Explorarea Datelor

Capitolul 4 – Analiză și Vizualizare RELAȚIA DINTRE PREPROCESAREA DATELOR ȘI EXPLORAREA DATELOR Preprocesarea datelor și explorarea datelor sunt etape complementare și interdependente ale procesului de analiză a datelor.

Preprocesarea datelor se concentrează pe pregătirea și curățarea datelor brute, prin activități precum eliminarea valorilor lipsă și a anomaliilor, transformarea datelor și eliminarea duplicatelor. Scopul este de a asigura că datele sunt curate și adecvate pentru analiză, prevenind rezultate inexacte sau înșelătoare.

Explorarea datelor (Exploratory Data Analysis - EDA) implică o analiză preliminară pentru a înțelege structura și relațiile din date, utilizând vizualizări și statistici descriptive. Această etapă ajută la formularea ipotezelor și identificarea tendințelor sau anomaliilor, ghidând analiza ulterioară.

Relația dintre ele este una de interdependență: preprocesarea este necesară pentru ca datele să fie ușor de explorat, iar explorarea poate dezvălui probleme care necesită ajustări suplimentare în preprocesare. Acest proces iterativ asigură pregătirea datelor pentru analiză și modelare ulterioară, conducând la rezultate mai precise și relevante.

Importanța Explorării Datelor

- Proces de analiză a datelor pentru identificarea modelelor, tendințelor și anomaliilor.
- Tehnici: statistice și grafice, tabele pivot, agregare.

 Primul pas esențial în analiza datelor.

Etapele Explorării Datelor

- 1. Încărcarea datelor
- 2. Curățarea datelor
- 3. Analiza descriptivă
- 4. Vizualizarea datelor
- 5. Identificarea outlier-ilor
- 6. Transformarea datelor
- 7. Verificarea ipotezelor

Curățarea Datelor

 Eliminarea erorilor, valorilor lipsă și duplicate.

• Îmbunătățirea calității și consistenței datelor.

Analiza Descriptivă

- Explorarea statisticilor esențiale: distribuție, tendință, variabilitate.
- Metode: tabele,
 histograme, boxploturi, grafice de dispersie.
- Scop: Oferirea unei imagini clare a datelor.

Vizualizarea Datelor

 Reprezentare grafică a datelor: clară și concisă.

• Tehnici: histograme, grafice de dispersie, heatmap-uri, boxploturi.

Detectarea
 valorilor anormale.

Identificarea Outlier-ilor

• Metode: boxploturi, Z-score, IQR, grafice de dispersie. Schimbarea formei datelor pentru analiză mai ușoară.

Transformarea Datelor

 Exemple: normalizare, escalare, aplicarea funcțiilor matematice.

Analiza Distribuției

- Examinarea formei şi proprietăților probabilistice ale datelor.
- Tipuri comune: normală, Poisson, binomială.
- Metode: histogramă, grafic de densitate, boxplot, Q-Q plot.

Corelația între Variabile

- Definirea relației între două variabile: pozitivă, negativă, nulă, parțială.
- Coeficienți de corelație: Pearson, Spearman, Kendall.

Vizualizarea Datelor (Detaliat)

- Tipuri de grafice:
- Bară
- Linie
- Scatter Plot
- Box Plot
- Hărți de densitate.

Agregare și Grupare

 Combinarea datelor pentru o imagine clară.

 Tehnici: grupare pe valori, intervale, pivotare, clustering.

Statistici Descriptive

 Măsuri de centralitate: medie, mediană, mod.

 Măsuri de dispersie: deviație standard, coeficient de variație. Aplicații Practice ale Explorării Datelor 1. Business Intelligence

2. Marketing

3. Cercetare

4. Healthcare

5. Finanțe

Exemplu Practic (Retail)

- Utilizare pentru înțelegerea comportamentului cumpărătorilor.
- Analiza vânzărilor, segmentarea clienților.

Concluzie

- Explorarea datelor ajută la înțelegerea detaliată a datelor.
- Primul pas esențial în analiza complexă și deciziile bazate pe date.