Fundamente de Big Data

Proiect

Termen limita de predare: cel puțin trei zile înaintea datei de examen anunțată în sesiunea de examen

Privire de ansamblu

Obiectivul acestui proiect este sa permită studenților sa experimenteze întreg ciclul de cercetare asociat studiului unei probleme de afaceri, pornind de la identificarea unei întrebări de cercetare interesante, obținerea setului de date, pregătirea acestuia pentru analiza, executarea sarcinii de analiza a datelor utilizând mai multe metode alternative, extragerea concluziilor si pregătirea unui raport de comunicare a rezultatelor.

Pregătirea sarcinii de lucru

Pentru acest proiect sugerăm studenților sa lucreze in RStudio, folosind limbajul de programare R. Studenții pot sa utilizeze si alte medii de programare dedicate R, sau să folosească Python ca si limbaj alternativ.

Pentru editarea raportului final, sugestia este sa folosiți Microsoft Word.

Studenții vor lucra in echipe de maxim câte 3, cu mențiunea ca toti studenți făcând parte dintr-o echipă trebuie sa cunoască în detaliu tot proiectul, incluzând aici toate experimentele executate in mediul de programare precum si concluziile desprinse de pe urma execuției acestor experimente. Proiectele vor fi susținute oral, si membrii echipei vor răspunde individual întrebărilor puse de examinatori. Fiecare student membru al unei echipe trebuie să fie capabil să răspundă la întrebări care vizează tot proiectul trimis spre evaluare. În funcție de modul in care studenții reușesc sa răspundă la întrebările puse pe parcursul susținerii, ei pot să contribuie pozitiv sau negativ la notarea finală a proiectului. Notarea finală este individuală pentru fiecare student, si nu pe proiect. Studenții care nu participă la susținerea proiectului, chiar dacă proiectul a fost predat pe Moodle, se consideră a fi absenți la examen.

Studenții înrolați la învățământ la distanță pot să lucreze și in echipe de 2 studenți sau individual.

Proiectul trebuie predat pe Moodle la termenul specificat, cu cel puțin **trei zile** înaintea datei alese pentru prezentarea la examen.

Instrucțiuni detaliate

In acest proiect aveți de îndeplinit mai multe sarcini.

1. Prima sarcina este sa identificați una sau mai multe întrebări de cercetare de interes pentru o audienta de afaceri.

2. Să identificați si obțineți un set de date care sa va permită sa răspundeți la întrebarea de cercetare pusa la pasul 1. Setul de date trebuie pregătit de așa maniera încât sa puteți aplica metode științifice de analiza a datelor pentru a obține modele alternative
3. Pentru întrebările de cercetare alese si setul de date pregătit in pasul 2 veți selecta mai multe metode de analiza a datelor si veți conduce experimente cu aceste metode pentru a crea modele alternative. Pe baza metodologiei de validare potrivită, veți determina metoda si modelul cel mai potrivit
4. Pe baza experimentelor realizate la pasul 3 veți extrage concluzii de business, încercând sa răspundeți cat mai clar si mai convingător la întrebările stabilite la pasul inițial.

Pentru punctele menționate anterior sunt de interes atât setul de date, codul sursa care conține experimentele executate cat mai ales raportul care prezinta întrebările si concluziile extrase in urma experimentelor.

Sfat: Puteți consulta secțiunea Resurse a acestui document pentru a desprinde idei posibile de întrebări de cercetare, precum si seturi de date relevante pentru acestea.

Structura raportului pe care trebuie sa îl realizați este următoarea:

- **Introducere.** Reprezintă prima secțiune a raportului. Aceasta trebuie sa furnizeze informații contextuale despre aria de cercetare aleasă, sa identifice in mod clar întrebările de cercetare alese, sa explice de ce aceste întrebări sunt relevante si importante si care este audienta care va beneficia de pe urma studiului, si daca întrebările au fost abordate in trecut, care sunt rezultatele altor studii răspunzând la aceleași întrebări sau lucrând pe același set de date. Pentru introducere, trebuie sa fiți concișI, si in general, sa finalizați introducerea in maxim o pagina. Daca este necesar mai mult spațiu, aveți liberate de expresie in acest sens.
- **Setul de date.** Folosiți câteva paragrafe ca să explicați setul de date utilizat. Precizați de ce acest set de date este relevant pentru întrebările de cercetare alese, descrieți pașii realizați pentru curățarea datelor sau preprocesarea acestora, precum si caracteristicile de baza ale datelor, înainte ca acestea sa intre in procesul de analiza. Puteți folosi tehnici de vizualizare a datelor si sa printați grafice relevante, cu mențiunea ca acestea trebuie explicate pentru a face raportul inteligibil.
- **Rezultate si discuții.** Aceasta reprezintă partea cea mai importanta a raportului. Trebuie sa prezentați in detaliu analiza realizata. Prezentați care sunt metodele de analiza alese, ce setări ați testat pentru aceste metode, care a fost strategia de validare selectata, ce rezultate ați obținut pentru metodele selectate si cu parametrii testați. Comparați rezultatele obținute, si pe baza acestora precizați care este metoda și modelul final considerat si cum răspunde acesta întrebărilor de cercetare descrise in prima parte a raportului. In aceasta parte a raportului, puteți motiva alegere rile realizate prin fraze de tipul: “am folosit metoda deoarece …”, etc. Aceasta secțiune trebuie sa fie o combinație de text, tabele si grafice. Este absolut necesar sa descrieți si interpretați rezultatele, nu doar sa le afișați in tabele sau grafice. De asemenea, trebuie sa discutați limitările acestor rezultate, daca credeți ca puteați

obține rezultate mai bune si ce anume v-a împiedicat in acest sens. Puteți sa creați subsecțiuni care sa structureze mai bine aceasta parte a raportului.

- **Concluzia.** Cel mult 2 paragrafe in care sa se sintetizeze întrebările de cercetare formulate precum si rezultatele obținute. Aceasta secțiune trebuie sa fie scurta si concise, însă nu trebuie sa supraliciteze (adică sa extragă concluzii mai puternice decât cele obținute din analiza si justificate in secțiunea precedenta)

Întregul raport va avea intre 10 si 15 pagini, va fi redactat cu fontul Times New Roman de 12 caractere si spațiere maxima de 1.2 intre rânduri.

Pe lângă raportul realizat, va trebui sa furnizați si setul de date, precum si codul / codurile sursa folosite pentru procesarea si analiza datelor.

Pe Moodle veți încărca o arhiva care va conține:

- Raportul cerut mai sus
- Setul de date
- Fișierele cu codul sursa folosit pentru procesarea si analiza datelor.

Exemple de întrebări de cercetare

- In mediul bancar, cat de bine putem sa identificam clienții care nu vor putea sa își ramburseze creditul luat?
- Cat de bine putem prezice vânzările unui magazin, pe un anume domeniu comercial?
- Cat de bine putem prezice succesul box-office al unui film?
- Putem identifica un grup de persoane care sa fie mai receptivi la o anumita forma de publicitate pentru un produs?
- Care sunt predictorii cei mai importanți pentru a caracteriza mișcarea de persoane pe piața muncii într-un domeniu particular?

Trimiterea proiectului

Termenul pentru trimiterea proiectului este: cel puțin trei zile înainte de data aleasă in sesiune pentru prezentare la examen

Veți trimite un fișier Zip care conține:

- Raportul in format Doc / PDF (neprotejate)
- Un director care sa conțină codul / codurile sursa
- Setul de date. Daca este un set de date foarte mare, puteți indica un link web de unde setul de date poate fi descărcat.

Atenție!!! Rapoartele vor fi verificate pentru similitudine folosind Turnitin. Proiectele cu grad mare de similaritate vor fi descalificate (nota 1 final).

Grila de notare – total 100 pct.

Următoarele criterii vor fi folosite pentru notarea proiectului (se acorda 10 puncte din oficiu):

- **Introducerea: 10 puncte**

– Daca se furnizează cititorului suficienta informație pentru a înțelege restul raportului

– Daca întrebările de cercetare sunt clar stabilite

– Relevanta respectiv importanta întrebărilor de cercetare. Daca contribuția propusa prin raport este clar prezentata

- **Setul de date (15 pct)**

– Daca datele culese sunt potrivite pentru a răspunde la întrebările de cercetare

– Daca datele sunt descrise corespunzător?

- **Rezultate si discuții (30 pct)**

– dacă analiza realizata este potrivita pentru a răspunde întrebărilor de cercetare alese

– dacă s-au ales metode potrivite de analiză, daca sarcinile de analiza au fost rulate corespunzător

– daca rezultatele obținute sunt interpretate corespunzător

– daca rezultatele prezentate sunt clare si aceasta prezentarea are o ordine logica, potrivita

– daca tabelele si graficele realizate au puterea de a informa asupra concluziilor si interpretărilor textuale

– daca sunt prezentate limitări ale studiului si munca adițională care se poate face pentru a obține rezultate si mai pertinente?

- **Concluzia (5 pct)**

– daca se furnizează un sumar scurt si concis potrivit pentru raport?

– daca concluziile sunt potrivite întrebărilor de cercetare alese si sunt susținute de analiza realizata

- **Codul sursa furnizat (30 pct)**

– daca codul sursa este complet, susține analiza, si poate fi rulat cu ușurința, pentru a reproduce rezultatele prezentate in raport

– cât de eficient este codul sursa (daca sunt taskuri duplicat care pot fi evitate)?

In final, calitatea scrisului contează. Deci încercați sa fiți sigur ca exprimările sunt clare si concise si se înțelege ceea ce doriți sa transmiteți, sa eliminați posibilele confuzii de interpretare.

Resurse

Mai jos aveți câteva seturi de date care pot fi folosite ca si sursa de date si întrebări de cercetare:

– Google's dataset search (https://toolbox.google.com/datasetsearch)

– Kaggle (https://www.kaggle.com/datasets)

– OpenML (https://www.openml.org)

– UCI ML (https://archive.ics.uci.edu/ml/index.php)

– KDNuggets (https://www.kdnuggets.com/datasets/index.html)