

MAHAN BUSINESS SCHOOL

MAHAN BUSINESS SCHOOL



مدیریت داده در هوش مصنوعی
مدرس: دکتر عماد کوشا
پاییز ۱۴۰۴

تغییر نقش مدیریت



آشنایی با افراد کلاس

امروز قراره یک سفر ۵ ساعته داشته باشیم. اما نگران نباشید - این یک کلاس تکنیکال نیست! قراره از زبان شما مدیران حرف بزنیم: تصمیم‌گیری، ریسک، و ROI



فهرست مطالب

فاز اول

بحران داده

- چرا ۸۷٪ پروژه‌های AI شکست می‌خورند؟
- معادله طلایی ROI
- تمرین Data Health Check سازمان شما

فاز دوم

کالبدشکافی داده

- انواع داده ها
- تشخیص گنج‌های پنهان سازمانی
- فریمورک اولویت‌بندی داده

فاز سوم

فرآیند تبدیل

- ۵ مرحله حیاتی: از Chaos به Clarity
- Normalization: چرا مهم‌ترین گام است؟
- کیس واقعی: Target Data Breach (2013)
- Point Quality Checklist

فاز چهارم

کارگاه تمرینی و نقشه راه

- Impact Demo
- Strategic Workshop
- Action Roadmap





سؤال اول: چرا ۸۷٪ پروژه‌های AI شکست می‌خورند؟



مدل‌های
ضعیف

بودجه کم

فناوری
قدیمی

تیم غیر حرفه
ای



آمار جهانی Gartner 2024

- سازمان‌ها به طور متوسط ۱۲.۹ میلیون دلار سالانه ضرر می‌کنند فقط به خاطر کیفیت پایین داده!
- ۴۷٪ تصمیمات استراتژیک بر اساس داده‌های نادرست گرفته می‌شود
- هزینه تمیزکاری داده بعد از شکست پروژه: ۱۰-۱۰۰ برابر بیشتر از قبل شروع



The Golden Equation

✗ $ROI_AI = Model\ Power \times Team\ Skill \times Budget$

✓ $ROI_AI = Model\ Power \times DQ$

Completeness

Accuracy

Consistency

Timeliness

Validity

Uniqueness



Real Case - Target Data Breach

کیس واقعی: فاجعه Target (۲۰۱۳)
When Bad Data Costs \$162 Million



دسامبر ۲۰۱۳ - Target (زنجیره خرده‌فروشی آمریکا)

● اتفاق:

- ۴۰ میلیون کارت اعتباری لو رفت
- ۷۰ میلیون رکورد مشتری سرقت شد
- هزینه‌ها: \$

- ۱۶۲ میلیون دلار غرامت مستقیم
- ۲۹۲ میلیون دلار ضرر غیرمستقیم
- CEO استعفا داد



سلامت داده سازمان شما

شماره سوال	سوال	بله	تا حدودی	خیر
۱	آیا داده‌های شما یک "منبع واحد حقیقت" دارند؟	☐	☐	☐
۲	آیا می‌دانید داده‌های شما کجا ذخیره می‌شوند؟	☐	☐	☐
۳	آیا سیاست Data Governance مکتوب دارید؟	☐	☐	☐
۴	آیا داده‌های شما به‌روز (Real-time/Daily) هستند؟	☐	☐	☐
۵	آیا تیمی مسئول کیفیت داده دارید؟	☐	☐	☐



فاز دوم – کالبدشکافی داده

مدیریت داده در هوش مصنوعی | مدرس: دکتر عماد کوشا



در سازمان شما چند نوع داده وجود دارد؟

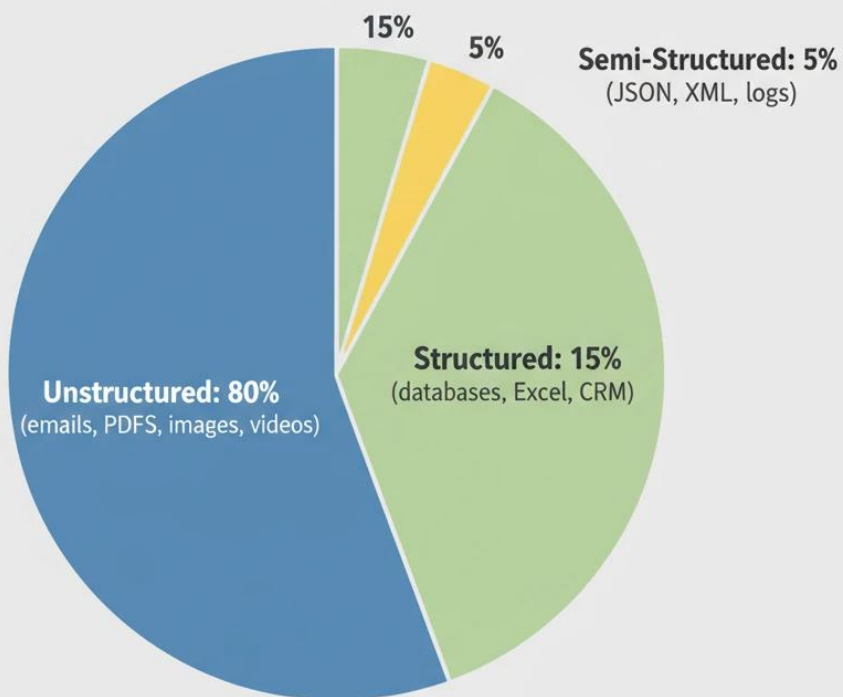


مدیریت داده در هوش مصنوعی | مدرس: دکتر عماد کوشا



دسته اصلی داده ها – دسته بندی اول: ساختار

📊 Data Distribution in Typical Organizations



نوع	ساختار	درک AI	مثال
Structured	✓	آسان	CRM, اکسل
Semi-Structured	?	متوسط	JSON, XML
Unstructured	✗	سخت	PDF, ویدیو



دسته اصلی داده ها – دسته بندی دوم: زمان

نوع	تعریف	کاربرد AI
Time Series	یک متغیر $\times$ زمان	پیش بینی روند
Cross-Sectional	چند متغیر $\times$ یک زمان	مقایسه
Panel Data	چند متغیر $\times$ چند زمان	علیت



دسته اصلی داده ها – دسته بندی سوم: مدالیت

نوع داده	تعریف	مثال کسب و کار	مدل / تکنیک AI	چالش اصلی	فرمت
Tabular	داده جدولی (ردیف/ستون)	CRM، فروش، موجودی	Random Forest، Neural Nets، XGBoost	کم	.csv، .xlsx
Text	متن طبیعی	ایمیل، نظرات، قرارداد، چت	GPT، BERT، NLP، Transformers	متوسط	.txt، .pdf، .docx
Image	تصویر ثابت	عکس محصول، اسناد اسکن، لوگو	YOLO، ResNet، CNN، Vision Transformers	زیاد	.jpg، .png، .tiff
Audio	صدا/گفتار	کال سنتر، میتینگ، پادکست	Wav2Vec، Whisper، Speech-to-Text	زیاد	.mp3، .wav، .flac
Video	تصویر متحرک	جلسات، نظارت، دموها	Action، Video CNNs، Recognition	خیلی زیاد	.mp4، .avi، .mov
Time Series	سیگنال زمان دار	سنسور، لاگ سرور، IoT، بورس	Prophet، LSTM، Temporal، ARIMA، CNNs	متوسط	.csv + timestamp
Graph	شبکه/رابطه	شبکه اجتماعی، زنجیره تأمین	Graph Neural Networks (GNN)	زیاد	.graphml، .gml
Geospatial	موقعیت مکانی	GPS، نقشه، مسیریابی	GIS، Spatial Analysis، Location Models	متوسط	.geojson، .shp
Multimodal	ترکیب چند مدال	محصول (عکس + توضیح + ویدیو)	Flamingo، CLIP، Multimodal Transformers	خیلی زیاد	ترکیبی



دسته اصلی داده ها - دسته بندی چهارم: منبع

منبع	نوع	مثال	دسترسی	کیفیت	هزینه	Use Case
Internal - Operational	عملیاتی داخلی	ERP، CRM، POS، انبار	✓ آسان	? بالا	\$ پایین	تحلیل عملکرد
Internal - Transactional	تراکنش های داخلی	فاکتورها، پرداخت ها، سفارشات	✓ آسان	? بالا	\$ پایین	تحلیل رفتار مشتری
Internal - Logs	لاگ سیستم ها	سرور، اپلیکیشن، امنیت	✓ آسان	? متوسط	\$ پایین	مانیتورینگ، امنیت
Internal - Documents	اسناد سازمانی	ایمیل، قرارداد، گزارش	? متوسط	? متوسط	\$ پایین	Knowledge Mining
External - API	سرویس های خارجی	Google، Twitter، نرخ ارز	? محدود	? بالا	\$ متوسط	Enrichment
External - Web Scraping	وب اسکرپ	قیمت رقبا، نظرات	● سخت	● پایین	\$ متوسط	رصد بازار
External - Open Data	داده عمومی	دولتی، آماری، علمی	✓ آزاد	? متوسط	\$ رایگان	بنچمارک
External - Purchased	خریداری شده	تحقیقات بازار، دیتابیس	? قراردادی	? بالا	\$ \$ \$ بالا	استراتژی
User-Generated	تولید کاربر	فرم، نظرسنجی، فیدبک	✓ کنترل شده	? متوسط	\$ متوسط	بهبود محصول
Sensor/IoT	سنسورها	دما، فشار، حرکت، موقعیت	✓ مستقیم	? بالا	\$ متوسط	Real-time Monitoring
Synthetic	مصنوعی/شبیه سازی	شبیه سازی، تولید مدل	✓ کامل	? وابسته به مدل	\$ پایین	آموزش مدل، تست



دسته اصلی داده ها – دسته بندی پنجم: حساسیت

سطح حساسیت	تعریف	مثال	دسترسی	الزامات قانونی	خطر نشت	راهکار امنیتی
Public	عمومی	آمار سایت، محصولات، قیمت عمومی	همه 🌐	ندارد ✕	کم 2	هیچ
Internal	داخلی سازمان	گزارش فروش، بودجه، استراتژی	کارمندان 🏢	ندارد ✕	متوسط 2	VPN ، Authentication
Confidential	محرمانه	اطلاعات مشتری، قراردادها، فرمول	افراد خاص 🛡️	قراردادی (NDA) ⚠️	بالاتر 2	Encryption, Access Control
Restricted	محدود شده	اطلاعات پزشکی، مالی، شناسایی	دسترسی خیلی محدود 🛡️🔑	قانونی (GDPR, HIPAA) ⚠️	بسیار بالا ●	End-to-End Encryption ، Anonymization
PII (Personal)	اطلاعات شخصی	نام، آدرس، تلفن، ایمیل	محدود 🛡️	GDPR، قوانین حریم خصوصی ⚠️	بسیار بالا ●	Pseudonymization, Consent Management
PCI (Payment)	اطلاعات پرداخت	کارت اعتباری، حساب بانکی	خیلی محدود 🛡️🔑	PCI-DSS ⚠️	بسیار بالا ●	Tokenization ، Secure Vault
PHI (Health)	اطلاعات سلامت	پرونده پزشکی، آزمایش	خیلی محدود 🛡️🔑	HIPAA ⚠️	بسیار بالا ●	De-identification ، Secure APIs
Trade Secret	راز تجاری	R&D، فرمول، الگوریتم اختصاصی	افراد کلیدی 🛡️🔑	قوانین مالکیت فکری ⚠️	بسیار بالا ●	Air-Gapped Systems, Legal Protection



جدول خلاصه: چه موقع از کدام جدول استفاده کنیم؟

سؤال کلیدی	جدول مرتبط	تصمیم‌گیری
چطور این داده رو پردازش کنم؟	جدول ۱: ساختار	انتخاب روش پیش‌پردازش
چه نوع تحلیلی می‌خوام؟	جدول ۲: بُعد زمانی	انتخاب مدل آماری
کدوم مدل AI مناسبه؟	جدول ۳: مدالیته	انتخاب معماری شبکه
از کجا داده بیارم؟	جدول ۴: منبع	تعیین استراتژی جمع‌آوری
چه کسی می‌تونه ببینه؟	جدول ۵: حساسیت	تعیین سیاست امنیتی و Governance



چارچوب اولویت‌بندی داده

$$\text{Priority Score} = (\text{Impact} \times \text{Accessibility}) / \text{Cost}$$

- تأثیر روی کسب و کار (۱-۱۰) = Impact
- سهولت دسترسی (۱-۱۰) = Accessibility
- هزینه تبدیل (۱-۱۰) = Cost

۱. قراردادهای فروش (PDF)

Impact: 9 | Accessibility: 6 | Cost: 7
Score = $(9 \times 6) / 7 = 7.7$

۲. ایمیل‌های پشتیبانی

Impact: 7 | Accessibility: 8 | Cost: 5
Score = $(7 \times 8) / 5 = 11.2$ ★

۳. ویدیوهای جلسات

Impact: 5 | Accessibility: 4 | Cost: 9
Score = $(5 \times 4) / 9 = 2.2$

۴. لاگ‌های سیستم (JSON)

Impact: 8 | Accessibility: 9 | Cost: 3
Score = $(8 \times 9) / 3 = 24.0$ 🏆



? چالش:

- میلیون‌ها ویدیو
- متادیتای ضعیف
- سیستم توصیه ضعیف

✓ راه حل:

1. تبدیل ویدیو → Structured Data

- ژانر، بازیگران، زمان
- تگ‌های خودکار
- 1. تحلیل رفتار کاربر
- چی می‌بینند؟
- کی pause می‌کنند؟

نتیجه: 📈

- افزایش ۲۵٪ در Engagement
- کاهش ۱۵٪ در Churn
- ارزش: \$+1B سالانه

! درس:

“Unstructured → Structured = طلا”

“کدام داده unstructured در سازمان شما می‌تواند طلا باشد؟”



تمرین تبدیل – پرامیت پیشنهادی

شما یک متخصص تحلیل داده و پردازش زبان طبیعی (NLP) هستید. وظیفه شما تبدیل ایمیل‌های شکایت مشتریان فروشگاه آنلاین به یک جدول ساختاریافته است.

دستورالعمل‌های دقیق:

مرحله 1: استخراج اطلاعات کلیدی

از هر ایمیل، اطلاعات زیر را استخراج کنید:

نام مشتری: نام کامل فرد (اگر ذکر نشده، بنویسید “ذکر نشده”)

شماره سفارش: کد سفارش با فرمت ORD-XXXXX

تاریخ سفارش/رویداد: تاریخ شمسی (اگر ذکر نشده، از متن استنباط کنید)

مبلغ (تومان): (مبلغ خرید) فقط عدد، بدون واحد)

محصول: نام محصول خریداری شده

نوع شکایت اصلی: یکی از این دسته‌ها انتخاب کنید:

تاخیر تحویل

محصول معیوب

عدم تطابق با توضیحات

مشکل پرداخت/استرداد

خدمات مشتری ضعیف

بسته‌بندی آسیب‌دیده

محصول گم شده

کد تخفیف/قیمت اشتباه

نقض قوانین/گزارنتی

احساس کلی: یکی از سه حالت:

بسیار منفی: زبان تند، تهدید به شکایت، استفاده از کلمات “کلاهبرداری”، “غیرقابل قبول”

منفی: ناراضی ولی مودب

خنثی: فقط بیان مشکل بدون احساس خاص

سطح اولویت: براساس شدت مشکل و احساس:



وبسایت های مجموعه داده

نوع داده	وبسایت پیشنهادی	مثال دیتاست
ساختاریافته (جدول)	Data.gov ، Kaggle	فروش فروشگاه، CRM، موجودی
غیرساختاریافته (متن)	MIMIC-IV، Kaggle	نظرات مشتری، ایمیل، گزارش
نیمه ساختاریافته (JSON/XML)	Data.gov ، GitHub	لاگ های سرور، API Response
پنل داده (Panel)	Kaggle، NYU Stern	داده های طولی چند شرکت/فرد
سری زمانی (Time Series)	UCI، Kaggle	قیمت سهام، دما، فروش روزانه
تصویر (Image)	ImageNet، Kaggle	عکس محصول، اسکن اسناد
صدا/ویدیو	Mozilla Common Voice، Kaggle	کال سنتر، پادکست



چرخه حیات در یک بانک

مرحله	مثال عملی	زمان
۱. جمع آوری	مشتری درخواست وام می دهد → فرم پر می کند	روز ۱
۲. ذخیره	اطلاعات در Core Banking System ذخیره می شود	روز ۱
۳. پردازش	سیستم Credit Scoring داده را تحلیل می کند	روز ۲-۳
۴. تحلیل	کارشناس تصمیم می گیرد: تأیید یا رد	روز ۴-۵
۵. آرشیو	بعد از ۷ سال، پرونده آرشیو می شود	سال ۷+



1. در مرحله جمع آوری:

1. آیا فیلدهای حیاتی اجباری هستند؟
2. آیا validation rule داریم؟ (مثلاً ایمیل باید @ داشته باشد)

2. در مرحله ذخیره:

1. آیا Backup روزانه داریم؟
2. آیا طبق GDPR/قوانین محلی عمل می کنیم؟

3. در مرحله پردازش:

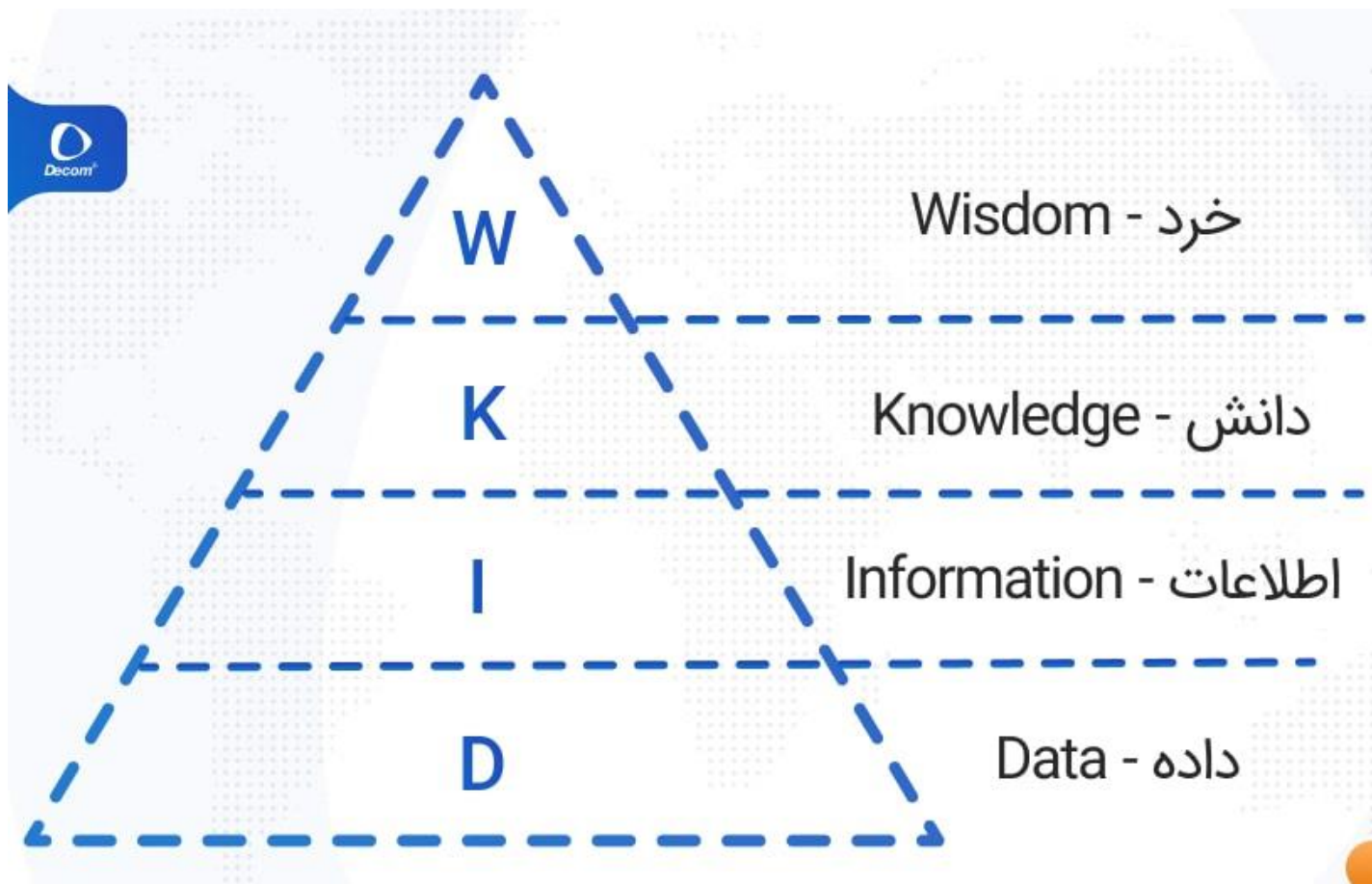
1. چه کسی مسئول تمیزکاری است؟
2. آیا documentation داریم؟ (چه تغییری روی داده انجام شد؟)

4. در مرحله تحلیل:

1. آیا نتایج قابل تفسیر هستند؟
2. آیا Bias چک شده؟

5. در مرحله حذف:

1. آیا قوانین Retention Policy داریم؟
2. آیا حذف قابل بازگشت است؟ (Soft Delete vs Hard Delete)





ویژگی	DATA	INFORMATION	KNOWLEDGE	WISDOM
سوال کلیدی	چه؟	چه چیزی؟	چرا؟	چه کار کنیم؟
مثال	500000	۵۰۰,۰۰۰ تومان	میانگین سبد خرید = K۵۰۰	تخفیف ۱۰٪ برای سبدهای بالای K۴۰۰
ارزش	پایین	متوسط	بالا	خیلی بالا
نیاز به AI	خیر	خیر	متوسط	بله
نقش مدیر	جمع‌آوری	سازماندهی	تحلیل	تصمیم‌گیری



مثال عملی اوپر	لایه
GPS راننده: (۴۰.۷۱۲۸° N, ۷۴.۰۰۶۰° W) در ساعت ۱۴:۲۳:۱۱	DATA
راننده در منهن، نیویورک است و در ترافیک گیر کرده	INFORMATION
منهن در ساعت ۲ بعدازظهر همیشه ترافیک دارد → الگو	KNOWLEDGE
تصمیم: قیمت Surge Pricing را ۱.۵x کنیم تا راننده‌های بیشتری جذب شوند	WISDOM



فاز سوم – فرآیند تبدیل



۵ مرحله تمیزکاری داده

مرحله	ورودی	عملیات	خروجی	زمان تقریبی*
۱	داده خام	شناسایی مشکلات	لیست مشکلات	10%
۲	داده با مشکل	حذف تکراری	داده Unique	15%
۳	داده Unique	پرکردن خالی	داده کامل	30%
۴	داده کامل	استانداردسازی	داده یکپارچه	25%
۵	داده یکپارچه	اعتبارسنجی	داده تمیز	20%



چک لیست مشکلات رایج

مشکل	مثال	شناسایی
Missing Values	فیلد تلفن خالی است	df.isnull().sum()
Duplicates	یک مشتری ۳ بار ثبت شده	df.duplicated()
Outliers	سن مشتری = ۲۵۰ سال!	Box Plot / Z-Score
Inconsistency	تاریخ: ۳۱/۱۰/۲۰۲۵ و ۳۱-۱۰-۲۰۲۵	Pattern Analysis
Typo	نام شهر: Tehrann	Fuzzy Matching



پرکردن مقادیر خالی

سه استراتژی:

۱- حذف (اگر فیلد حیاتی نباشد، اگر از دست دادن دیتا زیاد نباشد)

۲- پرکردن فیلد خالی

روش	مثال	کاربرد
Mean (میانگین)	سن خالی → جایگزین با میانگین کل سن‌ها	فیلدهای عددی
Median (میانه)	درآمد خالی → میانه (مقاوم به Outlier)	فیلدهای مالی
Mode (مد)	شهر خالی → پرتکرارترین شهر	فیلدهای دسته‌ای
Forward Fill	قیمت امروز خالی → کپی از دیروز	سری زمانی
Model-Based	پیش‌بینی با Regression	پیچیده

۳- نگه داشتن به عنوان یک ویژگی



استانداردسازی

نوع استانداردسازی	تعریف	مثال قبل	مثال بعد	کاربرد اصلی	ابزار/تکنیک
Format Standardization (فرمت)	یکپارچه‌سازی فرمت نمایش	31/10/2025 2025-10-31 Oct 31, 2025	همه: ۳۱-۱۰-۲۰۲۵	تاریخ، شماره تلفن، کد پستی	ISO 8601, Regex
Unit Standardization (واحد)	تبدیل به یک واحد مرجع	75 kg 165 lbs 75000 g	همه: ۷۵.۰۰ kg	وزن، طول، ارز، دما	Conversion Tables
Naming Standardization (نام‌گذاری)	یکسان‌سازی نام‌ها و برچسب‌ها	تهران Tehran TEHRA N Tehrān	همه: Tehran	شهر، استان، نام محصول	Fuzzy Matching, Mapping Tables
Case Standardization (حروف)	یکپارچه‌سازی بزرگ/کوچک	Ali ali ALI aLi	همه: ali (lowercase)	ایمیل، username، کدها	.lower() / .upper()
Statistical Standardization (آماري)	تبدیل به مقیاس یکسان	سن: ۶۰, ۴۵, ۳۰, ۲۵ درآمد: ۱۲۰M, ۸۰M, ۵۰M, ۳۰	Z-Score: -1.2, -0.5, 0.8, 1.5 -1.1, -0.3, 0.7, 1.4	Machine Learning, مقایسه متغیرها	Z-Score, Min-Max Scaling



استانداردهای رایج جهانی

کاربرد	مثال	استاندارد	حوزه
ثبت زمان سراسری	2025-10-31T14:30:00Z	ISO 8601	تاریخ و زمان
تراکنش‌های مالی	USD , EUR , IRR	ISO 4217	ارز
آدرس‌دهی جهانی	US , IR , GB	ISO 3166	کشور
چندزبانه‌سازی	en , fa , ar	ISO 639	زبان
طراحی UI/UX	#FF5733	Hex Code	رنگ
تماس بین‌المللی	+989121234567	E.164	شماره تلفن
ارتباطات الکترونیکی	user@domain.com	RFC 5322	ایمیل



مقایسه روش‌های Statistical Standardization

کاربرد	معایب	مزایا	محدوده خروجی	فرمول	روش
Machine Learning کلاسیک	- حساس به Outlier - نیاز به توزیع نرمال	- حفظ توزیع - شناسایی Outlier	$(-\infty, +\infty)$	$Z = \frac{X - \mu}{\sigma}$	Z-Score
Neural Networks	- بسیار حساس به Outlier	- محدوده مشخص - قابل تفسیر	$[0, 1]$	$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$	Min-Max
داده‌های مالی	- کمی پیچیده	- مقاوم به Outlier	نامحدود	$X' = \frac{X - Q_2}{Q_3 - Q_1}$	Robust Scaler
Sparse Data	- نادیده‌گیری میانگین	- حفظ علامت	$[-1, 1]$	$X' = \frac{X}{ X _{max}}$	Max Abs

نکته: μ = میانگین، σ = انحراف معیار، Q_1, Q_2, Q_3 = چارک‌ها



فاز سوم – فرآیند تبدیل



پایان

از توجه شما متشکرم