

عنوان ارائه:

Data Mining / داده کاوی

استاد:

دکتر کاملیا دهقانی

محقق:

امیر محمد ملاحبابائی

۱۴۰۴/۰۹/۰۱



هدف از *Data Mining* چیست؟

- هدف اصلی داده‌کاوی استخراج الگوهای مفید، ناشناخته و اقدام‌پذیر از میان حجم زیادی داده است.
- داده‌کاوی داده خام را به دانش قابل استفاده برای تصمیم‌گیری تبدیل می‌کند.
- به کشف روندهای پنهان، همبستگی‌ها و روابطی می‌پردازد که با روش‌های تحلیلی معمول قابل مشاهده نیستند.

چرا *Data Mining* مهم است؟

سازمان‌ها حجم عظیمی از داده را ذخیره می‌کنند، اما بدون تحلیل، این داده‌ها هیچ ارزشی ندارند. داده‌کاوی کمک می‌کند:

- افزایش بهره‌وری
- کاهش هزینه‌ها
- پیش‌بینی رویدادهای آینده
- شخصی‌سازی تجربه کاربر
- شناسایی ناهنجاری‌ها
- تصمیم‌گیری‌ها به جای حدس، مبتنی بر داده می‌شود.

اهداف کلیدی *Data Mining*

Prediction (پیش‌بینی): پیش‌بینی رفتار آینده مانند فروش، ریزش مشتری و تقاضا.

Clustering (خوشه‌بندی): گروه‌بندی داده‌های مشابه بدون برچسب‌های از قبل تعیین‌شده (تقسیم‌بندی مشتریان).

Classification (طبقه‌بندی): قرار دادن داده‌ها در دسته‌های مشخص مانند اسپم / غیر اسپم.

Association Rule Mining (استخراج قوانین وابستگی): پیدا کردن رابطه بین اقلام، مثل تحلیل سبد خرید.

Anomaly Detection (تشخیص ناهنجاری): تشخیص رفتار غیرعادی، مثل تراکنش‌های مشکوک.

Pattern Recognition (تشخیص الگو): شناسایی رفتارها و روندهای تکرارشونده.

مثال: amazon

چطور آمازون با داده‌کاوی ۳۵٪ فروشش را از الگوریتم *Recommendation* کسب می‌کند؟

آمازون از داده‌کاوی برای تحلیل میلیون‌ها رفتار مشتری استفاده می‌کند: تاریخچه جستجو، کلیک‌ها، خریدها و مدت زمانی که کاربر روی یک کالا می‌ماند.

سپس با ترکیب روش‌های قوانین وابستگی، خوشه‌بندی و پیش‌بینی، محصولاتی را پیشنهاد می‌دهد که احتمال خریدشان بالاست.

این کار نرخ خرید را بالا می‌برد و فروش را به شکل چشم‌گیری افزایش می‌دهد.

مثال: Walmart

پیش‌بینی لحظه‌ای تقاضا در والمارت

والمارت بیش از ۲,۵ پتابایت داده در هر ساعت متشکل از تراکنش‌ها، سنسورها و فعالیت آنلاین پردازش می‌کند.

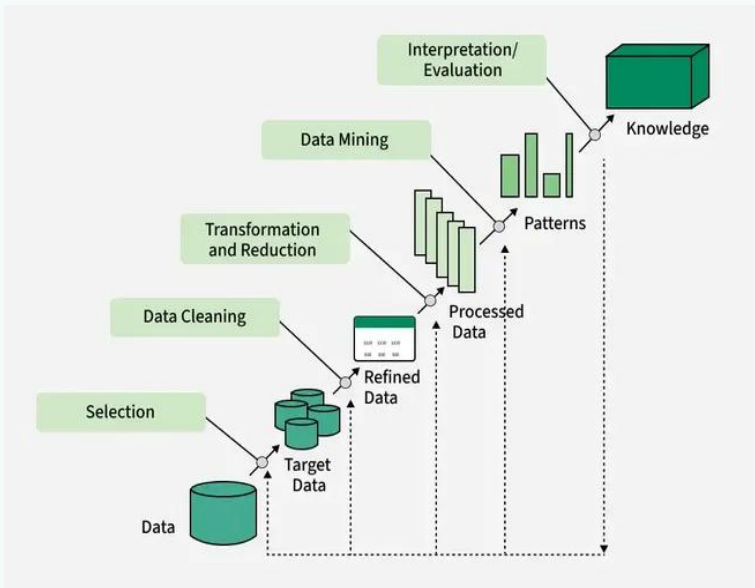
با ترکیب داده‌کاوی و تحلیل لحظه‌ای، قادر است:

- تقاضای محصولات را پیش‌بینی کند (مثلاً نزدیک طوفان ← افزایش خرید باتری و آب)
- موجودی کالا را بهینه کند
- الگوهای خرید منطقه‌ای را شناسایی کند
- لجستیک زنجیره تأمین را به‌صورت خودکار تنظیم کند

این مجموعه باعث کاهش هدررفت، جلوگیری از کمبود کالا و افزایش سود می‌شود.

Data Mining چیست؟

- داده‌کاوی فرآیند کشف الگوهای معنادار، روندها و بینش‌های ارزشمند از میان حجم بزرگی از داده‌هاست.
- داده‌کاوی یکی از مراحل اصلی در حوزه‌ی بزرگ‌تر KDD^1 (کشف دانش از پایگاه‌های داده) است.
- این حوزه ترکیبی از آمار، یادگیری ماشین و سیستم‌های پایگاه داده را به کار می‌گیرد.
- نقش اصلی آن تبدیل “داده” به “اطلاعات مفید” و “بینش اقدام‌پذیر” است.



Data Mining از کجا آمد؟

داده‌کاوی زمانی شکل گرفت که تحلیل داده‌های سنتی دیگر نمی‌توانست با حجم و پیچیدگی داده‌ها مقابله کند: رشد انفجاری داده‌ها - داده‌های پیچیده و چندبعدی - نیاز به تصمیم‌گیری لحظه‌ای - نیاز به اتوماسیون در کشف الگوها

مهم‌ترین نقاط تاریخی:

دهه ۱۹۸۰: توسعه پایگاه‌داده‌های رابطه‌ای

دهه ۱۹۹۰: ظهور الگوریتم‌های ماشین لرنینگ برای کشف الگو

دهه ۲۰۰۰ به بعد: Big Data، پردازش توزیع‌شده، ذخیره‌سازی ابری

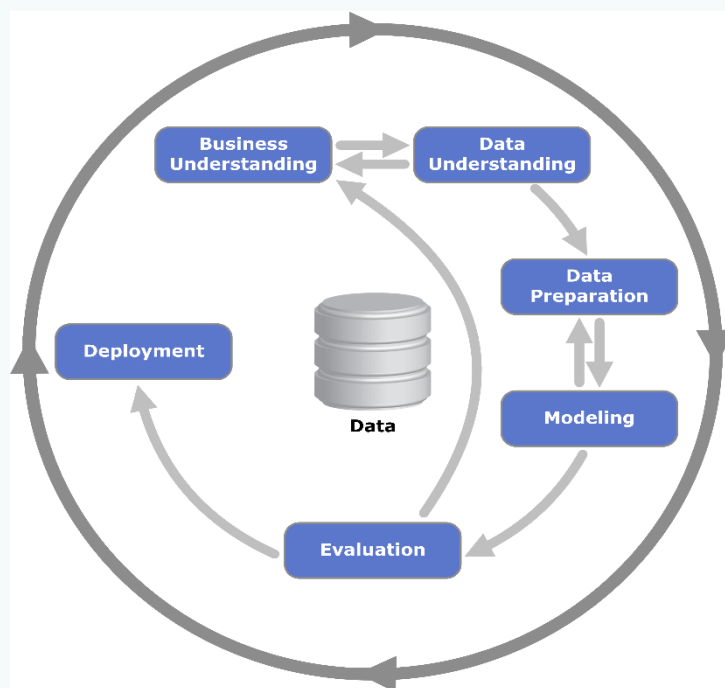
امروزه: ادغام با هوش مصنوعی، دیپ لرنینگ، و تحلیل لحظه‌ای

Data Mining در کجای چرخه‌ی داده قرار می‌گیرد؟

داده‌کاوی تنها یکی از مراحل چرخه‌ی بزرگ تحلیل داده است:

1. جمع‌آوری داده (Data Collection): سنسورها، تراکنش‌ها، لاگ‌ها، فعالیت وب
2. پاکسازی داده (Data Cleaning): حذف خطاها، داده‌های گم‌شده و نویز
3. یکپارچه‌سازی (Data Integration): ترکیب منابع مختلف
4. انتخاب ویژگی (Data Selection): انتخاب ویژگی‌های مرتبط
5. تبدیل داده (Data Transformation): نرمال‌سازی، کدگذاری، استخراج ویژگی
6. Data Mining: کشف الگوها و بینش‌ها
7. ارزیابی و تفسیر (Evaluation): اعتبارسنجی نتایج
8. دانش و تصمیم‌گیری (Knowledge): استفاده‌ی عملی از نتایج

فرآیند کلی *Data Mining*



مراحل معمول داده‌کاوی:

- 1) درک کسب‌وکار
- 2) درک داده‌ها
- 3) آماده‌سازی داده‌ها
- 4) مدل‌سازی
- 5) ارزیابی
- 6) استقرار

CRISP-DM رایج‌ترین استاندارد جهانی برای انجام پروژه‌های داده‌کاوی است.

مرحله ۱: درک کسب‌وکار

- شناسایی اهداف و نیازهای کسب‌وکار
- تبدیل اهداف کسب‌وکار به مسئله‌های داده‌کاوی
- تعریف معیارهای موفقیت

مثال: شرکت مخابرات می‌خواهد ریزش مشتریان را کاهش دهد.
مسئله داده‌کاوی: پیش‌بینی مشتریانی که احتمالاً خدمات را ترک می‌کنند.

مرحله ۲ و ۳: درک داده‌ها و آماده‌سازی داده‌ها

درک داده‌ها:

- جمع‌آوری داده‌ها (تراکنش‌ها، لاگ‌ها)
- بررسی داده‌ها با آمار و نمودار
- شناسایی داده‌های گمشده و مشکلات کیفیت

آماده‌سازی داده‌ها:

- پاک‌سازی داده‌ها (مدیریت داده‌های گمشده، حذف نویز)
- انتخاب و مهندسی ویژگی‌ها
- تبدیل داده‌ها (نرمال‌سازی، کدگذاری)

مرحله ۴: مدل سازی

- انتخاب الگوریتم‌ها
- آموزش مدل روی داده‌ها
- آزمایش و تنظیم چند مدل

مرحله ۵ و ۶: ارزیابی و استقرار

ارزیابی:

- بررسی عملکرد مدل با معیارهای Accuracy، Precision، Recall، F1، RMSE
- اطمینان از رسیدن مدل به اهداف کسب‌وکار

استقرار:

- ادغام مدل در سیستم‌های عملیاتی
- پایش عملکرد و به‌روزرسانی مدل
- مثال‌ها: سیستم‌های پیشنهاد دهنده، تشخیص تقلب، پیش‌بینی ریزش مشتری

مثال واقعی: **NETFLIX** - مدل‌سازی ترجیحات کاربران

- داده‌ها (**Data**): رفتار تماشای کاربران (فیلم‌ها و سریال‌های دیده شده، زمان مشاهده، جستجوها)
- مدل‌سازی (**Modeling**): استفاده از طبقه‌بندی (**Classification**) و خوشه‌بندی (**Clustering**) برای پیش‌بینی علایق کاربران
- ارزیابی (**Evaluation**): بررسی میزان دقت مدل در پیش‌بینی فیلم مورد علاقه کاربران
- استقرار (**Deployment**): سیستم Recommendation Engine که فیلم و سریال مناسب را به کاربران پیشنهاد می‌دهد

هدف: افزایش رضایت کاربران و مدت زمان مشاهده محتوا با ارائه پیشنهادهای دقیق و شخصی‌سازی شده.

معرفی الگوریتم‌ها

الگوریتم‌های داده‌کاوی مجموعه‌ای از روش‌ها هستند که به ما کمک می‌کنند از داده‌ها الگو، دانش و پیش‌بینی استخراج کنیم.

مهم‌ترین دسته‌بندی‌ها:

1. Classification (طبقه‌بندی)
2. Clustering (خوشه‌بندی)
3. Association Rules (قوانین انجمنی)
4. Regression (رگرسیون)
5. Anomaly Detection (تشخیص ناهنجاری)

Classification (طبقه‌بندی)

مفهوم:

- هدف: پیش‌بینی یک برچسب (Label) از پیش‌تعریف شده
- خروجی مدل ← Yes/No یا انواع دسته‌ها
- مثال‌ها: تشخیص ایمیل اسپم - پیش‌بینی ریزش مشتری - تشخیص بیماری‌ها

مهم‌ترین الگوریتم‌ها:

- Decision Tree (درخت تصمیم)
- Naive Bayes (بیز ساده)
- Random Forest
- SVM

مثال واقعی: بانک‌ها از طبقه‌بندی برای پیش‌بینی اینکه آیا یک مشتری وام را بازپرداخت می‌کند یا نه استفاده می‌کنند.

Clustering (خوشه‌بندی)

مفهوم:

- هدف: گروه‌بندی داده‌ها بر اساس شباهت‌ها
- بدون برچسب (Unsupervised)
- کاربرد: تقسیم مشتری‌ها به گروه‌های مشابه

مهم‌ترین الگوریتم‌ها:

- k-Means
- Hierarchical Clustering
- DBSCAN

مثال واقعی: نتفلیکس از خوشه‌بندی برای گروه‌بندی کاربران با رفتار مشابه مشاهده استفاده می‌کند سپس نتایج آن به سیستم پیشنهاد دهنده کمک می‌کند.

Association Rules (قوانین انجمنی)

مفهوم:

- هدف: کشف روابط بین آیتم‌ها در مجموعه داده
- خروجی: "وقتی X خریداری شود، Y هم احتمال خرید دارد"

مهم‌ترین الگوریتم‌ها:

- Apriori
- FP-Growth

مثال واقعی (معروف‌ترین): آنالیز سبد خرید آمازون

- اگر مشتری "لپ‌تاپ" بخرد $\leftarrow$ احتمال خرید "کول‌پد" و "ماوس" بالا می‌رود
- این تحلیل پایه اصلی «Customers Also Bought» است.

Regression (رگرسیون)

مفهوم:

هدف: پیش‌بینی یک مقدار عددی

کاربردها:

- پیش‌بینی قیمت
- پیش‌بینی مقدار فروش
- پیش‌بینی مصرف انرژی

مهم‌ترین الگوریتم‌ها:

- Linear Regression (رگرسیون خطی)
- Logistic Regression (لجستیک – برای Yes/No)
- Lasso / Ridge

مثال واقعی: اسنپ از رگرسیون برای پیش‌بینی تقاضای سفر در ساعات مختلف روز استفاده می‌کند.

Anomaly Detection (تشخیص ناهنجاری)

مفهوم:

- پیدا کردن داده‌هایی که غیرطبیعی، پرت یا مشکوک هستند
- کمک می‌کند حمله یا اشتباهات سیستم شناسایی شود

مهم‌ترین الگوریتم‌ها:

- Isolation Forest
- One-Class SVM
- Autoencoders در (Deep Learning)

مثال واقعی: مسترکارد از این روش برای تشخیص تراکنش‌های مشکوک استفاده می‌کند.

کاربردهای عملی *Data Mining*

داده‌کاوی تقریباً در تمام صنایع استفاده می‌شود.
هدف: تصمیم‌گیری هوشمند، افزایش سود، کاهش خطا و پیش‌بینی دقیق آینده.
در ادامه چند نمونه پرکاربرد در دنیای واقعی ارائه می‌شود.



کاربرد در تجارت الکترونیک

- تحلیل سبد خرید مشتریان
- پیشنهاد محصول بر اساس رفتار خرید
- تشخیص کاربران در معرض ترک خرید
- پیش‌بینی فروش آینده

مثال واقعی – آمازون:

- با تحلیل تاریخچه خرید، محصولات مرتبط را پیشنهاد می‌کند.
- باعث می‌شود بسیاری از خریدها افزوده‌شده (*Add-on*) باشند.

کاربرد در سرویس‌های محتوایی

- پیشنهاد سریال و فیلم بر اساس سابقه تماشا
- یافتن سلیقه کاربر با تحلیل الگوهای تماشا
- دسته‌بندی کاربران برای شخصی‌سازی تجربه

مثال واقعی – نتفلیکس:

با تحلیل زمان تماشا، توقف‌ها و نوع ژانر مورد علاقه، محتواهای مشابه پیشنهاد می‌دهد.

کاربرد در بخش مالی

- تشخیص تراکنش‌های مشکوک و تقلبی
- پیش‌بینی ریسک و اعتبار کاربران
- تحلیل رفتار مشتریان بانکی
- شناسایی مشتریانی که احتمال ترک بانک دارند

مثال واقعی:

بانک‌ها با الگوریتم‌های تشخیص تقلب، تراکنش‌های غیرعادی را در چند ثانیه مسدود می‌کنند.

کاربرد در حوزه سلامت

- پیش‌بینی بیماری‌ها بر اساس علائم و سابقه
- تحلیل داده‌های بزرگ بیمارستانی
- کمک به تشخیص زودهنگام
- کاهش خطاهای انسانی

مثال واقعی:

مدلهایی که احتمال ابتلا به دیابت یا بیماری قلبی را پیش‌بینی می‌کنند.

نمونه واقعی: پیشنهاد محصول مشابه

```
import pandas as pd
from sklearn.metrics.pairwise import cosine_similarity

data = {
    "User":      ["A","A","A","B","B","C","C","D","D"],
    "Product":   ["Laptop","Mouse","Headphone","Laptop","Headphone","Mouse","Keyboard",
                  "Laptop","Keyboard"]
}

df = pd.DataFrame(data)
user_item = pd.crosstab(df["User"], df["Product"])

similarity = pd.DataFrame(
    cosine_similarity(user_item.T),
    index=user_item.columns,
    columns=user_item.columns
)

def recommend(product, similarity_matrix, top_n=3):
    scores = similarity_matrix[product].sort_values(ascending=False)
    return scores.iloc[1:top_n+1]

print("Similar to 'Laptop':")
print(recommend("Laptop", similarity))
```

خروجی کد:

Similar to 'Laptop':

Keyboard 0.66

Headphone 0.40

Mouse 0.33

References / منابع

- Data Mining Concepts and Techniques - Jiawei Han and Micheline Kamber (2003)
- Data Mining Techniques in Customer Churn Prediction - Tsai, C.-F., & Lu, Y.-H. (2010)
- The applications of Data Mining - Masoud Kaviani (2019)
- ChatGPT (OpenAI) - for conceptual support.