HUR KAN BESLUTSPROCESSEN

STÖDJAS AV DATA MINING?

Examensarbete II, 10 poäng
Göteborgs Universitet
Institutionen för Informatik
Examination 010607
Av: Helena Gustafsson och Anna Larsson
Handledare: Faramarz Agahi
Examinator: Thanos Magoulas

Förord

Vi vill tacka vår handledare Faramarz Agahi för konstruktiva förslag och kunskap. Vi vill även tacka kursansvarig Thanos Magoulas för sitt stora engagemang och tålamod.

Vi vill också tacka de personer som tagit sig tid och varit med i vår undersökning.

Sist men inte minst, vill vi tacka oss själva för ett gott samarbete och att vi genomfört ett examensarbete till!

Göteborg den 7 juni 2001
Helena Gustafsson och Anna Larsson

© Helena Gustafsson och Anna Larsson, 2001

Abstract

In the complex world in which organisations exists is it not an easy task to make correct judgements of the environment. In most cases the competition is very keen. This situation makes great demands of having knowledge about your customers, suppliers and competitors to make decisions in favour of the organisation. Today there are tools that have the ability to discover new relations and patterns in the large amount of data that many organisations have stored for a considerable time. It is in this stage Data mining comes into the picture and if the technique is used properly can it in the long run give the organisation advantages in the competition. This essay is primarily focusing on how Data mining can support the decision-making process. It is treating which problems that can be solved with the support of Data mining and which the critical factors are when using Data mining and how they are handled. The purpose of the essay is to construct a realitymodel of how Data mining can support the decision-making process and the model is based on Macintosh framework. The essay's theory chapter is primarily based on Scott Morton and Macintosh framework. The empirical study consists of suppliers of Data mining solutions, owners of the systems and users. This method is used to get different aspects of Data mining. The main results of the study are that Data mining primarily can be used on problems with high knowledge availability and with high variation in knowledge need. Critical factors to consider when using Data mining are systemintegration the knowledge and competence of the users and the quality of the data.

© Helena Gustafsson och Anna Larsson, 2001

Sammanfattning

I den komplexa värld som organisationer befinner sig är det inte någon lätt uppgift att kunna göra korrekta bedömningar av omgivning. Konkurrensen är i de flesta fall mycket hård. Vilket ställer stora krav på att ha kunskap om bland annat sina kunder, leverantörer och konkurrenter för att fatta beslut som är till organisationens fördel. Idag finns det verktyg som har förmågan att hitta nya samband och mönster i den stora mängd data som många organisationerna samlat på sig under en längre tid. Det är här Data mining kommer in i bilden och som i förlängningen kan ge organisationen konkurrensfördelar om tekniken används rätt. Uppsatsen fokuserar främst på hur Data mining kan stödja beslutsprocessen. Den behandlar vilka problem som kan lösas med hjälp av Data mining samt vilka de kritiska faktorerna är vid Data mining användning och hur dessa hanteras. Syftet med examensarbetet är att utifrån Macintosh ramverk skapa en verklighetsmodell för hur Data mining kan stödja beslutsprocessen. Examensarbetets teorikapitel grundas främst på Scott Morton och Macintosh respektive ramverk. Den empiriska studien har innefattat leverantörer av Data mining, systemägare och användare. Detta för att fånga in olika aspekter kring Data mining. De slutsatser som vi kan dra från studien är att Data mining främst kan användas på problem med hög kunskapstillgänglighet och med hög variation i kunskapsbehovet. De faktorer som främst är viktiga att beakta är systemintegration, användarnas kunskap och kompetens samt kvalitén på datan.

© Helena Gustafsson och Anna Larsson, 2001

Innehållsförteckning

1 INLEDNING 4

1.1 BAKGRUND 4
1.2 PROBLEMFORMULERING 5
1.3 SYFTE OCH AVGRÄNSNING 6
1.4 DISPOSITION 7

2 TEORETISK REFERENSRAM 8

2.1 DATA, INFORMATION OCH KUNSKAP 8
2.2 BESLUT I SITT ORGANISATORISKA SAMMANHANG 9
Definitioner 9
Organisation 9
Beslut 10
Beslutsfattande 11
Två dimensioner 11
Tre dimensioner 13
Människan som beslutsfattare 15
Osäkerhet i beslutssituationen 15
2.3 DATA MINING 16
Definitioner 16
Konkreta exempel 17
Data mining i sitt organisatoriska sammanhang 19
Funktioner 21
Klassificering 22
Associering och Sekvensiering 22
Klustring 23
Prediktion 23
Uppskattning (Estimation) 23
Visualisering 24
Design metoder 24
Metoder 24
Tekniker 24
Verktyg 32
Arbetsmetodik vid Data mining projekt 36
CRoss-Industry Standard Process for Data mining 36
SEMMA 38
Kritiska framgångsfaktorer vid Data mining 38
2.4 DATA MINING OCH BESLUTSFATTANDE 39

3 METOD 41

3.1 VETENSKAPLIGT FÖRHÅLLNINGSSÄTT 41
3.2 VAL AV METOD 41
3.3 UNDERSÖKNINGENS UPPLÄGG 42
Datainsamling 42
Försökspersoner 42
Design av frågor 42

© Helena Gustafsson och Anna Larsson, 2001

Bakgrundsvariabler 43
Frågor som berör funktionell kvalité (Processer) 43
Frågor som berör infologisk kvalité (Aktörer) 44
Frågor som berör social kvalité (Design) 45
3.4 TILLVÄGAGÅNGSSÄTT 46
3.5 SVARSBEARBETNING 47
3.6 KVALITETSGRANSKNING 47
Reliabilitet *47*
Validitet *48*

4. ANALYS 49

4.1 DEMOGRAFI 49
4.2 EN FRÅGA OM PROCESSER 50
4.3 EN FRÅGA OM AKTÖRER 52
4.4 EN FRÅGA OM DESIGN 53

5. DISKUSSION 56

5.1 MODELL FÖR DATA MINING ANVÄNDNING 56
5.2 SLUTSATSER 59
5.3 FRAMTIDA FORSKNING 59
5.4 EGNA REFLEKTIONER 60

6. REFERENSER 61

6.1 BÖCKER 61
6.2 ARTIKLAR 63
6.3 ELEKTRONISKA KÄLLOR 63
6.4 OPUBLICERAT MATERIAL 64

© Helena Gustafsson och Anna Larsson, 2001

Figurförteckning

FIGUR 1:1 ÖVERGRIPANDE MODELL FÖR STUDIEN 6
FIGUR 2:1 LEVITTS MODELL MODIFIERAD AV GALIBRAITH (1977) 10
FIGUR 2:2 DEN RATIONELLA BESLUTSMODELLEN (EDLUND ET AL., 1999) 11
FIGUR 2:3 MACINTOSH (1994) MODELL ÖVER BESLUTSFATTANDE (MODIFIERAD AV FÖRFATTARNA) 12
FIGUR 2:4 SCOTT MORTONS RAMVERK FÖR BESLUTSFATTANDE (SCOTT MORTON, 1971) 13
FIGUR 2:5 KLUSTER FÖR PERSONALREKRYTERING (EGEN PRODUKTION) 18
FIGUR 2:7 DATA MINING I SITT SAMMANHANG (MAGOULAS, 2001) 19
FIGUR 2:6 KNOWLEDGE DISCOVERY I DATABASES PROCESSEN (FAYYAD ET AL., 1996) 21
FIGUR 2:8 BESLUTSTRÄD (DILLY, 1997) 25
FIGUR 2:9 NEURALA NÄTVERK (DILLY, 1997) 26
FIGUR 2:10 KOHONENS NETWORK (SPSS, 1999) 27
FIGUR 2.11 KLUSTER (DILLY, 1997) 28
FIGUR 2:12 REGELINDUKTION (SPSS TRAINING, 1999, MODIFIERAD AV FÖRFATTARNA) 29
FIGUR 2:13 GENETISKA ALGORITMER – SELEKTION (BERRY & LINOFF, 1997) 30
FIGUR 2:14 GENETISKA ALGORITMER – CROSSOVER (BERRY & LINOFF, 1997) 30
FIGUR 2:15 GENETISKA ALGORITMER – MUTATION (BERRY & LINOFF, 1997) 31
FIGUR 2:16 LOGISTISK REGRESSION (SPSS, 1999) 31
FIGUR 2.17 LINJÄR REGRESSION (FAYYAD, 1996) 32
FIGUR 2:18 ENTERPRISE MINERS ANVÄNDARGRÄNSSNITT (SAS, URL 3) 33
FIGUR 2:19 CLEMENTINE ANVÄNDARGRÄNSSNITT (URL 4) 35
FIGUR 2:20 DATA MINING PROCESSEN ENLIGT CRISP-DM (URL 1) 37
FIGUR 2:21 MACINTOCH (1994) MODELL MODIFIERAD AV FÖRFATTARNA 39
FIGUR 4:1 ORGANISATIONENS BEDÖMNING AV DATA MINING SYSTEMET 51
FIGUR 4:2 ANVÄNDARENS BEDÖMNING AV DATA MINING SYSTEMETS ANVÄNDARVÄNLIGHET 53
FIGUR 4:3 LEVERANTÖRERNAS SYN PÅ DATA MINING VERKTYG 55
FIGUR 5:1 MODELL FÖR DATA MINING ANVÄNDNING 57

Bilagor

Bilaga 1 Objektmodeller
Bilaga 2 Missivbrev

© Helena Gustafsson och Anna Larsson, 2001

1 Inledning

1.1 Bakgrund

Dagens hårda företagsklimatet karakteriseras av snabba förändringar i omvärlden. Organisationer måste utveckla nya metoder och strategier för att behålla och utveckla sin position på marknaden. För att organisationen skall överleva måste den anpassa sig till det som krävs i en dynamisk och föränderlig miljö. I denna kamp spelar informationsteknologi (IT) och information en viktig roll för att de beslut som fattas i organisationer skall göra dem framgångsrika. Under de senaste årtiondena har information, datorer och informationssystem blivit allt viktigare för att stödja organisationen så den kan uppnå sina mål. För att illustrera detta kan informationen liknas vid organisationens blod. Många aktiviteter i organisationen till exempel problemidentifiering och problemlösning samt beslutsfattande grundas på informationsunderlaget (Szysmanski, 1995). Varje dag fattas många beslut både enskilt och i grupp. För individen kan detta vara allt från att välja vilken mat man skall äta till middag eller att besluta hur pensionspengarna skall placeras. Oavsett om det är en individ eller en organisation som fattar besluten finns en önskan att kunna förutspå konsekvenserna. Många är nog lite avundsjuka på spåtantens förmåga att se in i sin kristallkula. Då det är ytterst få som besitter denna egenskap så får de flesta förlita sig på de möjligheter dagens informationsteknologi kan erbjuda.

Det har under de senaste åren blivit allt lättare och billigare att samla in stora datamängder på grund av den dramatiska hårdvaruutveckling. Kotler et al. (1999) poängterar dock att det är en sak att samla in data och det är en helt annan sak att använda den lönsamt. Dahlbom & Mathiasson (1993) menar att organisationer sätter ett stort värde på information och är villiga att betala mycket för den. Ofta är det så att när informationen samlats in och använts till en specifik uppgift faller den i glömska. Vidare menar de att organisationer hela tiden frågar efter mer information. Samtidigt så använder de inte informationen som de har på ett rationellt sätt. För att ligga i framkanten av utvecklingen vill organisationer förvandla stora mängder rådata till kunskap. Detta för att få kunskap om olika omvärldsfaktorer till exempel kunder, leverantörer, konkurrenter och marknader. En metod att åstadkomma detta är att använda sig av OnLine Analytical Processing (OLAP). OLAP ger användaren möjlighet att bevisa existerande hypoteser och som rapporterar vad som redan inträffat. Enligt Noonan (2000) är det en brist att inte kunna förutse vad som kommer att ske. Det är som att köra bil och bara titta i backspegeln. Konsekvenserna vid hög hastighet är att någonting lätt går fel. Schor (Olofsson, 2000) menar att OLAP inte är tillräckligt när det gäller att ge organisationerna nya infallsvinklar. Vidare fortsätter han med att det inte är förrän alla tänkbara kombinationer kan testas som de oväntade resultaten uppkommer vilket organisationen kan utnyttja för nya affärsmöjligheter. Målet med användningen av avancerade och skräddarsydda algoritmer är att kunna hitta nya samband. Detta kan i sin tur leda till nya idéer om hur organisationen skall hantera sina affärsprocesser. Grunden i hela konceptet är Data mining som kan vara ett hjälpmedel för beslutsfattaren och organisationen när det gäller att utvecklas framgångsrikt i de dynamiska förändringar och komplexa affärsmiljöer som finns. Nyckeln till framgångsrik Data mining är att integrera informationsinfrastrukturen och affärsprocesserna anser Putten (1998). Genom att lagra och återanvända Data mining processerna och resultaten ger det möjlighet att samla på sig kunskap så organisationen kan lära sig mer om exempelvis sina kunder (Putten, 1998).

Data mining är ett begrepp med idéer och tekniker som härstammar från olika fält, bland annat statistik, marknadsföring, artificiell intelligens, databas administration och forskning (Berry & Linoff, 1997). Det finns inget vedertaget svenskt ord för Data mining som genom direkt översättning från engelska språket skulle bli ”gruvbrytning”. Svenska datatermgruppen genom Urban Hurtig (Url 2) föreslår ”datautvinning”. Hans förklaring är att ”*termen datautvinning är bra eftersom den ger associationer till förädling. Datautvinning står nämligen för en samling metoder för att automatisera sökandet efter komplexa samband i stora datamängder och förädla slutresultatet till en graf, tabell eller liknande. Datautvinning är därför en bättre term än till exempel databrytning (förädlingsaspekten saknas) eller dataanalys (för diffust). Datautvinning har ett naturligt samband med datalager som är databassystem uppbyggda för att underlätta just datautvinning*”. Vi kommer i fortsättningen att använda oss utav den engelska beteckningen Data mining då det är en välkänd term.

Utvecklingen framskrider i snabb takt inom detta forskningsområde. De undersökningar som tidigare gjorts på detta området har flertalet varit beskrivningar av vad Data mining är. Under 1998 genomfördes en magisteruppsats av Andersson och Elfstöm vid Göteborgs Universitet vars syfte var att se hur två utvalda organisationer kan använda sig av Data mining. De undersökte även vilka faktorer som var kritiska vid implementering av ett Data mining system.

Med vår utbildningsbakgrund som systemvetare och med tidigare erfarenheter av Data mining genom en högskolekurs vid Högskolan Trollhättan/Uddevalla, väcktes vårt intresse för denna teknik. Det är en anledning till att vi väljer att fördjupa oss inom detta ämnesområde i form av en D-uppsats.

1.2 Problemformulering

Data mining leverantörer är mycket effektiva när det gäller att framhålla de stora fördelar som finns med tekniken. En brist är att det inte tydligt framgår hur organisationer kan stödjas i sina beslutsprocesser med hjälp av Data mining. Med anledning av detta resonemang finner vi det intressant att undersöka följande frågeställning.

Hur kan beslutsprocessen stödjas av IT i allmänhet och Data mining i synnerhet?

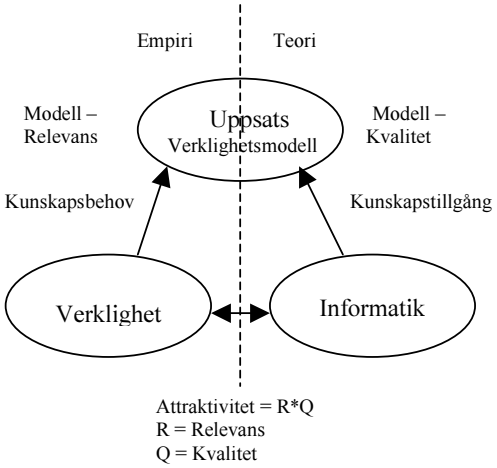
Huvudfråga har brutits ned i två delfrågor:

Delfråga I: Vilken typ av problem kan lösas med hjälp av Data mining?

Delfråga II: Vilka är de kritiska faktorerna vid Data mining användning och hur hanteras dessa?

1.3 Syfte och avgränsning

Syftet med examensarbetet är att skapa en verklighetsmodell utifrån vår teoretiska referensram. Denna verklighetsmodell bygger på Macintosh (1994) ramverk. Modellen skall sedan provas empiriskt. Vi kommer att reda ut den djungel av design metoder och de arbetsmetodiker som finns inom Data mining området. Målgruppen som vi främst vänder oss till är studenter inom det systemvetenskapliga området samt de organisationer har tankar att starta med Data mining eller som redan har den implementerad i sin verksamhet.



FIGUR 1:1 ÖVERGRIPANDE MODELL FÖR STUDIEN

På grund av den tid som står till förfogande har vi valt att avgränsa oss från att fördjupa kunskapen om de algoritmer som Data mining teknikerna är uppbyggda på.

1.4 Disposition

Nedan följer en kort beskrivning vilka huvuddelar som examensarbetet består av. Bilagor finns i slutet av examensarbetet.

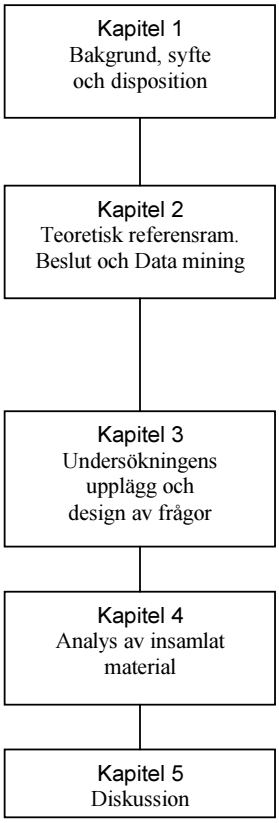
- Inledningen i kapitel ett avser att ge en bakgrund till problemområdet samt definiera syftet och frågeställningen i uppsatsen

- Kapitel två består av en teoridel som kommer att fungera som en referensram för examensarbetet. Här kommer begrepp och modeller som är väsentliga för uppsatsen att förklaras. En inventering av de design metoder som finns inom Data mining området kommer att presenteras. Anledningen till denna inventering är att skapa en samlingsplats då denna information inom litteraturen är utspridd.

- Kapitel tre behandlar metoden för den empiriska studien. Här presenteras designen av studien, urval, datainsamling och svarsbearbetning.

- I kapitel fyra analyseras empirin från den kvalitativa studien.

- I kapitel fem, diskussionsdelen förs en diskussion om de empiriska resultaten med återkoppling till teorin och problemformuleringen. Slutsatser dras tillsammans med några egna reflektioner samt några idéer kring fortsatt forskning inom området.



2 Teoretisk referensram

I teori kapitlet kommer en beskrivning väsentliga begrepp för examensarbetet. Den teoretiska referensramen kommer att ligga till grund för design av frågor vid den empiriska studien. Den kommer även att vara en utgångspunkt i resultat och diskussions avsnittet.

Studien kommer att grundas på tre ramverk. Leavitts modell (modifierad av Galbraith, 1977) behandlar organisationens beståndsdelar. Scott Mortons (1971) modell illustrerar hur beslutsfattande skall betraktas samt Macintosh (1994) modell som ger stöd vid klassifikation av beslut med hänsyn till problemets struktur och variation.

2.1 Data, information och kunskap

Data mining ingår i en process där data och information skall omvandlas till kunskap som skall ligga till grund vid beslutsfattande. Följande avsnitt avser att ge en kort beskrivning av dessa begrepp.

Enligt Dahlbom och Mathiassen (1995) är data, information och kunskap tre begrepp som starkt relaterar till varandra och gränserna emellan dem är otydliga. De menar att data är formaliserad representation av information som gör det möjligt att behandla eller kommunicera den. Szysmanskis (1995) definition av data lyder: *Numbers, letters, characters, or combinations thereof that can be entered and processed by a computer. Raw facts before processing.* Det som är data för en människa kan vara information för en annan.

Även om det kan vara svårt att definiera information så är vi väl medvetna när den saknas (Dahlbom & Mathiassen, 1995). Information är något som ger kunskap. Den främsta skillnaden mellan kunskap och information är att information är någonting vi får och ger medan kunskap är någonting vi har. Information är relaterat till kommunikation överföring eller spridning av kunskap. De viktiga egenskaper som Dahlbom och Mathiassen (1995) anser att informationen bör ha för att vara av god kvalité är att informationen skall vara relevant det vill säga anpassad till situationen. Den skall vara pålitlig och fri från redundans. Andra egenskaper som är värdefulla är konsistens, stabilitet, begriplig, tillgänglig och att informationen skall gå att återfinna. Szysmanskis (1995) definition av information är följande: *Data that have been processed into an organizied, usable form and are meaningful to there recipient for the task at hand.* Informationen kan karakteriseras genom ett antal attribut. Det ger användaren ett slags ramverk för bedömning av meningsfullheten och användbarheten. Accuracy, med noggrannhet menas om informationen är sann eller falsk. Relevance, med relevans menas om informationen är behövd och användbar i en specifik situation. Completness, med fullständighet menas att informationen berättar det du behöver veta för en situation. Timeliness, syftar till att informationen är tillgänglig när den behövs och att den inte är gammal när den erhålls. Det är mycket förödande att fatta beslut på gammal information. Cost-effitiviness, syftar till sambandet mellan fördelarna med att härleda informationen och kostnaden att producera den. Auditability, möjligheten att kunna verifiera informationen med andra ord möjligheten att kontrollera noggrannheten (exaktheten) och fullständigheten av informationen. Reliability, pålitlighet summerar hur väl informationen passar in på de övriga sex attributen (Szysmanski, 1995).

Enligt Aristoteles strävar alla människor enligt naturen efter kunskap. Kan människan veta något om framtiden eller om kontrafaktiska händelser (vad som skulle varit fallet om något annat än det som faktiskt inträffade hade inträffat)? Finns det några gemensamma drag för att något skall anses vara kunskap? Ett krav är att, allt man kan sägas veta måste vara sant, man kan inte veta något som är falskt. Det andra kravet är att man måste tro på det man säger sig veta. Man skall ha en föreställning om det man anser sig veta är sant. Ett tredje krav är att ha goda skäl att tro att något är sant. Ett exempel, om någon frågar "hur vet du det?" Jag kan bevisa det genom att göra en uträkning. I vissa fall kan en person ha en sann tro med goda skäl, men det ändå verkar helt otroligt att kalla det kunskap. Dessa fall kallas för gettierexempel efter den amerikanska filosofen, Edmund Gettier (Stjernberg,1999). Davenport (1998) menar att kunskap är information kombinerat med sammanhang, erfarenhet, reflektion och tolkning. Det är en förädlad form av information som kan appliceras vid bland annat beslutsfattande. Enligt Jönsson och Rehman (2000) har varje individ en egen makt att söka och lära. De kallar det som skapas och lagras inom en människa för kunskap och det som finns utanför för information. Lärandet representerar övergången mellan information och kunskap. Nonaka och Takeuchi (1995) anser att det finns två typer av kunskap, explicit (lätt att uttrycka och lära ut) och implicit (underförstådd och subjektiv). Det är den formaliserade och explicita kunskapen som idag finns lagrade i databaser hos organisationer. Den explicita kunskapen går att uttrycka i ord och siffror och kan kodifieras i ett formellt och systematiskt språk. Inom Data mining området kan mönster betraktas som kunskap om det överträffar intressanta gränser. Denna definition av kunskap är en ren användarorienterad och områdesspecifik och är förutbestämd genom vilka funktioner och gränser som användaren väljer (Fayyad et al., 1996).

2.2 Beslut i sitt organisatoriska sammanhang

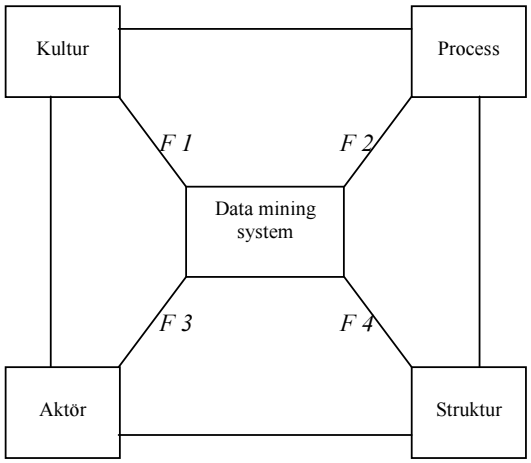
I detta avsnitt kommer begrepp som organisation och beslut att beskrivas. Vi kommer även att beskriva människan som beslutsfattare och vad som kännetecknar en beslutssituation. I samband med detta avsnitt presenteras de ramverk som studien kommer att grundas på.

Definitioner

I följande avsnitt presenteras en definition av organisationen enligt Leavitts modell som är modifierad av Galibrath (1977). Vidare kommer begreppet beslut att beskrivas.

Organisation

För att förstå en organisation, som är en abstrakt och komplex företeelse används en modell. Leavitts välkända modell ger hjälp och förståelse till att avgränsa och sätta informationen i sitt sammanhang. Förändras en av modellens beståndsdelar kommer det att påverka de övriga på grund av den stora integrationen dem emellan. Det går inte att studera tekniken utan att även vara medveten om strukturer och processer. Den valda modellen är en utgångspunkt när diskussion skall föras kring olika problem och hur de hänger ihop (Bakka et al.,1998). Vår studie kommer att grundas på Leavitts modell som är modifierad av Galibraith (1977), se figur 2:1. Den hjälper oss att klargöra beslutsprocessens natur och dess relation till Data mining. Detta klargörande baseras på process, struktur, kompetens, kultur och förhållandet med Data mining.



FIGUR 2:1 LEVITTS MODELL MODIFIERAD AV GALIBRAITH (1977)

Data mining systemet handlar om anskaffning, anpassning och utvecklingen de teknologier som finns för att stärka organisationens konkurrensförmåga. *Kultur* representerar ett komplext mönster av uppfattningar, idéer, värden attityder och beteenden som delas av organisationens medlemmar. Kulturen omfattar gemensamma förväntningar, värderingar och hållningsätt. Förhållandet mellan kultur och Data mining system (F1) är hur kulturen påverkar aktiviteter och processer i organisationen. Data mining systemet blir då en sorts kulturbärare. I modellen finns variabeln *process* som står för de uppgifter eller arbetsuppgifter som organisationen utför. Relationen mellan organisationens process och Data mining systemet illustreras ovan i förhållandet F2. Förhållandet karakteriseras av de processer som genererar information och de processer som utnyttjar information. Aktör är den kunskap och kompetens som ingår i organisationen. Relationen mellan aktör och Data mining system (F3) är den kunskap och kompetens som påverkar eller påverkas av Data mining systemet. Modellen består av en variabel som benämns *struktur*. Den beskriver organisationsutformning, arbetsfördelning, och de beslutstrukturer som finns i en organisation. Det vill säga styrning, kontroll och koordinering. Förhållandet mellan struktur och Data mining system (F4) belyser människans roll gentemot systemet.

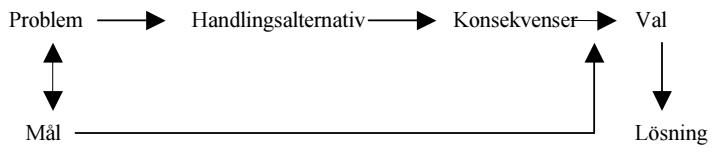
Beslut

Det finns två huvudtyper av beslut nämligen målbeslut och handlingsbeslut. Ett *målbeslut* kan förklaras som ett resultat från en omvandling av beslutsfaktorer till en kvantifierbar mängd det vill säga beslutskriterier. Ett *handlingsbeslut* kan förklaras som en produkt av beslutsprocessen för att välja ett handlingsalternativ. Ett handlingsbeslut innebär i praktiken att beslutsfattaren har ett önskat utfall som sedan jämförs (Ekman, 1970).

Beslutsteori är ett verktyg för analysering av situationer där beslutsfattare stöter på viktiga beslut. Enligt Sabini (1995) involverar viktiga beslut antingen *osäkerhet* eller en *konflikt* mellan positiva och negativa aspekter på en viss situation. Beslutsteorin hjälper beslutsfattare att bestämma hur de skall agera vid ett visst beslut. Inom beslutsteorin finns det två olika synsätt. Det normativa och det deskriptiva synsättet.

Det normativa synsättet skall ge en bild över hur beslut bör fattas för att ge optimala resultat (Agahi, 2000). Modellerna bygger enligt Edlund et al. (1999) på ideala förutsättningar om hur beslut bör fattas för att åstad komma bästa möjliga resultat. Modellerna bygger på orealistiska

antaganden. Ett exempel på det normativa synsättet är den rationella beslutsmodellen. Den förutsätter att beslutsprocessen tas i ett steg i taget i ett fördefinierat systematiskt förlopp. Mål formuleras innan handlingsalternativ tas fram. Edlund et al. (1999) jämför den normativa modellen med ett datorprogram. Informationen behandlas logiskt och med fördefinierade normer.



FIGUR 2:2 DEN RATIONELLA BESLUTSMODELLEN (EDLUND ET AL., 1999)

Det deskriptiva synsättet beskriver hur beslut fattas i verkligheten (Agahi, 1995). Modellerna hanterar de problem som uppstår vid beslutsfattande. Istället för högsta möjliga måluppfyllelse inriktar de sig på en acceptansnivå. Edlund et al. (1999) talar om satisfieringsmål och inte optimeringsmål. Dessa modeller är realistiska i avseende på de kostnader som hänger samman med att samla in information och definiera mål. Med anledning till detta är satisfiering en mer realistisk metod än optimering (Edlund et al., 1999).

Beslutsfattande

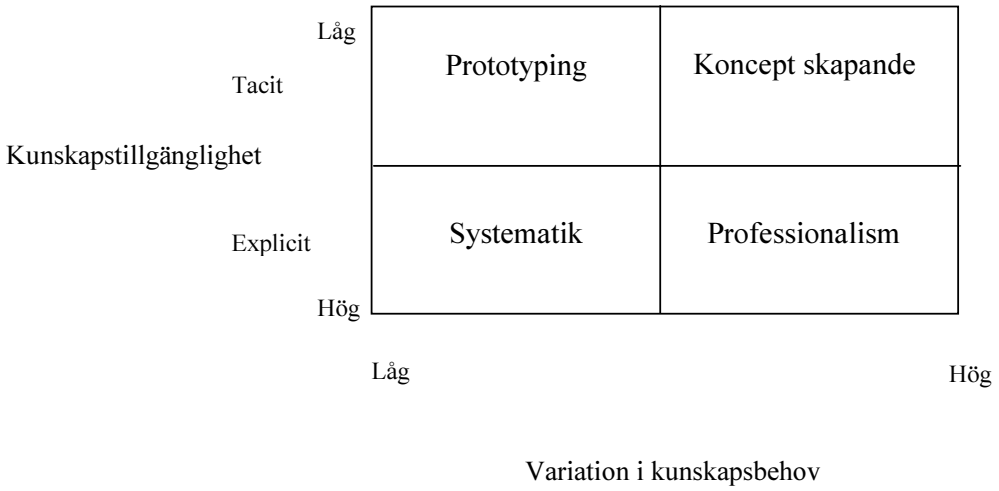
I de två nästkommande avsnitten kommer Macintosh (1994) ramverk presenteras. Det baseras på två dimensioner, kunskapstillgänglighet och variation i kunskapsbehov. Macintosh (1994) ramverk ger stöd vid klassificering av beslut med hänsyn till problemens struktur och variation. Ett annat ramverk som kommer att presenteras är Scott Mortons (1971). Det ramverket baseras på tre dimensioner beslutsnivå, beslutsprocess och problemstruktur. Ramverket hjälper oss att förstå beslutsfattande i organisationer (Scott Morton, 1971).

Två dimensioner

Macintosh (1994) modell hjälper till att klassificera respektive beslut med beaktande till problemens struktur och variation (se figur 2:3). Enligt Macintosh (1994) finns det två dimensioner av teknologier som är speciellt viktiga vid informationssystem: variation i kunskapsbehov och kunskapstillgänglighet. Med teknologi menar Macintosh (1994)*"the action that an individual performs on an object, with or without the aid of tools or mechanical devices, in order to make some change in that object"*. Denna definition betonar konverteringsprocessen (conversion process) som omvandlar indata till resultat. Variationen i kunskapsbehov är frekvensen av oväntade och ovanliga händelser som kan inträffa i konverteringsprocessen. Variationen kan anses vara hög när beslutsfattarna stöter på många ovanliga situationer med frekventa problem. Variationen är låg då det är få förväntningar och arbetsuppgifterna bedöms som repetetativa. Enligt Macintosh (1994) kan variationerna i arbetsuppgifterna vara från att sortera post på ett postkontor till ett jobb med en serie orelaterade problem. Den andra dimensionen (kunskapstillgänglighet) berör hur beslutsfattarna reagerar på de problem som uppstår i deras arbete. Om kunsksapstillgängligheten är hög följer beslutsfattarna ett objektivt mönster för att lösa problemen. Det kan innefatta lagrade procedurer såsom instruktioner, manualer, program och standarder. Här finns det välkända sätt att lösa problemen på. I kontrast till de välkända

arbetsuppgifterna finns de okända arbetsuppgifterna. Dessa karaktäriseras av att kunskapstillgängligheten är låg. När problemen uppstår är det svårt att finna den korrekta lösningen. Här finns inga lagrade procedurer att använda sig av, kunskapen är tacit (underförstådd). Beslutsfattarna får aktivt söka efter en lösning utöver de tillgängliga lagrade procedurerna. De får fatta beslut baserat på erfarenhet, intuition och bedömningar. Den slutgiltiga lösningen är inte ett resultat av beräkningar. Det finns heller ingen garanti för att den valda lösningen för problemet är den korrekta lösningen.

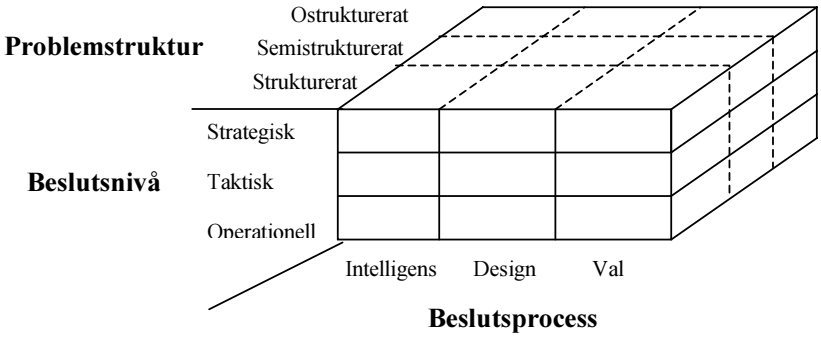
Utifrån dessa två dimensioner har Macintosh (1994) utkristalliserat fyra kategorier av teknologierna: systematik, prototyping, professionalism och konceptuell modellering. *Systematik* karakteriseras av att uppgifterna har lite variation och kunskapstillgängligheten är hög och strukturerad. Arbetsuppgifterna är rutinmässiga. *Prototyping* kännetecknas av förutsägbara strömmar av aktiviteter, men med en låg kunskapstillgänglighet. Här finns det inga lager med procedurer och tekniker att applicera på problemen som uppstår. Uppgifterna kräver omfattande träning och erfarenhet då beslutsfattarnas reagerar på problemet med hjälp av visdom, intuition och erfarenhet. *Professionalism* tenderar till att vara väldigt komplext då det finns ett stort spann av arbetsuppgifter. Det är väldigt sällan som professionalister möter identiska problem i arbetet. Ett exempel är att ingenjörer kan referera till böcker och tekniska manualer för att finna den korrekta formeln för problemet. En och samma formel kan vara till hjälp vid flera olika problem. Här finns ett antal analys program som kan användas som stöd. Det är dock svårt att förutse vilka problem som kan uppstå. *Konceptuell modellering* kan sägas vara motsatsen till systematisering. Variationen av kunskapsbehovet är stort och kunskapstillgängligheten är låg. Osäkerheten härskar på grund av den breda variationen av problem som stöts på i uppgifterna och oförmågan att förutse det bästa sättet att uppnå ett önskad resultat. Den konceptuella modelleringen kräver en stor analytisk förmåga. Ofta finner beslutsfattarna här inte en bra lösning utan flera stycken som kan accepteras.



FIGUR 2:3 MACINTOSH (1994) MODELL ÖVER BESLUTSFATTANDE (MODIFIERAD AV FÖRFATTARNA)

Tre dimensioner

Enligt Scott Mortons (1971) ramverk innefattar beslutsfattande tre olika dimensioner. Dessa visualiseras i figur 2:4. Ramverket är en kombination av Anthonys och Simons modeller. Den första dimensionen visar de olika beslutsnivåerna som en organisation består av. Den andra dimensionen beskriver de huvuddelar som beslutsprocessen består av. Den tredje dimensionen belyser de strukturer ett problem kan kategoriseras i.



FIGUR 2:4 SCOTT MORTONS RAMVERK FÖR BESLUTSFATTANDE (SCOTT MORTON, 1971)

Beslutsnivå

Anthony (1965) delar in beslutsfattande i tre nivåer: *strategisk*, *taktisk* och *operationell* vilket ställer krav på differentiering av informationsbehoven. Olika beslut ställer olika krav på informationens kvalité och funktion. Han definierar strategisk planering enligt följande: *"The process of deciding on objectives of the organization, on changes in these objectives, on the recources used to attain these objectives, and on the policies that are to govern the acquisation, use and disposition of these resources"*. Med objectives (mål) menas det som organisationen önskar att uppnå. Policies är riktlinjer vid val av det alternativ som bäst uppfyller de mål som organisationen har satt upp. Strategi är enligt Anthony (1965) att kombinera och bruka resurser. Taktisk kontroll är förenat med den pågående administrationen av verksamheten. Här sker både planering och kontroll. Taktisk kontroll definieras som: *"The process by which managers assure that recources are obtained and used effectively and efficently in the accomplishment of the organization's objectives"*. Definitionen påvisar tre idéer. Den första är att processen involverar chefer som får gjort saker genom att samarbeta med andra människor. Det andra är att processen ingår i en kontext av mål och policies som har satts upp i den strategiska planeringen. Det sista är att kriteriet för att bedöma besluten och aktiviteterna i processen är effektivitet. Operationell kontroll definieras som: *"The process of assuring that specific tasks are carried out effectively and efficiently"*. Det som skiljer denna nivå från den taktiska nivån är att här krävs ingen eller lite personlig bedömning. Alla aktiviteter och beslut har fördefinierade handlingsplaner. Här fokuseras enbart på verkställande medan på den taktisk nivån fokuseras både på verkställande och planering (Anthony, 1965) (Scott Morton, 1971).

Beslutsprocess

För att bättre försöka förstå beslutsprocessen skapade Simon 1960 en modell som bestod av tre huvudfaser: *intelligens, design* och *val* (Turban & Aronson, 1998). Beslutsprocessen

startar med *intelligensfasen* där verkligheten undersöks för att finna och identifiera problemen. Intelligens betecknar den sökprocess som görs i en beslutsituation eller problemområdet för att på det sättet försöka finna villkor som kräver beslut. Denna process inkluderar ett antal aktiviteter vars syfte är att föröka identifiera problemsituationen eller möjligheter. Intelligensfasen börjar med att identifiera organisationens mål för att därefter på ett objektivt sätt försöka bestämma om de uppfylls. Ett problem uppkommer oftast genom att saker och ting inte är som man önskar att det vore. Det missnöjet är skillnaden mellan vad vi önskar oss och vad vi vill skall hända eller inte. I denna fasen bör fokus läggas på att bestämma det problem som existerar och identifiera symptomen för att sedan mer exakt komma fram till huvudproblemet. Det kan vara svårt att definiera grundproblemet eftersom vår värld är mycket komplex och det kan råda oklarheter mellan orsak och verkan. Efter att ha identifierat problemet skall en modell konstrueras. Efter den grundliga utredningen är det möjligt att bestämma om ett problem verkligen existerar var det är lokaliserat och hur signifikant det är. Det är viktigt att därefter försöka klassificera problemet i definierbara kategorier. Ett sätt klassificera problemet är graden av struktur i problemet. När intelligens fasen har genererat en redogörelse över problemet kommer nästa steg i fasen. *Designfasen* innebär att innebär att inventera, utveckla och analysera möjliga handlingsalternativ och försöka konstruera en modell. Designfasen består av att generera, utveckla och analysera möjliga förlopp. I denna fas ingår även aktiviteter som att förstå problemet och prova möjliga lösningar. En modell av problemsituationen konstrueras, provas och valideras. När modellen skapas ligger utmaningen i att finna en balans mellan nivån på enkelheten i modellen och den verklighet den skall representeras. En modell består av resultatvariabler och beslutsvariabler. Resultatvariabeln visar nivån på effektivitet i systemet det vill säga hur väl systemet klarar de uppsatta målen. Beslutsvariabler beskriver alternativa handlings alternativ. Det är dessa variabler som beslutsfattaren kan bestämma över. Utvärderingen av vilket av de alternativ som finns beror på vilka kriterier som finns. Målet kan vara att optimera eller satisfiera. Till hjälp finns den normativa eller den deskriptiva modellen (se sidan 7). *Val fasen* är en utvärderingsprocess där ett av de möjliga alternativen väljs och implementeras. Denna process mynnar ut i ett förslag till lösning av modellen. Val fasen består av att söka, utvärdera och rekommendera en lämplig lösning av modellen. Det gäller att finna den lämpliga lösningen bland de som blev identifierade under designfasen. För normativa modeller kan dels en analytisk modell användas eller att jämföra alla alternativ med varandra. För deskriptiva modeller jämförs ett begränsat antal alternativ.

Problemstruktur

Simon (Turban & Aronson, 1998) delade in problemtyper i *strukturerade, ostrukturerade* och *semistrukturerade* problem. Ett strukturerat problem är då alla faserna (intelligens, design, val) är strukturerade. I ett strukturerat problem är tillvägagångssättet för att uppnå den bästa lösningen känd och det går att arbeta fram en standard. Här är målen kända, vanliga mål är kostnadsminimering och vinstmaximering. Den andra ytterligheten valde han att kalla för ostrukturerade eller lågprogrammerade problem. Ett problem är ostrukturerat när ingen av de tre faserna är strukturerade. I ett ostrukturerat problem används enligt Turban och Aronson (1998) den mänskliga intuitionen. Ett vanligt ostrukturerat problem är planering av nya tjänster, att anställa en beslutsfattare till exempel. Den tredje nivån som hamnar mellan dessa ytterligheter kallade Simon för semistrukturerade problem. Beslut är enligt Simon (Turban & Aronson, 1998) semistrukturerat om någon eller några men inte alla av faserna är strukturerade. Keen och Scott Morton (Turban & Aronson, 1998) menar att sätta upp en marknadsföringsbudget för kunder är ett exempel på ett semistrukturerat problem.

Ekman (1970) delar Simons tankegångar men reflekterar om det är lämpligt att utgå från klassificering av problemsituationen. Det borde vara så att om en beslutsfattare ofta möter en viss typ av problem desto enklare är lösa det. Vid problemsituationer som kommer tillbaka så utvecklas en rutin för hur problemet skall hanteras. Det vill säga vem och vilka personer som skall hantera de olika delaktiviteterna och på vilket sätt dessa skall handla. Med detta exemplet belyses att när en beslutsfattare löser ett specifikt problem och tar ett beslut för första gången så är det mindre strukturerat än om denne har gjort det antal gånger. Detta resonemang får även stöd av Adam et al. (1998).

Människan som beslutsfattare

Vem skall ha befogenhet att fatta det avgörande beslutet, det vill säga välja det handlingsalternativ som skall realiseras? Den som fattar beslutet är också ansvarig för det. Denna frågeställning blir aktuell vid delegering och decentraliseringsfrågor. Hammer och Champy (1995) menar att genom informationsteknologin har det hierarkiskt uppbyggda beslutsfattandet ersatts. Istället har beslutsfattande blivit en del av arbetsuppgifterna för alla medlemmar i organisationen. En grundregel som Ekman (1970) tar upp i sin litteratur är att *”beslutsaktiviteterna skall fördelas till beslutskomponenter med en informationsstruktur som motsvarar de krav som ställs”*. Med beslutskomponent menas en individ, en grupp människor eller en dator. Det finns ett flertal exempel och undersökningar som visar det inte är självklart att olika människor bedömer ett problem på samma sätt. En avgörande betydelse är deras organisatoriska roll. Ekman (1970) menar att en komponents funktionella, hierarkiska och geografiska placering i organisationen ger en definition av en komponents organisatoriska roll. Med komponent menas i detta sammanhang beslutsfattarna. Genom människans fem sinnen samlas informationen in om verklighetens olika situationer. Ibland utnyttjas även det sjätte sinnet, intuitionen.

För att en beslutsfattare skall kunna utvinna information ur en datamängd krävs att en tolkningsprocess utförs. Langefors Infologiska ekvation syftar till att förklara människans kognitiva förmåga som informationsbehandlare och informationsanvändare. Ekvationen säger att I = i(D,S,t) där I står för informationen som är resultatet av tolkningsprocessen (i) på Data (D) givet en viss tidsram (t) samt individens förkunskaper (S). En lyckad tolkning leder till ökat kunskapstillskott. Efter tolkningen ser formeln ut enligt följande: I = S’-S eller S’=S+I där informationen (I) är skillnaden mellan bilden som individen hade före meddelandet (S) och den bild som skapades efter tolkningen av meddelandet (S’) (Dahlbom, 1993). Magoulas och Pessi (1998) skriver att varje form av beslutsprocess är en beståndsdel av individens informationsbehandlingsprocess. De menar att beslut kan ses som en process där information om verklighetens tillstånd transformeras till information för handling. De skriver vidare att en individs verklighetsbild blir följden av tre grundläggande begrepp vilka är begreppskunskap (Sb), erfarenhet (Su) och till sist värderingar (Sv). Detta resulterar i en formeln som ser ut enligt följande: I = i(D, <Sb, Su, Sv>, t).

Osäkerhet i beslutssituationen

Edlund et al. (1999) uttrycker att anledningen till att det är svårt att fatta beslut är att det inte finns kunskap om vad som kommer att hända i framtiden. För att minimera osäkerheten i en beslutssituation vore det ultimata att samla in all information om man kan tro påverka utfallet

av beslutet. Ekman (1970) menar att detta är en utopi. Organisationer måste kompromissa mellan kostanden att producera informationen och beslutets nytta. Edlund et al. (1999) säger att minst tre olika typer av osäkerhet kan finnas i en beslutssituation. Beroende på vilken typ av osäkerhet som präglar en problemsituation kommer tillvägagångssättet att vara olika (Edlund et al., 1999; Edlund et al., 1993). Det finns tre olika typer av osäkerhet i en beslutssituation (Edlund et al., 1993). *Osäkerhet beträffande utfall*. Det är alltid svårt att försöka säga någonting om framtiden. Därför kan det också vara svårt att säga någonting om konsekvenserna av vad ett visst handlingsalternativ innebär. Ibland finns det faktorer med som inte kan kontrolleras till exempel valutakurser. Denna osäkerhet kan behandlas genom prognoser. Framtida förutsägelser kan aldrig förutsägas med full säkerhet. *Osäkerhet beträffande värderingar*. Ofta kan det råda konflikt mellan olika mål. Dessutom kan det råda delade meningar om vad som är viktigast. Denna problematik kan hanteras på så sätt att man tar ut ett mål i taget eller genom att mål prioriteras. Människors föreställning om vad som är viktigt förändras över tid. En organisations olika mål kan vara i konflikt med varandra. *Osäkerhet beträffande samband*. Ofta kan det vara svårt att ha kunskap om sambanden i organisationen. Ett beslut som berör en avdelning kan också påverka en annan. När det gäller sambanden i organisationen kan man tillsätta en projektgrupp. Edlund et al. (1993) menar att inom en organisation så påverkar ett beslut inom ett beslutsområde förutsättningarna för beslut inom andra beslutsområden. Kunskapen om dessa samband är begränsad och av praktiska skäl svåra att hänsyn till

2.3 Data mining

I detta avsnitt vill vi definiera begreppet Data mining samt ge konkreta exempel på olika användningsområden. Här kommer en inventering göras över de design metoder och arbetsmetodiker som ingår inom Data mining området presenteras.

Definitioner

Data mining som tidigare nämnts är ett begrepp med idéer och tekniker som härstammar från olika fält, bland annat statistik, marknadsföring, artificiell intelligens och databas administration och forskning (Berry & Linoff, 1997). Detta innebär att det i litteraturen finns en mängd lika ord som författare använder när de menar samma saker medan de använder samma ord när de menar olika saker. Data mining är inte en ensam teknologi utan en samling av verktyg som används för att få fram (extract) information från data.

Turban och Aronson (1998) ger följande definition: *”Data mining is a term used to describe knowledge discovery in databases, knowledge extraction, data archaeology, data exploration, data pattern processing, data dredging, information harvestion and software”*

Singh´s (1999) definition av Data mining är: *”Data mining is the process of extracting meaningful, yet unobvious, information from large databases”*

Debevoise (1999) definition lyder: *”Data mining is the process of extracting valid, previously occult data, and ultimately decoding that information from large databases, using it to evaluate buisness practices that lie beyond the boundaries of ordinary managment decisions”*.

Data mining kan även definieras som ”*undersökandet och analyserandet av automatiska eller semiautomatiska metoder av stora kvantiteter av data i syfte att upptäcka meningsfulla mönster och regler*” (Berry & Linoff, 1997) (Berry & Linoff, 2000).

Programföretaget SAS beskriver Data mining som en process för att söka efter okända mönster och trender i stora datamängder (Url 3). SPSS definition av Data mining är att det är en process för att upptäcka mönster och trender i stora datamängder för att få användbar information för beslutsfattande (Url 4).

Konkreta exempel

I litteraturen beskrivs ett antal områden som lämpar sig för Data Mining. Några av dessa är användningsområden där Data mining kan användas är inom bank och finans branschen. Inom bankbranschen kan tekniken användas för att förutspå bedrägerier, utvärdera risker, genomföra trendanalyser samt som hjälp vid direktmarknadsföring. Inom finansvärlden används det för att förutspå aktiepriser, options handeln och portfölj management. Sjukvården är ytterligare ett område där Data mining kan appliceras. Här används Data mining för till exempel utveckling av mediciner (Groth, 2000) (Fayyad et al.,1996). För att få en klarare uppfattning på vilket sätt Data mining kan appliceras i ett antal skilda organisationer beskrivs tre korta exempel.

Suppliers Relational Management (SRM)

Data mining kan användas som beslutsunderlag vid beslut om leverantörer. Det är användbart när en organisation skall besluta om vilka av de gamla leverantörerna som skall bort och vilka som skall vara kvar. Organisationen behöver ha data lagrat om leverantörerna angående de faktorer som anses vara viktigast. Exempel på sådana faktorer kan vara pris, produkt kvalité, leveranstider och tillgänglighet. Utifrån dessa faktorer kan organisationen göra analyser för att få fram ett beslutsunderlag om leverantörerna skall få vara kvar eller ej. Vid beslut om att organisationen skall anta en ny leverantör måste andra metoder vidtagas. Detta på grund av att ingen information om den nya leverantören finns lagrat sedan tidigare. Andra faktorer får då ligga till grund för dessa beslut.

Personalrekrytering

Data mining kan användas som hjälpmedel vid personalrekrytering. Vid beslut om vilken person som skall anställas kan Data mining vara en stor hjälp. Anta att varje sökande har studerats med hänsyn till 15 variabler. Varje variabel är en egenskap eller förmåga som bedömts mellan skala 1-10. Det kan vara svårt för en rekryterare att bedöma vilken person som är mest lämplig för tjänsten. Med hjälp av Data mining kan man dela in variablerna i kluster (grupperingar).

Rotated Component Matrix [a]

	Component			
	1	2	3	4
CV	,116	,830	,109	-,136
UTSEENDE	,440	,151	,399	,227
UTBILDN	6,399E-02	,128	7,142E-03	,928
ÄLSKVÄRD	,220	,245	,871	-8,15E-02
SJÄLVFTR	,916	-,107	,163	-6,50E-02
KLARSYNT	,863	9,709E-02	,255	1,819E-03
ÄRLIGHET	,219	-,242	,863	5,727E-04
SALESMAN	,910	,223	,103	-4,14E-02
ERFARENH	8,730E-02	,851	-5,46E-02	,211
DRIFTIGH	,800	,349	,156	-5,21E-02
AMBITION	,918	,159	,100	-4,11E-02
IQ	,811	,255	,331	,143
POTENTIA	,747	,326	,413	,224
ANGELGH	,440	,363	,534	-,524
LÄMPLIGH	,383	,797	7,608E-02	8,418E-02

Extraction Method: Principal Component Analysis.
Rotation Method: Varimax with Kaiser Normalization.
a. Rotation converged in 6 iterations.

FIGUR 2:5 KLUSTER FÖR PERSONALREKRYTERING (EGEN PRODUKTION)

Bilden ovan illustrerar fyra kluster där varje variabel har ett värde som visar variabelns tillhörighet till klustret. Vilken gräns som skall sättas som visar att variabeln tillhör ett kluster bestäms av bedömaren. I kluster nummer ett ligger självförtroende, klarsynthet, säljaregenskaper, driftighet, ambition, IQ och potential högt. Detta kluster skulle kunna få namnet *social kompetens* vilket innebär att en sökande med stort självförtroende ofta har en hög ambition och driftighet. Detta visar ovanstående modell genom att dessa variabler visar ett stark samband (högt korrelerade). Inom kluster nummer två ligger CV, erfarenhet och lämplighet hög. Detta kluster skulle kunna få namnet *praktisk erfarenhet* med samma förklaring som ovanstående resonemang. I kluster nummer tre ligger älskvärdhet och ärlighet högt. Denna klustergrupp skulle kunna benämnas *trevlig*. I kluster nummer fyra är det bara utbildning som ligger högt och kan därmed döpas till *utbildning*.

Nästa steg i Data mining processen kan vara att analysera de sökande utifrån dessa fyra kluster istället ifrån 15 olika variabler. Rekryteraren kan använda sig av en teknik till exempel Minnesbaserat resonemang (se sid. 25) som delar in de sökande i grupper som har likande poäng på de olika klustrena. Utifrån dessa kan rekryteraren se vilka som ligger högt inom ett visst kluster eller vissa kluster. Sedan är det rekryteraren som slutligen bestämmer vilken person som skall anställas. En sak som är viktig vid granskning av klustrena är att veta vad varje variabel står för.

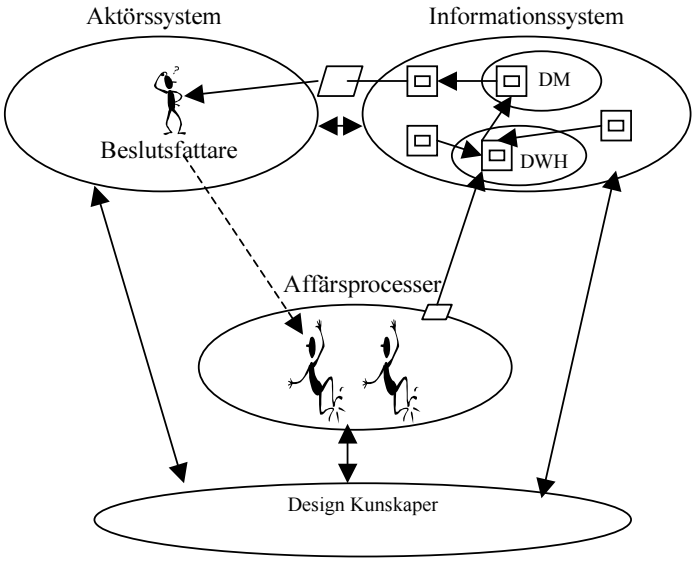
Kundprofilering hos British Telecommunications

Data mining användes av British Telecommunications för att skapa kundprofiler. Med hjälp av tekniken kunde organisationen identifiera kundernas benägenhet att göra inköp samt deras värde när de väl var kunder. Efter skapandet av kundprofiler började British Telecommunications att producera nya produkter med specifika kundgrupper som målgrupp. Resultatet som förväntades var att de skulle få en högre respons än tidigare på kampanjer, ökad avkastning på produkterna samt en större marknadsandel. Det första organisationen gjorde var att samla in relevant data gällande kunder, produkter, faktureringar och annan värdefull historisk data. Efter det började arbetet med att preparera datan. Det vill säga att data hämtades från befintliga system för att matchas, tvättas och formateras. Detta följdes av att tester och analyser utfördes på varje enskilt dataattribut. Detta för att avgöra den prediktiva

förmågan på dataattributet i jämförelse med kundens benägenhet att köpa en produkt. Det som framkom var att det geografiska läget korrelerade med responsen på kampanjer och benägenheten att göra inköp. Efter detta följde den kritiska delen med analysering och byggandet av en modell. För att göra detta använde sig British Telecommunication av tekniken beslutsträd (Se sid 20). Data mining användningen resulterade i detta fall till att responsen på British Telecommunications kampanj ökade med 100 procent.

Data mining i sitt organisatoriska sammanhang

Modellen nedan (figur 2:7) illustrerar Data mining i sitt sammanhang. Modellen består av fyra beståndsdelar som alla är viktiga att ta hänsyn till vid studier av Data mining. Detta beroende på att alla delar påverkar varandra.

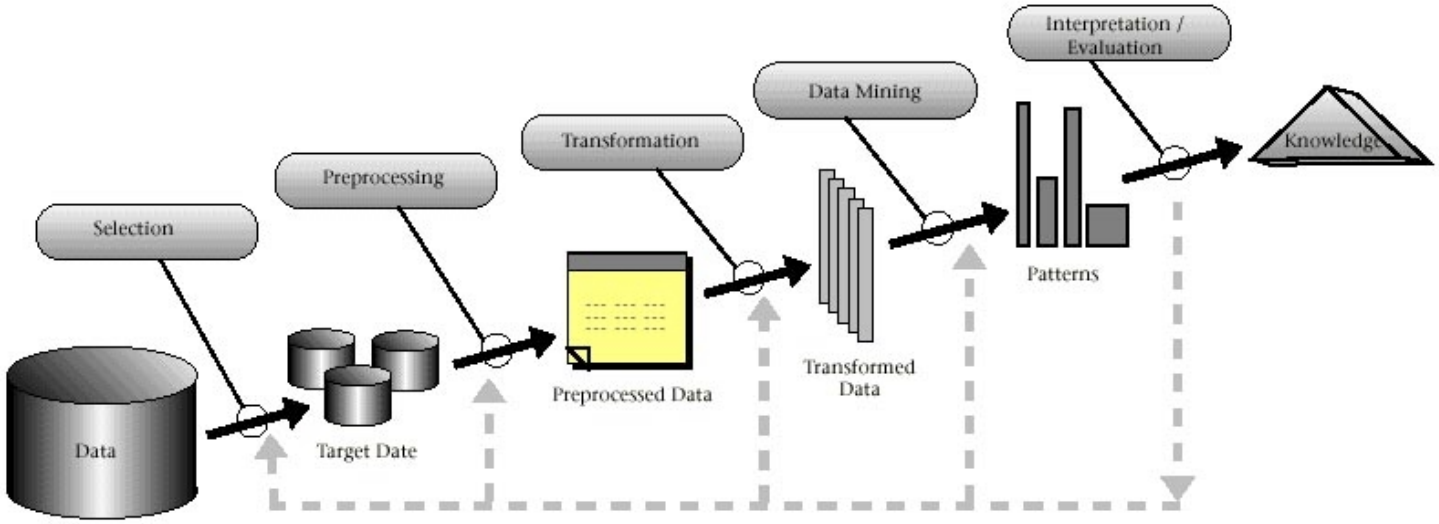


FIGUR 2:7 DATA MINING I SITT SAMMANHANG (MAGOULAS, 2001)

I organisationens affärsprocesser genereras data om verkligheten exempelvis kunder och leverantörer. Denna data lagras i organisationens olika informationssystem. Ett informationssystem består i korthet av tre delar. *Informationsbasen* kan definieras som en databas som innehåller information om händelser som har inträffat eller som kommer att inträffa. *Regelbasen* innehåller regler för hur informationen skall matas in i systemet, hur informationen skall presenteras och hur den skall lagras. *Begreppsbasen* hanterar kopplingen mellan ovanstående två delar det vill säga regelbasen och informationsbasen. Detta möjliggör att informationsbasen kan utföra sin avsedda uppgift och kommunicera med omgivningen (Magoulas, 2001). För till exempel en koncern där dotterbolagen använder sig av olika datasystem kan det vara mycket svårt att samla in relevanta data från dessa och på så sätt få en bra överblick. Det är idag inte helt ovanligt att det inom en organisation finns olika datasystem för ekonomiavdelningen, produktionen och lager, vardera med olika format på data. Här kan ett datalager (Data WareHouse, DWH) användas för att samla in data och sammanställa informationen från de olika systemen inom organisationen och från externa källor (Inmon, 1996). Proceduren att sammanföra data till gemensamt format från olika

databaser inom organisationen kallas ofta för att ”tvätta” datan. Det finns även möjlighet för organisationen att använda sig av dataskafferi, svenskans ord för engelskans Data mart. Dataskafferi skiljer sig från datalager med hänsyn till att dataskafferi ofta innehåller en del av informationen som finns i ett datalager. Dataskafferiet används för att möta specifika behov för en del av organisationen (Humphries et al., 1999) (Kachur, 2000). Dessa kan finnas utlokaliserade inom organisationen och med hjälp av något övergripande centralt system kan dessa sedan kopplas ihop för samkörning. Speciellt värdefullt blir ett datalager när det har varit i drift en tid och analyser kan göras med hänsyn till historiska data. När datan som finns i datalagret skall användas i Data mining systemet (DM) sker en integrering mellan dessa informationssystem. Vid integrering är det att viktigt att systemen talar samma språk vad det gäller information, regel och begreppsbas. Det skall också gå att härleda datan till dess ursprung. För att kunna härleda och kontrollera den data som finns i datalagret används metadata för att beskriva struktur och regler för innehållet. För att kunna möjliggöra en effektiv och optimal användning av Data mining behöver den data som samlats in och lagrats vara strukturerad i ett enhetligt format, som i ett datalager. Datalagret underlättar de avancerade sökningar och sammanställningar som behövs för beslutsstöd och analys inom en organisation. Datalagerkonceptet utvecklades allt som IT professionalister förstod att strukturen på data för transaktionsrapportering signifikant skilde sig från strukturen som krävs vid analysering av data. En organisations datalager innehåller idag ”snapshots” av innehållet i det operationella datalagret. När data har hämtats från datalagret till Data mining system där ett underlag analyserats fram träder akörens roll som beslutsfattare in. I aktörssystemet beaktas underlaget och ett eventuellt beslut fattas. Vanligtvis återkopplas dessa beslut till organisationens affärsprocesser. Beståndsdelen design kunskaper är relationen mellan aktör, informationssystem och affärsprocesser.

Betraktas Data mining som ett verktyg för att assistera aktörerna att finna användbar information i stora mängder data ingår det i ett sammanhang som litteraturen benämner som knowledge discovery in databases. Vi kommer i fortsättningen att använda oss av den engelska termen knowledge discovery in databases eller dess förkortning KDD då det inte finns någon lämplig svensk motsvarighet. Det är ett begrepp som myntades i slutet av 1980-talet. Definitionen av knowledge discovery in databases är: *The non-trivial process of identifying valid, novel, potentially useful, and ultimatelly understandable patterns in data*" (Fayyad et al.,1996). Begreppet fokuserar på processen från att samla in och lagra data för att i slutet av processen finna användbar och intressant kunskap. Varje steg i processen fungerar som input till nästa steg. I denna process är Data mining ett steg där syftet är att finna mönster i från data. Tanken bakom att Data mining ingår som en del i en större process är att motverka risken från att hitta meningslösa och ogiltiga mönster.



FIGUR 2:6 KNOWLEDGE DISCOVERY I DATABASES PROCESSEN (FAYYAD ET AL., 1996)

KDD är en iterativ process vilket betyder att vissa steg i processen ibland måste upprepas. Första steget i knowledge discovery processen är att förstå verksamhetsområdet och att identifiera målen för projektet. Anledningen är att det är viktigt att kunna identifiera den data som behöver användas för att kunna utföra projektet. I första steget ingår också att ta tillvara på relevant kunskap som erhållits genom tidigare projekt. Det andra steget innefattar att välja den datamängd som skall användas. Tredje steget består av att datan skall tvättas och prepareras. Här fattas det beslut om vilka strategier som skall användas när det gäller att ta hand om förlorad data. Genom att filtrera data så minimeras störande och avvikande värden. Det fjärde steget innebär att datan reduceras och transformeras. Reduceringen sker främst genom att de variabler som inte är relevanta sorteras bort. Beroende på vilket mål som användaren arbetar efter bestäms hur datan skall representeras. Femte steget består av att matcha målen som klargjordes i första steget i KDD processen med en passande Data mining funktion till exempel klassifikation eller regression. Det sjätte steget innehåller val av Data mining algoritm och val av metod som användaren skall söka efter mönster i data materialet. Sjunde steget karakteriseras av att söka efter mönster som är av intresse. Det åttonde steget är att tolka de funna mönstren. Genom till exempel visualisering av mönstret kan användaren bedöma resultatet. I vissa fall kan det vara nödvändigt att gå tillbaka till steg ett igen. Sista steget består av att använda den kunskap som erhållits genom att dokumentera rapportera till berörda parter samt informationssystem. I denna fas genomförs en kontroll av det nya resultatet mot tidigare resultat (Brand Gerritsen, 1998a) (Fayyad et al.,1996). För att organisationen skall kunna ta lärdom av Data mining processen och resultaten bör de återanvänds (Putten, 1998)(Fayyad et al., 1996).

Funktioner

I detta avsnitt kommer funktionerna inom Data mining att presenteras. Funktionerna kan betraktas som olika strategier för att analysera data (Groth, 2000). Strategier väljs utifrån vilket resultat som användaren önskar.

Klassificering

Klassificering används ofta inom Data mining för att bestämma kundprofiler, kvalitetsförutsägelser, kreditbeviljande och riktad marknadsföring (Söderström, 1997) (Brand & Gerritsen, 1998b). Angreppssättet går ut på att indela poster (exempelvis individer) i föredefinierade kategorier. Posterna kan analyseras med uppgifter om kön, ålder, kreditstatus och så vidare (Söderström, 1997). De objekten som skall klassificeras finns i databasen som *rader*. Klassifikationsfrågan är karakteriserad av väldefinierade klasser och ett träningsset bestående av preklassificerade exempel. Uppgiften är sedan att bygga en modell som kan appliceras på oklassificerad data för att klassificera den. Resultaten av klassificeringen blir i form av ja-nej resultat. Tekniker som lämpar sig väl för klassificering är beslutsträd och minnesbaserat resonemang. Länkanalys kan också vara användbart under vissa omständigheter (Berry & Linoff, 1997).

Associering och Sekvensiering

Associering och sekvensiering eller varukorgsanalys som det ibland benämns är mycket vanligt att användas för att bestämma kundernas köpbeteende. Andra användningsområden för associations och sekvensiering är att finna mönster i transaktionsdata och många organisationer samlar just denna datan. Att förstå mönster av beteende eller aktivitet kan ge värdefull insikt (Brand & Gerritsen, 1998c). Ett associations verktyg skulle till exempel finna regler som: när kunden köper en vara köper de också någon annan vara för 50 procent av fallen. En marknadsanalytiker kan använda varukorgsanalys för att planera kuponger och prisnedsättning av varor, produkt placeringar och korsförsäljning (crossmarketing).

Associations algoritmer kan bara användas på kategori indelad data. Vid användningen av icke kategoriserad data måste datan delas in i intervaller. En annan vanlig karakteristika av associations regler är produkt ordning. Den delar in liknande produkter i grupper. Sekvensiering påminner om associations tekniken. Skillnaden är att den lägger till tid och produceras regler som till exemepel att en kund som köper en tv, inom en månad köper en video. All data som används av associerings algoritm är gjorda av entiteter och attribut. Entiteten kan bli en varukorg och attributen alla saker som inhandlats. Datan som skall användas i associering behöver att vara i två format. Dessa kallas för horisontell och vertikal. I det horisontella formatet är det en rad för varje entitet och en kolumn för varje attribut. Ett problem med det horisontella formatet är att antalet entiteter kan bli ganska många. Ett annat problem är att ”schemat” för datan varierar. När en ny produkt läggs till i varukorgen eller när en produkt är kategoriserad på ett annorlunda sätt måste ”schemat” förändras för att lägga till eller omorganisera kolumnerna. Det vertikala formatet som är mer använt av Data mining applikationer eliminerar dessa risker genom att använda multipla rader för att lagra entiteter och en rad för varje attribut. Varje rad med sin entitet är kopplad med ett ID-nummer.

Vid varukorgsanalys är det lätt att finna slumpmässiga, triviala samband. Ett exempel är om man hittar ett samband som säger att 30 procent av de som köper mjölk också köper bröd. Detta samband behöver inte vara ett riktigt samband det kan vara så att 30 procent av inköpen i affären innehåller mjölk. Om så är fallet säger detta samband oss ingenting (Groth, 1999).

Varukorgsanalys baseras på ett belägg – övertygelse ramverk (support – confidence framework). *Belägg* definieras enligt Groth som minimum procent av transaktioner som innehåller A och B. *Övertygelse* definierar han som transaktioner som innehåller A som också innehåller B. Exempel: en organisation har en miljon transaktioner. Av dessa transaktioner innehåller 10 000 A och B. Belägg blir då 10 000/1 000 000 = 1 procent. Organisationen har

50 000 transaktioner som innehåller A. Av dessa innehåller 10 000 också B. Övertygelse blir då 10 000/50 000 = 20 procent.

Klustring

Klustring är en deskriptiv teknik som används för att dela upp en datamängd i grupper så att varje post som har likartade egenskaper förs till samma grupp. Fördelen med kluster är att man får en uppfattning om hur datamaterialet ser ut (Groth, 1999) (Brand & Gerritsen, 1998a). Exempel på områden där klustring används är enligt Söderström (1997) marknadssegmentering och att upptäcka bedrägerier. Han nämner också att klustring ofta används som beredning till någon annan form av Data mining. Till exempel så kan klustring vara det första steget för en marknadssegmentation. Kunder med liknande köpbeteende delas in i kluster för att se vilken sorts reklam som passar bäst till vilket kluster. Ett problem som kan uppstå med Data mining är att man hittar för många mönster som, i slutändan, inte bidrar med relevant information. För att lösa detta problem kan man tillämpa klustertekniken, en av de få teknikerna som använder sig av indirekt discovery-modellen (Berry & Linoff, 1997).

Klustring är enligt Brand och Gerritsen (1998a) en väldigt subjektiv teknik. Detta på grund av att ett avstånds mått måste anges. Klustringen är helt beroende på vilket mått som används. En konsekvens av detta kan vara att två användare som arbetar med samma data få helt olika resultat. Skillnaden mellan klustring och klassifikation är att klustring inte har några fördefinierade klasser.

Prediktion

Berry och Linoff (2000) är tveksamma till att prediktion (förutsägelse) är en enskild funktion. De hävdar att prediktion är det samma som klassifikation eller uppskattning förutom att raderna är klassificerade för att stämma överens med förutspådda framtida beteende eller uppskattade värden i framtiden. Den enda skillnad som finns är att vid användning av klassificering och uppskattning kan man inte gå tillbaka och se om resultatet stämde. Med prediktion är datan klassificerad enligt framtida beteenden och uppskattningar. Det enda sättet att få reda på om resultatet stämmer är att vänta och se vad som händer (Berry & Linoff, 2000). Någon av de tekniker som används till klassifiktaion och uppskattning kan anpassas till prediktion genom att använda träningsexempel där värdet av variablerna skall förutsägas. De historiska datan används för att bygga en modell som förklarar det aktuella observerade beteendet. När sedan denna modellen blir applicerad på aktuella indata är resultaten en förutsägelse om framtida beteende (Berry & Linoff, 2000).

Varukorgsanalys, minnesbaserat resonemang, beslutsträd och neurala nät passar till att användas till prediktion. Val av teknik beror på hur indatan ser ut och typen på värdet som skall förutsägas och hur viktigt det är med att prognosens tydlighet.

Uppskattning (Estimation)

Uppskattning arbetar med resultat som kan anta kontinuerliga värden (Groth, 1999). Funktionens syfte är att förutspå ett specifikt värde. Det används i fall där det förutspådda resultatet kan ta obegränsat antal möjliga värden. Funktionen är nära besläktat med klassifikation. Uppskattning är den vanligaste funktionen som används inom Data mining området. Den används för att bestämma kunders beteende, att förutsäga en affärsvinst, att upptäcka bedrägeri, uppskatta en familjs totala inkomst, uppskatta antalet barn i en familj och för att identifiera kandidater för medicinska procedurer (Brand & Gerritsen, 1998b). En teknik som passar bra för uppskattning är neurala nät.

Visualisering

Visualisering går ut på att beskriva en komplicerad databas. Detta görs för att få en förståelse för människor, produkter, processer eller det som producerade datan från början. Enligt Berry och Linoff (2000) lämpar sig visualisering när ett datamaterial skall beskrivas, men det kan vara svårt att alltid se bra saker.

Design metoder

Följande avsnitt kommer att presentera metoder och tekniker samt ett urval av de verktyg som finns inom Data mining området.

Metoder

Berry och Linoff (2000) skiljer på två typer av arbetsätt vid användningen av Data mining. Beroende på användarens avsikter finns det möjlighet att arbeta med direkt Data mining och då använda sig av en top-down metod. Det andra alternativet är att arbeta med indirekt Data mining genom att använda sig av bottom-up metoden.

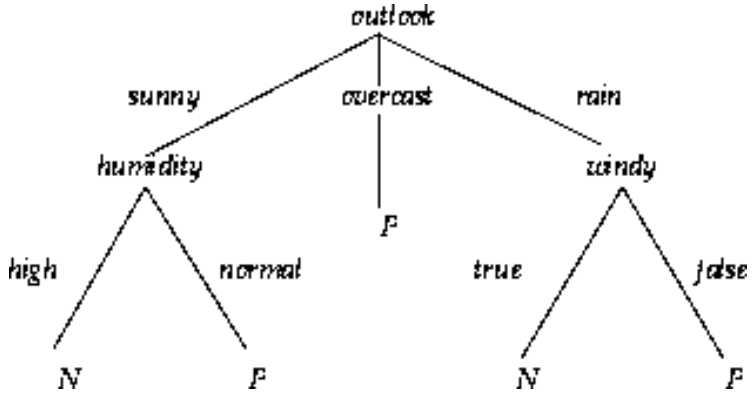
Direkt Data mining är en top-down metod. När direkt Data mining används lägger användaren inte någon vikt på hur modellen arbetar. Bästa möjliga resultat sökes. Metoden används när användaren vet vad den letar efter. Oftast används här prediktiv modellering. Samlad information om tidigare händelser används för att förutspå kommande. För att lyckas så bra som möjligt krävs mycket träningsdata. Dessa modeller är enligt Berry och Linoff (2000) aldrig 100 procent tillförlitliga. Indirekt Data mining är en bottom-up metod. Metoden används för att finna mönster i datan där användaren får avgöra resultaten är intressanta eller ej. Här har användaren chans att se hur teknikerna arbetar och hur den har kommit fram till resultatet.

Tekniker

Inom Data mining finns det en mängd olika tekniker som kan användas. Vi har i detta avsnitt valt att presentera de mest förekommande tekniker som finns i facklitteraturen. Detta för att vi vill ge en bred förståelse för vad Data mining är uppbyggt kring. Varje teknik kan i sig innehålla flera olika algoritmer. Dessa algoritmer kommer inte att presenteras i denna uppsats då vi har avgränsat oss från att gå djupare in på dessa.

Beslutsträd

Beslutsträd är en teknik som skapar en modell som både är deskriptiv och prediktiv. Tekniken heter beslutsträd på grund av att resultatet av modellen har en struktur som liknar ett träd. Visualiseringen av beslutsträdet ger en god insikt och översikt för att förstå modellen. Mycket på grund av detta har beslutsträd blivit en intressant och populär modell. Beslutsträd det är en maskinlärningsteknik som används till klassifikation. Rader och objekt är indelade i grupper baserade på ett attributs värde (Thuraisingham, 1999). Beslutsträd omfattar ett antal specifika algoritmer som inkluderar klassifikation och regressionsträd (CART), Chi-squared automatic interaction detection (CHAID).



FIGUR 2:8 BESLUTSTRÄD (DILLY, 1997)

Varje box i trädet representerar en nod. Topp noden kallas för rot noden. Ett beslutsträd växer från rot noden som sedan delar datan på varje nivå till att forma nya noder. Noder som är i ändarna av grenarna kallas för löv (leaf) noder och spelar en speciell roll när trädet är används för prediktion. Varje nod innehåller information om antalet instanser av den noden och om distributionen av värdet på beroende variabler. Instansen i roten är första split i det fallet splittas datan i två nya noder osv. Det primära utkastet av beslutsträdsalgoritmen är trädet självt. Träningsprocessen som skapar beslutsträdet kallas vanligen induktion och kräver mindre antal körningar genom ett träningsset. Detta gör algoritmen mindre effektiv än Naïve-Bayes algoritmer som kräver en körning, men signifikant mer effektiv än neurala nät vilket kräver stora antal körningar ibland tusentals. För att bli mer precis, antalet körningar som krävs för att bygga beslutsträdet är inte mer än antalet nivåer i trädet. Det finns inget förutbestämd antal nivåer fastän komplexiteten av trädet som mäts av djup och bredd generellt ökar om antalet oberoende variabler ökar (Brand & Gerritsen, 1998e).

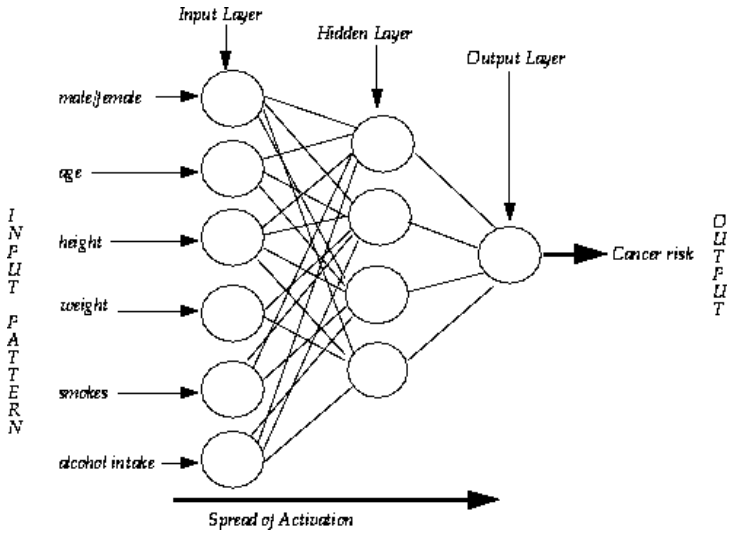
Även när beslutsträd inte används för prognoser kan tekniken förutse intressant deskriptiv information. När trädet växer är det en iterativ process vilket involverar att dela datan till gradvis mindre dataset. Varje iteration betraktar bara datan i en nod. Första iterationen betraktar rot noden som innehåller all data. Algoritmen börjar med att analysera datan för att hitta den oberoende variabeln som till exempel inkomst, civiltillstånd, som sedan används som delningsregel vilket resulterar i noder som är mest olika från varandra med respekt för den beroende variabeln. Det finns flera alternativa vägar för att mäta denna skillnad. Oavsett vilken mätmetod som används kräver alla metoder en korstabell mellan beroende variabler och varje oberoende variabel. En viktig karakteristika när trädet delar algoritmen är att den är snål (greedy). Snåla algoritmer tar beslut lokalt hellre än globalt. När beslutet tas att dela en nod så tittar inte algoritmen framåt i trädet för att se ifall ett annat beslut skulle producera ett bättre resultat. När en nod är delad görs samma process på nya noder vilket innehåller submängder av data i huvud noden. Variablerna blir analyserade och den bästa delningen väljs. Vid vilken punkt skall det växande trädet stoppas? Träd algoritmen har vanligen ett flertal stoppregler. Dessa regler baseras på flera faktorer till exempel maximum träd djup, minimum av element i en nod som överväger att delas eller är nära lika, minsta antalet element det måste vara i en ny nod. Reducering är en vanlig teknik som används för att göra träden mer generella. Men hur bestämmer reducerings algoritmen var trädet skall reduceras? Det är flera algoritmer men den som är mest tilltalande för att kontrollera stickprov med känd och verifierad relationer mellan oberoende och beroende variabler. Genom att jämföra prestationen av varje nod (som ett mått på deras noggrannhet) med deras subträd, kommer det

att bli tydligt vilka delningar som kommer att behöva reduceras för att uppnå högsta övergripande noggrannhet på kontroll stickprovet (Brand & Gerritsen, 1998e).

En av fördelarna med beslutsträd är att modellen gör det möjligt för en bredare grupp av användare att förstå. Beslutsträd har tydliga värden som både prediktiva som beskrivande modeller. Prognoser kan göras på ”fall till fall” genom att navigera i trädet. Efter att Data mining applikationen har skapat ett beslutsträd måste en affärsanalytiker undersöka modellen. Det är viktigt att undersöka beslutsträdet även om det har stoppregler. Detta för att avslöja noder eller subträd som är oönskade på grund av överträning eller innehåller regler som områdesexperten finner olämplig i sammanhanget. Modellens validitet måste också testas på ett oberoende dataset. När noggrannheten har blivit mätt på det oberoende dataset och bestämts för att accepteras är trädet eller dess regler redo för att användas som en prognos. Trädet bör testas med jämna mellanrum för att försäkra att dess önskvärda noggrannhet bevaras (Brand & Gerritsen, 1998e).

Neurala nätverk

Enligt Söderström (1997) är ett neuralt nätverk en teknik som söker mönster i en del av ett dataset för att bygga en modell som sedan kan appliceras på hela datasetet. Neurala nätverk baseras på en tidig modell av en mänsklig hjärna. Nätet är en matematisk funktion som skapar ett output baserat på ett set av input. Nätet använder sig inte av några deskriptiva komponenter i en neural nätverks modell utan den är enbart prediktiv. På grund av detta är nätets val svåra att förstå. Neurala nätverk kallas därför ibland för ”black box” teknologi (Söderström, 1997) (Brand & Gerritsen, 1998d) (Thuraisingham, 1999).



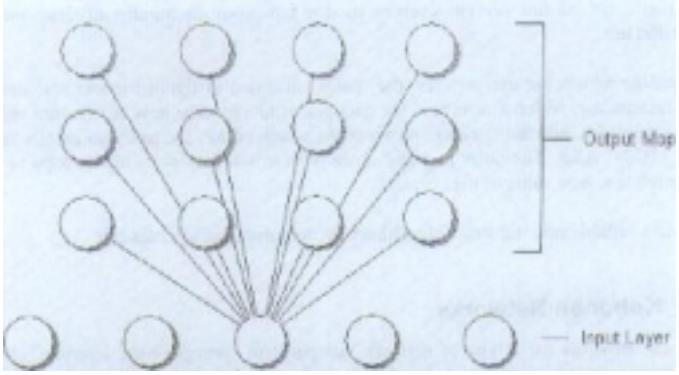
FIGUR 2:9 NEURALA NÄTVERK (DILLY, 1997)

Ett neuralt nätverk lär sig av tidigare erfarenheter och är användbar för att upptäcka icke kända relationer mellan ett set av indata och ett resultat. I ovanstående figur skall en risk för cancer uppskattas. Lagret med indata (input layer) består av sex variabler. I det dolda lagret (hidden layer) vägs varje variabel för att till sist producera ett resultat (output layer). Antalet indata variabler, noder i det dolda lagret, resultat samt ”vägnings” algoritmerna bestämmer komplexiteten på ett neuralt nätverk. En styrka med det neurala nätverket är dess förmåga att

skapa korrekta förutsägelser på komplexa problem. Tekniken används främst till uppskattning eller vid kontinuerliga värden. Två stora nackdelar med tekniken är att den är svår att förstå sig på då den enbart är prediktiv samt att det finns en stor risk för att nätet övertränas. Meningen med ett neuralt nätverk är att det skall tränas på en liten del av ett data set och finna önskvärda relationer. När ett neuralt finner det oönskade specifika relationer kallas det att nätet övertränas eller överanpassa. Detta problem kan testas genom att periodvis testa resultatet på träningsdata (Groth, 2000).

Kohonen networks

Kohonen networks är en teknik som utför en oövervakad inlärning, det vill säga att den inte har någon resultat att förutspå. Kohonen networks används för klustring eller segmentering av ett datamaterial baserat på mönster på indata fälten. Tekniken gör ett grundläggande antagande att kluster är formade från mönster som delar liknande egenskaper och grupperar därför liknande mönster i samma grupp. Kohonen networks är vanligtvis en eller två dimensionella rutnät eller arrayer av artificiella neuroner. Varje neuron är kopplad till varje indata och en vikt sätts på varje koppling. Det finns egentligen inget utdata lager dock kan Kohonen kartan innehållande neuronerna ses som ett resultat (SPSS, 1999).



FIGUR 2:10 KOHONENS NETWORK (SPSS, 1999)

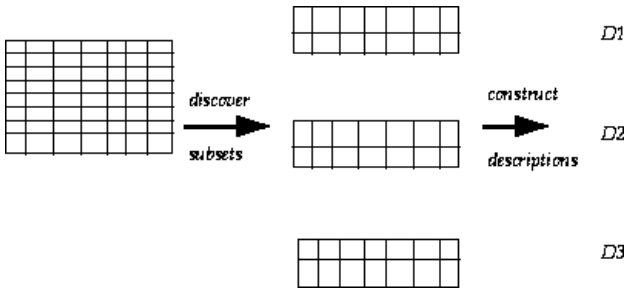
När en rad är presenterad i rutnätet är dess mönster jämförda med de artificiella neuronerna inom rutnätet. En rad representeras i bilden ovan som en cirkel. Den artificiella neuronen som har mest lika mönster med indatan ”vinner” datan. Detta medför att vikten av de artificiella neuronerna kan anpassas för att likna indata materialets mönster. Kohonen networks anpassar också vikterna på de artificiella neuroner i omgivningen med liknande mönster. Detta medför att den mest liknande neuronen och omgivningsnoderna flyttas till en lägre nivå, det vill säga till positionen av raden i indata lagret. Indata lagret benämns i ovanstående bild som input layer. Resultatet efter att datan har passerat genom nätverket en X antal gång skapas en karta (output map) som innehåller kluster av rader vilka svarar mot olika typer av mönster inom datamängden (SPSS, 1999).

Kohonen networks kan beskrivas som en låda som har blivit indelad i ett rutmönster av fack och en väska med bollar i flera olika färger. Ett barn har fått som uppgift att placera bollarna i facken så att de är organiserade utefter färg. Sammankopplade fack måste innehålla bollar av liknande färger och icke sammankopplade fack måste innehålla bollar av mycket olika färger. Den första bollen som barnet tar från väskan är blå. Bollen placeras slumpmässigt ut i ett av facken. Den andra bollen som plockas är rött vilket placeras i ett annat fack då färgen markant

skiljer sig från blått. Det fack som den röda bollen placeras i är inte sammankopplat med det fack som den blåa bollen placerats i. Den tredje bollen som plockas är orange. Barnet bestämmer sig för att orange liknar rött och på grund av detta placeras bollen i facket som innehåller den röda bollen. De omgivande noderna anpassar sig lite till detta och börjar dra till sig bollar som har en nyans av rött. Nästa boll som plockas är grön. Denna boll placeras någonstans mellan den röda och den blåa bollen. Detta medför att omgivande noder anpassas och kommer att attrahera bollar som är mer blåa och gröna. Denna process kommer att fortsätta tills alla bollar är organiserade in i kluster med karaktärsdrag som mest liknar blå, röd och grön (SPSS, 1999).

K-means (Cluster Detection)

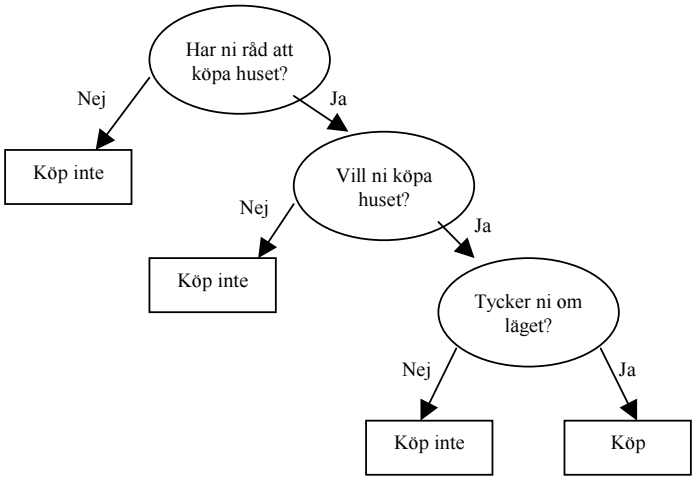
K-means är en teknik som används vid klustring. Tekniken bygger modeller på att den hittar rader med data som är liknande. Dessa grupper kallas för kluster. K-means är indirekt Data mining då målet är att finna förut okända likheter i data materialet (Berry & Linoff, 1997).



FIGUR 2.11 KLUSTER (DILLY, 1997)

Regel induktion

Regel induktion försäkrar att endast ett fält sätts som målvariabel och att alla datatyper är helt instansierade. Algoritmen delar datan i två delar och försöker att bygga en regel på en del av datan för att sedan testa regeln på resterande data. Denna process repeteras tills algoritmen hittar en regel som stämmer med en fördefinierad del av fallen. Det resulterar i ett beslutsträd som överensstämmer med träningsdatan. Då datamaterialet innehåller extremvärden eller unika värden så försöker algoritmen även att förutspå dessa. Detta kan resultera i att modellen har hög överensstämmelse med träningsdatan men låg överensstämmelse på okänt data (SPSS Training, 1999).



FIGUR 2:12 REGELINDUKTION (SPSS TRAINING, 1999, MODIFIERAD AV FÖRFATTARNA)

Ovanstående modell (figur 2:12) visar hur resultatet av Regelinduktion kan visualiseras. Resultatet i denna modell visas i form av ett beslutsträd. Varje faktor är delad upp i undergrupper och ”reglerna” kan utläsas som ”Om – Så” (”IF – THEN”). Ett exempel från bilden ovan skulle lyda enligt följande: *Om* ni inte har råd att köpa huset *Så* köp det inte.

Naïve-Bayes

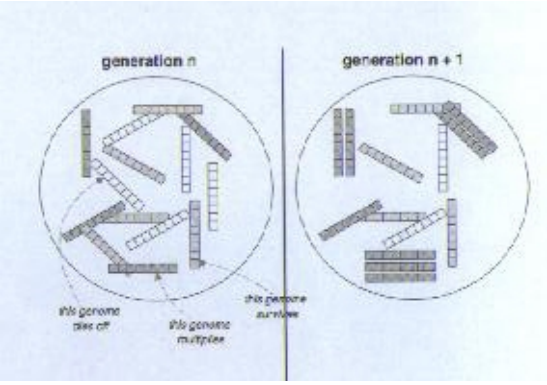
Naïve-Bayes är en teknik som begränsar resultaten till kategorisk data. Tekniken kan endast användas vid klassificering. Enkelheten och snabbheten gör tekniken till ett idealiskt verktyg för att undersöka datamaterialet (Brand, & Gerritsen, 1998b). Naïve-Bayes behandlar varje variabel oberoende och mäter effekten som olika värden har på resultatet. De oberoende resultaten kombineras för att göra en klassificering (Berry & Linoff, 2000).

Minnesbaserat resonemang

Minnesbaserat resonemang eller K-nearest neighbor som det också benämns skiljer sig från andra tekniker på så sätt att den inte har någon träningsfas. Det är modellen som är träningsfasen. Prognosen för nya fall görs genom att hitta grupper som har liknande fall. Den använder övervägande reslutat för det prognostiserade värdet (Brand & Gerritsen, 1998b). Det första steget i minnesbaserat resonemang är att identifiera liknande fall som det aktuella problemet. Därefter skall informationen från de fallen appliceras på det aktuella fallet. Tekniken finner grannar (neighbors) som är liknande det nya fallet och använder sig av grannarna för klassifikation och för att förutse framtida händelser. Minnesbaserat resonemang är inte beroende av ett viss format på raderna. Tekniken bryr sig endast om två funktioner, distansfunktionen och kombinationsfunktionen. Distansfunktionen tilldelar en distans mellan raderna och kombinationsfunktionen kombinerar resultatet av ”grannarna” för att komma fram till ett gemensamt resultat (Berry & Linoff, 1997). Vid användningen av tekniken är det tre detaljer som måste lösas. Det första är att välja rätt mängd med historiska poster som skall ingå i modellen. Det andra är att bestämma på vilket sätt de historiska posterna skall presenteras på. Det sista är att bestämma distans och kombineringsfunktionen samt hur många grannar en post skall ha.

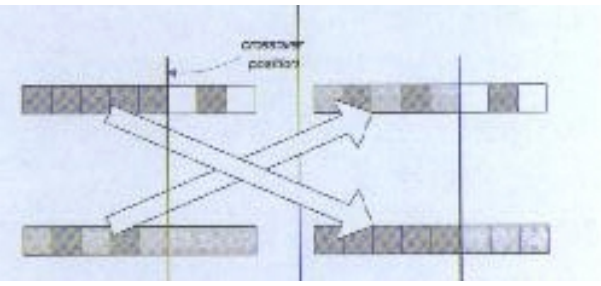
Genetiska algoritmer

Genetiska algoritmer använder sig av biologiska processer för att finna den optimala gruppen parametrar som beskriver en prediktiv funktion. Tekniken används för direkt Data mining (se sid. 20). En likhet som genetiska algoritmer har med klassisk statistik är att modellen som används måste vara känd i förväg. Genetiska algoritmer utvecklar successiva generations gener som progressivt blir mer och mer lämpliga. Målet är att optimera lämpligheten av generna i populationen. Inom naturvetenskapen är lämplighet när en organism överlever och reproducerar sig. Genetiska algoritmer består av följande steg: Det första är att identifiera gener och lämplighetsfunktioner för att sedan skapa en initial generation av gener. Därefter skall den första generationen modifieras genom att applicera selektion, crossover eller mutation. Modifieringen skall upprepas till dess att lämpligheten av populationen inte längre förbättras (Berry & Linoff, 1997).



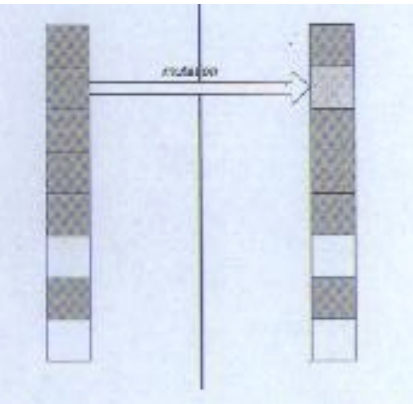
FIGUR 2:13 GENETISKA ALGORITMER – SELEKTION (BERRY & LINOFF, 1997)

Selektion är när storleken på populationen är konstant men förbättrar lämpligheten på följande generation (Berry & Linoff, 1997).



FIGUR 2:14 GENETISKA ALGORITMER – CROSSOVER (BERRY & LINOFF, 1997)

Vid crossover kombineras två gener. En crossover position bestämmer var generna skall brytas och återkombineras (Berry & Linoff, 1997).



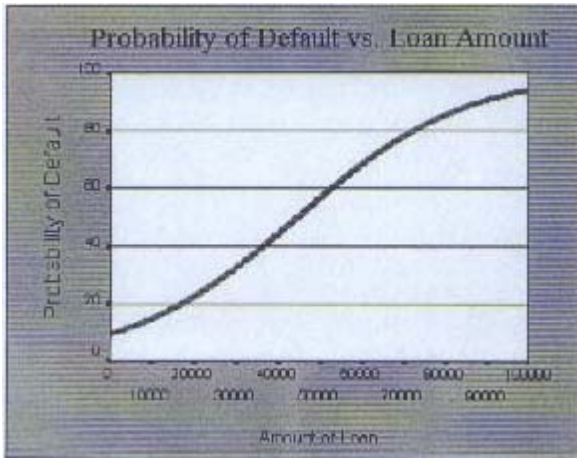
FIGUR 2:15 GENETISKA ALGORITMER – MUTATION (BERRY & LINOFF, 1997)

Mutation är då en ovanlig slumpmässig förändring sker i en gen. Detta medför att kännetecken som inte finns i originalet finns i den nya populationen (Berry & Linoff, 1997).

Genetiska algoritmer används inte lika omfattande inom Data mining som vissa andra tekniker. Detta tror Berry och Linoff (1997) beror på att Data mining fokuserar på klassificering och prediktion och inte optimering. De säger vidare att det är väldigt få produktpaket som innehåller tekniken. Genetiska algoritmer används för att förbättra neurala nätverk och minnesbaserat resonemang.

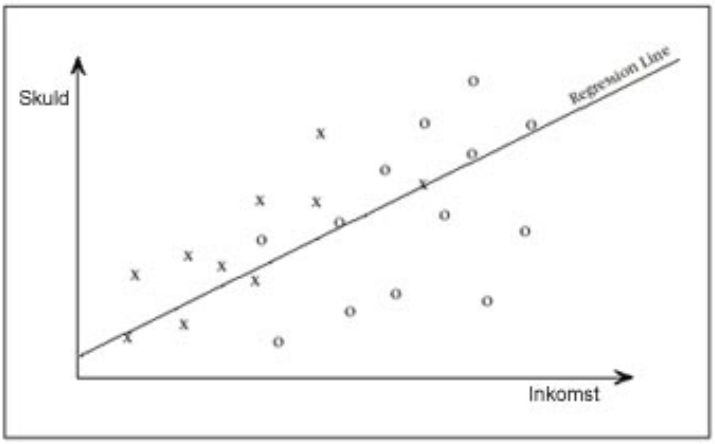
Regression

Regression betraktas oftast som en statistisk teknik och inte en Data mining teknik. Berry och Linoff (2000) anser att denna uppdelning är godtycklig. Enligt dem går Data mining först och främst ut på att skapa förutspående modeller. Detta är precis vad regression gör. Berry och Linoff (2000) tror att anledningen till att regression inte anses som en renodlad Data mining teknik är att den är svår att marknadsföra som en ledande framkants teknologi. Regression kombinerar alla indata variabler (oberoende variabler) till en formel för att förutspå en målvariabel (beroende variabel). Det finns två typer av regression, linjär och logistisk. Den vanligaste formen av regression är linjär regression (Berry och Linoff, 2000).



FIGUR 2:16 LOGISTISK REGRESSION (SPSS, 1999)

Logistisk regression (figur 2:16) uppskattar möjligheten för en viss händelse. Den använder sig av observerade faktorer parade med förekomster eller icke förekomster av händelsen för att skapa en modell över möjligheten av förekomsten under olika faktorer (Url 6).



FIGUR 2.17 LINJÄR REGRESSION (FAYYAD, 1996)

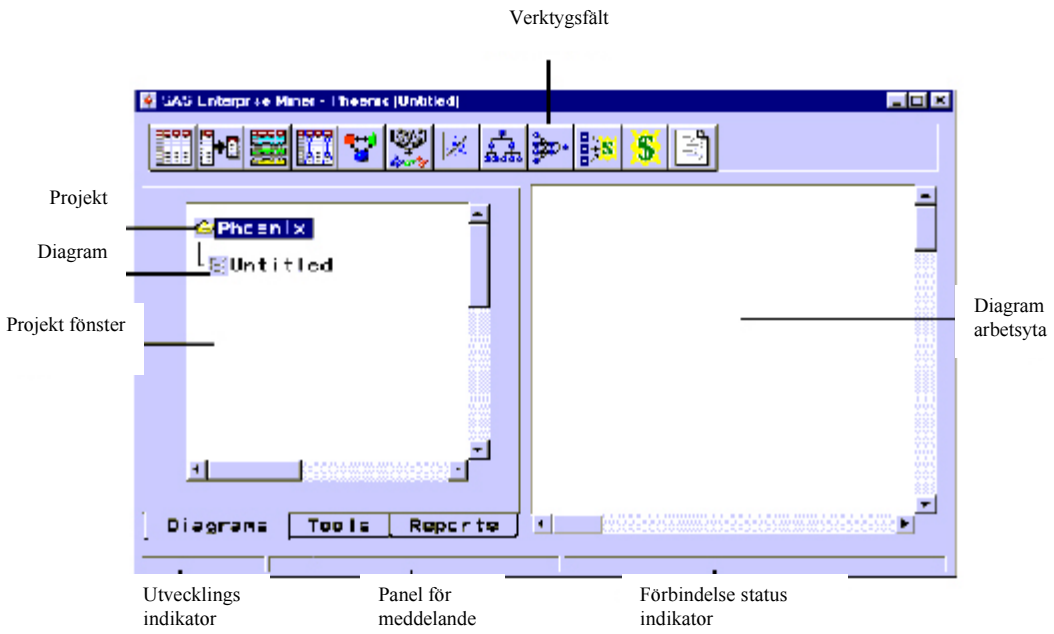
Linjär regression (figur 2:17) är när formeln skapar en rak linje som går genom en rymd av variabler. Regressionen skapar den mest lämpliga linjen (Berry och Linoff, 2000).

Verktyg

I detta avsnitt kommer vi att presentera Data mining verktyg. Syftet är att ge en kort orientering på ett urval av verktyg som finns på marknaden idag och vad dessa erbjuder.

SAS Enterprise Miner

SAS Enterprise Miner är en mjukvara med ett grafiskt användargränssnitt som enligt SAS (Url 3) är lätt att använda för affärsanalytiker likaväl som för kvantitativa experter. Användargränssnittet är uppbyggt på peka och klicka (point-and-click) gränssnitt. Gränssnittet är uppbyggt med välkända skrivbords objekt såsom verktygsfält, menyer, fönster och dialog rutor för att erbjuda båda grupperna ett komplett Data mining verktyg. Enterprise Miners fönster består av följande komponenter: *Projekt Navigator* som används för att managing projekt och diagram. *Diagram Workspace* som används för att bygga, editera och running processer flow diagram. *Verktygsfält* som består av ett urval (subset) av Enterprise Miners verktyg. Där finns bland annat de noder som är de mest använda. Användaren kan själv lägga till eller ta bort verktyg från verktygsfältet. *Noder* som är en samling ikoner som hjälper användaren att utföra stegen i Data mining processen.



FIGUR 2:18 ENTERPRISE MINERS ANVÄNDARGRÄNSSNITT (SAS, URL 3)

Enterprise Miner är uppbyggt på noder som användaren har som hjälp vid användningen. SAS har delat in noderna i olika delar beroende på vad noderna utför. Noderna kan sedan grupperas beroende på var i SEMMA processen (se under rubrik Processer) de används. Steg ett i SEMMA processen är att göra ett urval (Sampling) av det data material som skall användas. En av noderna som kan användas här är *Input Data Source Node*. Denna nod hjälper användaren att komma åt och query det data set som skall användas för Data mining projektet. Flera Input Data Source Noder kan användas i ett projekt för att definiera multipla källor av indata. En annan urvals nod är *Sampling* Noden. Denna nod hjälper till att extract ett urval indata källorna. Denna nod rekommenderas i fall där data seten är enormt stora då den hjälper till att minska tiden för att bygga en lämplig modell. Den sista noden i första steget är *Data Partition Noden.* Data Partition noden hjälper till att partitionera data setet efter urvalsprocessen innan konstruktionen av modellen.

Det andra steget i SEMMA processen är att undersöka (Explore) data materialet. Här erbjuder Enterprise Miner användarna fem stycken olika noder. Den första noden som vi skall ta upp är *Distribution Explorer Node*. Noden är ett avancerat visualiseringsverktyg som hjälper till att undersöka stora data material. Denna nod kan användas för att identifiera mönster och trender i materialet. Noden kan generera multidimensionella histogram för både diskreta och kontinuerliga variabler. Denna nod kan enligt SAS (Url 3) även användas i andra steg i SEMMA processen. *Multiplot Node* är en annan nod som används vid undersökandet av data materialet. Multiplot noden skapar automatiskt stapeldiagram och scatter plots av indata och mål variablerna i data materialet. En tredje nod som används i undersöknings fasen är *Insight Noden.* Med hjälp av denna kan användaren interaktivt undersöka och analysera data genom multipla grafer och analyser som är länkade genom flera fönster. *Association Noden* används för att utföra associering eller sekvensiering. Den sista noden i undersöknings stadiet är *Variable Selection Noden.* Denna nod kan hjälpa användaren med att reducera antalet indata

variabler genom att välja bort de variabler som inte är relaterade till målvariabeln. Noden kan snabbt identifiera indata variabler som är användbara för att förutsäga målvariabeln baserat på en linjär modell.

Det tredje steget i SEMMA modellen är att skapa, selektera och transformera variablerna för att fokusera på modellerings-selekterings processen. En nod som kan användas är *Data Set Attributes Node.* Noden används för att modifiera data setets attribut såsom namn, beskrivning och roller. Denna noden kan även användas vid modifiering av metadata urvalet som är associerat med data setet. En annan nod som kan användas vid modifierandet av data materialet är *Transform Variables Node.* Med hjälp av noden kan användaren skapa nya variabler som är transformeringar av existerande variabler. Transformeringar kan vara viktiga för att förbättra en modells lämplighet. *Filter Outliers Node* är en nod som kan användas för att lägga på ett filter på data materialet. Detta för att exkludera observationer som exempelvis ”outliers”. Om källmaterialet innehåller saknade värden kan *Replacement Noden* användas. *Clustering Node* utför klustring som kan användas för att segmentera en databas. Klustring delar in objekten i grupper. En annan nod som används för modifiering av data materialet är *SOM / Kohonen Noden.* Denna nod genererar själv organiserade kartor. I modelleringsfasen av SEMMA processen har användaren flera noder att tillgå. De noder som Enterprise Miner erbjuder är *Regression Node, Tree Node, Neural Networks Node, User-Defined Model Node* och *Ensamble Node*. För information om regression, beslutsträd och neurala nätverk se avsnittet tekniker. User-Defined Model Node gör det möjligt för användaren att generera bedömnings statistik genom att använda förutsagda värden från en modell byggd med *SAS Code Node* eller Variable Selection Node. Ensamble Noden används för att kombinera resultaten från multipla modellerings körningar för att skapa en integrerad modell. För att bedöma modellernas lämplighet har SAS ett antal bedömnings noder. *Assessment Node* tillhandahåller ett ramverk för att jämföra modeller och prediktioner från Ensamble noden, neurala nätverk, regression och träd noden. Det vanligaste kriteriet för att alla modellerings och prediktions verktyg är den förväntade och den aktuella vinsten som uppnås med hjälp av modellen. En annan bedömnings nod är *Score Nod*. Score noden gör det möjligt att hantera, editera, exportera och exekvera ”scoring kod” som är genererad av tränade modeller. Den används ofta som slutresultat av ett Data mining problem. En organisations markandsanalytiker kan använda score noden för att skapa en lista över de kunder som är troliga köpare av företagets produkter. Den sista bedömnings noden är *Reporter Node*. Med reporter noden kan användaren skapa rapporter av Data minings projektets process flow diagram. Noden assembles resultat från analyserna transformeras till en HTML rapport som kan ses med en webb läsare.

Den nyaste versionen av Enterprise Miner är version 4.1. I denna version har SAS lagt till nya noder och features såsom *länkanalys* och *minnesbaserat resonemang* som nämns tidigare i detta arbete. De har också lagt till en *Princomp/Dmneural* nod som är till för modellering som utför självständiga princip komponent analyser. Den förutspår binär och interval målvariabler genom att använda princip komponenter som indata för att passa att icke-linjära modeller. En annan ny nod är *SAS Process Monitor* som är en interaktiv grafisk display som tillåter användaren att stoppa träningsprocessen på regression neurala nätverk och SOM/Kohonen noderna. Ytterligare en ny nod SAS har lagt till är *Time Series* noden. Denna nod är en experiment nod. Dess syfte är att konvertera transaktionsdata till time series data och utföra trend och säsongsanalyser på intervall mål. *Two Stage Model* är också en ny nod i Enterprise Miner 4.1. Denna nod tillåter Data Miners att skapa modeller som kan förutsäga en kategori mål och ett intervall mål inom kategorin**.** Den sista noden som SAS har utökat med är *Tree Results Viewer*. Denna nod är liksom Time Series en experimentell nod. Noden förbättrar

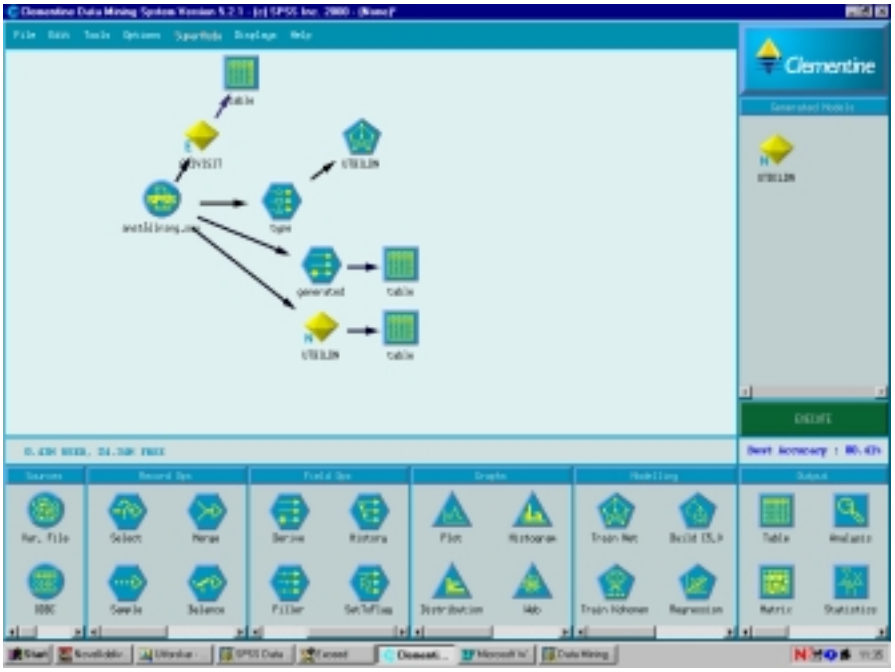
resultatets grafiska möjligheter i trädmodelleringsnoden med förbättrad kvalité och kundanpassad resultat display. Enterprise Miner, Version 4.1, innehåller också experimentella *Text Mining* noder som gör det möjligt att skapa term och dokument frekvens tabeller från text data, reducering av datans dimensioner, och utföra observation klustring för ett givet data set genom att identifiera primär och sekundär kluster som är baserade på sannolikhet.

Andra noder som Enterprise Miner erbjuder är *Group Proccessing Node, Data mining Databse Node, SAS Code Node, Control Point Node* och *Subdiagram Node.*

För mer information om Entprise Miner hänvisar författarna till SAS hemsida www.sas.com eller till artikeln Finding the Solution to Data mining.

Clementine

Clementine är programtillverkaren SPSS Data mining verktyg. Programmet har ett grafiskt gränssnitt men en drag och släpp metodik (drag and drop). Gränssnittet består av menyer, titel fält, rapport fönster, status fönster, palett för genererade modeller, paletter och en ström panel. Programmet är konstruerat så att användaren får bygga upp strömmar. För att bygga upp en ström drar användaren en nod (ikon) från paletten till ström panelen. Varje nod representerar en operation som kommer att utförs på datamängden. Noderna i paletten är grupperade med hänsyn till vilken typ av operation som den utför. En ström representerar ett flöde av data som går genom ett antal operationer till en destination i form av en tabell, en genererad modell eller diagram. En genererad modell skapas då någon av maskininlärnings teknikerna används. Modellerna lägger sig i paletten för genererade modeller.



FIGUR 2:19 CLEMENTINE ANVÄNDARGRÄNSSNITT (URL 4)

Det första som krävs vid byggandet av en ström är att en datafil läses in till Clementine. För att läsa in SPSS filer tillhandahåller programmet *Spss Import*. Vill användaren läsa in filer

med ASCII format finns *VAR. File* och *Fixed File* noderna att tillgå. Om det finns en Open Database Connectivity kan en mängd olika dataformat bland annat Excel, Ms Access och dBase importeras genom att *ODBC* noden används. När väl datamängden är importerad till Clementine kan noden *Table* användas för att titta på datan. Användaren har en nod vid namn Type som kan användas för att definiera varje fält i datasetet. När väl datan är inläst till Clementine finns ett antal noder för att utforska datans integritet. En av dessa noder är *Quality* noden. Quality noden används för att studera om det finns några saknade värden i materialet. En annan nod är *Distributions* noden som nyttjas då värden som är felaktiga eller inte ligger inom vissa gränser. Andra noder som kan användas för utforskning av datan integritet är *Histogram* noden eller *Statistics* noden. Histogram noden visar händelser för numeriska fält medan Statistics noden visar mer detaljerad information. För data manipulering finns tre noder att tillgå i Clementine. Dessa noder är *Select*, *Derive* och *Filter*. Select noden ger användaren möjlighet att välja eller eliminera en grupp rader baserat på ett specificerat vilkor. Derive noden kan användas för att modifiera värden på data eller skapa nya kolumner. För att ta bort alla oönskade kolumner kan Filter noden användas. För att upptäcka relationer i datan kan användaren utnyttja *Matrix*, *Plot* eller *Web* noden. Matrix noden utför korstabulering av två eller fler kolumner som visar hur ett värde i en kolumn relaterar till ett värde i en annan kolumn. Om användaren vill se detta grafiskt kan Web eller Plot noden användas. De modelleringstekniker som finnas att tillgå i Clementine är Neurala nätverk, Beslutsträd Regel induktion, Kohonen networks och Associerings regler. För vidare information om dessa tekniker hänvisas till avsnittet Tekniker.

För mer information om Clementine hänvisar författarna till SPSS hemsida www.spss.com eller till manualen SPSS Traning: Introduktion to Clementine.

Arbetsmetodik vid Data mining projekt

Nedan kommer två arbetsmetodiker över Data mining processen att beskrivas. Syftet är att visa de faser som ingår i ett Data mining projekt. En anledning till att det finns en arbetsmetodik är att säkra kvalitén på de resultat som genereras.

CRoss-Industry Standard Process for Data mining

Ett antal organisationer har gått samman i ett projekt för att utarbeta en standard för Data mining processen. Projektet heter CRoss-Industry Standard Process for Data mining (Url 1). Det har arbetats fram av bland andra programvaruföretaget SPSS och Daimler Chrysler. De inblandade hoppas att denna beskrivning av Data mining processen skall bli en standard för ”data miners” världen över.

CRISP-DM är en metodologi och modell över processen Data mining (SPSS Manual, 1999) (Url 1). Den är framtagen så att både nybörjare och experter skall kunna följa den i sitt arbete. Modellen hjälper användaren att undvika de vanligaste felen vid användningen av Data mining genom att beskriva varje steg i Data mining processen grundligt.

Skaparna av CRISP-DM menar att både stora som små Data mining projekt kommer att bli effektivare och mindre kostsamma genom att man följer en redan utprovad process. Detta är den första riktiga enhetliga metodologin som kommit om Data mining processen. Det finns de som anser att området Data mining har varit väldigt diffust. Nedan presenteras CRISP-DMs faser (Url 1).

CRISP-DMs faser

1. ***Förståelse för verksamheten (Business understanding)***
   I denna fas gäller det att få en förståelse för projektmålen och de behov som finns inom verksamheten. Utifrån denna kunskap formuleras problemformuleringar som skall lösas med hjälp av projektet.

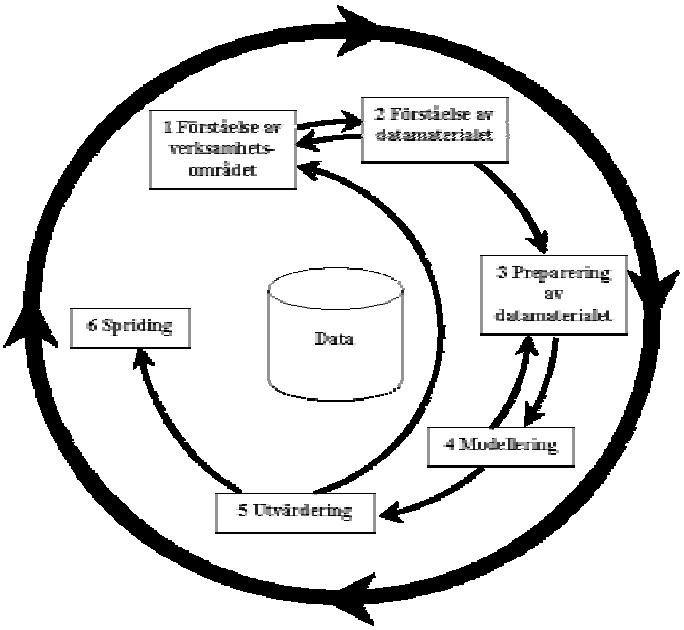
2. ***Förståelse för data (Data understanding)***
   Här gäller det att lära känna, identifiera kvalitetsbrister och få en inblick i data som hämtats.

3. ***Databearbetning (Data preparation)***
   Databearbetningen är en uppgift som ofta får upprepas flera gånger. Den består i att de tabeller, kolumner och rader väljs ut som skall matas in i modellerings verktyget. Här anpassas också data för de olika teknikerna som skall användas. Tvättning av data sker också så att materialet kan korrigeras då alla algoritmer inte kan användas på ofullständiga material.

4. ***Modellera***
   I denna fas väljs de tekniker som skall användas. Med hjälp av dessa tekniker skapas modeller. Före en modell skapas måste det avgöras hur validitet och kvalitet skall kunna testas.

5. ***Utvärdering av modell***
   Detta utförs för att säkerställa modellens relation till de projektmål som sattes upp från början. I slutet av denna fas bestäms om modellen är användbar och skall användas.

6. ***Implementering av resultat***
   Här måste data organiseras och presenteras så att verksamheten kan dra nytta av den.



FIGUR 2:20 DATA MINING PROCESSEN ENLIGT CRISP-DM (URL 1)

SEMMA

SEMMA är SAS Institutets metod för Data mining processen. SEMMA står för Sample, Explore, Modify, Model och Assess. Enligt Groth (1999) behöver användaren inte utnyttja alla steg i Data mining processen.

1. ***Urval (Sample)***
   Här sker ett urval av den totala datamängden. Detta görs när man är intresserad av generella mönster.

2. ***Undersökning (Explore)***
   Undersöker vilken typ av data som finns för att användaren skall få förståelse och idéer för vilka mönster och trender som kan bli resultatet från datamaterialet.

3. ***Modifiering (Modify)***
   Här skall man skapa, selektera och transformera variablerna för att fokusera på modellerings-selekterings processen.

4. ***Modellera (Model)***
   Här används de analytiska verktygen för att söka efter kombinationer av datan som tillförlitligt förutspår ett önskat resultat.

5. ***Bedömning (Assess)***
   Nu sker en process där användbarhet och tillförlitlighet av iakttagelserna utvärderas.

Kritiska framgångsfaktorer vid Data mining

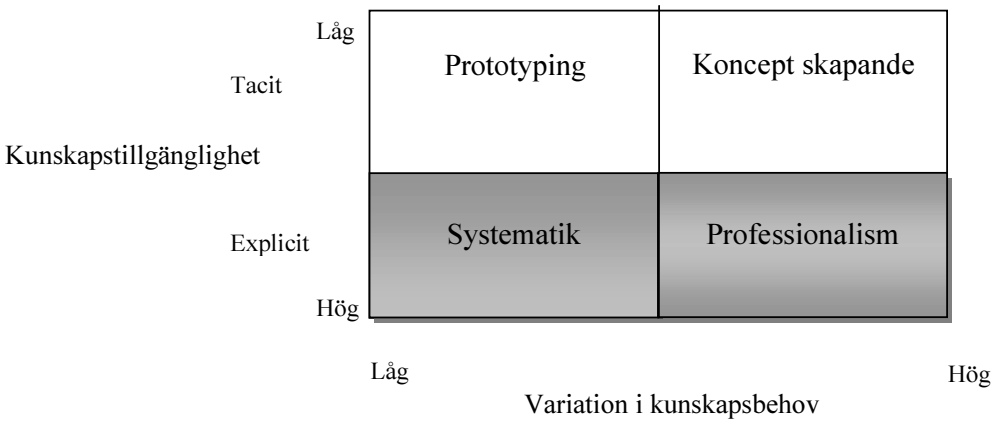
För att nå framgång med ett Data mining projekt krävs att stora mängder data insamlas samt att datan kan göras tillgänglig i ett datalager. Om data materialet tillgodoser en god täckning av möjliga resultat kan datamängder ner till ett par tusen poster användas. Dock måste användaren tänka på att för att mer komplexa relationer och mönster skall finnas krävs mer poster (SPSS, 1999). Datainsamlingen är ett kritiskt moment eftersom det är av vikt att relevant data samlas in. Vilken data som skall användas beror primärt på vilken fråga som Data mining skall svara på. Innan analyser kan påbörjas krävs det att data prepareras. Den data som används i Data mining projekt är en viktig faktor för att nå fram till fruktbara resultat. Erfarenheter visar att det tar cirka 80 procent av den totala tiden i ett Data mining projekt (Putten, 1998). En kritisk faktor som nämns i litteraturen är att dålig *datakvalité* leder till sämre resultat. Problemet kan belysas med talesättet ”Garbage In Garbage Out”. Med god datakvalité menar Groth (1999) att data inte skall vara redundant, samma data skall inte finnas lagrad på flera platser. Han nämner vidare att data skall vara korrekt och konsistent, det vill säga att värden i datamaterialet skall inte fattas eller vara felaktiga. Stavfel och inaktuell data är andra faktorer som leder försämrad datakvalité. En annan faktor som kan påverka datakvalitén är variation i definiering av termer. Ett exempel kan vara variablerna personnummer och IDnummer. Det är viktigt att användarna vet om variablerna motsvarar samma sak eller ej. *Gruppering* är en annan kritisk faktor som användaren bör ta hänsyn till vid användandningen. Många Data mining verktyg ger användaren möjlighet att gruppera datamaterialet och det finns olika metoder för att göra detta. Beroende på vilken metod som används påverkar resultatet av analyserna. En viktig faktor som Groth (1999) nämner är att

användaren måste ha en förståelse för Data miningens gränser. Trots att det inte är meningen att man som användare skall behöva specificera mellan vilka objekt man söker relationer och mönster skall det ändå finnas i åtanke att resultaten som genereras inte behöver vara korrekta. Det är av stor vikt att kunna förstå och tolka de resultat som Data mining processen genererar. Den viktigaste faktorn enligt SPSS (Url 4), är användarens kunskap och erfarenhet. Fayyad et al. (1996) menar att det är viktigt att användaren har kunskap om det verksamhetsområde där tekniken skall användas.

2.4 Data mining och beslutsfattande

Detta avsnitt syftar till att sammanfatta teorikapitlet genom placera in Data mining i Macintosh (1994) ramverk för se om och hur Data mining kan stödja beslutsprocessen. Genom Scott Mortons (1971) ramverk kommer en definition på beslutskvalité att presenteras.

Med den teoretiska referensramen som grund har Data mining placerats in i Macintosh (1994) ramverk (se sid.9). Ramverket hjälper oss att klassificera beslut med hänsyn till problemstruktur och variation.



FIGUR 2:21 MACINTOCH (1994) MODELL MODIFIERAD AV FÖRFATTARNA

I modellen (figur 2:21) ovan har vi valt att placera Data mining i de områdena där tekniken kan stödja beslutsprocessen, det vill säga systematik och professionalism. För att Data mining skall kunna stödja beslutsfattande krävs att det finns en hög grad av kunskapstillgänglighet. Nonaka och Takeuchi (1995) skiljer på tacit och explicit kunskap. När kunskapen betecknas som tacit är den ickeformaliserad och kan inte lagras i en databas. För att kunna använda Data mining tekniker är det en förutsättning att datan är strukturerad och formaliserad vilket kännetecknas av den explicita kunskapen. Data mining kan användas oberoende om variationen i kunskapsbehovet är lågt eller högt. Då variationen är låg är det få oväntade händelser som uppkommer under arbetets gång (Macintosh, 1994). Skall tekniken användas där variationen i kunskapsbehovet är lågt är en viktig aspekt att det inte kostar mer att producera informationen än vad den inbringar (Ekman, 1970).

Scott Mortons (1971) ramverk (se sid.10) hjälper oss att förstå beslutsfattande och vad som påverkar en beslutssituation. Från det ramverket kan följande formel för beslutskvalité skapas

(Magoulas, 2001): Beslutskvalité = $I*M*R_{ij}$. Där I = Information, M = Människa det vill säga den person som har kunskap och kompetens att fatta ett beslut, R = Regler, i = Fasen i beslutsprocessen och j = Beslutsnivå. Formeln illustrerar att beslutkvalitén är beroende av rätt information till rätt person vid rätt tidpunkt. Om någon av parametrarna är lika med noll blir beslutskvalitén lika med noll.

3 Metod

I metod kapitlet kommer en redogörelse över två vetenskapliga förhållningssätt. Därefter förs ett resonemang över lämpligt metodval. Vidare kommer undersökningens upplägg, tillvägagångssätt och svarsbearbetning att presenteras. Metodkapitlet avslutas med ett avsnitt om kvalitetsgranskning.

3.1 Vetenskapligt förhållningssätt

Enligt Hartman (1998) består vetenskapsteorin av två olika synsätt: det positivistiska och det hermanuetiska. I det positivistiska synsättet ägnar forskare sig helt åt att beskriva den observerbara världen. Hartman (1998) menar att positivism är en vetenskapsteori där kunskap består av teorier. Teorierna innehåller termer som refererar till mätbara företeelser och satser som anger relationerna mellan företeelserna. Vidare menar Hartman att det viktigaste sättet för att nå kunskap om verkligheten är att observera den. Observationerna kan antingen ske i naturliga situationer eller genom experiment. Det andra synsättet hermaneutiken är läran om tolkning. Här strävar forskaren efter en förståelse för människors livsvärld. Kunskap om människors livsvärld fås genom tolkning av människors beteende. Denna kunskap benämner Hartman som förståelse. Inom hermanuetiken ligger inte intresset i att veta hur världen *är* utan hur den *uppfattas* och *tolkas.*

3.2 Val av metod

Vid forskning finns det möjlighet att utforma studien utefter en kvalitativ eller kvantitativ metod. Vilken design av undersökning som väljs är beroende på hur problemet ser ut, vilka frågor det genererar samt vilket resultat som önskas (Merriam, 1994).

Den kvalitativa metoden kännetecknas av att den är flexibel, utvecklande och ostrukturerad. Detta kan jämföras med den kvantitativa designen som är förutbestämd och strukturerad. Urvalet vid ett kvalitativ angreppssätt kännetecknas av att vara litet och icke-slumpmässigt medan i det kvantitativa angreppssättet strävar efter ett stort, slumpmässigt och representativt urval. Målet i den kvalitativa designen är att få en förståelse och beskrivning av problemet medan den kvalitativa fokuserar på förutsägelse, kontroll, bevisning och hypotesprövning (Merriam,1994). Holme och Solvang (1999) anser att en kvalitativ ansats bör användas vid skapandet av referensramar, byggandet av teorier samt då forskaren vill ha ett totalperspektiv eller en fullständig förståelse. Den kvantitativa metoden bör utnyttjas då forskaren utifrån urvalet vill säga något om den grupp urvalet gäller eller visa på hur starka vissa samband är mellan olika företeelser. Den kvalitativa studien är en metod där subjektiva upplevelser spelar stor roll. Forskaren är det främsta instrumentet som används för att samla in och tolka data. Det är också här den främsta begränsningen ligger, i och med att forskaren själv väljer ut, registrerar och tolkar data kan det vara en svårighet att hålla sig helt objektiv (Backman, 1998).

Vid val av metod tog vi i beaktande att välja en metod som var relevant för vårt frågeställning och syfte samt gav möjlighet att samla in relevanta data. Syftet med uppsatsen är att skapa en

verklighetsmodell av Data mining användning. Med ovanstående resonemang om metodval som grund valde vi att arbeta utefter den kvalitativa metoden. Detta på grund av att vi strävar efter att förstå och beskriva hur Data mining används idag samt att vi avser att skapa en verklighetsmodell för Data mining användning.

3.3 Undersökningens upplägg

I följande avsnitt kommer vi att redogöra för den inledande delen av vår studie. Här presenteras hur vi gått tillväga för att samla in underlag till den teoretiska referensramen. Vi kommer även att beskriva hur vi designat frågorna till den empiriska studien samt vilka respondenter som skall deltaga i den.

Datainsamling

Det första steget i vår studie var att genomföra en omfattande litteraturstudie för att bilda oss en uppfattning om de teorier som existerar inom ämnesområdet. Insamling av sekundärdata förekom i form av lämplig facklitteratur. Vi tog del av informationskällor på Internet samt artiklar som hade anknytning till vårt undersökningsområde. Även organisations och bransch specifik information samlades in. Med hjälp av litteraturstudien kunde vi precisera en problemställning, hitta relevanta begrepp och avgränsningar för att sedan kunna gå in i en mer vetenskaplig fas av undersökningen. För att få en uppfattning av tidigare forskning beslutade vi oss för att botanisera bland tidigare undersökningar. Detta för att får inspiration och se hur andra författare har gått tillväga.

Primärdata kommer att samlas in genom ett antal intervjuer för att få en djupare förståelse av verkligheten. Det är den empirin som kommer att ligga till grund i analys och diskussions kapitlen.

Försökspersoner

Vid urval av försökspersoner i kvalitativa undersökningar är det viktigt med en spridning för att fånga de variationer och skiftningar som är av intresse. Vid en kvalitativ undersökning strävar forskaren efter att få en helhetsbild vilket kan gå förlorat om urvalet består av en alltför homogen grupp (Starrin & Svensson, 1994). I vårt fall kommer vi att intervjua tre kategorier av människor med anknytning till Data mining området denna triangulering syftar till att främja validiteten (Jensen, 1995). Den första kategorin är leverantörerna. Denna grupp består av organisationer som säljer Data mining lösningar. Den andra kategorin är de organisationer som använder Data mining tekniken i sin verksamhet. Den individ som avses att intervjuas i denna grupp är systemägaren. Den sista kategorin består av användare. Med användare menas den person som använder Data mining resultatet i sitt vardagliga beslutsfattande. Tanken med detta urval är att försöka fånga in olika aspekter samt möjligheten att betrakta Data mining ur olika perspektiv

Design av frågor

Genom vår omfattande litteratur studie som grund samt de ramverk som presenterades i teorikapitlet utkristalliserades de attribut som är viktiga för att uppnå en förståelse för beslutsfattande och Data mining (se bilaga 1). Utifrån dessa attribut formulerades explicita

frågor vilket borgar för en god validitet (Magoulas, 2001). De tre kategorier som ingår i studien representerar funktionell kvalité, infologisk kvalité och social kvalité. Med den funktionella kvalitén menas förhållandet mellan organisation och Data mining system, en fråga om processer. Den infologiska kvalitén skall hjälpa oss att undersöka sambandet mellan Data mining system och individ, en fråga om aktörer. Social kvalité symboliserar förhållandet mellan Data mining design och individ, en fråga om design. Genom Galibraiths (1977) modell erhålls också kultur samt kunskap och kompetens perspektivet. Det mätinstrument som vi avser att använda är telefonintervjuer. Anledningen till att denna form av intervju väljs är på grund av det geografiska avståndet till ett flertal av respondenterna Vid telefonintervjuerna kommer vi att använda oss av standardiserade frågor med en hög grad av strukturering. Med standardisering menas att samma frågor ställs till alla respondenter i respektive grupp. Strukturering innebär att vi vet vad vi vill fråga om och att allt handlar om det ämnet vi vill veta något om (Trost, 1993). Tanken med denna arbetsmetodik är att säkerställa reliabiliteten.

Bakgrundsvariabler

De bakgrundsvariabler som ingår i undersökningen finns med för att erhålla information om födelseår samt om det är en man eller kvinna. För att få en uppfattning om vilka personer som arbetar inom Data mining området och vad de har för referensramar ingår frågor om vilken utbildningsbakgrund, erfarenhet samt befattning de har. Vidare finns en fråga om hur många år respondenten arbetat med Data mining. Avsikten med den frågan är att få en uppfattning av den kunskap och kompetens den vederbörande har inom området. För att erhålla kulturperspektivet om än på ett ytligt plan ställs en fråga där respondenten får ge en kort presentation av organisationen och dess policy. Till aktörerna, det vill säga användarna, i undersökningen kommer ytterligare två frågor att ställas där de får berätta om sina arbetsuppgifter och vilket stöd de får av systemet. Syftet med den frågan är att få en bild av inom vilka beslutsområden de använder Data mining och hur de uppfattar att systemet stödjer dessa områden. De kommer även att få svara på hur deras arbetsuppgifter påverkas av tid och omgivning och hur dessa i sin tur påverkar systemet. Frågeställningen avser att belysa om användaren uppfattar att de handlar annorlunda jämfört om de inte hade haft Data mining systemet.

Frågor som berör funktionell kvalité (Processer)

- *Inom vilka beslutsområden använder Ni Data mining?*
  Meningen med denna fråga är att se på vilka slags problem som de anser att Data mining kan hjälpa dem. (Användbarhet)

- *Hur bestäms vilken data som skall lagras i systemets databasen?*
  Hur bestäms vilken data som är relevant att samla in och lagra. Det kan finnas en ekonomisk begränsning som gör att inte all data kan samlas in och lagras. Vilka områden involveras. (Relevans)

- *Hur ofta uppdateras datan i databasen?*
  Denna fråga skall ge svar på hur aktualiteten på beslutsunderlaget bibehålls så att beslut inte fattas på inaktuell data. (Aktualitet)

- *Hur övervakas transaktionerna mellan systemen för att försäkra sig om att inget går fel?*

Med denna fråga vill vi veta vad organisationen gör för att försäkra sig om att datan är konsistent när den används i Data mining processen. (Robusthet)

- *För vilket beslutsområde, processer i organisationen är Data mining tekniken tillgänglig?*
  Med denna fråga vill vi veta vem det är som tar del av beslutsunderlaget. Vi vill även veta på vilken nivå i företaget som Data mining används. (Informationsintegration)

- *Hur hanteras problemen med olika begreppsdefinitioner i transaktionerna mellan de olika systemen?*
  Med denna fråga vill vi veta hur kommunikationen sker mellan de olika systemen, att de pratar samma språk. Detta för att undvika feltolkningar. (Begreppsintegration)

- *Hur bedömer Ni Data Mining-systemet utifrån följande variabler på en skala mellan 1 – 5? Betyg fem betyder instämmer helt. Betyg ett betyder instämmer ej.*

  Med variabeln *fullständighet* menar vi om systemet ger en korrekt bild av verkligheten. Variabeln *jämförbarhet* syftar till att ta reda på hur Data mining systemet står sig i jämförelse med tidigare använda metoder. Med *härledbarhet* vill vi veta hur väl organisationen kan härleda data till källorna. För att undersöka hur väl systemet producerar korrekta resultat kommer vi att be respondenten betygsätta variabeln *noggrannhet.* Vidare vill vi att respondenten skall bedöma systemets *lämplighet.* Med det menar vi hur väl systemet klarar av de uppgifter som det är ämnat för i organisationen. *Tillmötesgående* avser att bedöma hur respondenten upplever att organisationen behöver anpassa sig till Data mining tekniken eller tvärtom. Variabeln *uppskattning* skall ge information om hur organisationen värderar systemet.

Frågor som berör infologisk kvalité (Aktörer)

- *Upplever Du att verktyget är anpassat efter egna specifika behov och önskemål?*
  Syftet med ovanstående frågeställning är att undersöka huruvida respondenten tycker att systemet uppfyller hans/hennes krav. (Individuell validitet)

- *Vem väljer ut vilken data som skall ingå i Data mining processen?*
  Meningen med denna fråga är att ta reda på hur de fastställer den individuella validiteten. Det vill säga vem bestämmer vilka data som är viktiga för enskilda beslutsfattare. (Relevans)

- *Vem kontrollerar och utvärderar resultaten?*
  Vi vill veta hur individen kan försäkra sig om att resultaten bygger på korrekta data. Vi vill veta hur de kontrollerar modellens relevans. (Kontroll)

- *Vem bestämmer över representationsform (hur resultaten skall presenteras)?*
  Med denna fråga vill vi veta om de kan visualisera resultaten efter vare enskild individs önskemål. (Klarhet)

- *Vem väljer vilken design metod som skall användas och hur väljs denna metod?*
  Det vi vill veta med denna frågan är hur användbart Data mining är för varje enskild individ. Om designen metoderna anpassas till varje enskilt fall. (Användbarhet)

- *Hur följer Du upp utvecklingen av systemet för att eventuellt utveckla och förfina processen?*
  Med denna fråga vill vi veta hur användarna lär sig av tidigare misstag och erfarenheter och på så sätt kan förfina och utveckla de vedertagna processerna till sina specifika problem. (Lärande)

- *Vad är det som du bestämmer över i systemutvecklingen?*
  Det vi vill veta är om användaren får möjlighet att medverka i systemutvecklings processen och i vilka delar. (Lärande)

- *Hur bedömer du användarvänligheten på Data mining systemet utifrån följande variabler på en skala mellan 1 –5? Betyg fem betyder instämmer helt. Betyg ett betyder instämmer ej.*

  Variabeln *förståelse* syftar till att undersöka hur väl respondenten förstår Data mining systemet. Med *inlärning* vill vi veta hur lätt användaren har att sätta sig in i systemet. När det gäller variabeln *klarhet* menar vi hur väl gränssnittet kommunicerar med användaren. *Anpassningsbarhet* syftar till att undersöka hur väl systemet kan anpassas efter användarens behov och önskemål. Variabeln *vänlighet* skall ge information om hur användaren upplever systemets totala användarvänlighet. Med s*upport* avser vi att undersöka hur väl systemet ger hjälp och stöd åt användaren.

Frågor som berör social kvalité (Design)

- *På vilket sätt är Data mining lönsamt för verksamheten?*
  Med denna fråga vill vi veta vad leverantörerna tycker att Data mining kan göra för verksamheten.(Lönsamheten)

- *Vilken social betydelse har systemets användning? Kommer någon genom beslutet att gynnas/drabbas direkt eller indirekt?*
  Syftet med ovanstående frågeställning är att ta reda på om Data mining systemet påverkar besluten som användaren tar och om dessa i sin tur påverkar någon annan. (Kultur,struktur, aktör och process)

- *Vad kan användaren bestämma över när det gäller hur resultaten skall visualiseras?*
  Det vi vill veta är om användaren föredrar att ett resultat skall visas på ett visst sätt påverkar det hur datan skall struktureras i databasen. (Anpassning)

- *Vilka säkerhetsproblem förekommer och hur hanteras dessa?*
  Syftet med ovanstående fråga är att få en bild av hur säkerheten hanteras i samband med analyser eftersom det kan vara känslig information och om det finns någon risk för att någon utomstående skall komma åt den analyserade datan? (Säkerhet)

- *Vilka kritiska faktorer har varit avgörande vid design av systemet?*
  Med denna frågan vill vi veta vad leverantörerna har gjort vid design av systemet för att säkerställa kvalitén på systemet. Att det är säkert ur en kvalitetssynpunkt det vill säga att systemet är att lita på, säkert att använda. (Säkerhet/ansvarighet))

- *Vilka ansvarigheter upplever Ni att hämta data från andra system. Finns det svårigheter? Vilka?*
  Om det finns svårighter att hämta data ifrån andra system vill vi veta vilka de svårighterna är och om leverantören tar ansvar för detta eller om det är organisationen. (Ansvarigheter)

- *Hur anser Ni att Data Mining-systemet uppfyller följande variabler på en skala från 1-5. Betyg fem betyder instämmer helt. Betyg ett betyder instämmer ej.*

  Med variabeln *effektivitet* avser vi att få information om hur leverantören bedömer effektiviteten på Data mining systemet. Med *flexibilitet* vill vi veta hur lätt systemet är att anpassa efter användarnas behov. När det gäller variabeln *fullständighet* menar vi om användaren kan utföra allt det vill utföra. *Snabbhet* syftar till att undersöka hur snabbt leverantörerna anser att systemet är. Variabeln *korrekthet* avser att bedöma hur väl systemet visar informationen på ett icke missvisande sätt. *Beslutskvalité* syftar till att undersöka en hur väl informationen kan anpassas till rätt information i rätt ögonblick. Med *kostnad/ekonomi* avser vi att bedöma hur kostnadseffektivt systemet är. Nyttan ställt mot kostnaden. *Samverkan* syftar till att undersöka hur väl Data mining systemet samverkar med andra system. Variabeln *stabilitet* avser att undersöka hur väl användaren kan lita på systemet att det inte går ner. När det gäller variabeln *underhåll* syftar den till att bedöma hur underhållsvänligt systemet är.

3.4 Tillvägagångssätt

Redan innan vi påbörjade examensarbetet hade vi kännedom om två stora Data mining leverantörer. Det föll sig därför naturligt att besöka deras hemsidor på Internet. Dels för att komma i kontakt med dem genom e-post och för att hitta referenser till organisationer som använder sig av deras Data mining lösningar. Då det var svårt att finna referenser till organisationer som använder sig av Data mining inom Sverige valde vi att först ta kontakt med leverantörerna. Tanken med tillvägagångssättet var att de möjligtvis skull kunna hjälpa oss med vidare kontakter.

Ett missivbrev formulerades där vi presenterade oss och beskrev syftet med undersökningen. En annan viktig del av brevet var att skriva var de kunde ta del av undersökningens resultat. Denna återkoppling gör det intressantare att deltaga då det en möjlighet att ta del av slutresultatet (Gellerstedt, 1997). Missivbrevet sändes till de utvalda organisationernas e-post för att på så sätt få hjälp att komma i kontakt med rätt personer (Trost, 1993). På grund av arbetets tidsram ringde vi efter cirka fem dagar upp de respondenterna som vi inte fått någon respons från. Under telefonsamtalet bokades tid för en telefonintervju. Intervjufrågorna sändes över via e-post till respondenterna för att de skulle få möjlighet att förbereda sig inför intervjun. Vid intervjutillfället tillfrågades respondenterna angående anonymitet. Denna fråga ställdes för att respondenten skulle känna sig säker i sin intervjusituation. Om respondenten får möjlighet att vara anonym kan bortfallet på känsliga frågor minska (Wäneryd, 1993). En överenskommelse om att respondenten skulle få ta del av arbetet innan publicering gjordes. Detta för att de skulle ha möjlighet att revidera. Vid intervjutillfället var vi båda närvarande. Intervjuerna spelades in på band under förutsättning att respondenten accepterade detta. I de fall det inte medgavs fördes anteckningar. Tanken med detta arbetsätt var att inte förlora värdefull information. Innan intervjun avslutades bad vi leverantörerna nämna organisationer i Sverige som använde sig av Data mining i sin verksamhet. En organisation nämndes. För att

hitta fler organisationer till vår underökning blev nästa steg att botanisera bland publicerade uppsatser om Data mining för att på så sätt få fler uppslag till tänkbara organisationer. Sammanlagt sände vi iväg elva stycken missivbrev till möjliga deltagare gällande systemägare och användare. Anledningen till att så många missivbrev skickades ut var för att garantera för att få det antal intervjuer vi önskade. Två organisationer svarade på brevet och ställde sig positiva till intervjun. Av de två organisationer som svarat var en intervju med en systemägare och en med en användare. Det bestämdes en tid för intervju och frågor till respektive systemägare och användare sändes via e-post. De organisationer som inte svarat på missivbrevet kontaktade vi efter ett antal dagar via telefon. Detta visade sig vara ett tidsödande moment. Svårigheten låg i att komma i kontakt med rätt person. Endast en av de uppringda organisationerna tackade ja till medverkan. I många fall avböjde organisationen med hänsyn till att tekniken inte användes i dagsläget eller tidsbrist. I övrigt genomfördes intervjuerna enligt den procedur som tidigare beskrivits.

3.5 Svarsbearbetning

Primärdatan som erhölls genom intervjuerna spelades in på band att inte gå miste om viktig information samt att vi skulle ha möjlighet att återgå till källmaterialet om så önskades. Efter varje intervju renskrev vi ordagrant det bandade samtalet i ett Word-dokument. Därefter konfiskerades inspelningsbandet. I de fall där intervjun inte spelades in med hänsyn till respondentens önskan renskrev vi de gjorda anteckningarna. Under intervjuerna och efterbearbetnings- processen deltog vi båda i arbetet. Efter intervjuerna diskuterade vi också igenom eventuella oklarheter. Det insamlade intervjumaterialet bearbetades genom att vi försökte finna teman och mönster. Vi var även uppmärksamma på likheter och skillnader i svaren från respondenterna.

3.6 Kvalitetsgranskning

I detta avsnitt tar vi upp hur vi har arbetat för att stärka kvalitén på studien. Här kommer begreppen reliabilitet och validitet förklaras. Dessa begrepp härstammar från den kvantitativa forskningen, men kan även appliceras på kvalitativa studier (Trost, 1993).

Reliabilitet

Reliabilitet eller tillförlitlighet som det också kan benämnas är ett mått på i vilken utsträckning en studie ger samma resultat vid olika tillfällen under i övrigt likartade omständigheter. I vårt fall kan det bli problematiskt då främst med tanke på att Data mining området är dynamiskt och utvecklas snabbt. Vi syftar inte till att försöka generalisera resultatet som i den traditionella forskningsmetoden (Merriam, 1994). När det gäller den kvalitativa undersökningen är forskaren främst ute efter att resultaten skall ha en mening, vara konsistenta och beroende. För att säkerställa reliabiliteten i vår undersökning samlade vi in material från många olika källor. Vid telefonintervjuerna användes standardiserade frågor med en hög grad av strukturering för att säkerställa reliabiliteten.

Validitet

Begreppet validitet kan brytas ner i två undergrupper vilka är inre validitet och yttre validitet. Med den inre validiteten så menas om vi mäter det vi avser att mäta. För att säkerställa den inre validiteten har vi som tidigare nämnt i metoddelen använt oss av triangulering det vill säga för att få problemen belysta ur olika perspektiv och horisontell granskning. Med det menar vi att det som var och en av oss kommit fram till granskas av den andre (Merriam, 1994). Denzin (Jensen, 1995) har identifierat olika former av triangulering. De former vi har applicerat på vår studie är metodtriangulering, datatriangulering och forskartriangulering. För metodtriangulering har vi valt att samla in data från tre olika perspektiv. Den empiriska datan har samlats in genom intervjuer med systemägare, användare och leverantörer. Detta arbetsätt gynnar validiteten (Jensen, 1995). När det gäller datatriangulering har data samlats in från olika källor såsom facklitteratur, företagsspecifik litteratur samt Internet. Forskartriangulering har använts då vi båda har närvarit under intervjutillfällena vilket är en stor fördel vid analysen av materialet. Det kommer även ske en deltagare kontroll där de personer som givit information till oss det vill säga respondenterna får ta dela av resultaten och därefter ta ställning till om det är trovärdigt. Den yttre validiteten behandlar ifall resultatet ifrån vår undersökningen kan vara generaliserbart till andra situationer (Merriam, 1994).

4. Analys

Analyskapitlet är uppdelat i fyra sektioner. Den första delen beskriver demografin av respondenterna. Den andra delen behandlar den funktionella kvalitén det vill säga förhållandet mellan organisation och individ. Den tredje delen, den infologiska kvalitén beskriver förhållandet mellan individ och systemet i detta fallet Data mining systemet. Den sociala kvalitén tas upp i den fjärde och avslutande delen som hanterar förhållandet mellan leverantörer och system. I analysen letar vi främst efter mönster och teman som hjälpa oss att se var i beslutsprocessen som Data mining kan appliceras. På vilka problem Data mining används samt vilka kritiska faktorer som upplevs vid användningen. Analysen kommer att ske utifrån varje enskild fråga i intervjun.

4.1 Demografi

I vår empiriska undersökning har vi genomfört fem telefonintervjuer med fem olika organisationer. Den organisation som önskat vara anonym i undersökningen har vi benämnt enligt det grekiska alfabetet första bokstav Alfa. Övriga organisationer som deltog i intervjuerna var SAS, Scandinavian Airlines System (SAS), SPSS och Volvia. Respondenterna bestod av två systemansvariga, en användare och två leverantörer.

Organisationen *Alfa* föredrar att vara anonym. De är en detaljistkedja inom dagligvaruhandeln. Den person vi intervjuade arbetade som marknadsanalytiker och hade en bakgrund som ekonom. Organisationen har haft Data mining som ett prov projekt under några månader.

SAS är en leverantör av Data mining produkten Enterprise Miner. De arbetar mot stora och medelstora företag. SAS grundades av Jim Goodnight som arbetade med genetisk forskning. Företaget har funnits i Sverige sedan 1986 och har runt hundra anställda här. Deras syfte är att hjälpa företag att förstå affärsnyttan av all data som genereras i olika system. Vi intervjuade en kvinnlig produktansvarig (Product Manager) för Data miningen som tillhör marknadsavdelningen. Hon har arbetat med Data mining i fem år och har en utbildningsbakgrund som filosofie kandidat examen i statistik.

Scandinavian Airlines System (SAS), är ett skandinaviskt flygbolag som till 50 procent ägs av staten och till 50 procent ägs av privata investerare. De flyger över ett hundratal destinationer och ingår i en allians med bland andra United Airlines, Lufthansa, Air Canada och Thai Airways. Organisationen har ett antal branding values som de vill förknippas med. Organisationen vill att kunden skall uppfatta dem som omtänksamma, innovativa och offensiva. Visionen inom SAS, är att de skall göra skandinaverna stolta över sitt flygbolag. Den person vi intervjuade har två befattningar dels som chef för SAS Internetsupport och dels som manager för customer experience. Han har arbetat med Data mining för att ta fram statistik och underlag för den operativa verksamheten i cirka tre år. Med Web mining har han arbetat i cirka tre till fyra månader. Den teoretiska utbildningsbakgrund han har är en ekonomisk gymnasieutbildning samt interna kurser inom SAS. Han upplever att systemet påverkar organisationens sätt att arbeta.

SPSS är leverantör av Data mining produkten Clementine. Det är ett amerikanskt programvaruföretag som grundades för 30 år sedan och har funnits i Sverige sedan1984. Totalt har företaget 1600 anställda varav 30 stycken i Sverige. Traditionellt har det kopplats ihop med statistikverktyg till den akademiska världen men även sjukvård och forskning. På senare år har företaget förknippats med andra inriktningar där deras verktyg kan passa framför allt inom Data mining. Syftet är att sälja Data mining lösningar till större företag där Data mining är kärnan i det hela. Deras inriktning är Data mining och analysverktyg. Den person vi intervjuade var en manlig Sales Engineer. Hans arbetsuppgifter består av att förklara för kunderna vad verktyget kan åstadkomma för dem. Han är även med och utbildar kunderna på verktygen samt konsulterar med kunderna i samband med Data mining. Han har arbetat med Data mining i 2,8 år och har en magisterexamen i statistik.

Volvia är ett försäkringsbolag med cirka 220 anställda. Aktiviteterna som består i att sälja motorförsäkringar till bilarna Volvo, Renault och Landrover. Försäkringarna säljs dels till privatpersoner men även till företag även lastvagnsförsäkringar till en del åkerier. Målsättningen är att har rätt pris (premie) till rätt kund. Vi intervjuade en person som arbetade som statistiker och dels som systemutvecklare. Han har en teoretisk utbildning med inriktning mot nationalekonomi och statistik.

4.2 En fråga om processer

Nedan analyseras förhållandet mellan organisation och Data mining system. Analysen sker utifrån varje fråga med tillhörande attribut.

Användbarhet
De beslutsområdena som förekom var inom marknadsföring och då främst kundsegmentering. En organisationen använde sig av regelbundna kampanjer för att locka till sig nya kunder. Tanken med Data mining var att analysera karaktären på de personer som nappade på erbjudandet. Resultaten skall användas för att hitta liknande personer som inte svarat på kampanjen för vidare bearbetning.

Relevans
Vid insamlandet av data nämnde en organisation att en begränsning är att de måste ta hänsyn till regler och lagar. En av respondenterna nämnde att de inom organisationen förekom en intern etisk policy som reglerade insamlandet. En annan aspekt på den datan som kan användas för Data mining är att den är tillgänglig. Det vill säga att datan finns lagrat i något system inom organisationen. En synpunkt som framkom var vikten av ett välorganiserat datalager som är byggt utifrån Data mining användningen. Sist nämndes att insamlandet skall vara beroende av vilken frågeställning som skall besvaras.

Aktualitet
En av organisationerna uppdaterade datan i databasen en gång i månaden varav en tabell innehållande aktuella kunder uppdaterades en gång i veckan. Datan lagrades i ett datalager som supportade större delen av verksamheten. Den andra organisationen uppdaterade databasen en gång per vecka.

Robusthet
Transaktionerna sker mellan de operativa systemen och datalagret. Datalagret har ett antal inbyggda funktioner som går i Batchmiljö och där finns det enskilda system som övervakar uppdateringarna. Detta för att systemet skall vara stabilt. Går en uppdatering inte igenom får en jouransvarig ett e-mail om detta på sin terminal. En organisation beskrev sin övervakning mellan transaktionerna med checkpoints där överföringarna mellan systemen var automatiska och där det skickades en signal om något gått snett. Ett påpekande var att det sällan går att övervaka allting utan man får välja det som man tror är viktigast.

Informationsintegration
De processer i organisationen som Data mining tekniken avsågs för var främst försäljningsprocessen men även analysverksamheten. En organisation avsåg att använda Data miningen inom marknadsavdelningen för kundsegmentering.

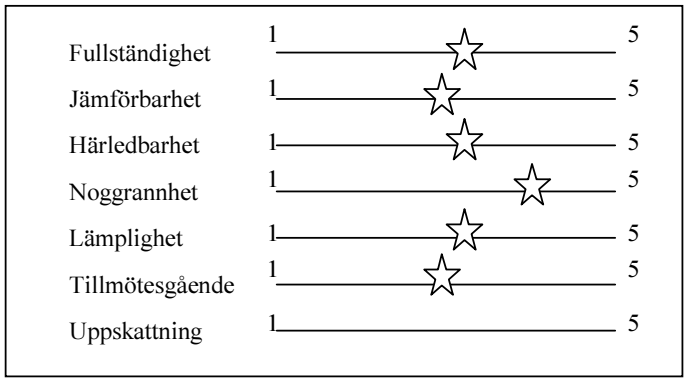
Begreppsintegration
Problematiken kring begreppsdefinitioner mellan olika system har en organisation tagit itu med vid införandet av ett datalager. En respondent uttryckte sig att:

- Begreppsproblematiken är inte något specifikt Data mining problem, utan ett problem som funnits länge.

En annan respondent menade att det kan vara problematiskt med begreppsdefinitioner då det inte finns ett datalager som grund. Om så är fallet får man bygga broar mellan systemen med till exempel separata Data marts.

- Det handlar om snickeriverkstad av ganska stora mått.

Vid intervjutillfället fick respondenterna möjlighet att ge betyg på hur väl de tyckte att Data mining systemet uppfyllde ett antal variabler. Nedan visas en modell (figur 4:1) över resultatet från vår empiri. Resultaten är subjektiva bedömningar och visar åt vilket håll organisationens åsikt om Data mining systemet lutar. Har stjärnan en placering på linjen med en dragning åt siffran fem visar det på ett "högt betyg", exempelvis visar studien att noggrannhet uppfattas som hög. Anledningen till att stjärnan fattas på linjen uppskattning är att respondenterna inte ansåg sig kunna göra en rättvis bedömning. Detta på grund av att Data mining användningen inte varit i bruk tillräckligt länge för att ge en rättvis bedömning.



FIGUR 4:1 ORGANISATIONENS BEDÖMNING AV DATA MINING SYSTEMET

Övriga kommentarer som framkom vid intervjutillfället var att för att Data miningen skall lyckas är det viktigt att användaren har kännedom om verksamheten, statistik och analys kompetens samt IT-kunskap. En respondent uttryckte att: - *det finns inte och kommer inte finnas ett verktyg som bara ger ett rakt svar utan man måste förstå hur man kommit fram till det.* En annan aspekt som lyftes fram var att det tar mycket tid i anspråk att bygga och preparera data för att kunna använda det i Data mining arbetet. Detta lyfts inte alltid fram i rätt proportion enligt en av respondenterna.

4.3 En fråga om aktörer

Här analyseras förhållandet mellan användaren och Data mining systemet. Analysen sker utefter varje fråga med tillhörande attribut.

Individuell validitet
På frågan hur respondenten upplevde om verktyget var anpassat efter egna specifika behov och önskemål blev svaret nej. Den största anledningen var att möjligheten att själv administrera systemet kunde varit högre. Användaren kunde inte anpassa rapporterna beroende på uppgift.

Relevans
När det gäller vem som skall välja viken data som skall ingå i processen är detta beroende av om organisationen använder sig av ett standardsystem eller ett egenutvecklat system. När det gäller det egenutvecklade systemet så är det en central enhet inom organisationen som styr de data som kan ingå i processen. De lokala cheferna hämtar information från denna enhet. Vid användningen av standardsystem så har man inte möjlighet att utveckla det verktyget varefter det uppkommer nya förutsättningar.

Kontroll
De framtagna resultaten rapporteras till en ledningsgrupp som utvärderar dessa.

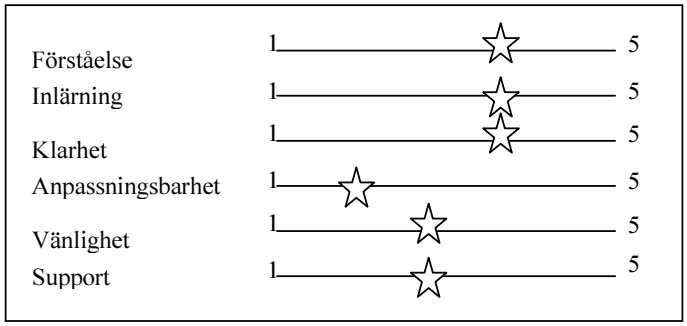
Klarhet
Den person som bestämmer hur resultaten skall presenteras är användaren. Hänsyn tags till de krav som finns från de enheter som skall ta del av resultaten.

Användbarhet
De inom organisationen om väljer vilken designmetod som skall användas är beroende på om det är ett egentutvecklat system eller ett standardsystem. När det gäller de egenutvecklade system är det en dataavdelning inom organisationen som väljer designmetod utifrån funktionella krav. Dessa krav kommer ifrån kommersiella enheter och frontlinjen i organisationen. De funktionella kraven påverkar även val av design metod, vid användning av standardsystem. En viktig faktor som organisationen (inköpsavdelningen) väger in är att det ur ekonomisk synvinkel är ett bra köp.

Lärande
För att utveckla och förfina processen använder sig respondenten av ett felrapporterings-system för att rapportera till drift och förvaltning. Det finns även ett system för nyutveckling.

Lärande
Den roll som användaren har i systemutvecklingen är mer som en funktionell kravställare än rollen av att bestämma. De funktionella kraven ställs med utgångspunkt från kundens krav och förväntningar.

Vid intervjutillfället fick respondenterna möjlighet att ge betyg på hur väl de tyckte att Data mining systemet uppfyllde användarvänligheten genom att betygsätta ett antal variabler. Nedan visas en modell (figur 4:2) över resultatet från vår empiri. Resultaten är subjektiva bedömningar och visar åt vilket håll användarens åsikt om Data mining systemets användarvänlighet lutar. Har stjärnan en placering på linjen med en dragning åt siffran fem visar det på ett "högt betyg", exempelvis visar studien att förståelsen, inlärning och klarhet är hög.



FIGUR 4:2 ANVÄNDARENS BEDÖMNING AV DATA MINING SYSTEMETS ANVÄNDARVÄNLIGHET

Övriga kommentarer som framkom vid intervjutillfället var respondentens definition av Data mining: - *Jag tycker att Data mining är att få fram så mycket data som möjligt för att kunna ställa de rätta frågorna och fatta de nödvändiga besluten. Detta för att lägga upp strategier för hur man skall sälja och hur man skall planera framtida resurser.*

4.4 En fråga om design

Här analyseras förhållandet mellan Data mining systemet och leverantörerna av Data mining lösningar med det empiriska materialet som grund. Analysen presenteras utefter varje enskild fråga och dess tillhörande attribut.

Lönsamhet
Det som vi kunde utläsa var att Data mining används för att förutse potential, risk och behov. Detta sker främst inom Customer Relational Management (CRM) området, men även inom kvalitetsoptimering för att öka lönsamheten. CRM är mycket kopplat till marknadsaktiviteter såsom att hitta nya kunder som är rätt kunder det vill säga där det finns en potentiell lönsamhet. För befintliga kunder kan företagen använda sig av Data mining för att kunna korsförsälja det vill säga sälja fler produkter eller tjänster mot rätt kunder. Data mining kan även utnyttjas för att öka merförsäljningen vilket innebär att företagen kan sälja mer av samma produkt till exempel utnyttjandet av kontokort. Tekniken ger även möjlighet till att

kunna förutse risken att förlora en kund för att på det sättet kunna bearbeta kunden med rätt erbjudanden för att undvika kundflykt. Organisationer som är aktiva i hela kundprocessen kan öka livstidsvärdet, att en kund stannar längre, genom att erbjuda fler produkter och tjänster och på det viset få ökad lönsamhet. De organisationer med en stor kundstock, någon som kan köra verktyget och tillämpa det kan göra en kraftig segmentering. Detta för att effektivt bearbeta rätt kunder med rätt erbjudande. Ett konkret exempel som gavs var att ett företag genom att effektivisera och hitta rätt målgrupp för sina utskick lyckades få upp svarsfrekvensen med över 100 procent. Det gav företaget i slutändan en minskad utgift genom att de kunde minska sina utskick. När det gäller kvalitetsoptimering är det kopplat till tillverkningsprocessen. Där används det för att till exempel förutse risken att en produkt inte skall hålla sig inom ramen för kvalitén när det gäller dimensionerna. Organisationen vill minimera risken för att få dåliga produkter som inte kan säljas.

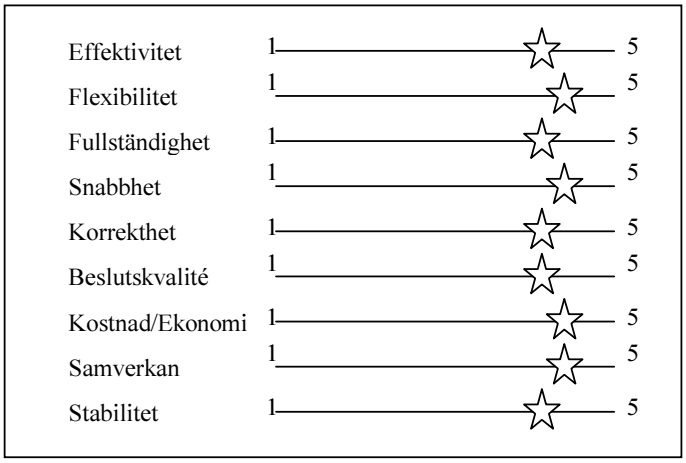
Användbarhet
För att en organisation skall kunna använda sig av Data mining tekniken krävs framför allt att det finns någon som kan köra programvaran. Tekniken gör så att företag får mer kunskap för att kunna agera på rätt information. De i organisationen som får mest nytta av kunskapen är marknadsavdelningen och ofta hela säljorganisationen. Denna del av organisationen har krav på sig att sälja utan att veta hur och mot vilka de skall rikta sig mot. Det resultat som framkommer av analyserna kan leda till att organisationen skall agera på ett nytt sätt och det omstrukturera hela organisationens verksamhet. I det traditionella sättet att arbeta med kampanjhantering ingår det många personer. Finns det ett verktyg som kan automatisera mycket av den processen kan många personer bli överflödiga. Ett konkret exempel som nämndes var en stor investmentbank som arbetat på ett produktorienterat sätt. Med hjälp av Data mining förändrades hela organisationsstrukturen till att arbeta segmentorienterat. Detta innebär att en person är ansvarig för ett visst kundsegment och skall utveckla det.

Anpassning
Det fanns en samstämmighet att det skulle finnas många möjligheter för användaren att kunna visualisera resultatet. Främst genom att grafiskt visa modellen genom till exempel lift diagram. I de fall där användaren är mer analytisk inriktad och statistiker skall det finnas möjlighet att se de exakta koefficienterna bakom varje variabel. Detta för att kunna se hur viktiga eller oviktiga variablerna är i modellen.

Säkerhet
Data mining som begrepp innebär att organisationen använder sig av historisk data för att kunna förutse framtiden. Detta betyder att organisationen först måste samla på sig data vilket gör att Data mining analysen inte sker i något operativt system. Detta medför att människor utifrån inte kan komma åt känslig information som skall användas för Data mining analys. När det gäller säkerhet vid användningen är att det inte är lämpligt att den som arbetar med att analysera data har tillgång till personuppgifter. Ett sätt är att använda sig av fiktiva ID-nummer när det gäller att identifiera och avidentifiera sitt data. Identifieringen kan ske genom att skicka upp ID-numren för att matcha mot personnummer. En annan aspekt på säkerhet är det hur leverantörerna har designat sitt system för att säkerställa kvalitén. Genomgående var att systemen var designade så att användaren skulle kunna arbeta efter vedertagen arbetsmetodik till exempel SEMMA eller CRISP-DM (se sid 33). Det är viktigt att kunna koppla systemet till ett datalager så att en hel analysmiljö erhålls för att det inte skall bli ett självstående analysverktyg.

Ansvarighet
Det finns egentligen inga svårigheter med att hämta in data från andra databaser. Förutsättningen är dock att organisationen har lagrat sin data i en välstrukturerad databas. Det är många företag som inte inser vikten av att ha en smart lagringsstruktur. I de fall där organisationen inte har en välstrukturerad databas och begreppssamordning kan det vara svårigheter för analytikern att förstå datamaterialet. Eftersom ett begrepp kan betyda en sak i den ena databasen men något helt annat i den andra databasen.

Vid intervjutillfället fick respondenterna möjlighet att ge betyg på hur väl de tyckte att Data mining systemet uppfyllde ett antal variabler. Nedan visas en modell (figur 4:3) över resultatet från vår empiri. Resultaten är subjektiva bedömningar och visar åt vilket håll leverantörernas åsikt om Data mining verktygen lutar. Har stjärnan en placering på linjen med en dragning åt siffran fem visar det på ett "högt betyg", exempelvis visar studien att effektiviteten är hög.



FIGUR 4:3 LEVERANTÖRERNAS SYN PÅ DATA MINING VERKTYG

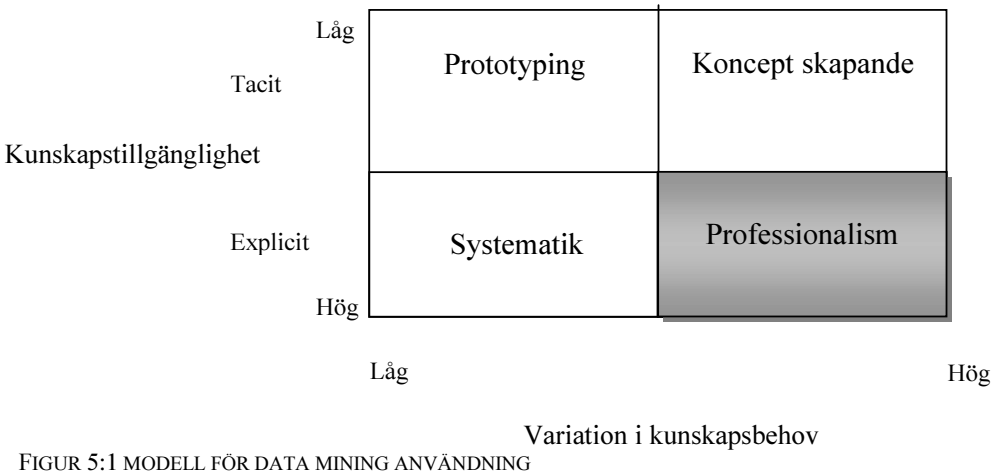
5. Diskussion

Genom den empiriska undersökning har värdefull information samlats in, bearbetats och analyserats. I detta kapitel kommer en diskussion att föras kring vilka problem som kan lösas med hjälp av Data mining samt vilka kritiska faktorer som finns vid användningen. Diskussionen kommer att anknyta till de teoretiska modeller som har presenterats i teori kapitlet samt de empiriska resultaten som har presenterats. Ett avsnitt med slutsatser redovisas och därefter kommer ett avsnitt med förslag till framtida forskning. Avslutningsvis kommer vi att delge våra reflektioner.

Människan har under alla tider använt sin förmåga för att utveckla och förbättra sin situation. När den första datorn såg dagens ljus var dess främsta syftet att stödja människan med att göra komplicerade beräkningar. Därifrån har datorn och dess möjligheter utvecklats till att underlätta även i andra situationer. Under lång tid har datorn använts i organisationer för att samla in och lagra information (Dahlbom & Mathiassen, 1993). Datorns förmåga att samla in data ligger långt över människans kapacitet att analysera den (Faayad et al., 1996). Allt eftersom konkurrensen hårdnar på marknaden krävs att organisationer utvecklar nya metoder och strategier. På grund av detta har det växt fram ett behov att använda den data som samlats in till något nyttigt. I inledningen nämner Noonan (2000) bland annat att det inte räcker i dagens läge att veta vad som hänt. Det gäller att kunna göra korrekta förutsägelser om framtiden för beslutsfattande. I denna process är Data mining ett verktyg som kan stödja beslutsfattandet.

5.1 Modell för Data mining användning

Vilken typ av problem kan lösas med hjälp av Data mining? I teorikapitlet presenterades Macintosh (1994) modell. Den modellen har varit utgångspunkten när vi klassificerat Data mining användningen utefter problemens struktur och variation. Utifrån teorin kom vi fram till att Data mining kan stödja beslutsprocessen vid hög kunskapstillgänglighet och oberoende på variation i kunskapsbehov. Det som i modellen benämns systematik och professionalism. När så är fallet framhåller litteraturen att Data mining kan appliceras på många beslutsområden. Det empiriska materialet visar att organisationer använder Data mining uteslutande inom professionalism området (se figur 5:1). De användningsområden som tekniken används för är marknadsaktiviteter och då främst segmentering. Ett annat område som tekniken används inom är planering av resurser till exempel personal.



FIGUR 5:1 MODELL FÖR DATA MINING ANVÄNDNING

När det gäller arbetsuppgifter med repetativ karaktär finns det redan välbeprövade verktyg som ger användaren det tillräckligt stöd för att utföra sina sysslor. Olika rapporteringssystem till exempel OLAP, kan ge information om historiska händelser som kan appliceras på liknande problem i framtiden. En anledning till varför Data mining används inom området professionalism tror vi är att det finns möjligheter att göra stora kostnadsbesparingar speciellt inom marknadsaktiviteter. Ett exempel som framkom under intervjuerna var en organisation som med hjälp av Data mining hittat ”rätt” målgrupp och ökat svarsfrekvensen på sina utskick betydligt. En annan orsak till att Data mining idag används nästan uteslutande för marknadsaktiviteter tror vi är att leverantörernas kunskap och erfarenhet är mest utvecklad inom detta område. Detta leder till att leverantörerna har mer underlag att sälja in en Data mining lösning eftersom de kan hänvisa till tidigare framgångsrika projekt.

Vilka är de kritiska faktorerna vid Data mining användning och hur hanteras dessa? I teorikapitlet presenterades Scott Mortons (1971) ramverk. Ramverkets uppgift var att klargöra beslutsprocessen och vad som påverkar en beslutsituation. Ur den modellen kunde en formel för beslutskvalité presenteras. För att beslutskvalité skall uppnås krävs att rätt information når rätt person vid rätt tidpunkt (Magoulas, 2001).

Beslutskvalité= $I*M*R_{ij}$

För att ge Data mining de bästa förutsättningarna för att tillfredsställa organisationen ständiga behov av tillförlitligt beslutsunderlag måste hänsyn till formeln för beslutskvalité beaktas. Formeln genererar ett antal kritiska faktorer som vi kommer att diskutera. Den första kritiska faktorn är information. Det vilar ett stort ansvar på organisationen att samla in rätt data till sitt datalager. Vilken data som skall användas beror främst på vilken fråga som Data miningen skall svara på (Putten, 1998). Detta fick även stöd i den empiriska undersökningen. En aspekt som framkom under intervjun var att även om organisationen har en frågeställning och vet vilka data som skall samlas in kan det finnas hinder för datainsamlingen. De begränsningar som nämndes var Sveriges rikes lagar och regler. I vissa fall har organisationen även en intern etisk policy som skall respekteras. När vi betraktar resonemanget ur ett kvalitetsperspektiv innebär detta att viktig samband och mönster inte kan upptäckas. Detta kan leda till att beslutskvalitén påverkas negativt. Ur ett etiskt perspektiv är det dock ett skydd för att värna

om den personliga integriteten. Noonan (2000) och Groth (2000) påpekar att för att få bra resultat krävs bra datakvalité på insamlad data. Ett uttryck som ofta nämns i litteraturen för att illustrera detta är garbage in- garbage out. För att erhålla ett bra resultat från Data mining krävs att den data som analyseras är aktuell. Det är också viktigt att det finns möjlighet att härleda den data som används vid Data mining analysen. Av den anldningen är det vässentligt att kunna hantera begreppintegrationen. Enligt Söderström (1997) bör uppdatering av datalagret ske varje natt för att upprätthålla aktualiteten på data. Vår studie visar att datalagret uppdateras en gång i veckan eller en gång i månaden. Vår tolkning är att det är en kostnadsfråga för organisationen. Ekman (1970) menar att organisationen måste göra en avvägning mellan hur aktuell data skall vara och kostanden för att producera den. Vi tolkar ovanstående faktorer som en problematik kring informationsintegration.

En kritisk faktor som Putten (1998) belyser är att data skall göras tillgänglig för att användas i Data mining projekt. En förutsättning för tillgänglig data är att organisationen har ett datalager där data finns strukturerad. Under intervjuerna framkom det tydligt att om det inte finns ett datalager i organisationen som är uppbyggt utefter Data mining användningen är det ett mycket tidsödande moment att strukturera data. Vår tolkning är att organisationen uppfattar tillgänglighet som att data finns lagrat i något system inom organisationen. Från dessa system kan data hämtas för att lagras i ett datalager. Vår uppfattning är att i de fall där organisationen har ett befintligt datalager som uppdateras kontinuerligt uppfattas tillgängligheten inte som något problem. När organisationen lagrar data i databaser utan att ha Data mining i åtanke upplever de problematiken kring datalager som ett Data mining problem. Det vill säga att problematiken att samordna och strukturera data kopplas samman med Data mining användning. Organisationen upplever att detta problem tillkommer enbart vid Data mining användning. De upplever också att leverantörerna ibland inte framhåller denna problematik i rätt proportioner. Leverantörerna och de organisationer med ett strukturerat datalager upplever inte att detta problem är specifikt för Data mining användningen. Denna problematik har funnits i organisationen sedan länge och är mer relaterad till problematiken kring datalager. Fayyad et al. (1996) ser omvandlingen från data till kunskap som en process i vilket Data mining är ett steg utav flera. Där Data mining innebär att söker efter mönster och samband som i nästa steg skall tolkas. En lösning är att diskutera Data mining i ett KDD perspektiv för att belysa att Data mining är en del av en större process. I den processen har organisationen många åtagande och problemställningar att ta itu med innan det verkliga Data mining arbetet kan påbörjas. Vår uppfattning är att det i dagsläget inte går med automatik att samla in data från olika funktioner i organsiationen till exempel försäljning och inköp. Vi upplever det som att Data miningsystemet är som en isolerad ö. Här anser vi att det ligger en utmaning hos utvecklare av Data mining lösningar att främst lösa problematiken med begreppsintegration.

För att gå vidare i formeln för beslutskvalité är människan en kritiska faktor. Fayyad et al. (1996) menar att det är viktigt att användaren har kunskap om det verksamhetsområde där tekniken skall användas. Den empiri som samlats in visar tydligt att användaren behöver kunskap dels inom sitt verksamhetsområde men även kunskaper inom statistikområdet. Detta är något som vi inte tycker lyfts fram tillräckligt i facklitteraturen. Där kan det emellanåt målas upp en bild att det inte krävs några specifika förkunskaper vilket vi anser är förkastligt.

Den sista kategorin av kritiska faktor i formeln för beslutskvalité är regler. Med regler menas fasen i beslutsprocessen och på vilken beslutsnivå i organisationen som beslutet skall fattas. Turban och Aronson (1998) menar att med Scott Mortons ramverk som utgångspunkt stödjer Data mining Intelligens fasen som består av att samla in information och början av Design

fasen, vilket innebär att alternativen utvärderas. Det kommer i slutändan alltid vara en människa i form av beslutsfattare som avgör vilket beslut som skall fattas. Vilket även en respondent i underökningen poängterade. Nämligen att det kan utvecklas sofistikerade analysverktyg men det går aldrig att komma ifrån att det är en människa som skall kunna förstå, tolka resultatet och handla därefter. Vår åsikt är att Data mining inte automatiskt leder till att hitta de lönsammaste kunderna utan är ett samspel mellan teknik och kompetens.
Det är främst på den strategiska nivån i organisationen som Data mining används. Detta tolkar vi som ett samband mellan vilka problemtyper som beslutsfattaren kommer i kontakt med på respektive nivå i organisationen. I Macintosh (1994) ramverk placerar vi som tidigare nämnts Data mining inom professionalism området. Främst för att det genom vår empiriska studie har framkommit att det är många marknadsaktiviteter som tekniken appliceras på. Anthony (1965) menar att på den strategisk nivån fattas de beslut om organisationens mål och inriktning. Vi anser att nyutveckling av tjänster och produkter samt segmentering är beslut som hanteras på strategisk nivå. Att det är främst på denna nivån i organisationen som Data mining används tror vi beror på att de flesta problem som skall fattas här kan klassificeras som ostrukturerade. En annan anledning tror vi är att Data mining systemet är utvecklad med dessa problemtyper i åtanke.

5.2 Slutsatser

Utifrån vår teoretiska referensram och det empiriska undersökningsmaterialet kan vi dra ett antal slutsatser. De slutsatser som vi presenterar här är relevanta utifrån vår huvudfråga. Dessa slutsatser dras under förutsättning att vår studie är korrekt.

Examensarbetets huvudfråga var:

Hur kan beslutsprocessen stödjas av IT i allmänhet och Data mining i synnerhet?

I dagsläget stödjer Data mining beslut som enligt Macintosh (1994) modell klassificeras enligt proffesionalism. Det vill säga problem som har hög kunskapstillgänglighet och låg variation i kunskapsbehov.

För att erhålla beslutskvalité har vi i vår studie kommit fram till att det är mycket viktigt att lösa problematiken kring systemintegration. Det gäller främst att kunna integrera Data mining systemet med organisationens befintliga system. Viktiga faktorer att ta hänsyn till vid utformning av Data mining systemet är begreppsintegration och informationsintegration. Andra faktorer att ta hänsyn till är kunskap och kompetens hos organisationens aktörer. En annan viktig faktor är kvalitén på den data som skall användas vid Data mining analysen.

5.3 Framtida forskning

Förslagsvis så kan det var intressant att förena Data mining med Knowledge Management området. Tanken är att organisationen skall lära sig och ta till var på den kunskap som genereras dels genom Data mining processerna samt genom de resultat som slutligen presenteras.

5.4 Egna reflektioner

Under arbetets gång har vi erhållit många tankar och reflektioner kring detta ämnesområde. Några av de reflektioner som vi tycker är intressanta vill vi delge läsaren. Det har under de senaste åren talats mycket om Data mining. Våra erfarenheter från den empiriska studien visar att det är mycket svårt att komma i kontakt med organisationer i Sverige som använder denna teknik. Vi tror att det kan bero på att utvecklingen inte har kommit så långt som den gjort i andra länder, främst USA. Vi upplevde att de organisationer som använder sig av Data mining i Sverige inte vill tala om sina erfarenheter på grund av konkurrensskäl. Detta kan betyda att när organisationen har hittat ett koncept som passar dem så är Data mining värdefull för verksamheten. Tekniken kan ge fördelar gentemot konkurrenter vilket då kan leda till att organisationens lönsamhet ökar.

En annan erfarenhet som vi vill uppmärksamma är användarrollen. Genom den empiriska undersökningen framkom det att det finns olika grupper av användare. De som använder Data mining verktyget och analyserar resultatet, de som tar del av det analyserade materialet och en grupp som både använder verktyget och fattar beslut på resultaten. Detta ställer stora krav på leverantörerna att utveckla ett anpassningsbart och användarvänligt gränssnitt. Den empirisk studien visar att här finns det mycket att förbättra. Det beror på att användare har specifika behov och krav.

I de otaliga böcker som vi tagit del får vi ett intryck av att Data mining är mycket bra och mycket enkelt att använda. Det skulle vi vilja påstå att det är en sanning med modifikation. Har organisationen ett datalager som är strukturerat med Data mining användningen i åtanke kombinerat med kunskap och erfarenhet av verksamheten och inom statistik området finns det en god grund för Data mining användning. Både litteraturen och empirin visar att det finns en potential för Data mining området som kan ge organisationen konkurrensfördelar. Vi tror att det är viktigt att närma sig tekniken med respekt. Att hämta data från olika databaser beskrivs i en mening i facklitteraturen men i verkligheten kan det vara något helt annat. Om data inte är strukturerad i ett datalager är det en lång väg att gå innan det verkliga Data mining arbetet kan påbörjas.

6. Referenser

6.1 Böcker

Anthony, R. (1965), *Planning and Control Systems: A FRAMEWORK FOR ANALYSIS*, Boston, Harvard University

Backman, J. (1998), *Rapporter och uppsatser*, Lund, Studentlitteratur

Bakka, F. J. och Fivelsdal, E. och Lindqvist, L. (1988), *Organisationsteori: struktur, kultur och process*er, Malmö, Liber

Berry, M.J.A. och Linoff, G. (1997), *Data mining Techniques: For Marketing, Sales and Customer Support*, New York, John Wiley & Sons, Inc.

Berry, M.J.A. och Linoff, G. (2000), *Mastering Data mining*, New York, John Wiley & Sons, Inc.

Dahlbom, B. Ed (1993), *Essays of Infology*, Department of Information systems, Göteborgs Universitet

Dahlbom, B. och Mathiassen, L. (1993), *Computers in context*, Oxford, Blackwell

Debevoise, N. T. (1999), *The Data warehouse method*, New Jersey, Prentice Hall Inc.

Edlund, P-O. och Högberg O. och Leonardz, B. (1993), *Beslutsmodeller i praktisk tillämpning*, Lund, Studentlitteratur

Edlund, P-O. och Högberg, O. och Leonardz, B. (1999), *Beslutsmodeller: redskap för ekonomisk argumentation*, Lund, Studentlitteratur

Ekman, B. (1970), *Beslutsrationalisering*, Uppsala, Almqvist & Wiksell Boktryckeri AB

Galbraith, J.R. (1977), *Organizational Design*, Addison-Wesley

Gellerstedt, M. (1997), *Statistiska metoder för kvalitetsutveckling*, Lund, Studentlitteratur

Groth, R. (1999), *Data mining: Building Competitive Advantage*, New Jersey, Prentice Hall Inc.

Hammer, M. och Champy, J. (1995*), Reengineering the corproation a manifesto for business revolution*, Finland, Werner Söderström Oy

Hartman, J.(1998), *Vetenskapligt tänkande*, Lund, Studentlitteratur

Holme, I. M. och Solvang, B. K. (1999), *Forskningsmetodik – om kvalitativa och kvantitativa metoder*, Lund, Studentlitteratur

Humphries, M. och Hawkins, M. W. och Dy, M. C. (1999), *DATAWAREHOUSING – Architecture and Implementation*, Upper Saddle River, Prentice Hall PTR

Inmon, W.H. (1996)*, Buildning the data warehouse*, USA, John Wiley & Sons, Inc

Jensen, K. (1995), *Kvalitativa metoder för samhälls- och beteendevetare*, Lund, Studentlitteratur

Jönsson, B. och Rehman, K. (2000), *Den obändiga söklusten*, Stockholm, Brombergs förlag

Kachur, R. J. (2000), *Data Warehouse Management Handbook*, New Jersey, Prentice Hall

Kotler, et al. (1999), *Principles of Marketing*, New Jersey, Prentice Hall

Macintosh, N. B. (1994), *Management Accounting and Control Systems*, John Wiley & Sons Ltd.

Magoulas, T. och Pessi, K. (1997), *Strategisk IT-management*, Göteborg, Vasastadens Bokbinderi

Manual från SPSS (1999), *SPSS Traning: Introduktion to Clementine*, UK: SPSS Ltd.

Merriam, B. S. (1994), *Fallstudien som forskningsmetod*, Lund, Studentlitteratur

Morton, Michael S. Scott. (1971), *Management Decision Systems Computer-based Support* for Decision Making, Boston, Harvard University

Nonaka, I. och Takeuchi, H. (1995), *The knowledge creating company*, New York, Oxford University Press

Putten, P. (1998), *Data mining in direct marketing databases, A collection of Essays on complexity and managment*, World Sientific Publishing co Pte Ltd, Singapore

Sabini, J. (1995), *Social Psychology*, New York, W. W. Norton & Company, Inc.

Singh, H. (1999), *Interactive data warehousing*, Upper Saddler River, Prentice Hall

Starrin, B. och Svensson, P-G. (1994), *Kvalitativ metod och vetenskapsteori*, Lund, Studentlitteratur

Szymanski, R. och Szymanski, P. och Pulschen, D. (1995), *Computers & Information system*, Upper Saddle River, Prentice Hall

Söderström, P. (1997), *"Data warehouse" Datalager Verksamhet, Metod, Teknik*, Lund, Studentlitteratur

Thuraisingham, B. (1999), *Data mining Technologies, Techniques, Tools and Trends,*

USA, CRC Press LLC

Trost, J. (1993), *Kvalitativa intervjuer*, Lund, Studentlitteratur

Turban, E. och Aronson, J. (1998), *Decision Support Systems And Intelligent Systems*, New Jersey, Prentice Hall

Wäneryd, B et al. (1993), *Att fråga?*, Örebro, Statistiska centralbyrån

6.2 Artiklar

Adam, et al. (1998), *A framework for the classification of DSS usage across organizations*

6.3 Elektroniska källor

Brand, E. och Gerritsen, R. (1998a), *Data mining and Knowledge Discovery* - DBMS Online, http://www.dbmsmag.com/9807m01.html

Brand, E. och Gerritsen, R. (1998b), *Classification and Regression* - DBMS Online, http://www.dbmsmag.com/9807m04.html

Brand, E. och Gerritsen, R. (1998c), *Association and Sequencing* - DBMS Online, http://www.dbmsmag.com/9807m03.html

Brand, E. och Gerritsen, R. (1998d), *Neural Networks* - DBMS Online, http://www.dbmsmag.com/9807m06.html

Brand, E. och Gerritsen, R. (1998e), *Decision Trees* - DBMS Online, http://www.dbmsmag.com/9807m05.html

Davenport, et al. (1998), *Successful knowledge management project*, Sloan Management review, EBSCOhost, http://www.epnet.com/ehost/login.html

Dilly, R. (1997), *"Data mining Notes"*, University of Belfast. http://www.pcc.qub.ac.uk

Fayyad, et al. (1996), *From Data Mining to Knowledge Discovery in Databases*, http://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf

Noonan, J. (2000), *Data Mining Strategies*, http://www.dmreview.com/portal_ros.cfm?NavID=91&EdID=2367&PortalID=9

Olofsson, K. (2000-06-16), *Business Intelligence ser nya möjligheter*, Computer Sweden, nr11, http://nyheter.idg.se/display.pl?ID=000616

Stjernberg, F. (1999), *Vad kan man veta?*,
www.hum.vxu.xe/publ/humanetten/nummer5/art9913.htm

Url 1: CRISP-DM, http://www.crisp-dm.org/

Url 2: Hurtig, U., *Svenska datatermgruppen*,
http://www.nada.kth.se/dataterm/fos-lista.html#f72

Url 3: SAS, http://www.sas.com

Url 4: SPSS, http://www.spss.com

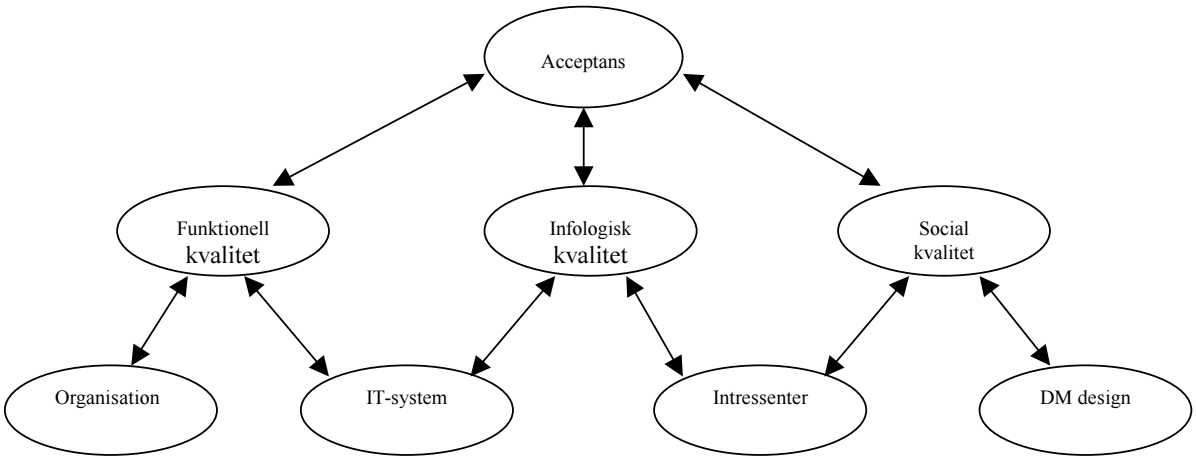
6.4 Opublicerat material

Agahi, F. (2000) Föreläsningsanteckningar Beslutstödssystem, Göteborgs Universitet

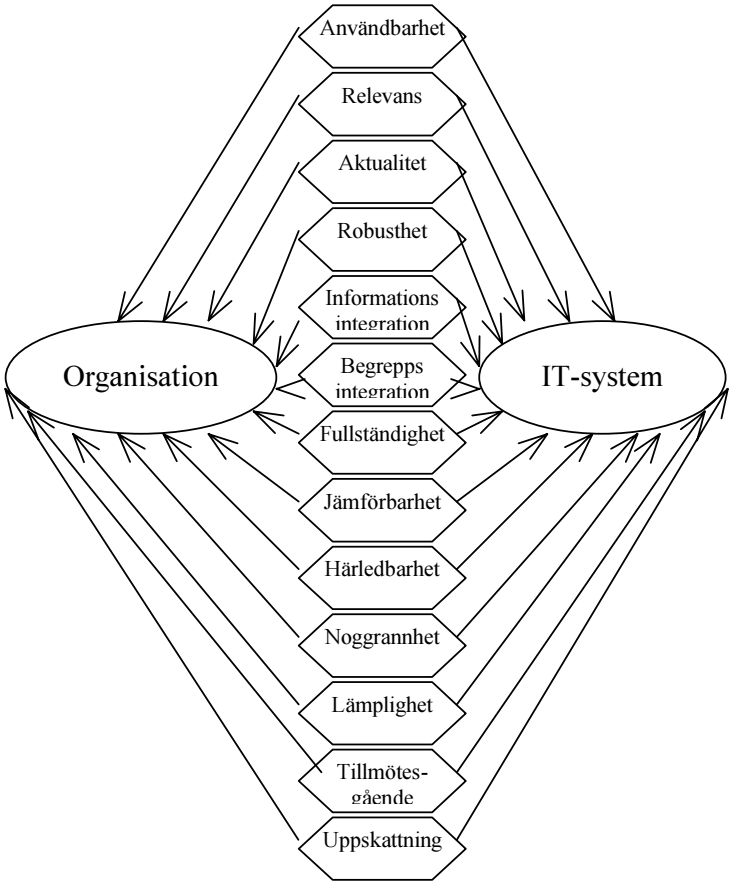
Gellerstedt, M. (2000) Föreläsningsanteckningar Data mining, Högskolan
Trollhättan/Uddevalla

Magoulas, T. (2001) Föreläsningsanteckningar Examensarbete II, Göteborgs Universitet

Övergripande objektmodell

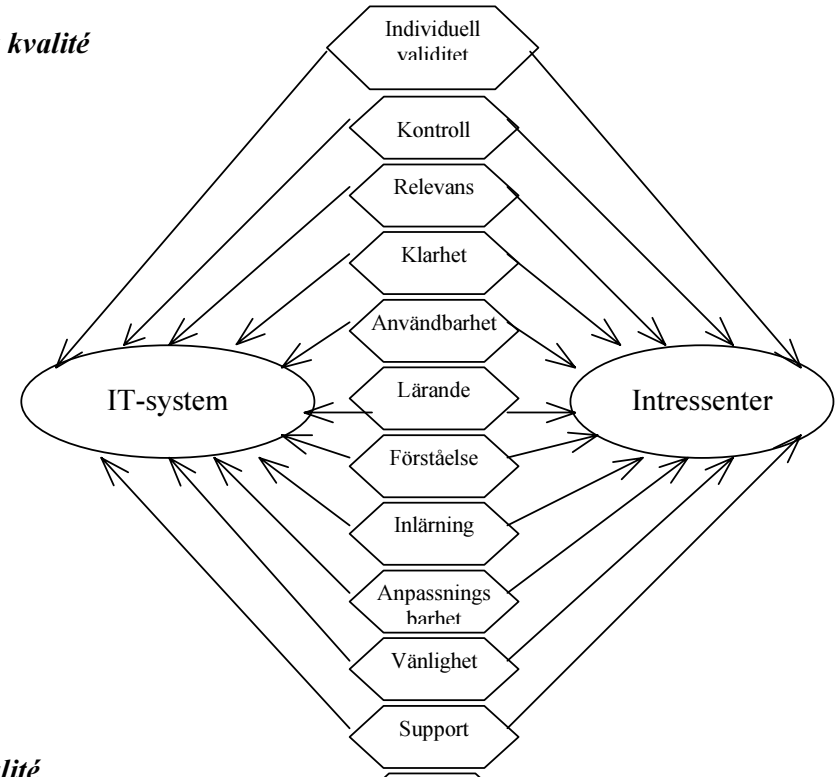


Funktionell kvalité

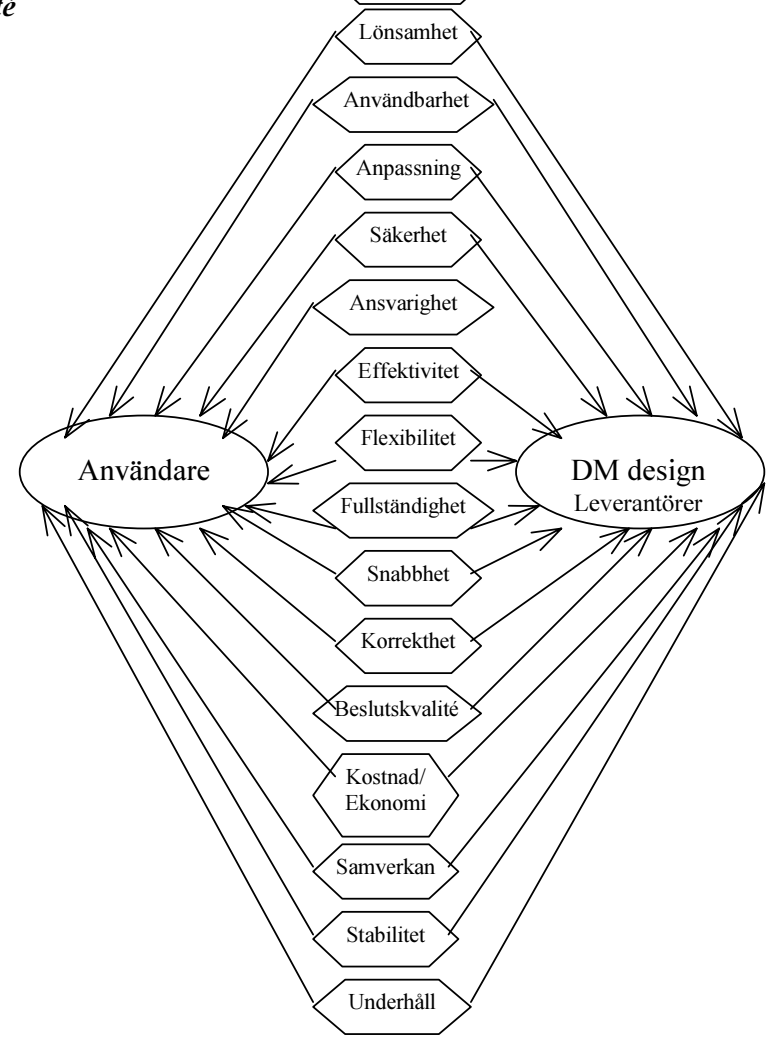


© Helena Gustafsson och Anna Larsson, 2001

Infologisk kvalité



Social kvalité



© Helena Gustafsson och Anna Larsson, 2001

Göteborgs Universitet
Institutionen för Informatik
Viktoriagatan 13
405 30 Göteborg

Organisationens namn

Data mining och beslutskvalité

Er organisation har blivit utvald till att deltaga i en undersökning för att kartlägga användningen av Data mining. Syftet med studien är att skapa en modell av hur Data mining kan stödja beslutsprocessen .

Vi önskar att komma i kontakt med en person som är kunnig inom Data mining området vid kontoret i Göteborg för en kort intervju på cirka 30 minuter. Vi behöver hjälp med en e-post adress eller ett telefonnummer till vederbörande person för vidare kontakt. Detta kan sändas till vår e-post adress: anna_helena00@hotmail.com

Ert deltagande är av stor betydelse för vår kartläggningen som är ett steg i vårt examensarbete i Informatik vid Göteborgs Universitet. Vi är därför mycket tacksamma för Er medverkan.

Resultatet av studien kommer Ni att ha möjlighet att ta del av på Internet i slutet av juni 2001. Adressen är http://www.handels.gu.se/epc/html/front.html

Med vänliga hälsningar

Anna Larsson och Helena Gustafsson

För mer information kontakta oss gärna!

Anna Larsson Tfn:
Helena Gustafsson Tfn:

© Helena Gustafsson och Anna Larsson, 2001