Framside

Institutt for datateknologi og informatikk

Eksamensoppgåve i TDT4300 Datavarehus og datagruvedrift

Eksamensdato: 29. mai 2020

Eksamenstid (frå-til): 09:00 – 13:00

Hjelpemiddelkode/Tillatne hjelpemiddel: A / Alle hjelpemiddel tillatne

Fagleg kontakt under eksamen:
Tlf.: 41 44 04 33

Teknisk hjelp under eksamen: NTNU Orakel
Tlf: 73 59 16 00

ANNAN INFORMASJON:

Gjer deg opp dine eigne meiningar og presiser i svara dine kva for føresetnadar du har lagt til grunn i tolking/avgrensing av oppgåva. Fagleg kontaktperson skal berre kontaktast dersom det er direkte feil eller manglar i oppgåvesettet.

Lagring: Svara dine i Inspera Assessment vert lagra automatisk. Jobbar du i andre program – hugs å lagre undervegs.

Juks/plagiat: Eksamen skal vere eit individuelt, sjølvstendig arbeid. Det er tillate å bruke hjelpemiddel. Alle svar vert kontrollert for plagiat. *Du kan lese meir om juks og plagiering på eksamen her*.

Varslingar: Dersom det oppstår behov for å gje beskjedar til kandidatane medan eksamen er i gang (f.eks. ved feil i oppgåvesettet), vil dette bli gjort via varslingar i Inspera. Eit varsel vil dukke opp som en dialogboks på skjermen i Inspera. Du kan finne att varselet ved å klikke på bjølla i øvre høgre hjørne på skjermen. Det vil i tillegg bli sendt SMS til alle kandidatar for å sikre at ingen går glipp av viktig informasjon. Ha mobiltelefonen din innan rekkevidde.

Vekting av oppgåvene: Som vist i oppgavesettet. Alle deloppgåver innanfor ei oppgåve tel likt.

OM LEVERING:

Svara dine vert levert automatisk når eksamenstida er ute og prøven stenger, under føresetnad av at du har svart på minst ei oppgåve. Dette skjer sjølv om du ikkje har klikka «Lever og gå tilbake til Dashboard» på siste side i oppgåvesettet. Du kan opne og redigere svara dine så lenge prøven er open. Dersom du ikkje har svart på nokon av oppgåvene ved prøveslutt, blir ingenting levert.

Trekk frå eksamen: Ønskjer du å levere blankt/trekke deg, gå til hamburgermenyen i øvre høgre hjørne og vel «Lever blankt». Dette kan ikkje angrast sjølv om prøven framleis er open.

Tilgang til svara dine: Du finn svara dine i Arkiv etter at sluttida for eksamen er passert.

1

Oppgåve 1 – Modellering og OLAP – 20 %

a. Sykkelmat (SM) leverer mat frå restaurantar til kundar i fleire byar. Kvar restaurant har eit sett med rettar dei tilbyr, og når kunden har bestilt maten på nett blir den levert med sykkelbod til kunden kort tid etterpå. SM ønskjer eit datavarehus som kan brukast til å analysere og optimalisere tenesta.

Eksempel på analysar ein skal vere i stand til å gjere mot datavarehuset:

- Totalt tal på leveransar per dag
- Totalt tal på leveransar per restaurant per dag
- Gjennomsnittleg pris på kvar levering
- Tal på kundar per by som bestilte mat 12. april 2020

Skildringa er litt upresist formulert og det er ein del av oppgåva å velje ut det som skal vere med. Vi er først og fremst ute etter at du skal vise modelleringsprinsippet for datavarehus. Forklar kort eventuelle føresetnader du finn det nødvendig å gjere.

Lag eit stjerne-skjema for denne case-beskrivelsen. Sidan det ikkje er høve til å levere figurar, ber vi om at de i staden leverer tabellane og deira attributtar, og angjev kva som er fakta- og kva som er dimensjonstabellar. Eksempel på korleis angje tabell: TabellA(key, attrib_a, attrib_b).

b. Gjeve en kube med dimensjonar og tilhøyrande konsepthierarki:

Time(day-month-quarter-year)
Item(item_name-brand-type)
Location(street-city-province_or_state-country)

Gå utifrå følgjande materialiserte kuboidar:

1) {*year, brand*}
2) {*year, item_name, street*}
3) {*item_name, country*} where *year = 2006*

Gjeve følgjande OLAP-spørjing: {*item_name, city*} med vilkår "*year = 2006*"
Kva materialiserte kuboidar kan brukast til å prosessere spørjinga? Grunngje svaret.

Skriv svaret ditt her...

Format

Words: 0

Maks poeng: 20

2 **2**

Oppgåve 2 – Klynging og klyngingsvalidering – 30 %

PointID	X	Y

P1	2	4
P2	2	5
P3	2	10
P4	2	11
P5	2	16
P6	3	3
P7	3	10
P8	3	11
P9	4	3

a. Gjeve eit datasett som vist i tabellen ovanfor, der første kolonne er punkt-identifikator, og kolonnene X og Y er numeriske verdiar. Utfør klynging ved hjelp av DBSCAN på dette datasettet, gjeve *MinPts*=4 (inkl. eige punkt) og *Eps*=3 (inkl. punkt som har distanse 3). Bruk Manhattan-distanse som avstandsmål.

b. Gjeve datasettet under, der vi allereie har utført klynging (basert på X og Y) og enda opp med 3 klynger (jfr. tilhøyrande klynge-identifikator for kvart punkt). Rekn ut Silhouett-koeffisienten for punkt P2.

PointID	X	Y	ClusterID
P1	1	1	C1
P2	2	1	C1
P3	2	4	C2
P4	2	5	C2
P5	4	1	C3
P6	5	1	C3

Skriv svaret ditt her...

Format | B I U x_2 x^2 I_x | ↩ ↪ | Ω | ∑ | ABC | ⛶

Words: 0

Maks poeng: 30

3 **3**

Oppgåve 3 – Klassifisering – 20 %

Nr	A	B	C	D	*Klasse*
1	L	K	S	2	*J*
2	H	F	S	4	*J*
3	H	T	H	4	*J*
4	L	F	S	2	*N*
5	L	F	H	5	*N*
6	H	T	G	2	*N*
7	L	F	S	6	*N*
8	L	K	G	4	*N*
9	H	T	S	4	*J*
10	L	F	S	5	*N*
11	L	K	H	7	*N*
12	H	F	G	9	*J*
13	L	K	G	2	*N*
14	L	F	H	1	*J*
15	L	F	H	7	*N*

Som ein del av ein større applikasjon ønskjer vi å kunne predikere klasse (*J* eller *N*) basert på inndata der kvar post består av eit sekvensnummer og attributta A, B, C, og D, jfr. tabellen ovanfor.

Gå utifrå at vi skal bruke *avgjerdstre* som klassifiseringsmetode. Vi bruker då data i tabellen over som treningsdata. Vi bruker *Gini index* som mål for ureinheit (”impurity”).

Oppgåve: Målet med klassifiseringa er å kunne predikere ”Klasse”. Rekn ut $GAIN_{split}$ for splitting på (1) ”*A*” og (2) ”*C*”. Kva for ein av disse splittingane ville du valt for å starte opprettinga av avgjerdstreet? Grunngje svaret.

Skriv svaret ditt her...

Format

Words: 0

Maks poeng: 20

4 **4**

Oppgåve 4 – Assosiasjonsreglar – 30 %

TransactionID	Element
T1	BF
T2	ABCDFH
T3	ABF
T4	ABFH
T5	ADEF
T6	ABFH
T7	ABDEFH
T8	AGH

a. Gå utifrå handlekorg-data som er gjeve ovanfor. Bruk *apriori-algoritmen* til å finne alle frekvente elementsett med minimum støtte på 50 % (dvs. *minimum support count* er 4). Bruk $F_{k-1} \times F_{k-1}$-metoden for kandidat-generering.
b. Eit av dei frekvente elementsetta er ABH. Finn alle assosiasjonsreglar basert på dette settet, gjeve konfidens på 75 % (det er ikkje nødvendig å bruke apriori til å finne assosiasjonsreglane, men vis korleis konfidens vert rekna ut for kvar av kandidatreglane som er basert på ABH).

Skriv svaret ditt her...

Format | B I U x_2 x^2 T_x | Ω | ∑ | ABC

Words: 0

Maks poeng: 30