Norges teknisk-naturvitenskapelige universitet
Institutt for datateknikk og informasjonsvitenskap

EKSAMENSOPPGAVE I FAG TDT4300 – DATAVAREHUS OG DATAGRUVEDRIFT

Faglig kontakt under eksamen: Kjetil Nørvåg og Trond Aalberg

Tlf.: 41440433/97631088

Eksamensdato:	**9. juni 2012**
Eksamenstid:	**09.00-13.00**
Tillatte hjelpemiddel:	**D: Ingen trykte eller håndskrivne hjelpemiddel tillatt. Bestemt, enkel kalkulator tillatt.**

Språkform: Bokmål

Sensurdato: 30. juni 2012

Oppgave 1 – Datavarehus – 20% (alle deler teller likt)

a) Forklar begrepene OLTP (Online Transaction Processing) og OLAP (Online Analytical Processing). Legg vekt på å få fram hva som er forskjellig mellom systemer for det ene eller det andre mht. egenskaper og bruk.

b) Forklar hva som menes med ordene som er understreket i følgende definisjon av datavarehus: ”A datawarehouse is a **subject-oriented**, **integrated**, **time-variant**, and **nonvolatile** collection of data.”

c) Forklar begrepet datakube og hva som menes med cuboids.

d) Gitt en datakube med 5 dimensjoner, hvor mange cuboids kan vi generere fra denne (eller si at den inneholder)?

e) Hva menes med datakube-operasjonene slice, dice, rollup og drill-down?

Oppgave 2 – Modellering – 20%

Du skal lage et datavarehus over trafikkulykker i Norge for å kunne undersøke hvilke veistrekninger som det er mest samfunnsnyttig å utbedre eller sette ned farten på etc. Vi skal bare se på direkte kostnader ved ulykkene og ikke bry oss om personskader etc. Datagrunnlaget kommer fra forskjellige forsikringsselskap og inneholder:

- hvor (dato) og når ulykken skjedde (gate og by, eller for eksempel veistrekning og fylke).
- data om fører (vi er mest interessert i alder til fører og om vedkommende var beruset eller ikke)
- type forsikring på bilen og forsikringsutbetalingen.

Datagrunnlaget er litt upresist formulert og det er en del av oppgaven å velge ut det som er mulig å få med, eller tenke ut en måte å uttrykke det som fakta om ulykkene. Vi er først og fremst ute etter at du skal vise modelleringsprinsippet for datavarehus. Forklar kort eventuelle forutsetninger du finner det nødvendig å gjøre.

a) Lag et stjerne- eller snøflak-skjema for denne case-beskrivelsen.

b) Lag to forskjellige konsepthierarkier (fritt valgte dimensjoner).

c) Skriv et eksempel-query i mdx hvor du genererer en todimensjonal tabell (som gir et to-dimensjonalt svarsett) som viser forsikringsutbetalinger for månedene i 2011 for forskjellige fylker (her antar vi at du har konsepthierarkier som lar deg bruke disse kategoriene).

Oppgave 3 – Klynging – 30 % (10% på a og b, 5% på c og d)

X	Y
1	11
1	9
1	5
1	2
6	7
11	7

a) Forklar hierarkisk klynging, og forskjellen på MIN-link og MAX-link.
b) Gitt et to-dimensjonalt datasett som vist i tabellen til høyre. Utfør hierarkisk agglomerativ klynging på dette datasettet, og vis det resulterende dendrogrammet. Spesifiser om du bruker MIN-link eller MAX-link.
c) Gi 4 årsaker til at vi ønsker å evaluere klynginger (klynge-validitet).
d) Forklar hvordan man kan finne hva som er passende antall klynger *K* ved klynging vha. *K*-means.

Oppgave 4 – Assosiasjonsregler – 20 %

Anta handlekurv-data til høyre. Bruk apriori-algoritmen for å finne hvilke assosiasjonsregler som gjelder, gitt at minimum støtte er 50 % (dvs. *minimum support count* er 3) og konfidens er 70 %.

TransaksjonsID	Element
T1	A,C,D
T2	B,C,E
T3	A,B,C,E
T4	B,E
T5	B,C
T6	A,C,D,E

Oppgave 5 – Klassifisering – 10 %

a) Beskriv Hunt's algoritme.
b) Forklar *underfitting* og *overfitting* i kontekst av beslutningstre.