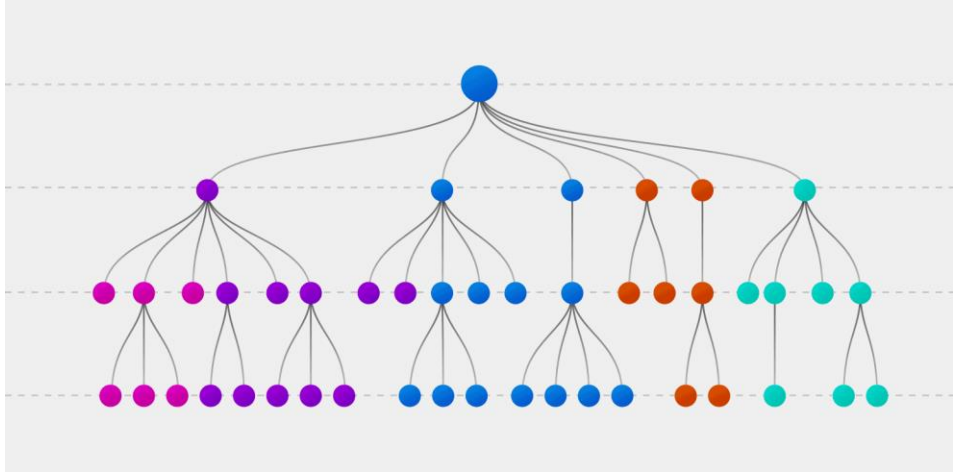
Unsupervised Approaches

- Equal-width discretization
- Equal-frequency discretization
- K-means discretization

Supervised Approach

- Discretization with correlation (ChiMerge Alg.)
- Discretization with decision trees

این رویکرد بصورت هدایت شده بر اساس ساخت درخت تصمیم، با عمق 2 الی 4 تنها بر اساس ویژگی مورد نظر و کلاس هدف، اقدام به دسته بندی مقادیر عددی ویژگی می نماید.



تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

تبدیل داده ها

❑ تجمیع و فشرده سازی (Aggregation)

تجمیع داده ها با ترکیب چند رکورد داده و خلاصه سازی آنها با اهداف زیر انجام می شود:

- **تغییر مقیاس و زاویه نگاه به داده ها**

بطور مثال با تجمیع داده های مشتریان بانک به تفکیک هر شعبه می توان زاویه نگاه تحلیل را از سطح مشتری به سطح شعب تغییر داد.

- **افزایش ثبات و پایداری در الگوها**

بطور مثال با تجمیع داده های بدست آمده از تردد هر خیابان به سطح هر محله می توان به الگوهای پایدارتری از تردد شهری دسترسی پیدا کرد.

- **کاهش داده ها**

بطور مثال تجمیع داده های تراکنشی بانک و خلاصه سازی آنها برای هر مشتری می تواند حجم داده های اولیه را به مقدار زیادی کاهش دهد.

فرآیند داده کاوی

تبدیل داده ها

❑ تجمیع و فشرده سازی (Aggregation)

برای عملیات تجمیع داده ها دو مفهوم اساسی را باید در نظر داشت:

1) انتخاب فیلد کلیدی (Key Field)

یک مجموعه داده شامل انواع مختلفی از فیلدها می باشد که عموما برخی از آنها ماهیت "کد شناسه" دارند؛ بطور مثال، کد مشتری، کد سفارش، کد فروشگاه و ...که هر کدام از آنها نشان دهنده سطح ریزدانگی در داده ها می باشد.

نکته 1: با داشتن مجموعه داده در پایین ترین سطح ریزدانگی، می توان با عملیات تجمیع، داده ها را به سطح ریزدانگی بالاتر تغییر داد، اما بصورت برعکس این امکان وجود نخواهد داشت.

نکته 2: فیلدهای کلیدی عموما از نوع اسمی هستند، اما الزامی برای آن وجود ندارد.

نکته 3: در برخی از تحلیل ها، امکان ترکیب چند فیلد کلیدی برای ایجاد مجموعه داده های مناسب وجود دارد. بطور مثال، ترکیب کد مشتری و تاریخ، منجر به ریزدانگی داده ها در سطح تراکنش های روزانه هر مشتری می شود

فرآیند داده کاوی

تبدیل داده ها

❑ تجمیع و فشرده سازی (Aggregation)

برای عملیات تجمیع داده ها دو مفهوم اساسی را باید در نظر داشت:

2) انتخاب تابع خلاصه سازی

توابع خلاصه سازی، همان شاخص های خلاصه سازی شناخته شده در آمار هستند:

مجموع، میانگین، مد، تعداد رکورد، حداقل و حداکثر،

میانه، چارک های اول و سوم.انحراف معیار، واریانس و ...

Region	Category	Metrics	Revenue	Units Sold
Central	Books		$251,279	15,715
	Electronics		$5,307,467	15,278
	Movies		$241,405	15,500
	Music		$1,112,782	75,809
Mid-Atlantic	Books		$210,320	13,023
	Electronics		$22,565,830	65,424
	Movies		$208,237	13,299
	Music		$194,436	13,063
Northeast	Books		$2,048,826	127,017
	Electronics		$8,563,073	24,974
	Movies		$395,246	25,355
	Music		$368,269	24,817

با انتخاب چند تابع خلاصه سازی به ازای مقادیر یکتای فیلدهای کلیدی ، یک بردار جدید از داده های فشرده شده ایجاد می شود. این بردارها نشان دهنده رفتار در سطح ریزدانگی مورد نظر می باشد.

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

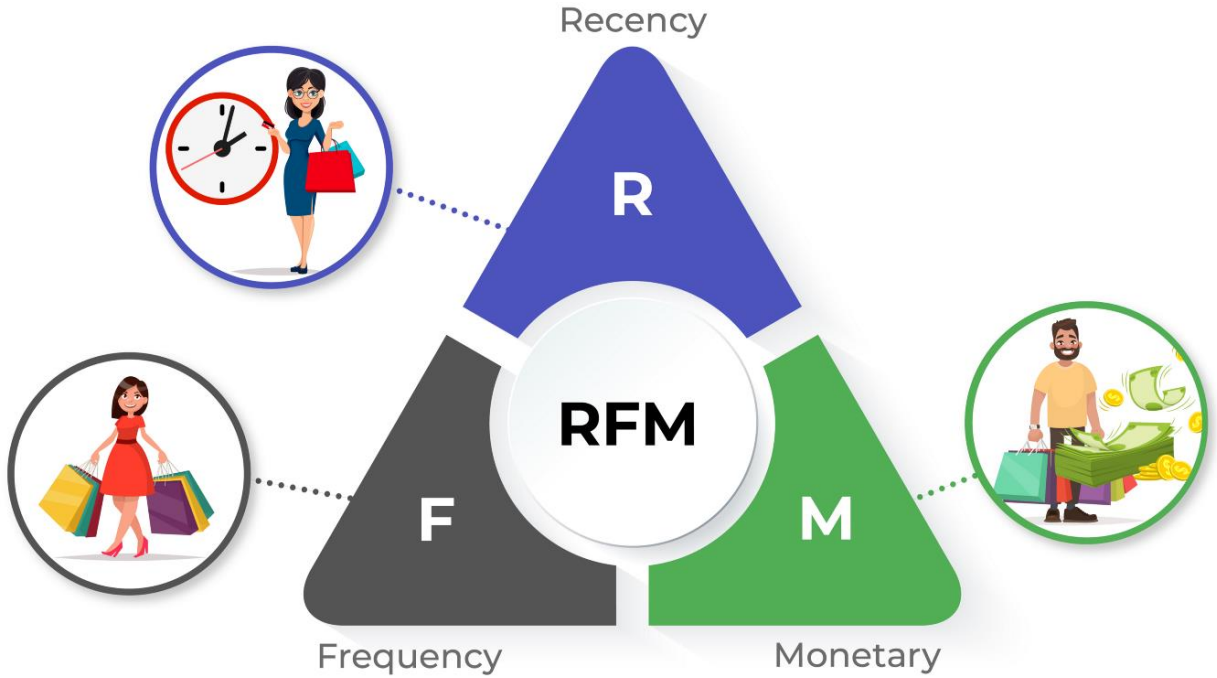
تبدیل داده ها

❑ تجمیع و فشرده سازی (Aggregation)

تحلیل RFM (Recency, Frequency, Monetary)

یک روش تحلیلی در بازاریابی محسوب می شود که با کمی کردن ارزش مشتری به رتبه بندی مشتریان و تعیین بازار هدف کمک می کند.

- شاخص Recency: فاصله از آخرین خرید مشتری
- شاخص Frequency: تعداد خرید مشتری
- شاخص Monetary: جمع مبلغ خرید مشتری



با تجمیع و فشرده سازی داده های تراکنشی سفارشات خرید، می توان شاخص های RFM را استخراج نمود و با گسسته سازی آن برای هر مشتری بردار رفتاری آن را ایجاد نمود.

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

تبدیل داده ها

❑ تجمیع و فشرده سازی (Aggregation)

تحلیل RFM (Recency, Frequency, Monetary)

داده های تراکنشی خام

CID	Date	Item	Price
1	99/01/07	z	100
2	99/01/07	a	50
3	99/02/15	a	50
1	99/02/25	d	200
2	99/03/05	s	20
3	99/03/17	z	100
2	99/03/22	d	200
2	99/03/28	s	20
2	99/04/07	b	150
3	99/04/07	s	20
.	.	.	.
.	.	.	.

داده های تجمیعی

CID	Recency	Frequency	Monetary
1	10	28	1850
2	3	35	750
3	35	3	170
4	12	18	400
5	5	1	20
6	45	1	1000
7	120	24	1600
8	17	22	580
9	155	18	1360
10	24	5	120
.	.	.	.
.	.	.	.

بردار داده های رفتاری

CID	Recency	Frequency	Monetary
1	4	5	5
2	5	5	3
3	3	1	2
4	4	3	2
5	5	1	1
6	2	1	4
7	1	4	5
8	4	4	3
9	1	3	4
10	3	1	1
.	.	.	.
.	.	.	.

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

تبدیل داده ها

❑ هموارسازی (Smoothing)

روش های هموارسازی به منظور حذف یا کاهش اثرات تصادفی از جریان داده هاست.

در واقع بخشی از پراکندگی و تغییرات در داده ها ناشی از اثرات غیرتصادفی و کنترل پذیر (متاثر از ویژگی های دیگر) صورت می گیرد و بخشی نیز بر اساس اثرات تصادفی و غیر قابل کنترل منجر به ایجاد نویز می شود.

دو رویکرد عمده در هموار سازی:

- **رویکرد محلی (Local):** هموارسازی بر اساس داده های موجود در همسایگی – داده های اثرگذار در هموارسازی : تعداد محدود
- **رویکرد سراسری (Global):** هموارسازی بر اساس کل داده ها و الگوهای موجود – داده های اثرگذار در هموارسازی : همه داده ها

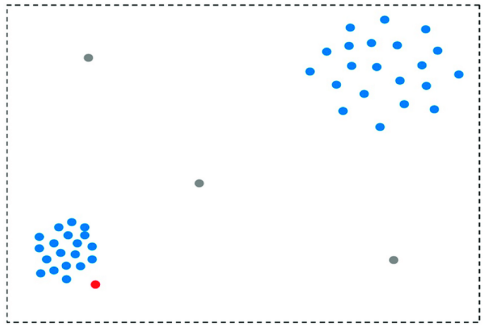
فرآیند داده کاوی

تبدیل داده ها

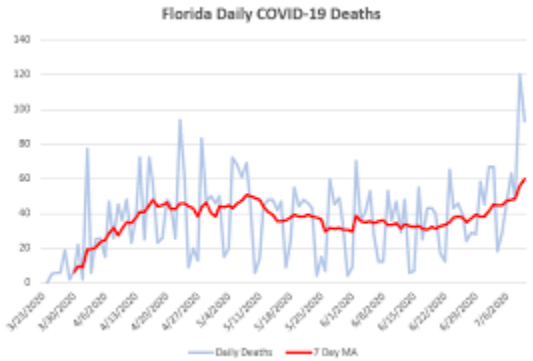
❑ هموارسازی (Smoothing)

روش های رایج در هموارسازی بر اساس نوع داده ها نیز می تواند در سه نوع زیر دسته بندی شود:

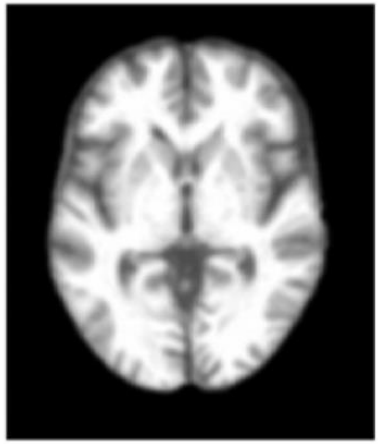
هموار سازی داده های سطری
Record Data



هموار سازی داده های زمانی
Time Series Data



هموار سازی داده های مکانی
Spatial Data



تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

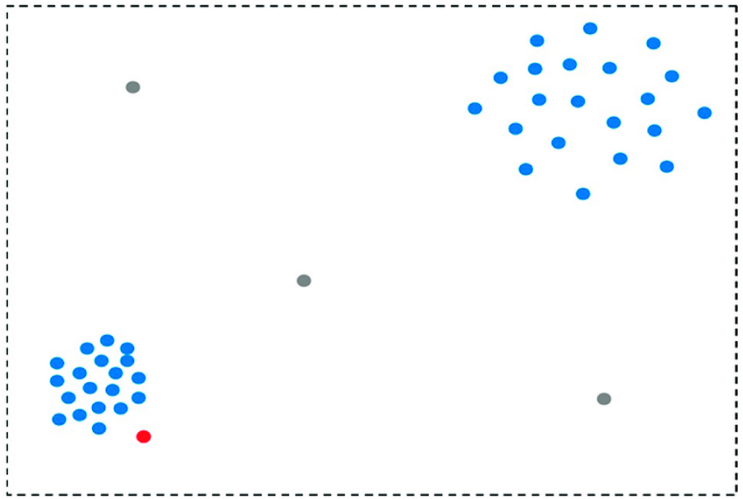
فرآیند داده کاوی

تبدیل داده ها

❑ هموارسازی (Smoothing)

- **هموارسازی داده های سطری**

هموارسازی داده های سطری شامل برخورد با ناسازگاری ها، داده های پرت و کاهش نویز های موجود در داده ها می باشد:



- روش های شناسایی و برخورد با داده های پرت (Outlier Detection)
- روش های گسسته سازی (Binning)

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

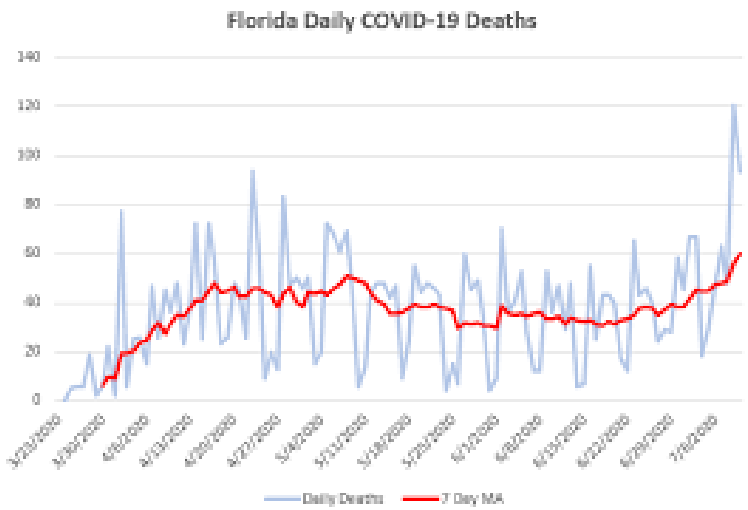
فرآیند داده کاوی

تبدیل داده ها

❑ هموارسازی (Smoothing)

- **هموارسازی داده های زمانی**

هموارسازی داده های زمانی جهت کاهش اثر نویز و داده های پرت در مطالعه روندهای زمانی بکار می رود تا تمرکز تحلیل بر الگوی کلی روند باشد.



- **هموارسازی نمایی ساده** (Simple Exponential Smoothing)

- **هموارسازی میانگین متحرک ساده** (Simple Moving Average Smoothing)

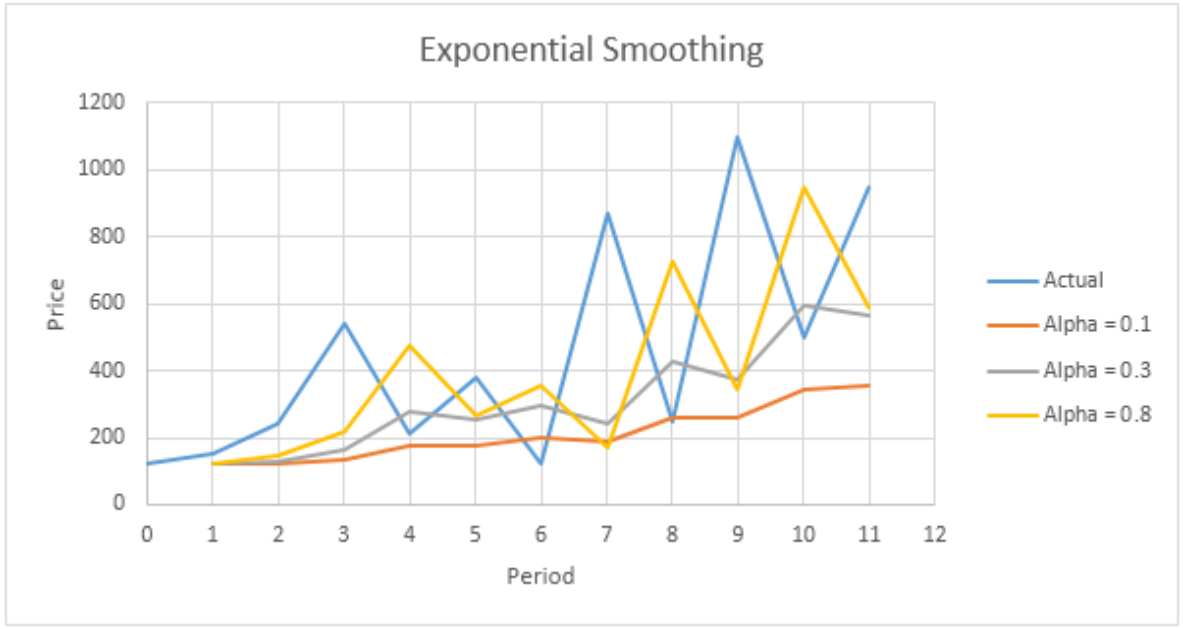
تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

❑ هموارسازی (Smoothing)

- هموارسازی داده های زمانی
  - نمایی ساده (Simple Exponential)

اگر بردار اعداد اولیه را X در نظر بگیریم، بردار مقادیر هموار شده S بر اساس ضریب هموارسازی α با رابطه زیر محاسبه می شود:

$$s_0 = x_0$$
$$s_t = \alpha x_t + (1 - \alpha) s_{t-1}, \quad t > 0$$
$$0 < \alpha < 1$$



تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

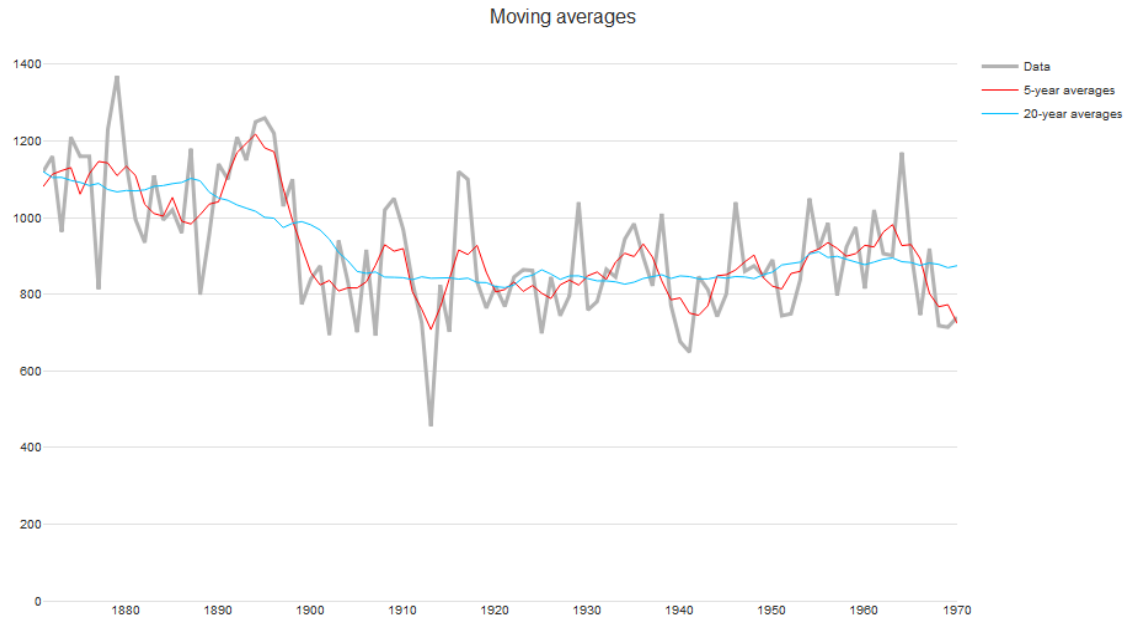
فرآیند داده کاوی

تبدیل داده ها

❑ هموارسازی (Smoothing)

- هموارسازی داده های زمانی
  - میانگین متحرک ساده (Simple Moving Average)

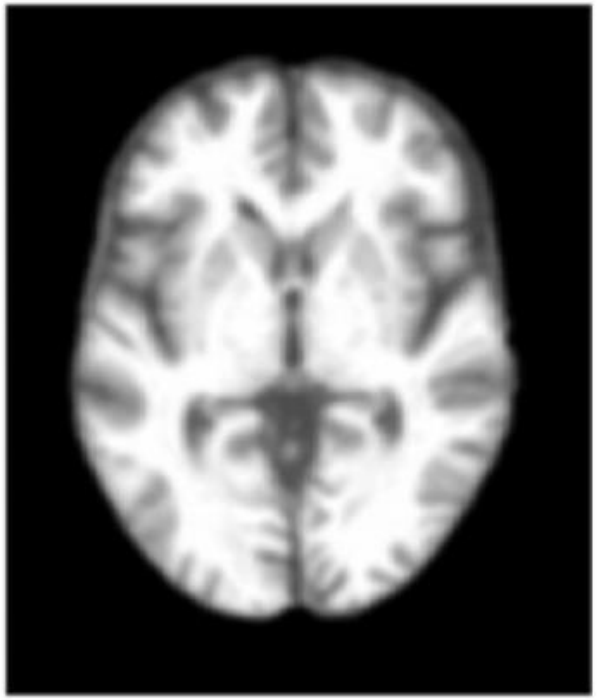
اگر بردار اعداد اولیه را X در نظر بگیریم، بردار مقادیر هموار شده S بر اساس پارامتر K با رابطه زیر محاسبه می شود:



$$s_k = \frac{x_{n-k+1} + x_{n-k+2} \ldots + x_n}{k}$$

$$s_k = \frac{1}{k} \sum_{i=n-k+1}^{n} x_i$$

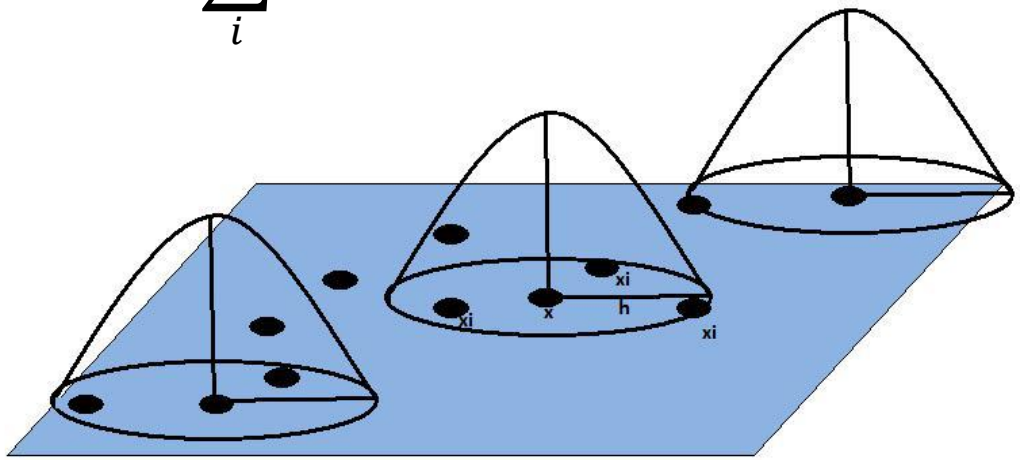
❑ هموارسازی (Smoothing)



- هموارسازی داده های مکانی
  - هموارسازی کرنل (Kernel Smoothing)

در این روش با تعیین تابع کرنل K_h (گوسی / کوادراتیک / ...) هموارسازی مقادیر اولیه بر اساس برآیندی از فاصله نقاط همسایه با رابطه زیر بدست می آید:

$$\widehat{\lambda_h}(x) = \sum_i K_h(x - x_i)$$



تولید محتوا: زهرا ذوالقدر

daychegroup

daychegroup

dayche.com | گروه دایکه

Dimension Reduction

and Sample Selection

کاهش ابعاد و انتخاب نمونه

dayche.com . گروه دایکه

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

کاهش ابعاد و انتخاب نمونه

❑ چرا سراغ روش های کاهش ابعاد می رویم؟

کاهش ابعاد در واقع یک تبدیل و نگاشت داده از فضای با ابعاد بالا به فضای با ابعاد پایین محسوب می شود.

این روش با حذف داده های نامرتبط (Irrelevant Data) و ازبین بردن افزونگی داده ها (Data Redundancy) اهداف زیر را محقق می سازد.

- کاهش حجم داده ها
- افزایش کارایی الگوریتم ها
- افزایش تفسیرپذیری مدل ها

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

کاهش ابعاد و انتخاب نمونه

❑ انتخاب ویژگی (Feature Selection)

در این رویکرد، با انتخاب زیرمجموعه مناسبی از ویژگی های مرتبط و حذف مابقی ویژگی ها، کاهش داده صورت می گیرد.

از آنجا که در این روش ماهیت ویژگی ها تغییر نمی کند، در اغلب مسائلی که نیاز به تفسیر نتایج وجود دارد این روش استفاده می شود.

گام اول در انتخاب ویژگی بر اساس دانش زمینه ای و نظر افراد خبره، حذف دادهای نامرتبط با مسئله هست.

گام دوم در انتخاب ویژگی بر اساس گزارش کیفی داده ها، بررسی و اقدام در شرایط زیر است:

- حذف ویژگی های دارای مقادیر گمشده بسیار زیاد
- حذف ویژگی های دارای واریانس بسیار کم
- حذف ویژگی های کیفی دارای کلاس های زیاد

گام سوم در انتخاب ویژگی بر اساس تحلیل همبستگی بین ویژگی ها، بررسی وجود همخطی (Collinearity) بین داده ها است تا نسبت به حذف ویژگی یا انتخاب استراتژی مناسب (تبدیل/استخراج ویژگی/...) اقدام لازم صورت گیرد.

تولید محتوا: زهرا ذوالقدر

فرآیند داده کاوی

کاهش ابعاد و انتخاب نمونه

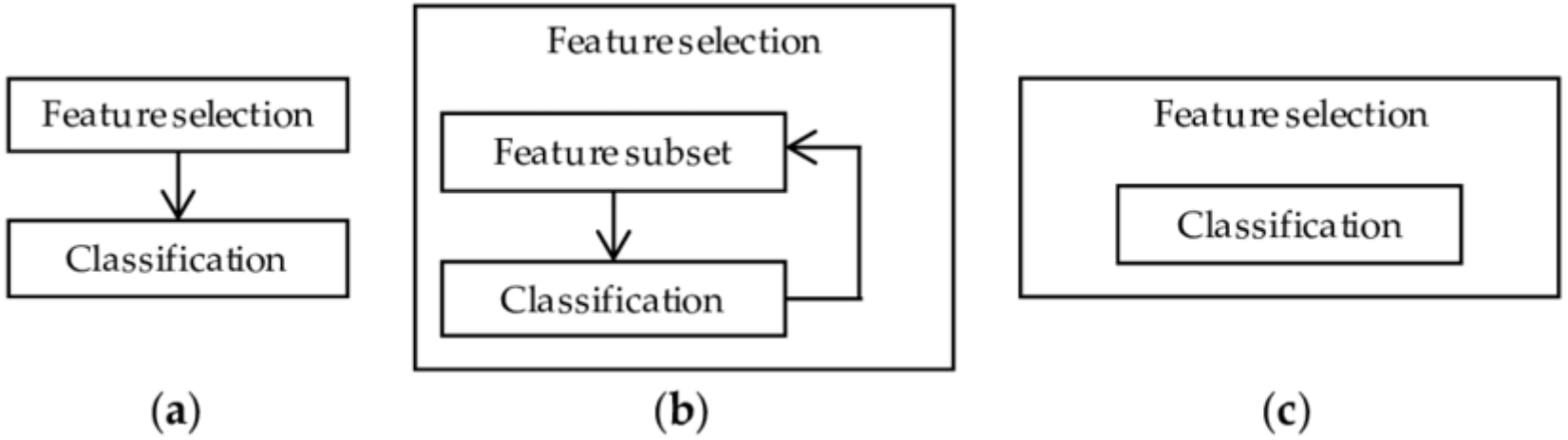
❑ انتخاب ویژگی (Feature Selection)

گام چهارم در انتخاب ویژگی بر اساس رابطه با فیلد هدف (رویکرد با نظارت)، استفاده از رویکردهای زیر می باشد:

روش فیلتر
Filter Method

روش بسته بند
Wrapper Method

روش توکار
Embedded Method



تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

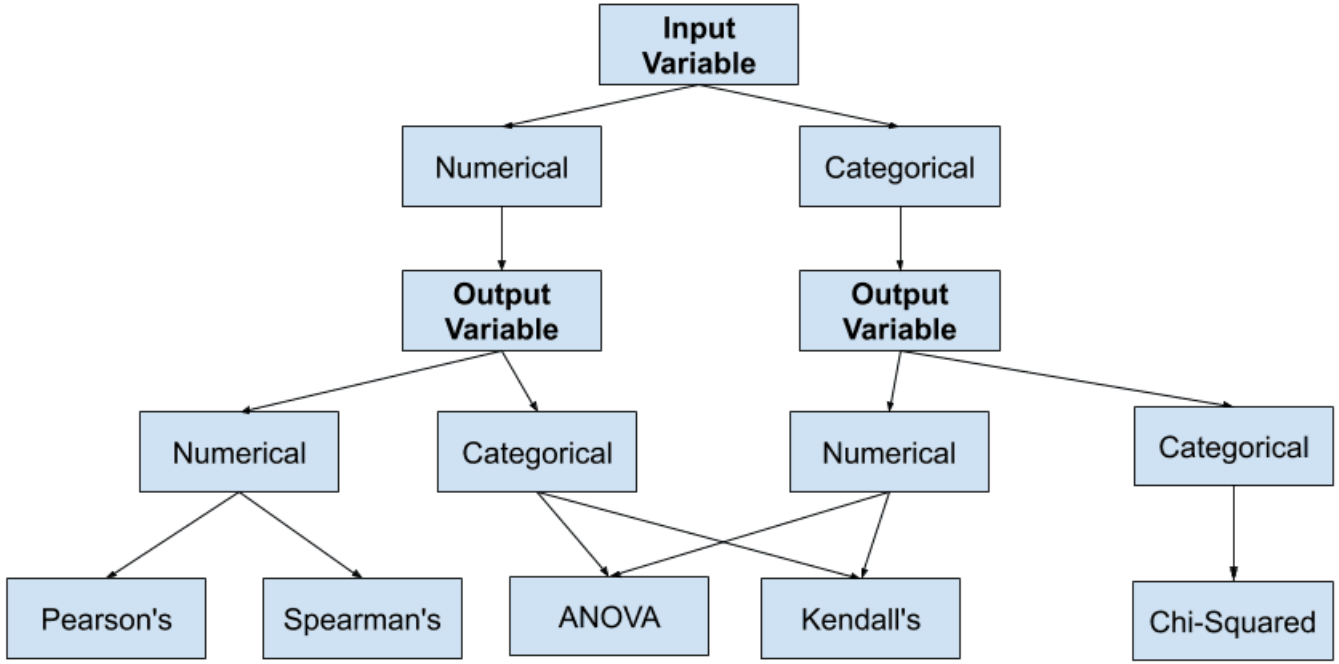
کاهش ابعاد و انتخاب نمونه

❑ انتخاب ویژگی (Feature Selection)

- روش فیلتر (Filter Method)

مبتنی بر روش های آماری (Statistical Method)

رویکرد اول در این روش، با ارتباط سنجی آماری بین هر کدام از ویژگی ها با فیلد هدف (مستقل از مدل)، به انتخاب زیرمجموعه مناسبی از ویژگی های مرتبط می پردازد.



نقاط قوت

- سادگی محاسباتی و سریع

نقاط ضعف

- مشکل افزونگی داده را حل نمی کند
- اثرمتقابل بین ویژگی ها بررسی نمی شود

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

کاهش ابعاد و انتخاب نمونه

- **انتخاب ویژگی (Feature Selection)**
  - روش فیلتر (Filter Method)

مبتنی بر روش های آماری (Statistical Method)

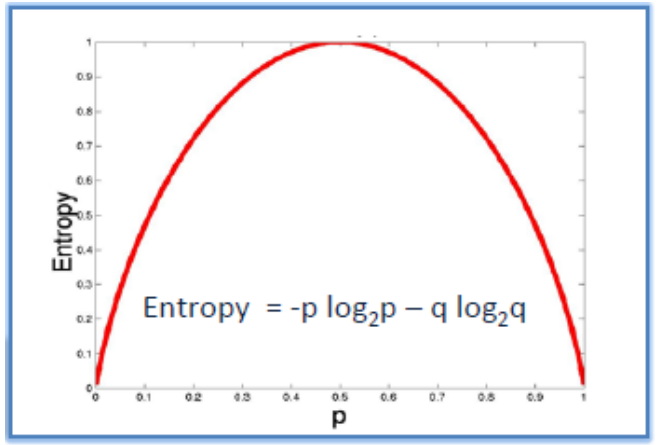
دیگر شاخص های رایج در بررسی ارتباط:

  - شاخص آنتروپی (Entropy)

$$E(T) = \sum_{i=1}^{c} -p_i \log_2 p_i$$

Play Golf	
Yes	No
9	5

Entropy(PlayGolf) = Entropy (5,9)
= Entropy (0.36, 0.64)
= - (0.36 $\log_2$ 0.36) - (0.64 $\log_2$ 0.64)
= 0.94



Entropy = -0.5 $\log_2$0.5 – 0.5 $\log_2$0.5 = 1

  - شاخص بهره اطلاعاتی (Information Gain)

$$Gain(T, X) = Entropy(T) - Entropy(T, X)$$

G(PlayGolf, Outlook) = **E**(PlayGolf) – **E**(PlayGolf, Outlook)
= 0.940 – 0.693 = 0.247

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

کاهش ابعاد و انتخاب نمونه

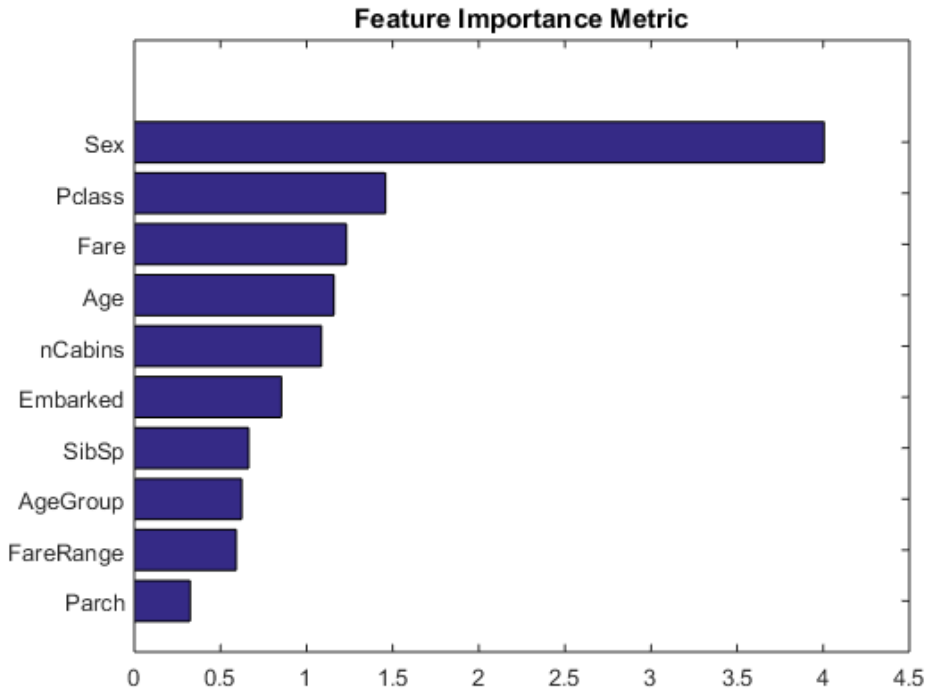
❑ انتخاب ویژگی (Feature Selection)

- روش فیلتر (Filter Method)

مبتنی بر اهمیت ویژگی: رویکرد دیگر در روش فیلتر، محاسبه و رتبه بندی اهمیت هر ویژگی برای هر مسئله می باشد. اندازه گیری اهمیت هر ویژگی با روش تحلیل حساسیت انجام می شود:

- تحلیل حساسیت

در مرحله اول با ساخت مدل روی همه ویژگی ها **میزان خطا** محاسبه می شود و سپس در هر مرحله یکی از ویژگی ها حذف شده و با مقایسه میزان خطای جدید با مقدار اولیه، میزان اهمیت آن ویژگی برآورد می شود.



نکته مهم: در روش مبتنی بر اهمیت ویژگی، به علت بررسی اثر متقابل و همبستگی بین ویژگی ها در ساخت مدل ها، اثر افزونگی داده ها توسط مدلها کنترل می شود.

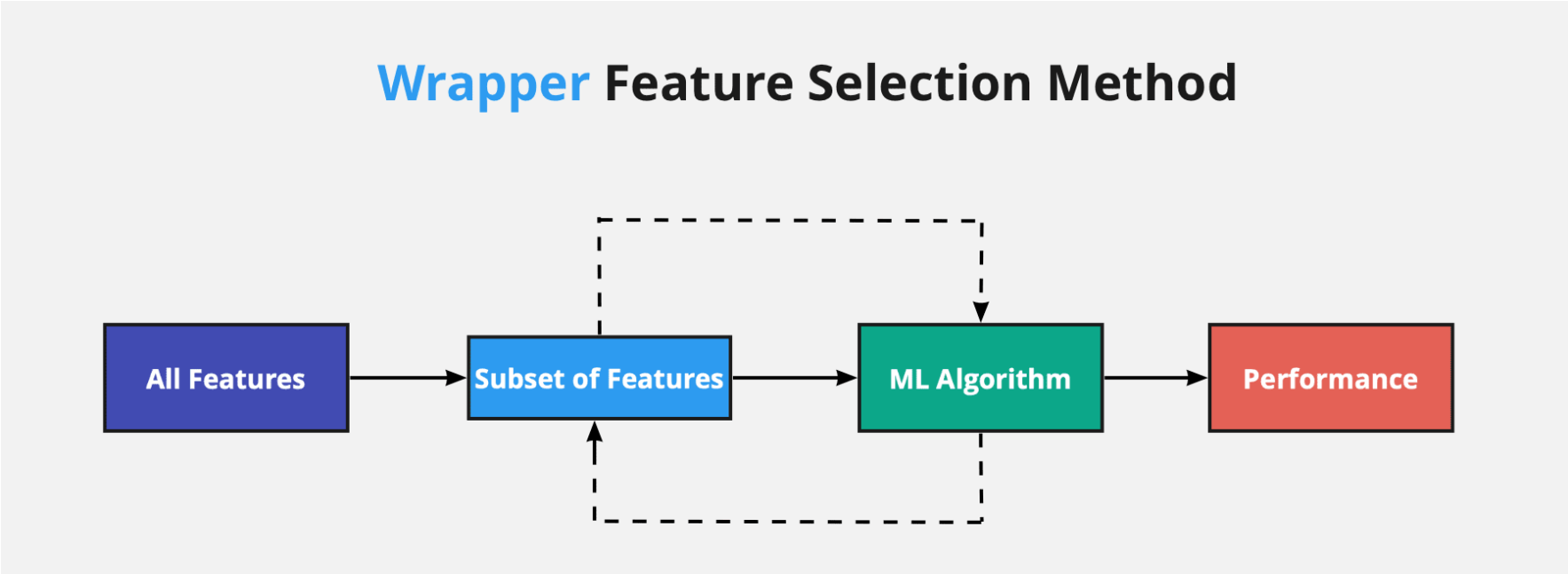
فرآیند داده کاوی

کاهش ابعاد و انتخاب نمونه

❑ انتخاب ویژگی (Feature Selection)

- روش بسته بند (Wrapper Method)

این روش با ساخت تعداد زیادی مدل پیش بینانه روی زیرمجموعه های مختلفی از ویژگی های ورودی، و ارزیابی عملکرد آنها بهترین زیرمجموعه از ویژگی های ورودی را انتخاب می کند.



فرآیند داده کاوی

کاهش ابعاد و انتخاب نمونه

❑ انتخاب ویژگی (Feature Selection)

- روش بسته بند (Wrapper Method)

قابلیت بهینه سازی در انتخاب ویژگی های موثر با رویکردهای زیر:

- استفاده از الگوریتم های Search
- رویکرد Forward Selection
- رویکرد Backward Elimination
- رویکرد Step-wise Selection (Bidirectional Elimination)

نکات قوت

- بررسی همبستگی و اثر متقابل بین ویژگی ها
- کنترل افزونگی داده ها
- تعیین تعداد بهینه از ویژگی های موثر

نکات ضعف

- پیچیدگی محاسباتی بیشتر نسبت به روش فیلتر
- مستعد بیش برازش به علت استفاده از مدل پیش بینانه

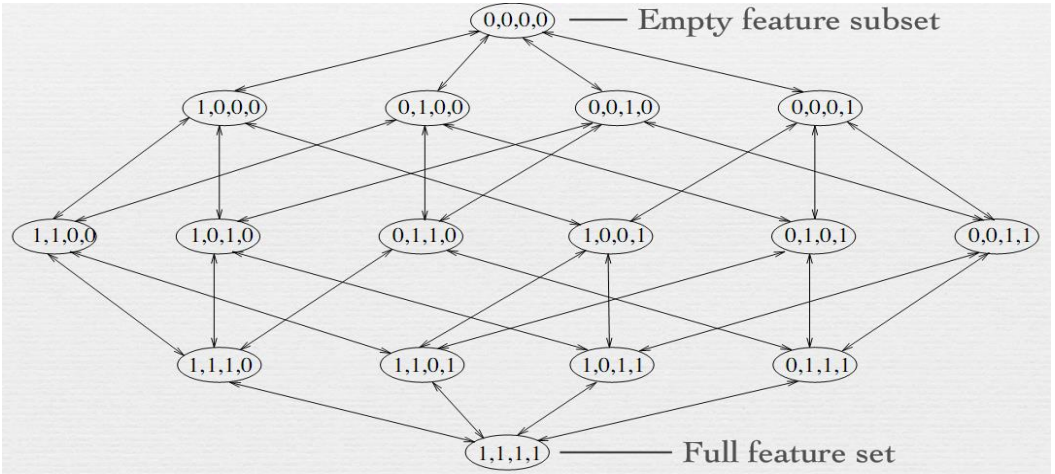
تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

کاهش ابعاد و انتخاب نمونه

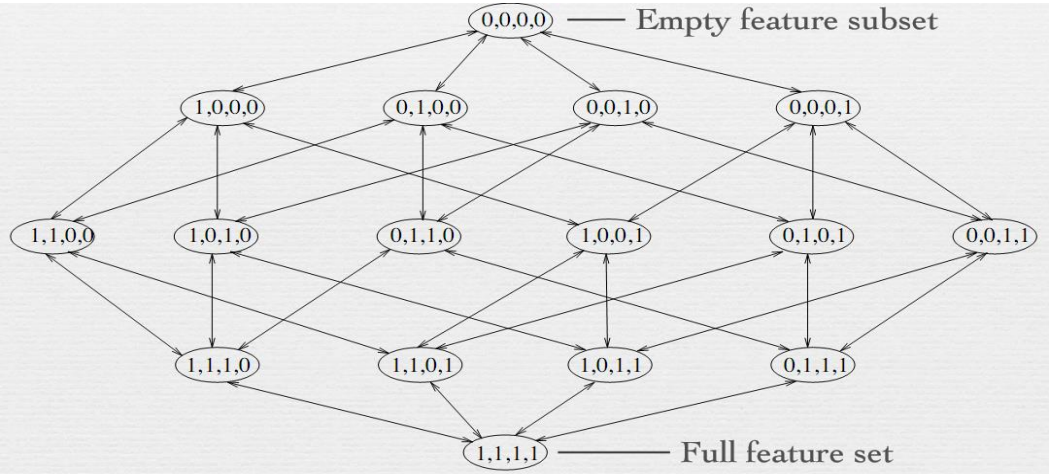
- **انتخاب ویژگی (Feature Selection)**
  - روش بسته بند (Wrapper Method)
    - استفاده از الگوریتم های Search

- Hill- climbing (Greedy stepwise)
- Best-first search
- Genetic Algorithm
- Ant Colony
- ...



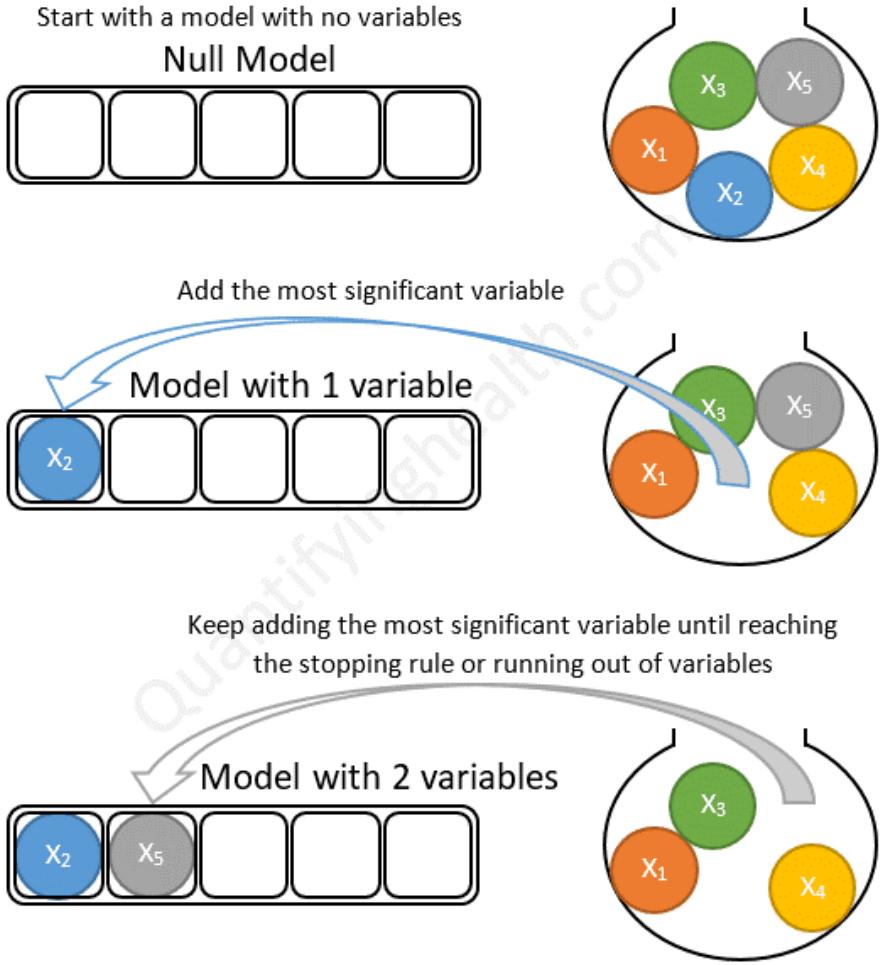
- **انتخاب ویژگی (Feature Selection)**
  - روش بسته بند (Wrapper Method)
    - استفاده از الگوریتم های Search
      - Hill- climbing (Greedy stepwise)
      - Best-first search
      - Genetic Algorithm
      - Ant Colony
      - ...

1. Let $v \leftarrow$ initial state.
2. Expand v: apply all operators to v, giving v's children.
3. Apply the evaluation function f to each child w of v.
4. Let $v' =$ the child w with highest evaluation $f(w)$.
5. If $f(v') > f(v)$ then $v \leftarrow v'$; goto 2.
6. Return v.



فرآیند داده کاوی

کاهش ابعاد و انتخاب نمونه

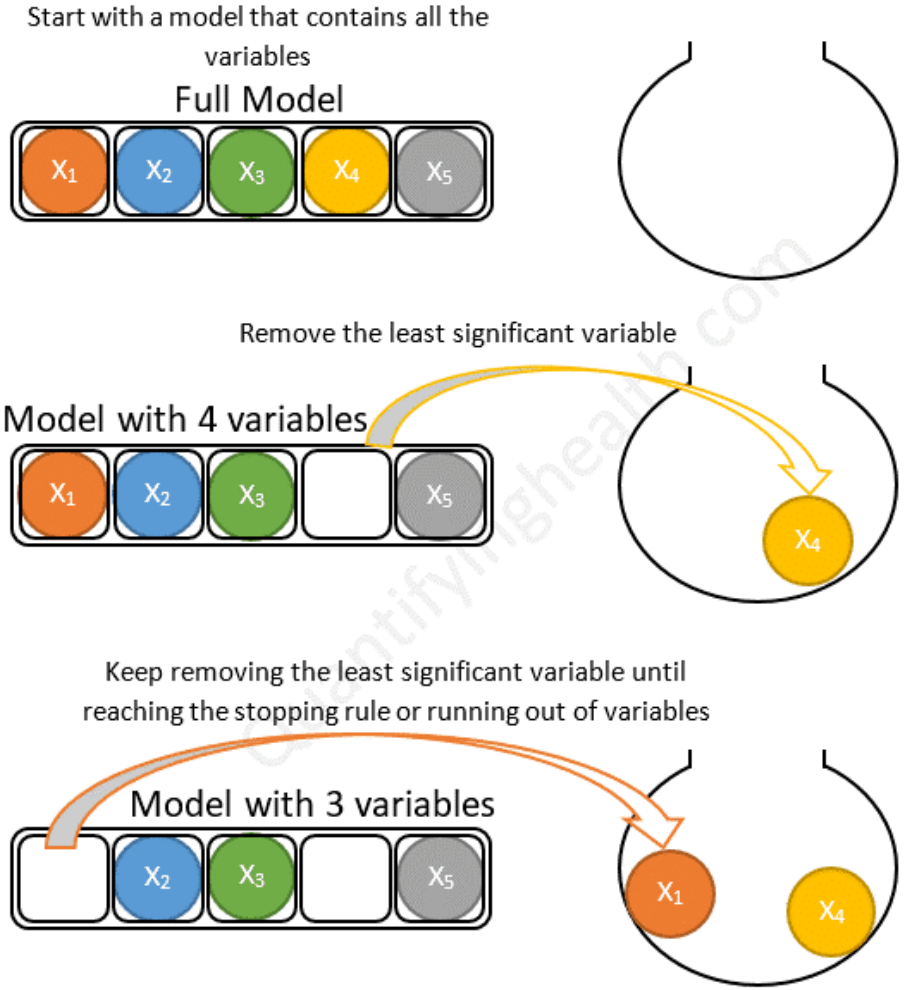


- **انتخاب ویژگی (Feature Selection)**
  - روش بسته بند (Wrapper Method)
    - رویکرد Forward Selection
      - در این رویکرد، ابتدا مدل اولیه بدون هیچکدام از ویژگی ها در نظر گرفته می شود.
      - سپس در گام بعدی ویژگی دارای بیشترین ارتباط با فیلد هدف وارد مدل می شود و بهبود کیفیت مدل ارزیابی می شود.
      - در گام بعدی به شرط حضور ویژگی اول، بهترین ویژگی دوم از نظر ارتباط، به مدل اضافه شده و کیفیت مدل بررسی می شود.
      - این گام ها تا شرط توقف (معنادار نبودن ارتباط ویژگی های باقیمانده) و یا ورود تمام ویژگی ها ادامه می یابد.

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

کاهش ابعاد و انتخاب نمونه



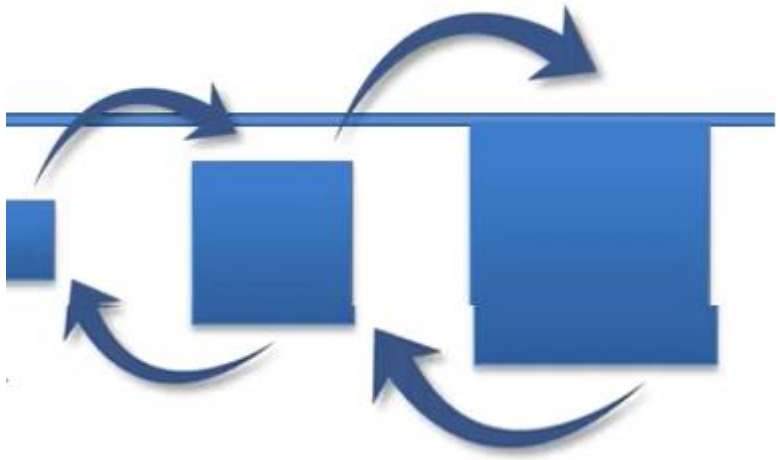
- انتخاب ویژگی (Feature Selection)
  - روش بسته بند (Wrapper Method)
    - رویکرد Backward Elimination
      - در این رویکرد، ابتدا مدل اولیه شامل تمام ویژگی ها در نظر گرفته می شود.
      - سپس در گام بعدی ویژگی دارای کمترین ارتباط با فیلد هدف حذف می شود و معنادار نبودن کاهش کیفیت مدل ارزیابی می شود.
      - در گام بعدی ضعیفترین ویژگی دوم از نظر ارتباط، از مدل حذف شده و کیفیت مدل بررسی می شود.
      - این گام ها تا شرط توقف (معنادار بودن ارتباط ویژگی های باقیمانده) و یا خروج تمام ویژگی ها ادامه می یابد.

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

کاهش ابعاد و انتخاب نمونه

- **انتخاب ویژگی (Feature Selection)**
  - روش بسته بند (Wrapper Method)
    - رویکرد Step-wise Selection (Bidirectional Elimination)



این رویکرد بصورت هیبریدی از دو رویکرد قبل استفاده می شود. بدین صورت که در گام اول، یک ویژگی بر اساس رویکرد Forward وارد مدل می شود و در گام دوم بر اساس شرایط رویکرد Backward امکان حذف آن از مدل مورد بررسی قرار می گیرد. این گام ها تا جاییکه ویژگی جدیدی شرایط ورود نداشته باشد و هیچکدام از ویژگیها شرایط حذف از مدل را نداشته باشند ادامه پیدا می کند.

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

❑ انتخاب ویژگی (Feature Selection)

- روش توکار (Embedded Method)

در این روش انتخاب ویژگی در فرآیند ساخت مدل ادغام شده است.

این روش همانند روش فیلتر، سریع بوده و ویژگی های مرتبط با مسئله را شناسایی می کند و مانند روش بسته بند، با بررسی ویژگی ها در کنار هم از افزونگی داده ها جلوگیری می کند و با تعیین تعداد بهینه از ویژگی های موثر از بیش برازش شدن مدل نیز جلوگیری می کند.

- lasso regression or L1 regularization
- ridge regression or L2 regularization
- elastic nets or L1/L2 regularization

استفاده از این روش در مدلسازی با ابعاد بالا و تعداد ویژگی های زیاد منجر به ساخت مدلهای تنک (Sparse Models) می شود.

جزئیات این روش و متدهای آن در مباحث پیشرفته کورس یادگیری ماشین پوشش داده می شود.

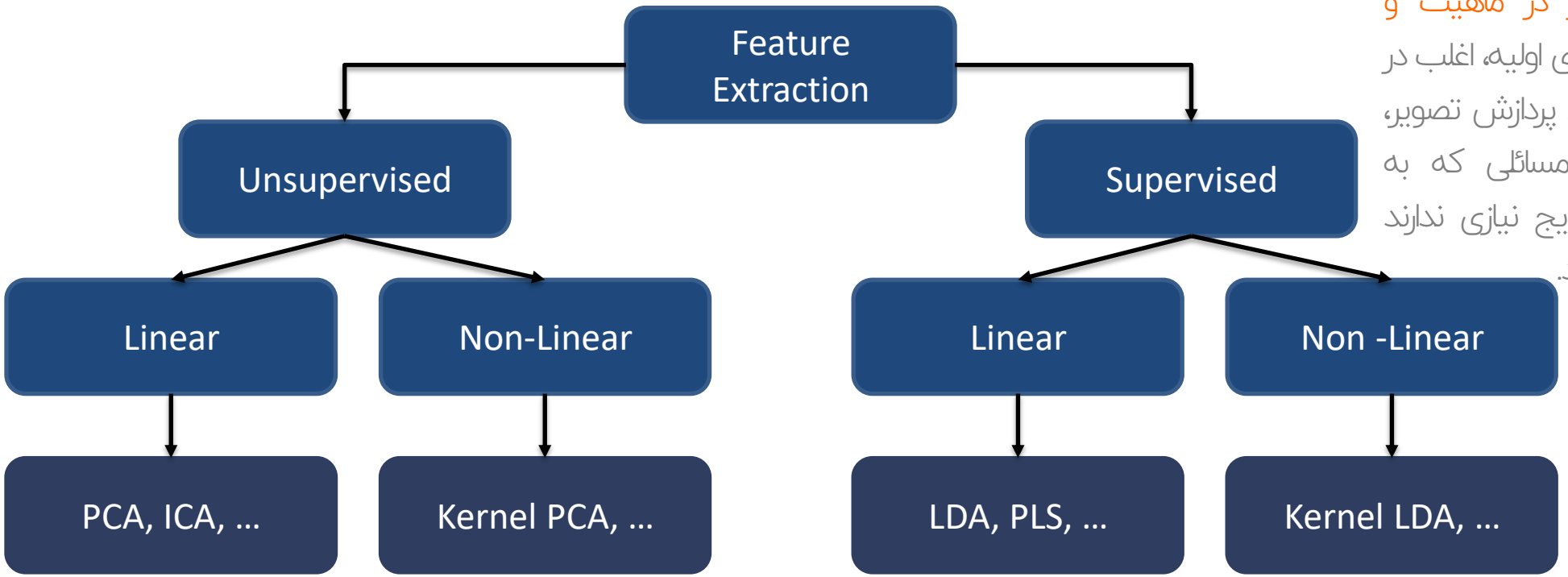
فرآیند داده کاوی

کاهش ابعاد و انتخاب نمونه

❑ استخراج ویژگی (Feature Extraction)

در تکنیک های استخراج ویژگی، با ایجاد ترکیب های خطی از ویژگی های اولیه، فضای n – بعدی داده ها به فضای کوچکتری نگاشت داده می شود، بطوریکه ویژگی های ساخته شده جدید نماینده خوبی از داده های اولیه باشد.

به علت تغییر در ماهیت و مقادیر ویژگی های اولیه، اغلب در مسائلی همچون پردازش تصویر، سیگنال و یا مسائلی که به تفسیرپذیری نتایج نیازی ندارند استفاده می شود.



تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

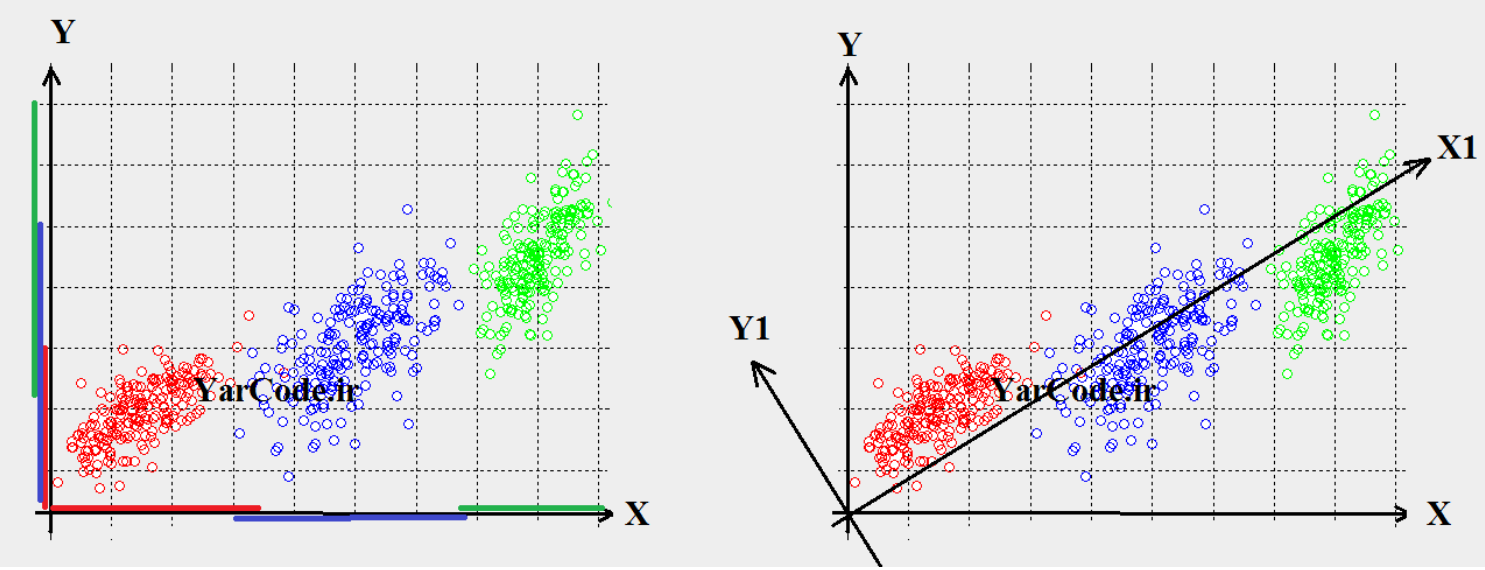
فرآیند داده کاوی

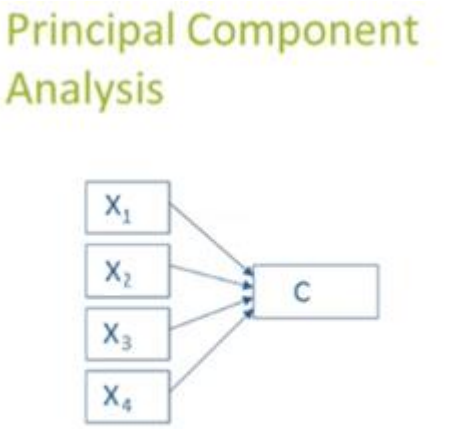
کاهش ابعاد و انتخاب نمونه

❑ استخراج ویژگی (Feature Extraction)

○ تحلیل مولفه های اصلی (Principle Component Analysis)

تحلیل مؤلفه‌های اصلی در تعریف ریاضی یک تبدیل خطی متعامد است که داده را به دستگاه مختصات جدید می‌برد به‌طوری‌که بزرگترین واریانس داده بر روی اولین محور مختصات، دومین بزرگترین واریانس بر روی دومین محور مختصات قرار می‌گیرد و همین‌طور برای بقیه.





تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

❑ استخراج ویژگی (Feature Extraction)

- تحلیل مولفه های اصلی (Principle Component Analysis)

مراحل محاسبه مولفه های اصلی

1. نرمالسازی داده ها (Z-Score)

Original data

X	Y
2	3
4	5
6	5
6	7
7	8
5	8

Z-Score calculation

X	Y
(2-5)/1.633	(3-6)/1.826
(4-5)/1.633	(5-6)/ 1.826
(6-5)/1.633	(5-6)/ 1.826
(6-5)/1.633	(7-6)/ 1.826
(7-5)/1.633	(8-6)/ 1.826
(5-5)/1.633	(8-6)/ 1.826

=

Scaled data

X	Y
-1.837	-1.643
-0.612	-0.548
0.612	-0.548
0.612	0.548
1.225	1.095
0	1.095

- ❑ **استخراج ویژگی (Feature Extraction)**
  - تحلیل مولفه های اصلی (Principle Component Analysis)

مراحل محاسبه مولفه های اصلی

2. محاسبه ماتریس کوواریانس A

As we know after Z scaling, the mean of X and Y becomes 0, i.e. $\bar{X} = 0, \bar{Y} = 0$

So, the covariance formula becomes:

$$Cov(X,X) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})}{n-1} = \frac{\sum_{i=1}^{n} X_i^2}{n-1}$$

$$= ((-1.837)^2 + (-0.612)^2 + (0.612)^2 + (0.612)^2 + (1.225)^2 + 0)/5$$

$$= 1.2$$

$$Cov(Y,Y) = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(Y_i - \bar{Y})}{n-1} = \frac{\sum_{i=1}^{n} Y_i^2}{n-1}$$

$$= ((-1.643)^2 + (-0.548)^2 + (-0.548)^2 + (0.548)^2 + (1.095)^2 + (1.095)^2)/5$$

$$= 1.2$$

$$Cov(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1} = \frac{\sum_{i=1}^{n} X_i Y_i}{n-1}$$

$$= ((-1.837)(-1.643) + (-0.612)(-0.548) + (0.612)(-0.548) + (0.612)(0.548) + (1.225)(1.095) + 0)/5$$

$$= 0.939$$

We know, $Cov(Y,X) = Cov(X,Y)$

We know covariance matrix of X, Y is

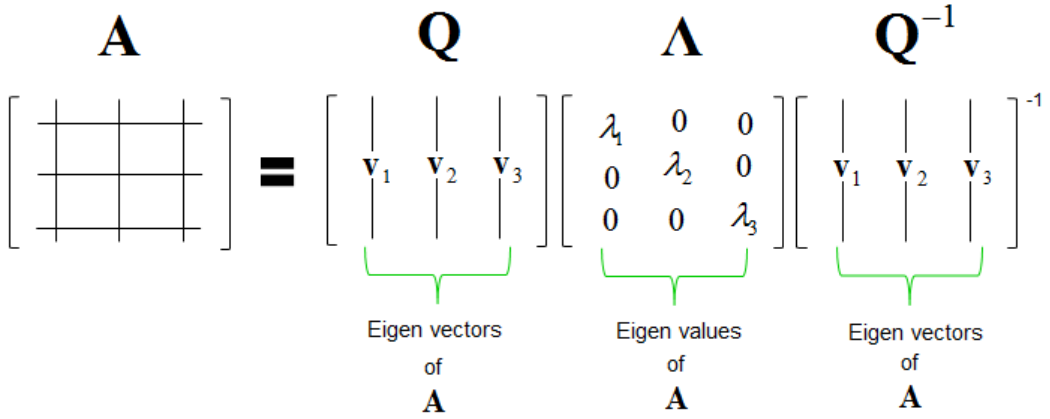
$$\text{Covariance Matrix} = \begin{bmatrix} Cov(X,X) & Cov(X,Y) \\ Cov(Y,X) & Cov(Y,Y) \end{bmatrix}$$

$$= \begin{bmatrix} 1.2 & 0.939 \\ 0.939 & 1.2 \end{bmatrix}$$

❑ استخراج ویژگی (Feature Extraction)

- مروری سریع بر مفاهیم جبر خطی مورد نیاز

$$A\vec{v} = \lambda\vec{v}$$



$$A = \sum_{i=1}^{p} \lambda_i \, v_i \, v_i'$$

$$|A - \lambda I| = 0$$

$$trace(A) = \sum_{i=1}^{p} \lambda_i$$

$$v_i' v_i = 0$$

❑ استخراج ویژگی (Feature Extraction)

- تحلیل مولفه های اصلی (Principle Component Analysis)

مراحل محاسبه مولفه های اصلی

3. محاسبه مقادیر ویژه (λ) ماتریس کوواریانس ($|A - \lambda I| = 0$)

if λ is the Eigen value of given matrix A then

$$| A - \lambda I | = 0$$

Now, the Eigen value of our covariance matrix will be

$$A - \lambda I = \begin{bmatrix} 1.2 & 0.939 \\ 0.939 & 1.2 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1.2 & 0.939 \\ 0.939 & 1.2 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}$$

$$= \begin{bmatrix} 1.2-\lambda & 0.939 \\ 0.939 & 1.2-\lambda \end{bmatrix}$$

So, $\begin{vmatrix} 1.2-\lambda & 0.939 \\ 0.939 & 1.2-\lambda \end{vmatrix} = 0$

$$\Rightarrow (1.2-\lambda)(1.2-\lambda) - (0.939) \times (0.939) = 0$$

$$\Rightarrow 1.44 - (2 \times 1.2)\lambda + \lambda^2 - 0.8817 = 0$$

$$\Rightarrow \lambda^2 - 2.4\lambda + 0.5583 = 0$$

Solving for λ we get $\lambda = 2.139$ and $\lambda = 0.261$

So, two Eigen values are 2.139 and 0.261

تولید محتوا: زهرا ذوالقدر

daychegroup

daychegroup

dayche.com | گروه دایکه

❑ استخراج ویژگی (Feature Extraction)

- تحلیل مولفه های اصلی (Principle Component Analysis)

مراحل محاسبه مولفه های اصلی

4. مرتب سازی بزرگ به کوچک مقادیر ویژه و انتخاب مقادیر ویژه منتخب
5. محاسبه بردارهای ویژه (v) متناظر با مقادیر ویژه منتخب ($Av = \lambda v$)

we know, $Av = \lambda v$. Considering $\begin{bmatrix} e \\ f \end{bmatrix}$ as the Eigen vector,

$$\begin{bmatrix} 1.2 & 0.939 \\ 0.939 & 1.2 \end{bmatrix}\begin{bmatrix} e \\ f \end{bmatrix} = 2.139 \times \begin{bmatrix} e \\ f \end{bmatrix}$$

$$\Rightarrow \begin{bmatrix} 1.2e + 0.939f \\ 0.939e + 1.2f \end{bmatrix} = \begin{bmatrix} 2.139e \\ 2.139f \end{bmatrix}$$

So, using 1st row, $1.2e + 0.939f = 2.139e$

$\Rightarrow 0.939f = 0.939e$

$\Rightarrow e = f$

or using 2nd row, $0.939e + 1.2f = 2.139f \Rightarrow e = f$

Here the ratio is important, hence we can consider a vector $(1, 1)$ which corresponds to (e, f).

For the vector to be Eigen vector the length of the vector should be 1.

Let's find the length of $(1, 1)$, which is nothing but distance of the point $(0, 0)$ and $(1,1)$ in two-dimensional space.

$$Using\ Pythagoras\ theorem,\ distance = \sqrt{1^2 + 1^2} = \sqrt{2}$$

Now if we divide the vector, represented as matrix $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ by $\sqrt{2}$, its length will become 1 and we will get our desired Eigen vector.

$$Eigen\ Vector = \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix} = \begin{bmatrix} 0.7071 \\ 0.7071 \end{bmatrix}$$

تولید محتوا: زهرا ذوالقدر

daychegroup

daychegroup

گروه دایکه | dayche.com

Scaled data

X	Y
-1.837	-1.643
-0.612	-0.548
0.612	-0.548
0.612	0.548
1.225	1.095
0	1.095

- **استخراج ویژگی (Feature Extraction)**
  - تحلیل مولفه های اصلی (Principle Component Analysis)

مراحل محاسبه مولفه های اصلی

6. محاسبه مقادیر بردار مولفه اصلی

با رابطه (Transpose of Eigen vector) X (Feature vector)

$$\begin{bmatrix}0.7071\\0.7071\end{bmatrix}^T \times \begin{bmatrix}-1.837\\-1.643\end{bmatrix}$$

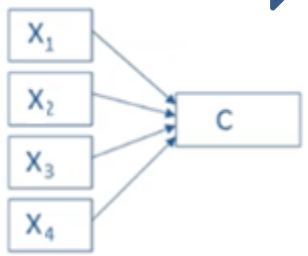
$$= [0.7071 \;\; 0.7071]\begin{bmatrix}-1.837\\-1.643\end{bmatrix}$$

$$= 0.7071 \times (-1.837) + 0.7071 \times (-1.643)$$

$$= -(1.2989 + 1.1617)$$

$$= -2.461$$

$$Hence, \begin{bmatrix}-1.837\\-1.643\end{bmatrix} \; is\; projected\; to \; = \; -2.461$$



$C = w_1(X_1) + w_2(X_2) + w_3(X_3) + w_4(X_4)$

Projected Data
-2.461
-0.820
0.046
0.820
1.640
0.774

تولید محتوا: زهرا ذوالقدر

daychegroup

daychegroup

dayche.com | گروه دایکه

فرآیند داده کاوی

کاهش ابعاد و انتخاب نمونه

- **استخراج ویژگی (Feature Extraction)**
  - تحلیل مولفه های اصلی (Principle Component Analysis)

فرضیات تحلیل مولفه های اصلی

- این روش ذاتا برای داده های کمی قابل استفاده است.
- بین داده های ورودی همبستگی خطی وجود دارد.

خصوصیات تحلیل مولفه های اصلی

- بردار ویژه جهت محورهای جدید را با طول ثابت یک تعیین می کنند و مقادیر ویژه اندازه و بزرگی واریانس داده ها رو در جهت تعیین شده مشخص می کند.
- مولفه های ایجاد شده با هدف حفظ بیشترین اطلاعات و پوشش حداکثر واریانس ایجاد می شود.
- مولفه های ایجاد شده متعامد و ناهمبسته هستند. (استقلال مولفه ها تضمین نمی شود)

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

کاهش ابعاد و انتخاب نمونه

❑ استخراج ویژگی (Feature Extraction)

- محدودیت های روش PCA

Solution:	Solution:	Solution:	Solution:
LDA	**Kernel PCA**	**ICA**	**Rotation**
مولفه های ایجاد شده، تفکیک پذیری در سطح فیلد هدف انجام نمی دهد و امکان اختلال در تفکیک پذیری دارد.	در صورت وجود ارتباط غیر خطی در داده ها، امکان حفظ حداکثر واریانس در ابعاد کوچکتر وجود ندارد.	در صورت عدم وجود فرض توزیع نرمال در ویژگی های ورودی، استقلال بین مولفه ها تضمین نمی شود.	تعریف مولفه ها بر اساس ترکیب خطی از تمام ویژگی ها، منجر به ایجاد مقادیر پیچیده و غیرقابل تفسیر می شود.

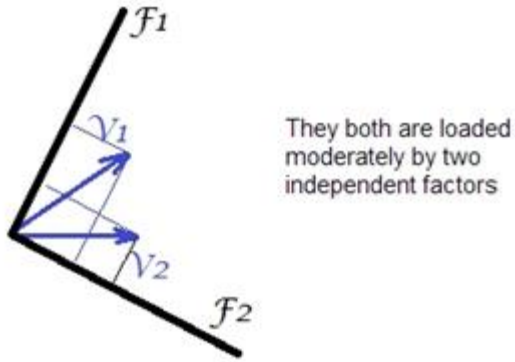
تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

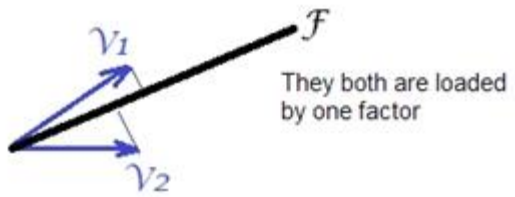
فرآیند داده کاوی

کاهش ابعاد و انتخاب نمونه

Item	Component			
	1	2	3	4
SAS16	0.832			
SAS15	0.735			
SAS17	0.694			0.362
SAS12	0.683		0.326	
SAS11	0.670			
SAS10		0.695		0.319
SAS7		0.684	0.304	
SAS9		0.630		0.388
SAS8		0.615		
SAS13		0.478	0.312	
SAS14		0.472		0.413
SAS6	0.390	0.465	0.412	
SAS3			0.750	
SAS1			0.655	
SAS4			0.641	
SAS2			0.608	
SAS5		0.490	0.569	
SAS19			0.320	0.687
SAS20				0.664
SAS18				0.598

Extraction Method: Principal Component Analysis
Rotation method: Varimax with Kaiser normalization





- **استخراج ویژگی (Feature Extraction)**
  - چرخش تحلیل مولفه های اصلی (Rotated PCA)

با چرخش محورهای مختصات می توان وزن ویژگی ها در ترکیب خطی مولفه ها را تغییر داد و از این طریق تفسیر پذیری بهتری از مولفه های خطی داشت.

انواع چرخش های رایج در PCA

- چرخش Varimax

واریانس بین مقادیر وزن ویژگی ها را بیشینه می کند و در نتیجه یک مولفه با تعداد کمی از ویژگی های مهم توضیح داده می شود.

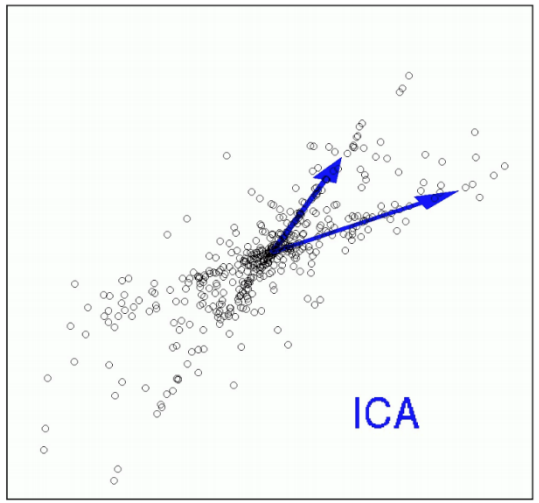
- چرخش Quartimax

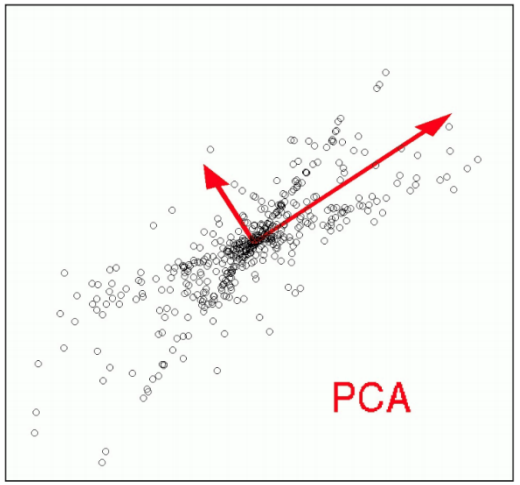
بر عکس روش قبلی، تعداد فاکتورهای لازم برای توضیح هر ویژگی را کم می کند.

❑ استخراج ویژگی (Feature Extraction)

○ تحلیل مولفه های مستقل (Independent Component Analysis)

تحلیل مؤلفه‌های مستقل در تعریف ریاضی یک تبدیل خطی غیر متعامد است که داده را به دستگاه مختصات جدید می‌برد به‌طوری‌که ترکیب های خطی ایجاد شده، مستقل از هم باشند.



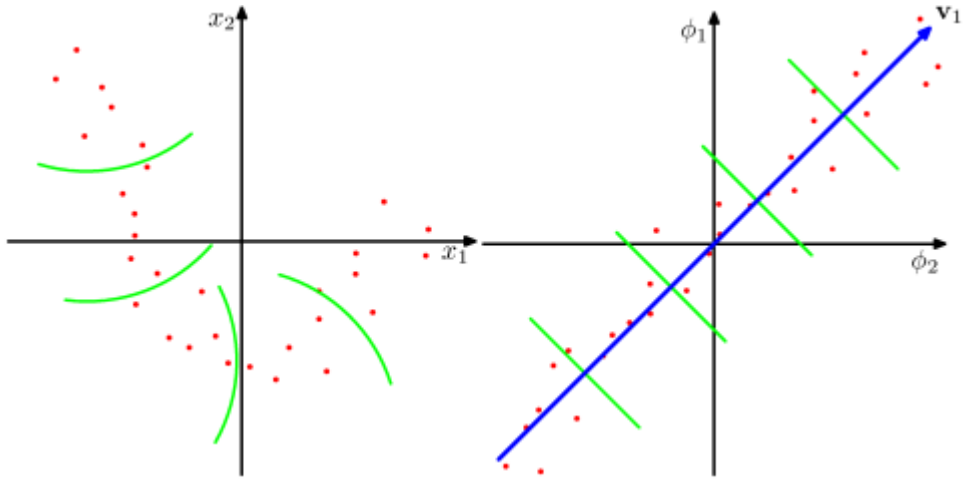


در بسیاری از مسائل مانند تشخیص نویز در سیگنال های صوتی و یا تفکیک صداهای مختلف در فایلهای صوتی و ... که جداسازی منابع و سورسهای مستقل اهمیت بالایی دارد، استفاده از روش ICA برای استخراج ویژگی مناسب هست.

❑ استخراج ویژگی (Feature Extraction)

- تحلیل مولفه های اصلی غیر خطی (Kernel PCA)

در صورتی که ارتباط بین ویژگی ها غیرخطی باشد، با نقض فرض مهم خطی بودن ویژگی ها، عملکرد PCA دچار اختلال خواهد شد و امکان بیشینه کردن مقدار واریانس در ابعاد کوچکتر از بین می رود.



راهکار: استفاده از توابع تبدیل کرنل جهت نگاشت داده های غیرخطی به فضای متفاوت به منظور برقراری ارتباط خطی در فضای جدید.

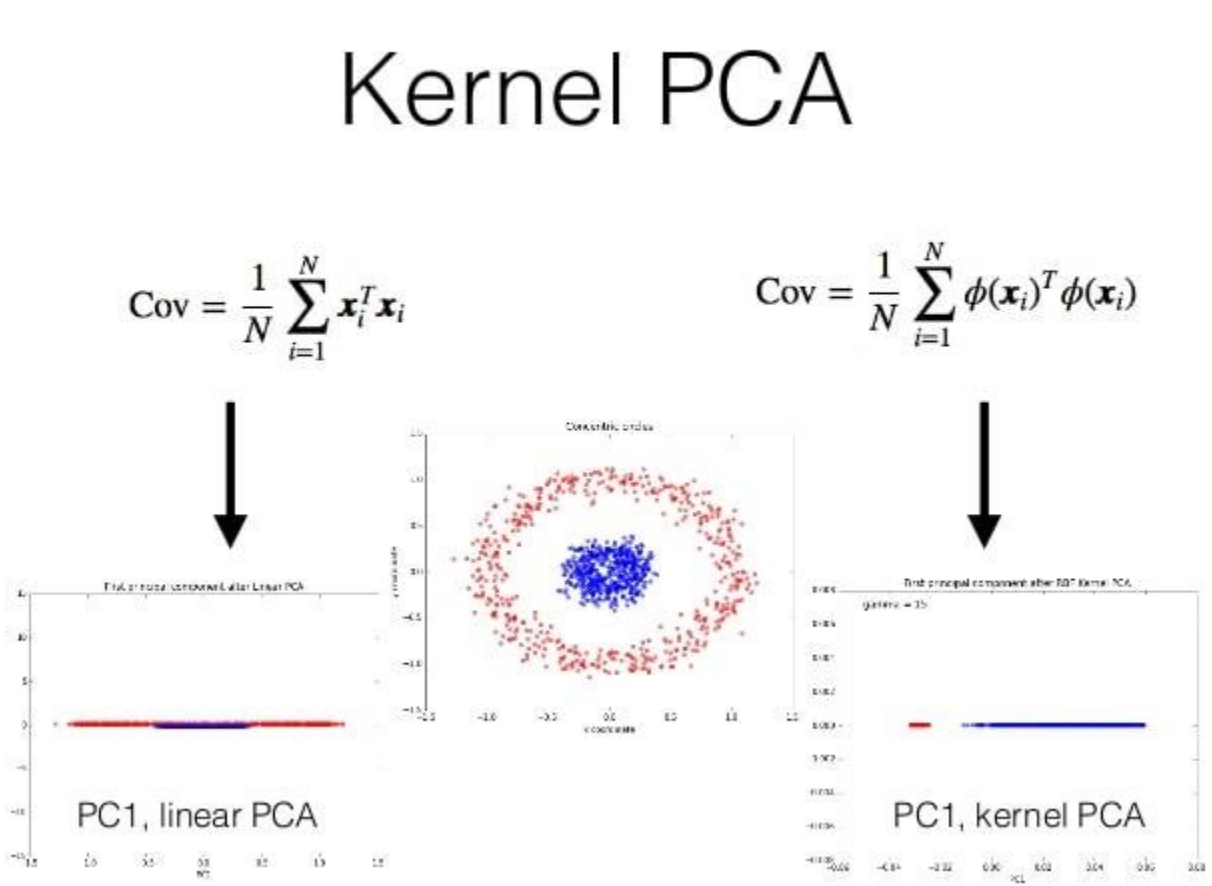
Kernels	Formula
linear	$k(x,y) = x.y$
sigmoid	$k(x,y) = \tanh(ax.y + b)$
polynomial	$k(x,y) = (1 + x.y)^d$
RBF	$k(x,y) = \exp(-a\|x-y\|^2)$
exponential RBF	$k(x,y) = \exp(-a\|x-y\|)$

❑ استخراج ویژگی (Feature Extraction)

- تحلیل مولفه های اصلی غیر خطی (Kernel PCA)

مشکل دیگری که در روابط غیر خطی فضای داده های اصلی امکان وقوع دارد این است که، به علت ماهیت غیر نظارتی روش PCA و عدم اطلاع از وضعیت برچسب های فیلد هدف، در نگاشت داده ها به ابعاد جدید، منجر به همسان سازی داده های با کلاسهای متفاوت شده و در نتیجه باعث کاهش عملکرد مدلهای رده بندی شود.

راهکار: استفاده از توابع تبدیل کرنل جهت نگاشت داده های غیرخطی به فضای متفاوت به منظور برقراری ارتباط خطی در فضای جدید.



تولید محتوا: زهرا ذوالقدر

daychegroup

daychegroup

dayche.com | گروه دایکه

فرآیند داده کاوی

کاهش ابعاد و انتخاب نمونه

❑ استخراج ویژگی (Feature Extraction)

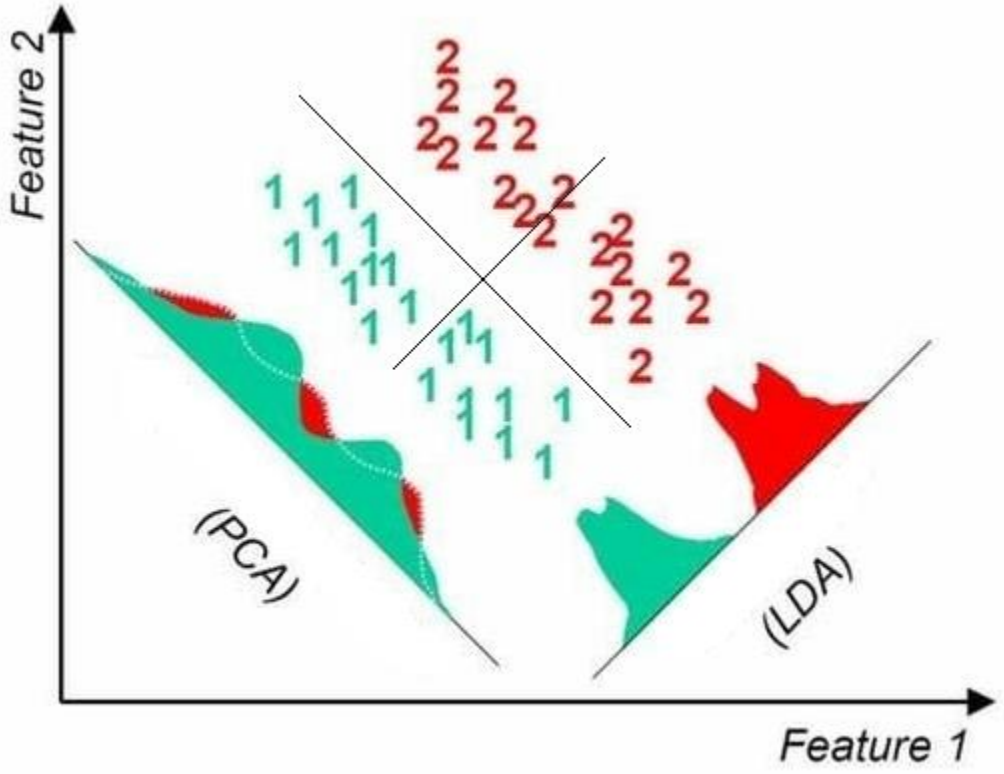
- تحلیل ممیزی خطی (Linear Discriminative Analysis)

روش LDA از نوع یادگیری با نظارت بوده و بر خلاف روش PCA روش کاهش بعد را در جهتی انجام می دهد که نگاشت داده ها روی آنها بیشترین تفکیک پذیری بین کلاسهای هدف ایجاد نماید.

از این روش به نام آنالیز خطی فیشر نیز نام برده می شود.

درصورت رابطه غیر خطی بین ویژگی ها:

راهکار: استفاده از توابع تبدیل کرنل جهت نگاشت داده های غیرخطی به فضای متفاوت به منظور برقراری ارتباط خطی در فضای جدید.



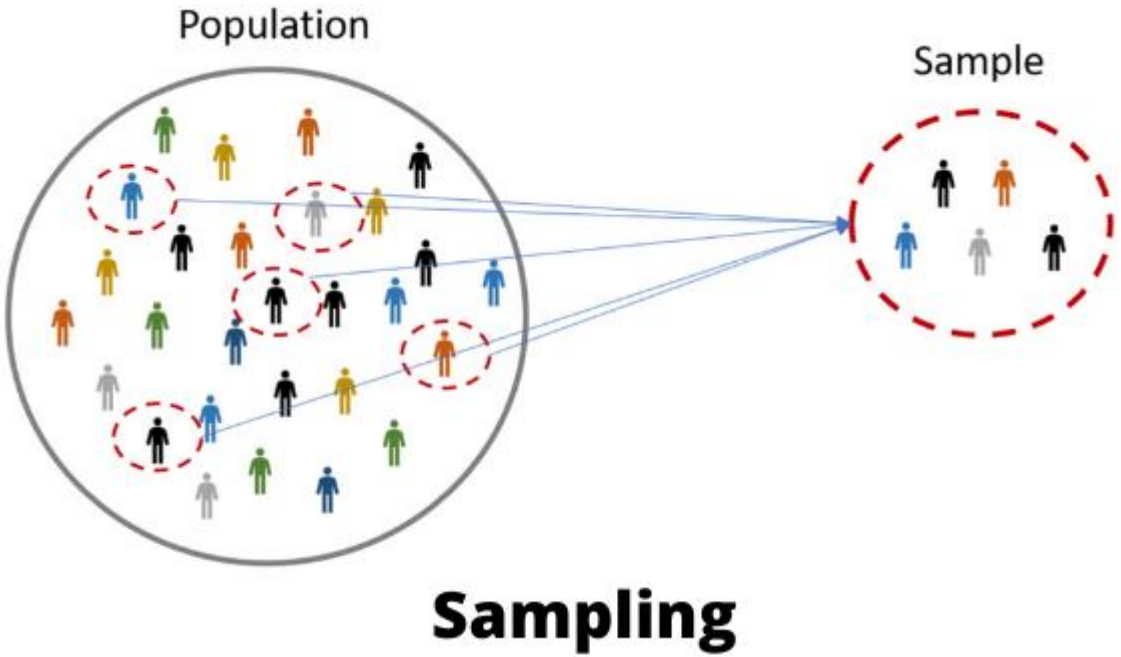
تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

کاهش ابعاد و انتخاب نمونه

❑ نمونه گیری (Sampling)

نمونه گیری یک روش آماری کم هزینه برای کاهش داده هاست تا بر اساس انتخاب زیرمجموعه ای از رکوردهای داده، نماینده مناسبی از داده ها را در حجم کمتر ایجاد نماید.



- سرعت بیشتر اجرای الگوریتم ها
- نیاز به منابع محاسباتی کمتر
- تمرکز بر الگوهای اصلی و معنادار

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

کاهش ابعاد و انتخاب نمونه

❑ نمونه گیری (Sampling)

فرآیند انتخاب نمونه را می توان به دو دسته کلی تقسیم کرد:

انتخاب نمونه بدون جایگذاری

در این حالت هر نمونه فقط یک بار شانس انتخاب دارد.

انتخاب نمونه با جایگذاری

در این حالت هر نمونه بیش از یک بار شانس انتخاب دارد.

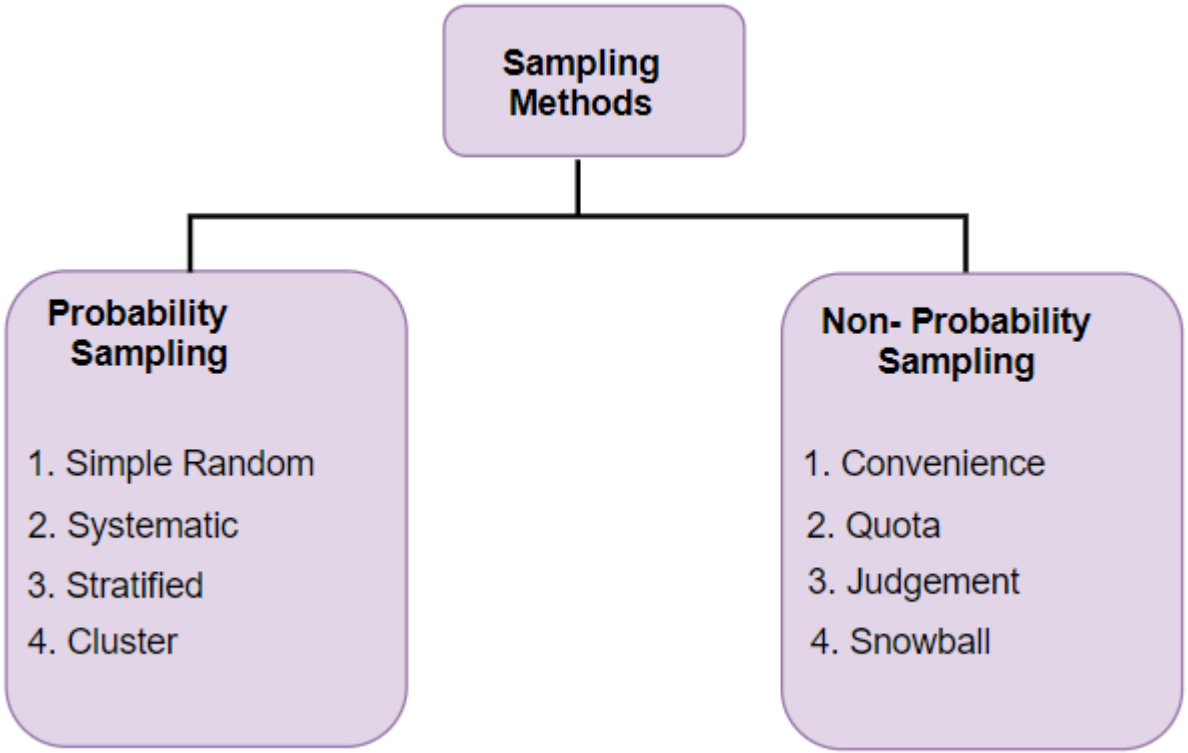
تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

کاهش ابعاد و انتخاب نمونه

❑ نمونه گیری (Sampling)

روش های نمونه گیری را می توان به دو دسته کلی تقسیم کرد:



- **نمونه گیری احتمالی**

در این رویکرد نمونه های موجود در جامعه، دارای احتمال انتخاب برابر هستند. به همین دلیل این رویکرد شانس خوبی برای انتخاب نمونه ای که به خوبی از جامعه نمایندگی کند را داراست.

- **نمونه گیری غیر احتمالی**

در این رویکرد نمونه های موجود در جامعه، دارای احتمال انتخاب نابرابر هستند، و این موضوع می تواند منجر به انتخاب نمونه ای شود که قابلیت تعمیم پذیری خوبی نداشته باشد.

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

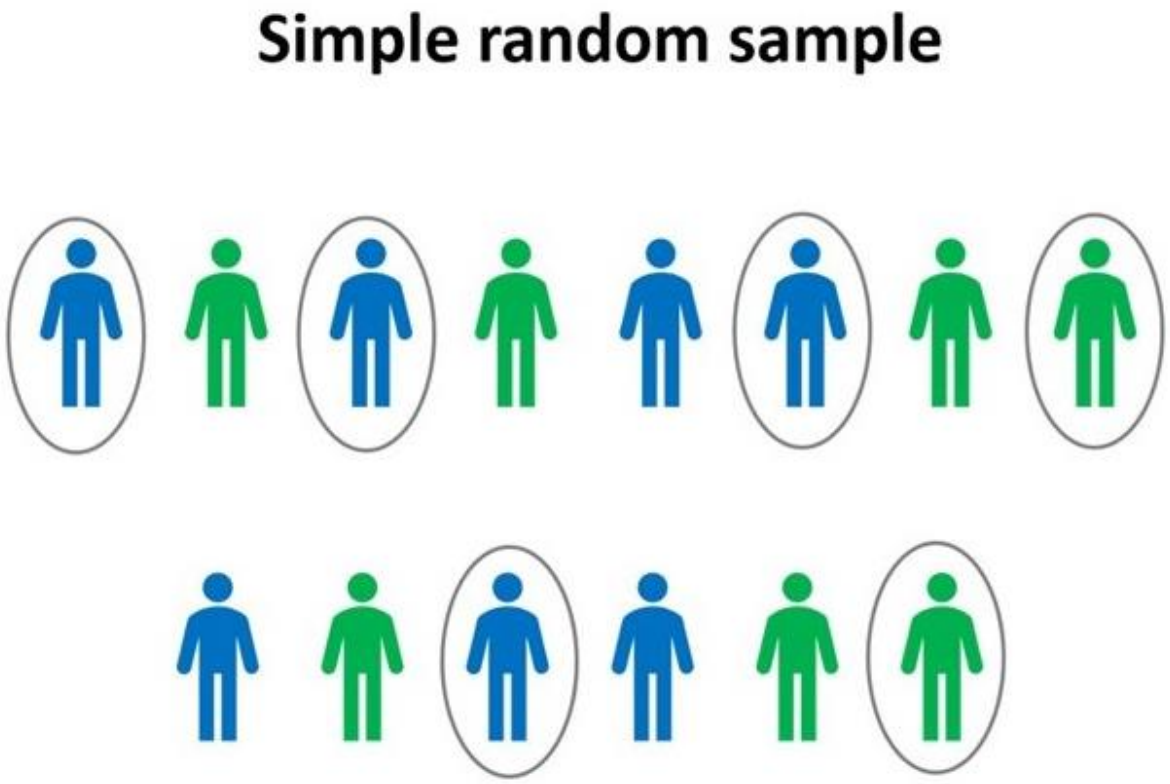
کاهش ابعاد و انتخاب نمونه

❑ نمونه گیری (Sampling)

- نمونه گیری تصادفی ساده

در نمونه گیری تصادفی ساده، تمام نمونه ها شانس برابر انتخاب شدن دارند و میزان بایاس (اریبی) در انتخاب نمونه ها بسیار کم می باشد.

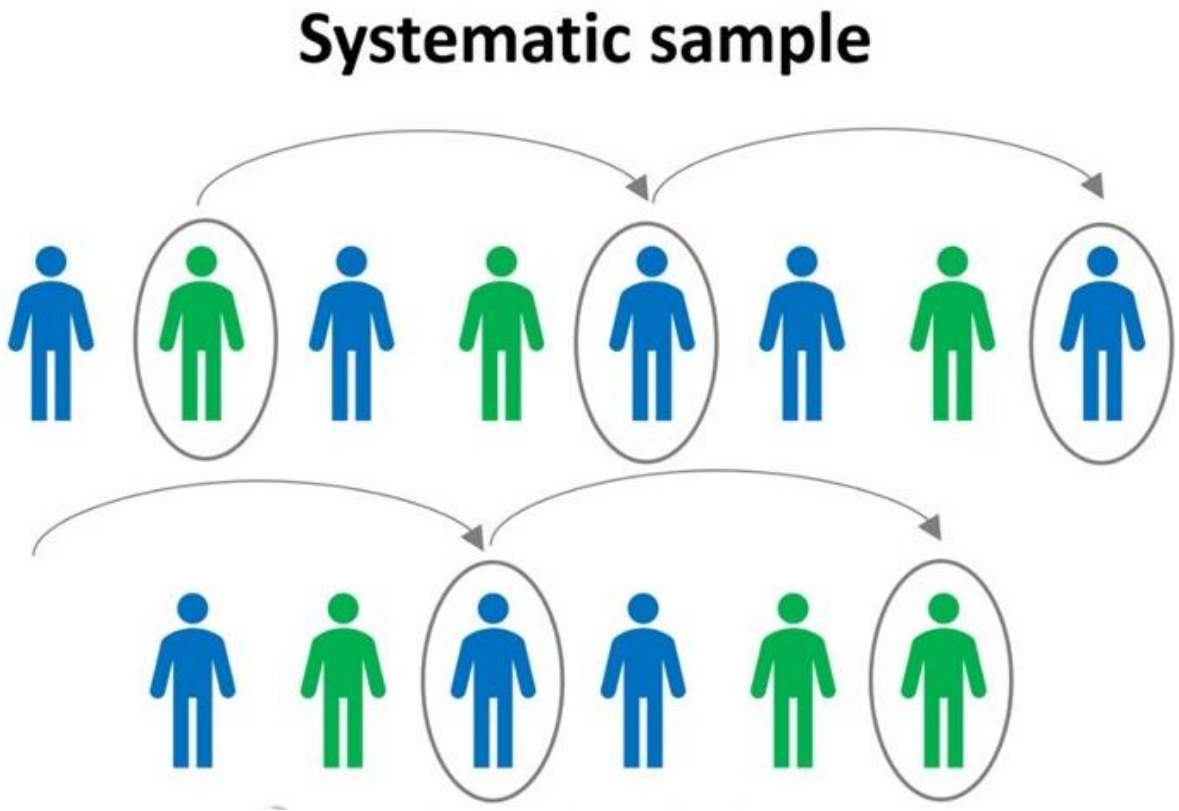
بطور مثال اگر بخواهیم از بین 50 نفر عضو یک صندوق خانوادگی، 3 نفر را به تصادف برای ارائه وام انتخاب کنیم، معادل تولید انتخاب تصادفی 3 عدد طبیعی در بازه 1 تا 50 می باشد.



تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

❑ نمونه گیری (Sampling)

○ نمونه گیری سیستماتیک



اگر تعداد کل نمونه های یک جامعه را N در نظر بگیریم و بخواهیم به تعداد n نمونه از آن انتخاب کنیم، در این طرح کافیست ابتدا شماره رکوردها را مرتب کنیم و در بازه شماره رکورد اول تا شماره رکورد k - ام یک نمونه تصادفی انتخاب شود و سپس با گام های به فاصله k تمامی رکوردهای دیگر نیز انتخاب گردند.

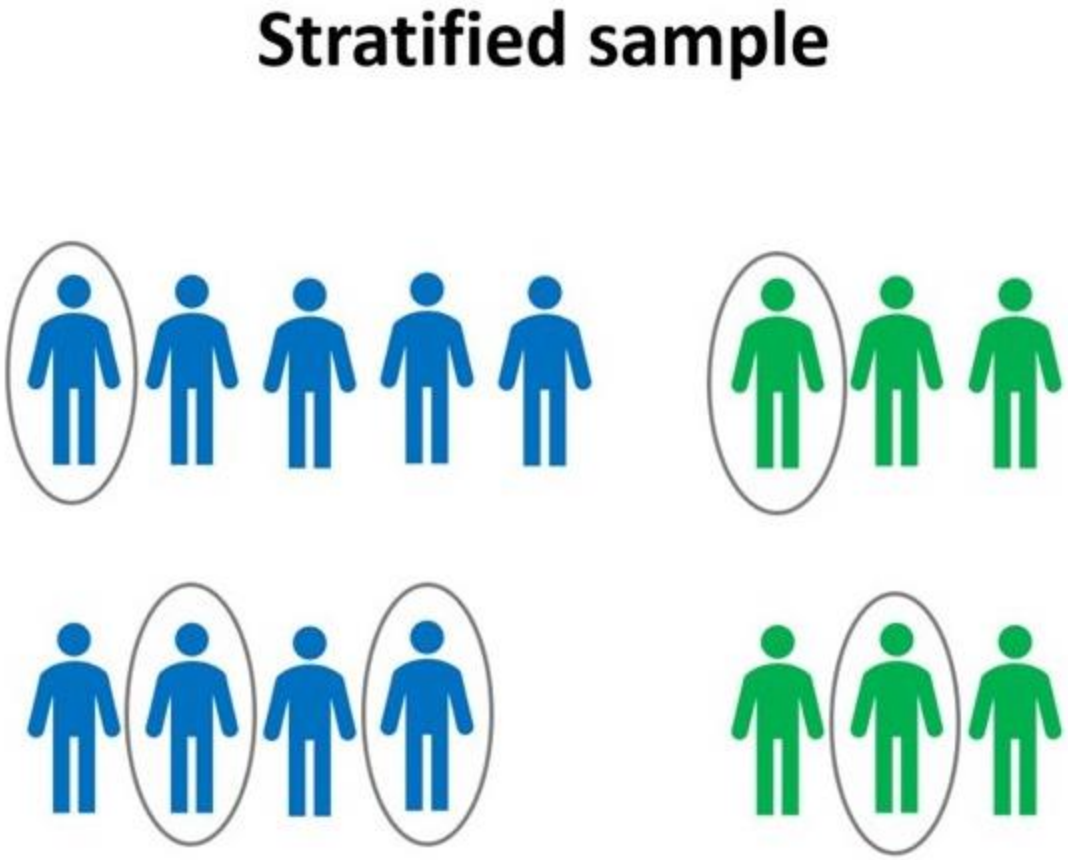
$$k = \frac{N}{n}$$

فرآیند داده کاوی

کاهش ابعاد و انتخاب نمونه

❑ نمونه گیری (Sampling)

- نمونه گیری طبقه ای



در نمونه گیری طبقه ای، کل جامعه به چند گروه یا طبقه تقسیم می شود بطوریکه می دانیم شاخص های مورد تحقیق در این گروه ها دارای تفاوت می باشند. بنابراین در اجرای طرح نمونه گیری طبقه ای، از هر طبقه بصورت جداگانه بر اساس روش های تصادفی ساده یا سیستماتیک نمونه برداری می شود.

بطور مثال برای بررسی شاخص های موثر بر ارتقای شغلی کارمندان یک سازمان، انتخاب تصادفی در گروه زنان و مردان بر اساس نسبت آنها در کل سازمان انجام می شود.

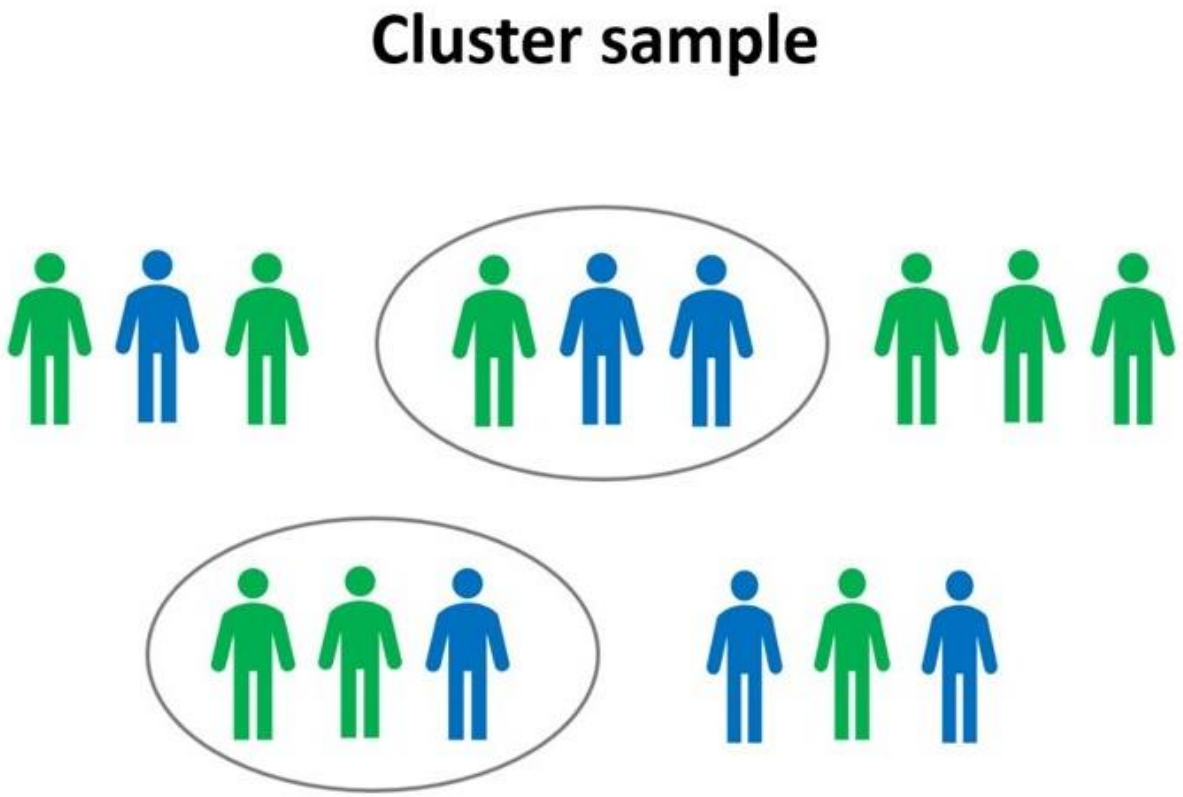
تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

کاهش ابعاد و انتخاب نمونه

❑ نمونه گیری (Sampling)

- نمونه گیری خوشه ای



در نمونه گیری خوشه ای، مشابه قبل کل جامعه به چند گروه یا خوشه تقسیم می شود، با این تفاوت که می دانیم شاخص های مورد تحقیق در این گروه ها دارای توزیع مشابه می باشند. بنابراین در اجرای این طرح کافیست بصورت تصادفی چند خوشه انتخاب شده و نمونه های آن مورد تحلیل قرار گیرند.

بطور مثال برای بررسی میزان موفقیت دانش آموزان مدارس دولتی یک شهر، با انتخاب چند مدرسه، وضعیت تحصیلی دانش آموزان بررسی و تحلیل گردد.

Data Integration

یکپارچه سازی داده ها

dayche.com . گروه دایکه

دای‌کِه DAYCHE

فرآیند داده کاوی

یکپارچه سازی داده ها

❑ یکپارچه سازی داده ها

عموما داده های مورد نیاز در حل یک مسئله در منابع متفاوت و جداگانه ثبت و نگهداری می شوند. یکپارچه سازی منابع مختلف داده یکی از اولین اقدامات در جهت حل مسئله و پیشبرد فرآیند داده کاوی می باشد.

توجه: بایستی توجه شود کلیه روش ها و اقداماتی که تاکنون جهت آماده سازی داده ها گفته شده پس از یکپارچه سازی داده ها انجام می شود.

ادغام کردن داده ها
MERGE

چسباندن داده ها
APPEND

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

فرآیند داده کاوی

یکپارچه سازی داده ها

❑ یکپارچه سازی داده ها

MERGE

Dataset 1

name	age
Nick	18
Tom	25
Jennifer	19
Janet	34

+

Dataset 2

name	sex	country
Nick	male	USA
Tom	male	France
Jennifer	female	Spain
Janet	female	Germany

→

Merged dataset

name	sex	country	age
Nick	male	USA	18
Tom	male	France	25
Jennifer	female	Spain	19
Janet	female	Germany	34

APPEND

Dataset 1

name	age
Nick	18
Tom	25
Jennifer	19
Janet	34

+

Dataset 3

name	age
Sam	22
Sara	31
Mary	28

→

Appended dataset

name	age
Nick	18
Tom	25
Jennifer	19
Janet	34
Sam	22
Sara	31
Mary	28

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه

❑ یکپارچه سازی داده ها

Inner Join

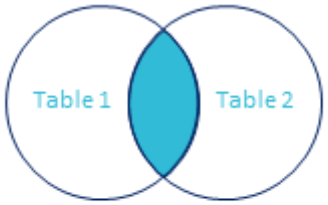


Table 1

id	name
1	Sayak
2	Alex
3	Sameer
4	Rick

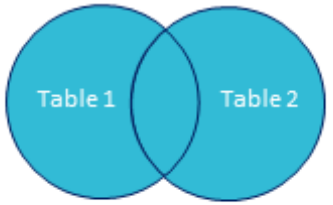
Table 2

id	stream
1	CS
1	IT
2	ECE
9	ECE

Inner Join

id	name	stream
1	Sayak	CS
1	Sayak	IT
2	Alex	ECE

Full Outer Join



Left Table

id	name
1	Sayak
2	Alex
3	Sameer
4	Rick

Right Table

id	stream
1	CS
1	IT
2	ECE
9	ECE

Full Join

id	name	stream
1	Sayak	CS
1	Sayak	IT
2	Alex	ECE
3	Sameer	
4	Rick	
9		ECE

تولید محتوا: زهرا ذوالقدر

daychegroup

daychegroup

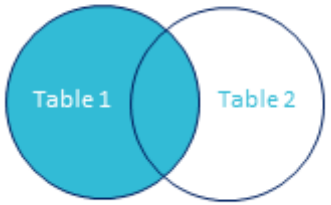
dayche.com | گروه دایکه

فرآیند داده کاوی

یکپارچه سازی داده ها

❑ یکپارچه سازی داده ها

Left Join



Left Table

id	name
1	Sayak
2	Alex
3	Sameer
4	Rick

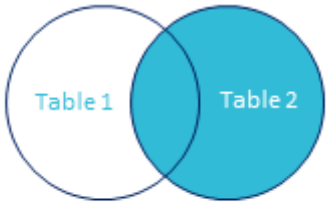
Right Table

id	stream
1	CS
1	IT
2	ECE
9	ECE

Left Join

id	name	stream
1	Sayak	CS
1	Sayak	IT
2	Alex	ECE
3	Sameer	
4	Rick	

Right Join



Left Table

id	name
1	Sayak
2	Alex
3	Sameer
4	Rick

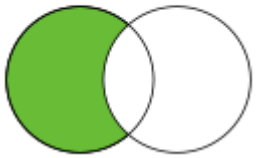
Right Table

id	stream
1	CS
1	IT
2	ECE
9	ECE

Right Join

id	name	stream
1	Sayak	CS
1	Sayak	IT
2	Alex	ECE
9		ECE

ANTI JOIN



Left Table

id	name
1	Sayak
2	Alex
3	Sameer
4	Rick

Right Table

id	stream
1	CS
1	IT
2	ECE
9	ECE

Anti Join

id	name
3	Sameer
4	Rick

تولید محتوا: زهرا ذوالقدر
daychegroup
daychegroup
dayche.com | گروه دایکه