

خلاصه کتاب مبانی و روش‌های داده‌کاوی

انتشارات: دانشگاه پیام نور نویسندگان: سیده فاطمه نورانی، سیمین قاسمی، جعفر پورامینی

مقدمه و کلیات

داده‌کاوی به عنوان فرآیند استخراج دانش از داده‌های حجیم، نقشی حیاتی در تصمیم‌گیری‌های هوشمند ایفا می‌کند. این کتاب با رویکردی آموزشی، مفاهیم پایه‌ای و روش‌های عملیاتی این حوزه را تشریح می‌کند.

فصل اول: آشنایی با مفاهیم داده‌کاوی

- تعریف:** استخراج الگوهای پنهان، معتبر و قابل فهم از پایگاه‌های داده بزرگ.
- فرآیند KDD:** شامل مراحل پاک‌سازی داده، یکپارچه‌سازی، انتخاب، تبدیل، داده‌کاوی، ارزیابی الگو و نمایش دانش.
- انواع داده‌ها:** داده‌های رابطه‌ای، انبار داده، داده‌های تراکنشی و داده‌های پیچیده (جریانی، متنی و وب).

فصل دوم: پیش‌پردازش داده‌ها

- پاک‌سازی:** مدیریت مقادیر مفقود (Missing Values) و نویز.
- یکپارچه‌سازی:** ترکیب داده‌ها از منابع مختلف و رفع تداخلات.
- کاهش داده:** استفاده از روش‌هایی مانند تحلیل مولفه‌های اصلی (PCA) برای کاهش ابعاد.
- تبدیل داده:** نرمال‌سازی (Min-Max, Z-score) و گسسته‌سازی.

فصل سوم: انبار داده و تحلیل OLAP

- مفهوم انبار داده:** مجموعه‌ای موضوع‌گرا، یکپارچه، متغیر با زمان و غیرقابل تغییر.
- مدل‌سازی داده:** طرح‌های ستاره‌ای (Star Schema)، دانه‌برفی (Snowflake) و منظومه ستاره‌ای.
- عملیات OLAP:** شامل Roll-up، Drill-down، Slice و Dice.

فصل چهارم: استخراج الگوهای مکرر و قوانین انجمنی

- الگوریتم Apriori:** شناسایی مجموعه‌آیتم‌های مکرر بر اساس خاصیت زیرمجموعه.
- الگوریتم FP-Growth:** روشی کارآمدتر بدون نیاز به تولید کاندیدا با استفاده از درخت FP.
- معیارهای ارزیابی:** حمایت (Support) و اطمینان (Confidence).

فصل پنجم: طبقه‌بندی و پیش‌بینی

- **درخت تصمیم:** استفاده از الگوریتم‌هایی مانند ID3 و C4.5 بر اساس شاخص بهره اطلاعات (Information Gain).
- **بیز ساده (Naive Bayes):** بر پایه تئوری احتمال بیز با فرض استقلال ویژگی‌ها.
- **شبکه‌های عصبی:** مدل‌سازی توابع پیچیده با استفاده از لایه‌های ورودی، پنهان و خروجی.
- **ارزیابی مدل:** ماتریس اغتشاش (Confusion Matrix)، دقت (Accuracy)، فراخوانی (Recall) و معیار F1.

فصل ششم: تحلیل خوشه‌بندی

- **روش‌های افرازی:** الگوریتم k-Means و k-Medoids.
- **روش‌های سلسله مراتبی:** روش‌های تجمعی (Agglomerative) و تقسیمی (Divisive).
- **روش‌های مبتنی بر چگالی:** مانند DBSCAN برای شناسایی خوشه‌های با شکل نامنظم.

فصل هفتم: کاربردها و روندهای آینده

- تحلیل شبکه‌های اجتماعی، داده‌کاوی متنی، تشخیص تقلب و کاربرد در پزشکی و تجارت الکترونیک.
- چالش‌های حریم خصوصی و داده‌کاوی روی داده‌های عظیم (Big Data).