Laborator1.

Aspecte teoretice

Curățare date

Principalele obiective ale curățării datelor sunt:

1. Înlocuire (sau eliminare) valorile lipsă,
2. Netezirea datelor zgomotoase
3. Ștergere sau identificare a datelor outliers
4. Valori NULL: Unele atribute au voie să conțină o valoare NULL.

În aceste cazuri, valoarea stocată în baza de date (sau valoarea atributului din setul de date) trebuie să fie ceva de genul „Nu se aplică” și nu o valoare NULL.

1. Valoarea care lipsește- lucru se poate face în mai multe moduri:

1.1. Completare manual- opțiune nu este posibilă în majoritatea cazurilor datorită volumului uriaș al seturilor de date care trebuie curățate.

1.2. Completare cu o valoare (distinctă de altele) - „Nu este disponibil” sau „necunoscut”.

1.3. Completare cu o valoare care măsoară tendința central. Ex: mean, median or mode.

1.4.Completare cu o valoare care măsoară tendința centrală,

1.5. Cea mai probabilă valoare, dacă această valoare poate fi determinată, prin arbori de decizie, maximizarea așteptărilor (EM), Bayes etc.

2. Zgomotul poate fi definit ca o eroare sau variație aleatorie într-o variabilă măsurată ([Han, Kamber 06]).

Wikipedia definește zgomotul ca o expresie usuală pentru cantități recunoscute de variație inexplicabilă într-un eșantion.

Pentru îndepărtarea zgomotului, se pot utiliza câteva tehnici de netezire:

1. Regresie

2. Binning

Binning poate fi utilizat pentru netezirea unui set ordonat de valori. Netezirea se face pe baza valorilor vecine. Există doi pași:

1. Împărțirea datelor ordonate in câteva seturi de date(cos) egale ca dimensiune

2. Netezirea pentru fiecare coș: valorile dintr-o coș sunt modificate pe baza unor caracteristici ale coșului: average, median, mod

outlier

este o valoare a atributului îndepărtată numeric de restul datelor.

Valorile extreme pot fi uneori valori corecte: de exemplu, salariul directorului general al unei companii poate fi mult mai mare decât toate celelalte salarii. Dar, în majoritatea cazurilor, valorile extreme sunt și trebuie gestionate ca zgomot.

Outliers trebuie identificați și apoi eliminați (sau înlocuiți, ca orice altă valoare zgomotoasă), deoarece mulți algoritmi de extragere a datelor sunt sensibili la valori superioare.

De exemplu, orice algoritm care folosește media aritmetică (una dintre ele este k-means) poate produce rezultate eronate, deoarece media este foarte sensibilă la valori superioare

Utilizarea IQR: valori mai mari de 1,5 x IQR sub Q1 sau peste Q3 sunt valori externe potențiale.

Boxurile pot fi utilizate pentru a identifica aceste valori exterioare (boxploturi)

sunt o metodă pentru reprezentarea grafică a dispersiei datelor).

Utilizarea abaterii standard: sunt de asemenea valori care sunt mai mult de două abateri standard față de media pentru un atribut dat outliers potențiali.

Clustering. După gruparea unui anumit set de date unele puncte sunt în afara oricărui grup (sau departe)

departe de orice centru de cluster.

II. Integrare Date

Integrarea datelor înseamnă fuzionarea datelor din diferite surse de date într-un set de date coerent.

Principalele activități sunt:

- Integrarea schemei
- Eliminare duplicatele și redundanța
- Manevrare inconsecvenței

Trebuie să se identifice traducerea fiecărei scheme sursă în schema finală (problema identificării entității) Subprobleme:

Același lucru se numește diferit în fiecare bd

Exemplu: ID-ul clientului poate fi numit Cust-ID, Cust #, Custid, CID în diferite surse.

Lucruri diferite sunt denumite cu același nume în surse diferite.

Exemplu: pentru datele angajaților, atributul „Oraș" poate înseamnă într-o sursă orașul unde se află și într-o altă sursă orașul de naștere

Date Duplicate:

aceleași informații pot fi stocate în multe surse de date. Fuzionarea lor poate provoca uneori duplicări ale informațiilor respective: ca atribut duplicat (același atribut cu nume diferite este găsit de mai multe ori în rezultatul final) sau ca instanță duplicat (același obiect / entitate se găsește de mai multe ori în baza de date finală).

Aceste duplicate trebuie identificate și eliminate..

Date redundant

Redundanță: - informații pot fi deduse / calculate.

- De exemplu, vârsta poate fi dedusă de la data nașterii, salariul anual poate fi calculat din salariul lunar și alte bonusuri înregistrate pentru fiecare angajat.
- Redundanța trebuie eliminată din setul de date înainte de a rula algoritmul de extragere a datelor

• Depozitele de date existente este permisă o anumită redundanță.

Aspecte practice

Exercitiu curatarea datelor

Executare Tableau Prep – primele 3 video tutoriale

Tabelul de mai jos contine datele culese de proprietarul unui magazin din formularele inmanate clientilor sai.

1. Identificati problemele existente in cadrul acestei relatii.
2. Oferiti o schema a relatiei cu restrictii de integritate care sa duca la eliminarea problemelor
3. Ce alte reguli ati adauga pentru evitarea acestor erori?

CUSTNO	NAME	PHONE NO	ISSUEDATE	SEX	BIRTHDATE	PURCHASED
10	CHRISTINE I HAAS	3978	01/01/05	F	08/24/05	52750
20	MICHAEL L THOMPSON	3476	02/12/05	M	02/02/88	41250
30	DANIEL S SMITH	961	02/23/05	M	11/12/79	19180
50	SALLY A KWAN	4738	03/03/05	F	05/11/81	38250
60	JOHN B GEYER	6789	03/24/05	M	09/15/65	40175
70	N/A	6423	04/05/05	M	07/07/85	32250
90	EVA D PULASKI	7831	04/11/05	F	05/26/53	36170
100	EILEEN W HENDERSON	5498	05/05/05	F	05/15/81	29750
110	THEODORE Q SPENSER	972	05/16/05	M	12/18/56	26150
120	VINCENZO G LUCCHESSI	3490	05/30/05	M	11/05/69	46500
130	SEAN O'CONNELL	M	06/19/05	M	10/18/82	29250
140	DAVID BROWN	4501	06/19/05	M	05/29/81	27740
150	DOLORES M QUINTANA	4578	07/07/05	F	09/15/65	23800
160	HEATHER A NICHOLLS	1793	07/07/05	F	01/19/86	28420
170	BRUCE ADAMSON	4510	07/26/05	M	05/17/87	25280
170	ELIZABETH R PIANKA	3782	07/28/05	F	14/24/55	22250
190	MASATOSHI J YOSHIMURA	2890	08/07/05	M	01/05/51	24680
200	MARIA L PERES	9001	08/15/05	F	05/26/53	27380
210	MARILYN S SCOUTTEN	1682	08/17/05	F	02/21/89	21340
220	JAMES H WALKER	2986	08/29/05	M	06/25/52	20450
230	DAVID BROWN	4501	09/11/05	M	05/29/81	27740
240	WILLIAM T JONES	942	09/12/05	M	02/23/53	18270
250	JENNIFER K LUTZ	672	09/14/05	F	03/19/00	29840
260	JAMES J JEFFERSON	2094	09/15/05	M	05/30/75	22180
270	SALVATORE M MARINO	3780	09/15/05	M	03/31/54	28760
280	DANIEL S SMITH	961	09/30/05	M	11/12/79	19180
290	SYBIL P JOHNSON	8953	09/30/05	F	10/05/76	17250
300	MARIA L PEREZ	9001	09/30/05	F	05/26/53	27380
310	ETHEL R SCHNEIDER	8997	09/30/05	F	03/28/76	26250
320	JOHN R PARKER	4502	10/10/05	M	07/09/86	15340
330	PHILIP X SMITH	2095	10/11/05	H	10/27/76	17750
340	MAUDE F SETRIGHT	3332	10/30/05	F	04/21/71	15900
360	RAMLAL V MEHTA	9990	11/21/05	M	08/11/72	19950
380	WING LEE	2103	12/05/05	M	07/18/81	25370000
400	JASON R GOUNOT	5698	12/05/05	M	05/17/66	23840

Instalare

https://www.tableau.com/products/prep/download

Please forward these instructions to your students:

Business Intelligence

- Download the latest version of Tableau Desktop and Tableau Prep Builder here

- Click on the link above and select "Download Tableau Desktop" and "Download Tableau Prep Builder". On the form, enter your school email address for Business E-mail and enter the name of your school for Organization.
- Activate with your product key: TCZE-1EB9-6190-77FD-E011
- Already have a copy of Tableau Desktop installed? Update your license in the application: Help menu -> Manage Product Keys