Framside

Institutt for datateknologi og informatikk

Eksamensoppgave i TDT4300 Datavarehus og datagruvedrift

Eksamensdato: 29. mai 2020

Eksamenstid (fra-til): 09:00 – 13:00

Hjelpemiddelkode/Tillatte hjelpemidler: A / Alle hjelpemidler tillatt

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

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

ANNEN INFORMASJON:

Gjør dine egne antagelser og presiser i besvarelsen hvilke forutsetninger du har lagt til grunn i tolkning/avgrensing av oppgaven. Faglig kontaktperson skal kun kontaktes dersom det er direkte feil eller mangler i oppgavesettet.

Lagring: Besvarelsen din i Inspera Assessment lagres automatisk. Jobber du i andre programmer – husk å lagre underveis.

Juks/plagiat: Eksamen skal være et individuelt, selvstendig arbeid. Det er tillatt å bruke hjelpemidler. Alle besvarelser blir kontrollert for plagiat. *Du kan lese mer om juks og plagiering på eksamen her.*

Varslinger: Hvis det oppstår behov for å gi beskjeder til kandidatene underveis i eksamen (f.eks. ved feil i oppgavesettet), vil dette bli gjort via varslinger i Inspera. Et varsel vil dukke opp som en dialogboks på skjermen i Inspera. Du kan finne igjen varselet ved å klikke på bjella øverst i høyre hjørne på skjermen. Det vil i tillegg bli sendt SMS til alle kandidater for å sikre at ingen går glipp av viktig informasjon. Ha mobiltelefonen din tilgjengelig.

Vekting av oppgavene: Som vist i oppgavesettet. Alle deloppgaver innenfor en oppgave teller likt.

OM LEVERING:

Besvarelsen din leveres automatisk når eksamenstida er ute og prøven stenger, forutsatt at minst én oppgave er besvart. Dette skjer selv om du ikke har klikket «Lever og gå tilbake til Dashboard» på siste side i oppgavesettet. Du kan gjenåpne og redigere besvarelsen din så lenge prøven er åpen. Dersom ingen oppgaver er besvart ved prøveslutt, blir ikke besvarelsen din levert.

Trekk fra eksamen: Ønsker du å levere blankt/trekke deg, gå til hamburgermenyen i øvre høyre hjørne og velg «Lever blankt». Dette kan ikke angres selv om prøven fremdeles er åpen.

Tilgang til besvarelse: Du finner besvarelsen din i Arkiv etter at sluttida for eksamen er passert.

1

Oppgave 1 – Modellering og OLAP – 20 %

a. Sykkelmat (SM) leverer mat fra restauranter til kunder i flere byer. Hver restaurant har et sett med retter de tilbyr, og når kunden har bestilt maten på nett blir den levert med sykkelbud til kunden kort tid etterpå. SM ønsker et datavarehus som kan brukes til å analysere og optimalisere tjenesten.

Eksempel på analyser man skal være i stand til å gjøre mot datavarehuset:

- Totalt antall leveranser per dag
- Totalt antall leveranser per restaurant per dag
- Gjennomsnittlig pris på hver levering
- Antall kunder per by som bestilte mat 12. april 2020

Beskrivelsen er litt upresis og det er en del av oppgaven å velge ut det som skal være med. Vi er først og fremst ute etter at du skal vise modelleringsprinsippet for datavarehus. Forklar kort eventuelle forutsetninger du finner det nødvendig å gjøre.

Lag et stjerne-skjema for denne case-beskrivelsen. Siden det ikke er anledning til å levere figurer, ber vi om at dere i stedet leverer tabellene og deres attributter, og angir hva som er fakta- og hva som er dimensjonstabeller. Eksempel på hvordan angi tabell: TabellA(key, attrib_a, attrib_b).

b. Gitt en kube med dimensjoner og tilhørende konsepthierarki:

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

Anta følgende materialiserte kuboider:

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

Gitt følgende OLAP-spørring: {*item_name, city*} med vilkår "*year = 2006*"
Hvilke(n) materialiserte kuboider kan brukes til å prosessere spørringen? Begrunn svaret.

Skriv ditt svar her...

Format | B I U x_2 x^2 | I_x | | | | | Ω | | | ∑ |

Words: 0

Maks poeng: 20

2 **2**

Oppgave 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. Gitt et datasett som vist i tabellen ovenfor, der første kolonne er punkt-identifikator, og kolonnene X og Y er numeriske verdier. Utfør klynging ved hjelp av DBSCAN på dette datasettet, gitt *MinPts*=4 (inkl. eget punkt) og *Eps*=3 (inkl. punkt som har distanse 3). Bruk Manhattan-distanse som avstandsmål.

b. Gitt datasettet under, der vi allerede har utført klynging (basert på X og Y) og endt opp med 3 klynger (jfr. tilhørende klynge-identifikator for hvert punkt). Regn 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 ditt svar her...

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

Words: 0

Maks poeng: 30

3 **3**

Oppgave 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 en del av en større applikasjon ønsker vi å kunne predikere klasse (*J* eller *N*) basert på inndata der hver post består av et sekvensnummer og attributtene A, B, C, og D, jfr tabellen ovenfor.

Anta at vi skal bruke *beslutningstre* som klassifiseringsmetode. Vi bruker da data i tabellen over som treningsdata. Vi bruker *Gini index* som mål for urenhet (”impurity”).

Oppgave: Målet med klassifiseringen er å kunne predikere ”Klasse”. Regn ut $GAIN_{split}$ for splitting på (1) ”*A*” og (2) ”*C*”. Hvilken av disse splittingene ville du valgt for å starte opprettingen av beslutningstreet? Begrunn svaret.

Skriv ditt svar her...

Format

Words: 0

Maks poeng: 20

4 **4**

Oppgave 4 – Assosiasjonsregler – 30 %

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

a. Anta handlekorg-data som er gitt ovenfor. 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. Et av de frekvente elementsettene er ABH. Finn alle assosiasjonsregler basert på dette settet, gitt konfidens på 75 % (det er ikke nødvendig å bruke apriori til å finne assosiasjonsreglene, men vis hvordan konfidens blir regnet ut for hver av kandidatreglene som er basert på ABH).

Skriv ditt svar her...

Format

Words: 0

Maks poeng: 30