Datalogisk Institut
Aarhus Universitet

December 2003

Den Århusianske Fødselskohorte

Projekt i Datamining & Machine Learning

Søren Tjagvad Madsen, Datalogisk Institut, Århus Universitet
Jakob Hjort, Afdeling for Folkesundhed, Århus Amt

Indhold

1 Machine Learning

Denne rapport beskriver nogle forsøg med at lave machine learning på udvalgte attributter i den århusianske fødselskohorte. Machine learning er en bred vifte af metoder, der beskæftiger sig med at lade maskinen tage ved lære af de eksempler den præsenteres for. I dette projekt er machine learning blevet brugt som en klassificerings-metode. Maskinen kan opbygge en mere eller mindre kompliceret model baseret på data fra fødselskohorten og det er så tanken at denne model kan generalisere til at udsige noget om hidtil usete eksempler.

Projektet går ud på at opbygge modeller ud fra de tilgængelige data, som på baggrund af oplysninger om en vordende moder kan sige noget om det kommende barn. Ydermere forsøger jeg at finde frem til hvilke informationer om moderen, der er afgørende for barnet.

Ofte er statistiske metoder benyttet til at gennemskue store datamængder og afprøve hypoteser. Machine learning lægger sig også op af statistiske modeller, men ideen er at give maskinen en mere autonom rolle i opdagelsen af hypoteser.

Den århusianske fødselskohorte er venligst udlånt af Perinatal Epidemiologisk Forskningsenhed, Skejby Sygehus. Kun en delmængde af oplysningerne i databasen er blevet analyseret i denne rapport. Personfølsomme data er blevet maskeret.

Rapporten beskriver de overvejelser, der er gjort ved analysen af datasættet. Med et begrænset udvalg af attributter er det klart, at man ikke skal forvente at kunne se den endegyldige sammenhæng mellem en moders fysik og barnets ve og vel. Dette ville nok heller ikke være tilfældet om man inddrog samtlige attributter i databasen – mange genetiske anliggender kan ikke medtages i spørgeskemaet, som ligger til grund for undersøgelsen. Det er dog muligt i den forsimplede udgave af datasættet at se visse (velkendte) sammenhænge, såsom at rygning under graviditeten nedsætter fødselsvægten.

Den benyttede machine learning software er hovedsageligt WEKA fra The University of Waikato, introduceret i [IHW00]. Endvidere er der også benyttet en neural netværkspakke, SNNS (Stuttgart Neural Network Simulator, University of Stuttgart).

2 Datasættet

Fødselskohorten er blevet til på baggrund af 13 års indsamling af data fra gravide kvinder og oplysninger om deres børn umiddelbart efter fødslen. Databasen indeholder ingen oplysninger om barnets videre vækst og de komplikationer, der måtte opstå senere hen. Der er i alt 52875 cases. Kvinderne har udfyldt spørgeskemaer, vedrørende deres fysiske og psykiske tilstand gennem

graviditeten. På samme måde er der anført oplysninger om de nyfødte.

Fødselskohorten indeholder i den ubeskårede version ca. 60 spørgsmål af forskellig svarmulighedsomfang – numeriske såvel som med forskellige kategorier af svarmuligheder. Spørgeskemaerne har ændret udseende omkring 10 gange gennem årene, så nogle spørgsmål er blevet omformulerede, hvilket komplicerer at få konsistente data fra alle perioder. Ligeledes findes der et skema for hvert barn om fødslens forløb og barnets tilstand. Ikke alle kvinder har udfyldt hele skemaet, hvilket giver anledning til nogle såkaldte *missing values*. De er en slags støj i datasættet, som nok skaber en smule unøjagtigheder i mine beregninger.

Jeg har i fællesskab med DataManager Jakob Hjort udvalgt nogle attributter, som vi havde mistanke om, ville kunne findes korrelerede. Ved nogle af disse har vi for overskuelighedens skyld forsimplet svarmulighederne til ja/nej spørgsmål (se 2.1). De numeriske data er blevet diskretiseret ligeledes for at forsimple virkeligheden og øge beregneligheden. Der findes machine learning teknikker der tager numeriske data som input, men her har jeg valgt at se på udelukkende diskrete data. I alt er der blevet eksperimenteret med 12 input features (oplysninger om mødrene) og 4 klassificerings features (oplysninger om børnene).

Vi har i denne udvælgelsesprocess afskåret nogle oplysninger, der potentielt kunne være vigtige for vort foretagende. I den ganske aldeles maskinelle verden uden nogen form for baggrundsviden må man ideelt tage så mange features i brug, som man har til rådighed, for at give computeren størst mulighed for at opdage nye ting. Samtidig er der dog en fare for, at man kommer til at indføre unødig kompleksitet og støj i modellen. Men det pæne ved at lade maskinen selv finde frem til hvilke attributter, der er en forbindelse imellem er, at den ikke er forudindtaget, men blot konkluderer ud fra de eksempler, den præsenteres for. En kompleks model vil således i princippet kunne indfange nye sammenhænge, som man måske ikke selv ville være kommet i tanke om. Der kan måske eksistere netop én kombination af flere faktorer, der tilsammen giver stort udslag på det, der undersøges for.

Men en sådan model vil kræve et omfang, som langt overgår dette projekts rammer. Derfor begiver vi os i stedet i kast med at se på de mere grundlæggende og mere indlysende sammenhænge.

2.1 De tilgængelige informationer

Her følger en beskrivelse af de udvalgte features, som der er blevet eksperimenteret med. Af oplysninger om moderen haves:

2.1.1 Input features – informationer om moderen

- Moders prægravide vægt. Angivet i kg. Benyttes til at udregne BMI.
- Moders højde. Angivet i cm. Benyttes til at udregne BMI.

- Behandlet for blærebetændelse under graviditeten. Et ja/nej spørgsmål.
- Behandlet for Nyrebækkenbetændelse under graviditeten. Et ja/nej spørgsmål.
- Alvorlige sygdomme i nærmeste familie (også arvelige). Et ja/nej spørgsmål. Der skelnes i denne opgave ikke mellem hvilken og hos hvem.
- Medicin indtaget indenfor de sidste 3 mdr. forud for graviditeten. Et ja/nej spørgsmål. Der skelnes i denne opgave ikke mellem hvilken og hvor meget.
- Medicin indtaget oftere end én gang pr. uge under graviditeten. Der skelnes ikke mellem hvilken og hvor meget.
- Allergier – penicillin, jod, plaster ell. andet. Et ja/nej spørgsmål. Der skelnes íkke mellem hvilke eller hvor alvorligt.
- Antal cigaretter pr. dag op til graviditeten. Inddeles i 5 kategorier.
- Antal cigaretter pr. dag under graviditeten. Inddeles i 5 kategorier.
- Alkoholindtag i genstande pr. uge. Er inddelt i 9 kategorier.
- Forbrug af hash, speed, kokain el. andet under graviditeten. Et ja/nej spørgsmål. Der skelnes ikke mellem hvad og hvor meget.

Af moders vægt og højde har jeg udregnet:

- Moders prægravide Body Mass Index (BMI). Inddeles i kategorier.

Denne feature benyttes som et samlet udtryk for moders prægravide vægt (kg) og højde (m) (idet BMI=vægt/højde^2) og erstatter i mine forsøg disse to attributter.

Den nærmere beskrivelse af inddelingen i kategorier findes i (2.2).

2.1.2 Output features – information om barnet

- Barnets vægt. Angives i gram. Inddeles i 10 kategorier.
- Hvorvidt barnet overlever 7 døgn.
- Om barnet bliver overført til neonatal afdeling (afdeling for nyfødte).
- Gestationstid. Graviditetens længde i uger. Inddeles i kategorier.

Graviditetens længde har stor betydning for de andre output-attributter. I mine eksperimenter har jeg også forsøgt at bruge den som input-attribut (se kap. 3).

2.2 Diskretisering af datasættet

Det har været hensigtsmæssigt at inddele de numeriske data i nogle kategorier. Ved ja/nej spørgsmålene (blærebetændelse, nyrebetændelse, arvelige sygdomme, medicin inden, medicin under, allergier, hash, overlevelse, neonatal) og de allerede kategoriserede (alkohol) er det meget lige for.

I appendix A side 19 er alle attributterne og de værdier, de kan antage beskrevet. Fordelingen af eksemplerne for alle attributter kan også beses.

I de numeriske attributter (BMI, cigaretter før, cigaretter under, barnets vægt, gestationstid) har jeg truffet nogle valg. Generelt er der to muligheder for dette. Den første er at forudbestemme nogle lige store intervaller, som værdierne derefter inddeles i. Dette kaldes *equal-interval binning* jf. [IHW00, p. 240]. Dette har jeg gjort ved rygningsattributterne, nemlig kategorierne: 0 cigaretter, 1-5 cigaretter, 6-10 cigaretter, 11-15 cigaretter samt flere end 16 cigaretter (se kap. A.2.2 og A.2.3). Denne måde at opdele på gør, at der kan være ret stor forskel på, hvor mange eksempler der falder i hver kategori (se figur 5 og 6).

Også gestationstiden er inddelt således, men ikke med helt lige store intervaller: 20-29 uger, 30-34 uger, ugevis op til 42 uger og over 43 uger (se histogrammet i figur 8 s. 24 samt mine inddelinger derunder).

BMI er inddelt på en anden måde, nemlig således at der kommer omtrent lige mange eksempler i hver af de 10 kategorier. Dette lader sig gøre ved, at de intervaller, som eksemplerne inddeles i, har forskellig vidde. For at finde disse intervaller har jeg sorteret alle eksemplerne efter BMI og udpeget grænserne ved hver tiendedel af alle eksempler. Denne metode kaldes *equal-frequency binning* eller *histogram equalization* jf. [IHW00, p. 240]. Forskellen på histogrammerne kan ses ved sammenligning af figur 3 og figur 4 på side 19 og 20. Fordelen ved denne inddeling er, at den kan skelne bedre og mere finkornet mellem eksempler, der ligger tæt op af hinanden. Man kan på denne måde lettere tvinge en classifier til ikke blot altid at gætte på den klasse med flest eksempler i (som jo statistisk set er den mest fornuftige at gætte på).

Attributten vedr. barnets vægt er også inddelt efter denne histogramudlignings metode. Se figur 9 og figur 10 på side 25 og 26.

3 Eksperimenterne

Eksperimenterne går ud på at tage output-værdierne en efter en og undersøge, om der kan 'machine learnes' en model, der kan klassificere eksemplerne med en vis grad af tilfredshed. Ud fra nogle af disse modeller kan man dernæst se på hvilke attributter, der har den største betydning for udfaldet (feature selection).

Jeg vil eksperimentere mest med forudsigelsen af den nyfødtes vægt. Fødselsvægten er et godt mål for barnets tilstand. Overførsel til neonatal afdeling er også blevet behandlet mere grundigt.

Barnets overlevelse vil kun blive behandlet kort, da der er meget få eksempler at tage i betragtning. Et forsøg på at udglatte uligheden er dog blevet gennemført.

Gestationstiden har det ligeledes været svært at forudsige, men jeg ikke kigget på på dens værdi som input feature.

3.1 Algoritmerne

Jeg har benyttet flere machine learning algoritmer fra WEKA-pakken til at bygge modellerne med. De fleste er simple, men klarer sig alligevel godt. De mere avancerede løber dog også ind i problemer.

Endelig har jeg benyttet en neural netværks pakke SNNS som har fået lov til at bruge en del længere processeringstid.

Desværre har den store datamængde afskåret muligheden at afprøve visse algoritmer på hele sættet. Der er således kun blevet eksperimenteret på med delmængder af datasættet med algoritmerne der benytter support vector machines og case based learning.

De simple algoritmer er først blevet evalueret gennem stratified 10-fold crossvalidation. Det er en standard evalueringsmetode [IHW00, p. 126], der opdeler datasættet i 10 omtrent lige store dele (folds) på en sådan måde at alle klasser er repræsenteret i alle folds. Herefter trænes der efter tur på 9 af de 10 folds, og den sidste benyttes til at teste på (udregne klassificeringskraft). Den gennemsnitlige fejl på hver af de 10 test er resultatet af testen.

Alle algoritmer er også blevet testet ved at at splitte eksemplerne op i to mængder: 2/3 til at træne på og 1/3 til at validere på (hvor klassificeringsegenskaberne måles). Det neurale netværk benytter præcist den samme opdeling af datasættene som WEKA-algoritmerne (der er præcist de samme eksempler i opdelingerne), så en direkte sammenligning af resultaterne er mulig.

3.1.1 ZeroR

Den mest primitive algoritme. Den foreslår altid simpelthen bare den største klasse – den som den fandt flest eksempler i i træningseksemplerne. Dvs. at hvis ZeroR skal klassificere ud fra oplysninger om barnets vægt, vælger den altid kategorien 3950-4200 gram, da den indeholder flest eksempler (se histogrammet på figur 10 side 26). Dette klassificerer 10.69% af alle tilfældende korrekt. ZeroR kan ikke bruges til ret meget mere end at teste andre algoritmer op imod. Alle andre algoritmer skal helst klare sig bedre. Ellers er der intet brugbart lært.

3.1.2 OneR

Denne algoritme er stadig meget simpel, men klarer sig ret godt. OneR (One Rule) genererer et decision tree af dybde 1 og giver dette træ tilbage i form af regler.

For hver antagelig værdi af hver attribut køres ZeroR algoritmen, som udpeger hvilken klasse, hver værdi i denne attribut bedst karakteriserer. En samlet error-rate udregnes for hver input-attribut ud fra disse oplysninger. Den input-attribut med den laveste error-rate udpeges til at være den mest betydningsfulde, og klassificeringsreglerne fra denne attribut benyttes.

F. eks. ser de resulterende regler fra OneR, der skal klassificere fødselsvægten ud fra BMI, cigaretter før og under, alkohol og gestationstid således ud:

```
grav_lgd_uger:
        20-29_uger      -> 500-2780_g
        30-34_uger      -> 500-2780_g
        35_u    -> 500-2780_g
        36_u    -> 500-2780_g
        37_u    -> 500-2780_g
        38_u    -> 2780-3060_g
        39_u    -> 3060-3250_g
        40_u    -> 3650-3800_g
        41_u    -> 4200->_g
        42_u    -> 4200->_g
        43->_u  -> 4200->_g
        ?       -> 500-2780_g
(10098/52495 instances correct)
```

Det viser sig, at gestationstiden er den mest betydningsfulde. Den alene har en korrekt klassificering på 19.24%. Spørgsmålstegnet betegner den mest fremherskende værdi for missing values i gestationstiden.

OneR er specielt velegnet til at opstille en ordning af attributterne mht. klassificeringskraft, og dette viser sig at være et vigtigt redskab i denne rapport.

3.1.3 J48 decision tree learner

Næste trin i rækken er en mere elaboreret decision tree algoritme. Den brancher også på hver attribut og på hver værdi, de kan antage. Der udregnes et information gain, der fortæller, hvor godt en given attribut separerer træningseksemplerne ud fra deres klassificering. Output er et dybere decision tree. Træet beskæres i de dele, hvor der kun falder få eksempler. Jeg har benyttet WEKA's J48 algoritme. Et output kunne se således ud:

```
J48 pruned tree
------------------
```

```
cigaretter_under = 0_cig
|   bmi = 0-18.93_bmi: 2780-3060_g (428.07/362.85)
|   bmi = 18.93-19.81_bmi: 3250-3400_g (451.68/391.79)
|   bmi = 19.81-20.45_bmi: 3650-3800_g (457.38/396.57)
|   bmi = 20.45-21.1_bmi: 3400-3520_g (466.2/408.42)
|   bmi = 21.1-21.79_bmi: 4200-> (511.49/448.53)
|   bmi = 21.79-22.57_bmi: 3650-3800_g (495.48/431.49)
|   bmi = 22.57-23.52_bmi: 3950-4200_g (492.9/418.42)
|   bmi = 23.52-24.85_bmi: 4200-> (511.86/442.28)
|   bmi = 24.85-27.34_bmi: 4200-> (533.58/459.28)
|   bmi = 27.34->_bmi: 4200->_g (448.76/368.18)
cigaretter_under = 1-5_cig
|   bmi = 0-18.93_bmi: 3250-3400_g (46.11/36.86)
|   bmi = 18.93-19.81_bmi: 3250-3400_g (41.46/34.04)
|   bmi = 19.81-20.45_bmi: 3250-3400_g (42.71/31.47)
|   bmi = 20.45-21.1_bmi: 2780-3060_g (46.73/35.69)
|   bmi = 21.1-21.79_bmi: 3250-3400_g (44.3/37.99)
|   bmi = 21.79-22.57_bmi: 3250-3400_g (37.78/31.51)
|   bmi = 22.57-23.52_bmi: 3250-3400_g (38.27/30.93)
|   bmi = 23.52-24.85_bmi: 3950-4200_g (33.41/25.25)
|   bmi = 24.85-27.34_bmi: 2780-3060_g (28.7/23.67)
|   bmi = 27.34->_bmi: 3950-4200_g (34.45/27.23)
cigaretter_under = 6-10_cig: 2780-3060_g (517.27/426.05)
cigaretter_under = 11-15_cig: 2780-3060_g (172.08/141.67)
cigaretter_under = 16->_cig: 500-2780_g (66.34/55.09)

Number of Leaves  :     23

Size of the tree :      26
```

Her er kun inkluderet attributterne BMI, cigaretter under og klasserne af barnets vægt. Klassificeringen står efter kolon'et. Eksemplet er kørt på 1/10 af alle data.

3.1.4 NaiveBayes

Denne algoritme baserer sig betingede sandsynligheder. Den klassificerer et nyt eksempel ved at tildele det den mest sandsynlige klasse givet eksemplets input-attributter. Det naive ligger i at algoritmen antager at inputattributterne er indbyrdes uafhængige, hvilket mindsker udregningskompleksiteten af de betingede sandsynligheder. Den resulterende klasse v_{NB} udregnes givet input-attributterne $a_1, a_2 \ldots a_n$ således:

$$v_{NB} = argmax_{v_j \in V}[P(v_j) \prod_i P(a_i|v_j)]$$

V er mængden af mulige klasser for output-attributten. $P(v_j)$ er sandsynligheden for v_j (frekvensen af v_j). $P(a_i|v_j)$ er sandsynligheden for attribut a_i givet v_j. Se også [Mit97, p. 177].

Mitchell rapporterer gode egenskaber for denne algoritme – ofte på højde med decision trees og neurale netværk. [Mit97, p. 154].

3.1.5 Neuralt Netværk

Et neuralt netværk (se [Mit97, p. 81]) (NN) er en kompleks model, der iterativt optrænes ved at blive præsenteret for eksempler mange gange. Netværket har en stor repræsentationskapacitet. Et neuralt netværk er i stand til tage mange faktorer i betragtning på en gang. Det er således muligt at repræsentere både de hyppige eksempler såvel som de mere sjældne.

Da alle attributterne i dette projekt er blevet diskrete er det ligetil at lave en passende indkodning – en input-attribut tildeles det antal input-units som det antal værdier attributten kan antage. En indkodning af attributten BMI giver altså alene 10 input units. En indkodning af BMI og rygning under graviditeten giver $10 + 5$ input units osv. Når netværket præsenteres for et eksempel, påvirkes de input units, der svarer til eksemplets kategorier med værdien 1, og de andre kategorier med værdien 0.

I denne indkodning er der altså tilpas stor forskel på indkodninger af eksempler, hvilket skulle øge netværkets muligheder for at skelne mellem dem. Fortolkningen af det trænede netværks output-units sker ved at vælge den unit som giver den højeste output-værdi, og resultatet bliver dermed den kategori som unit'en repræsenterer. Jeg har kun eksperimenteret med en output-kategori af gangen.

Netværket er et 3 lags feed forward net. Jeg har benyttet standard backpropagation (se [Mit97, p. 97]) til at træne netværket med. Netværket har mellem 6 og 16 knuder i det skjulte lag.

Et neuralt netværk giver umiddelbart ingen forklaringer i form af f.eks. regler om, hvorfor den klassificerer, som den gør.

3.2 Forudsigelse af fødselsvægt

Jeg har kørt algoritmerne på alle datasættets 52875 eksempler, med de missing values det indeholder. Alle 11 input features er fra starten med (BMI, betændelserne (2), sygdomme i familien, medicin (2), allergier, rygningsattributterne (2), alkohol og hash), men også gestationstiden. Dette giver følgende resultater:

Korrekt klassificeret	ZeroR	OneR	J48	NaiveBayes	NN(16 hid)
Crossvalidation	10.69 %	19.17 %	18.71 %	19.93 %	–
Percentage split – Train	10.66 %	19.15 %	33.22 %	20.25 %	20,47 %
Percentage split – Test	10.75%	19.40 %	18.77 %	20.22 %	19,79 %

ZeroR gætter hele tiden på, at barnet vejer 3950-4200 gram. OneR opstiller samme model som vist på side 6. J48 opstiller et ret stort decision tree. Dette træ er den bedste model for træningseksemplerne – hele 33.22 % – men desværre generaliserer det ikke så godt, og korrektheden på testsættet bliver noget mindre. Dette er et godt eksempel på *overfitting* – det opbyggede

decision tree er simpelthen for specialiseret i netop træningsdatasættet. J48 brancher i dybde 1 efter gestationstiden, i dybde 2 hovedsageligt efter, hvor mange cigaretter der ryges under graviditeten, men også et sted ved BMI i stedet. Det neurale netværk er trænet i ret kort tid. Det stabiliserer sig hurtigt, hvorefter det ikke er lykkedes at træne det til at give bedre resultater (se error graph på bilag B.1.1 side 27). NaiveBayes klarer sig bedst af alle.

Algoritmerne ser ud til at være enige om, at det kan lade sig gøre at klassificere omkring 19-20 % af alle instanserne korrekt ud fra de givne oplysninger. Eftersom grænserne for de 10 vægtklasser som børnene klassificeres ind i er udregnet således, at der falder ca. 10 % i hver kategori, afslører de 19-20 %, at der i hvert fald er tillært noget viden af disse algoritmer.

Fjerner man gestationstiden fra input-attributterne ser situationen noget anderledes ud:

Korrekt klassificeret	ZeroR	OneR	J48	NaiveBayes	NN(20 hid)
Crossvalidation	10.69 %	12.07 %	–	13.19 %	–
Percentage split – Train	10.75 %	12.76 %	23.27 %	13.68 %	14.16 %
Percentage split – Test	10.57 %	12.11 %	12.15 %	13.19 %	12.60 %

Nu klarer algoritmerne sig ikke ret meget bedre end ZeroR, og igen klarer NaiveBayes sig bedst. Det ser altså ud til, at vi langt fra er i stand til at lave en fejlfri fødselsvægtsforudsigelse ud fra de data og algoritmer, som er med i forsøget.

Men hvad gør algoritmerne forkert? Man kan danne sig et indtryk af, hvor fejltagelserne sker ved at kigge på en såkaldt *confusion matrix*. Den beskriver for hver klasse, hvad algoritmen har klassificeret eksemplerne som. Jeg viser her en confusion matrix for J48 algoritmen:

```
=== Confusion Matrix ===

   a   b   c   d   e   f   g   h   i   j   <-- classified as
 170 193 170 145  93  96 160  96 229 300 |   a = 500-2780_g
 143 232 196 155  84  98 171  83 288 282 |   b = 2780-3060_g
 135 229 185 173  86  96 175  90 290 366 |   c = 3060-3250_g
 121 209 174 152  94  89 180 101 275 337 |   d = 3250-3400_g
 110 176 154 170  66 121 189 105 313 364 |   e = 3400-3520_g
  89 155 141 137  83  81 191 115 330 392 |   f = 3520-3650_g
 107 155 153 168 110  99 187  95 349 379 |   g = 3650-3800_g
  89 130 116 133  86  93 210 101 291 398 |   h = 3800-3950_g
  96 121 130 161 113  88 195 118 358 507 |   i = 3950-4200_g
  68  79 102 127  87  78 177 113 342 570 |   j = 4200->
```

J48 har kategoriseret 170 af 500-2780 grams-børnene korrekt, men 193 af dem som 2780-3060 grams-børn osv. De korrekte eksempler står på diagonalen fra øverste venstre hjørne til nederste højre. Det ser ret kaotisk ud, og man kan se, at der er store problemer med blot at klassificere nogenlunde tæt på det rigtige. Eneste trøst er, at der står to-cifrede tal i nederste venstre

hjørne i stedet for tre-cifrede – der er kun 68 af de børn, der fødes i den tungeste vægtklasse, der bliver kategoriseret i den letteste!

Konklusionen er, at der findes mange modsigende eksempler i datasættet, og det er derfor svært for algoritmerne at lære noget meget generelt.

Men lad os alligevel kigge lidt på hvilke informationer, som algoritmerne bedømmer ud fra. OneR brancher på attributten BMI:

```
bmi:
        0-18.93_bmi     -> 2780-3060_g (kat 1)
        18.93-19.81_bmi -> 3250-3400_g (kat 3)
        19.81-20.45_bmi -> 3250-3400_g (kat 3)
        20.45-21.1_bmi  -> 3650-3800_g (kat 6)
        21.1-21.79_bmi  -> 3950-4200_g (kat 8)
        21.79-22.57_bmi -> 3650-3800_g (kat 6)
        22.57-23.52_bmi -> 3950-4200_g (kat 8)
        23.52-24.85_bmi -> 4200->_g    (kat 9)
        24.85-27.34_bmi -> 4200->_g    (kat 9)
        27.34->_bmi     -> 4200->_g    (kat 9)
        ?       -> 2780-3060_g
(6599/52495 instances correct)
```

Kategorierne har jeg tilføjet – de passer med figur 10 på side 26. Denne model antyder, at der er sammenhæng mellem moderens BMI og barnets vægt, således at jo større moder, jo større bliver barnet. Sammenhængen er ikke helt entydig, men dog til at få øje på.

J48 brancher i dybde 1 på cigaretter under graviditeten, og først dybde 2 på BMI (alle steder), og i dybde 3 på en del flere attributter: cigaretter før graviditeten, hash og alkohol, medicin under graviditeten og allergier. Disse oplysninger kan – selv om de kun lader sig generalisere i 12 - 13 % af tilfældene – alligevel godt undersøges nærmere. I hvert fald skal vi se lidt nærmere på rygning.

Fjerner vi endnu en attribut fra datasættet, nemlig BMI, kan vi tvinge OneR algoritmen til at branche på en anden attribut. Denne måde at anvende algoritmen på er altså i mindre grad som en klassificeringsalgoritme, og i større grad for at få et indtryk af hver enkelt attributs klassificeringskraft. På denne måde findes der også frem til en ordning af attributterne ud fra denne klassificeringskraft.

```
cigaretter_under:
        0_cig   -> 3950-4200_g        (kat 8)
        1-5_cig -> 2780-3060_g        (kat 1)
        6-10_cig        -> 500-2780_g (kat 0)
        11-15_cig       -> 500-2780_g (kat 0)
        16->_cig        -> 500-2780_g (kat 0)
        ?       -> 4200->_g           (kat 9)
(6511/52495 instances correct)
```

Det ses at, disse regler er i stand til at kategorisere næsten lige så mange instanser korrekt (6511/52495 instanser), som når der branches på BMI

(6599/52495 instanser). Her ses det tydeligt, at bare der ryges en lille smule, nedsnættes fødselsvægten. Ryges der lidt mere, tilhører fødselsvægten den aller laveste kategori. J48 var i stand til at kombinere de to attributter, og brancher altså først på rygning under graviditeten, og dernæst på BMI.

Fortsættes denne metode med hele tiden at fjerne den vigtigste klassificeringsattribut kommer denne rækkefølge:

```
cigaretter_før:
        0_cig   -> 3950-4200_g
        1-5_cig -> 3950-4200_g
        6-10_cig        -> 2780-3060_g
        11-15_cig       -> 2780-3060_g
        16->_cig        -> 500-2780_g
        ?       -> 2780-3060_g
(6477/52495 instances correct)

alkohol_under:
        alc_0   -> 2780-3060_g
        alc_1   -> 500-2780_g
        alc_2   -> 2780-3060_g
        alc_3   -> 3650-3800_g
        alc_4   -> 3950-4200_g
        alc_5   -> 3950-4200_g
        alc_6   -> 3950-4200_g
        alc_7   -> 4200->_g
        alc_8   -> 3950-4200_g
        ?       -> 2780-3060_g
(5858/52495 instances correct)

medicin_før:
        yes     -> 500-2780_g
        no      -> 3950-4200_g
        ?       -> 3060-3250_g
(5699/52495 instances correct)

beh_for_blærebet:
        yes_blæ -> 500-2780_g
        no_blæ  -> 3950-4200_g
        ?       -> 2780-3060_g
(5686/52495 instances correct)

allergier:
        yes     -> 3650-3800_g
        no      -> 3950-4200_g
        ?       -> 2780-3060_g
(5684/52495 instances correct)

arvelige_sygd:
        yes_arv -> 4200->_g
        no_arv  -> 3950-4200_g
        ?       -> 2780-3060_g
(5682/52495 instances correct)
```

```
medicin_under:
        yes     -> 500-2780_g
        no      -> 3950-4200_g
        ?       -> 2780-3060_g
(5669/52495 instances correct)

hash:
        yes     -> 3650-3800_g
        no      -> 3950-4200_g
        ?       -> 2780-3060_g
(5660/52495 instances correct)

beh_for_nyre_bæk_bet:
        yes_nyre        -> 3250-3400_g
        no_nyre -> 3950-4200_g
        ?       -> 2780-3060_g
(5655/52495 instances correct)
```

Det, der springer i øjnene er, at både alkohol, medicin før og under graviditeten og en blærebetændelse kan føre til en fødselsvægt der ligger i den mindste kategori (500-2780 gram). Det kan også se ud som om at rygning før graviditeten er minsker fødselsvægten. Det kan vi ikke konkludere noget om, uden at vide om det er de samme kvinder som har røget under graviditeten, der er årsag til denne regel.

Til at visualisere resultaterne henledes opmærksomheden til figur 1, side 13. Denne anskueliggører at kvinder med lav BMI får flest af de 'lette' børn, mens kvinderne med høj BMI får flest af de 'tunge' børn. Endvidere ses at kvinder der er føder tidligt får flest af de små børn.

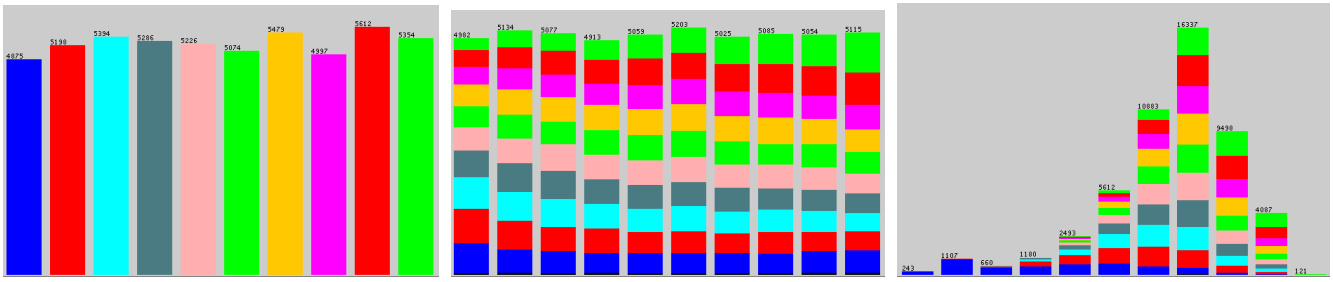
3.3 Forudsigelse af overførsel til neonatal afdeling

Denne gang er der kun to muligheder mht. klassificering. At gestationstiden også er en afgørende faktor for om barnet skal i særlig behandling efter fødslen kan man gøre sig klart ved at se på histogrammet i figur 2 side 13.

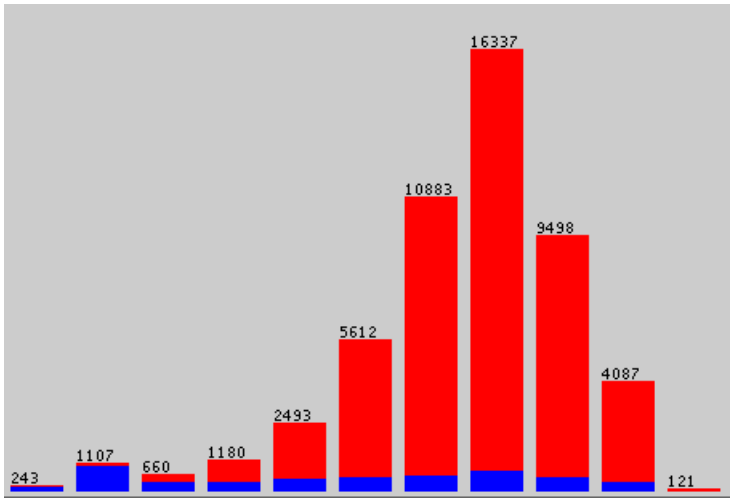
Det er klart at for tidligt fødte oftest skal have særlig behandling. Jeg har igen gennemført test af algoritmerne med og uden gestationstiden som input-attribut. Først med:

Korrekt klassificeret	ZeroR	OneR	J48	NaiveBayes	NN(12 hid)
Crossvalidation	90.08 %	92.12 %	92.09 %	92.06 %	–
Percentage split – Train	90.21 %	92.19 %	92.33 %	92.15 %	92.63 %
Percentage split – Test	89.81 %	91.96 %	91.93 %	91.92 %	91.80 %

ZeroR forudsiger at ingen skal overføres. OneR forudsiger at alle graviditeter der er kortere end 36 uger ender med et barn på nyfødts-afdelingen. J48 brancher også på graviditetslængden, og sender alle der føder inden uge 35 afsted, og alle efter uge 35 slipper. Men hvis graviditeten er på 35 uger,



Figur 1: Fordelingen af barnets vægt-kategorier (a) i forhold til henholdsvis BMI (b) og graviditetslængde (c).



Figur 2: Histogram over gestationskategorierne. De blå andele angiver børn der overføres til neonatal afdeling.

så er det næst mest afgørende afgørende om der har været hash (mm.) inde i billedet. Hvis moderen ikke har røget hash kommer barnet helt sikkert – ifølge træet – på neonatal afdeling hvis moderen enten er ikke-ryger eller røg over 11 cigaretter om ugen før hun blev gravid. Det er svært at forstå logikken i dette træ, og det klarer sig da også lidt dårligere på testsættet end OneR.

Det neurale netværk viser denne gang mere villighed til at blive optrænet. Men jo mere der trænes, og jo bedre det passer på træningsdatasættet, desto dårligere passer det på testsættet. Netværket er trænet i 100 generationer med træningssættet inden testen blev udført, og derfor er fejlen endnu ikke så stor. Træningskurven kan beses i figur B.1.2 side 27.

Igen er alle algoritmerne ret enige om hvor godt det kan lade sig gøre at forudsige denne attribut.

Uden gestationstiden bliver det igen lidt sværere:

Korrekt klassificeret	ZeroR	OneR	J48	NaiveBayes	NN(12 hid)
Crossvalidation	90.08 %	90.08 %	–	90.08 %	–
Percentage split – Train	90.03 %	90.03 %	90.03 %	90.03 %	90.13 %
Percentage split – Test	90.18 %	90.18 %	90.18 %	90.18 %	90.14 %

Alle algoritmerne er enige om at ingen kommer på neonatal afdeling. Det neurale netværk gætter dog lidt forskelligt, men mest ved siden af, og er derfor heller ikke så god. Jeg konkluderer at gestationstiden er langt den vigtigste af de attributter, som jeg har til rådighed.

Jeg har prøvet at sætte en IB1 algoritme (one-nearest-neighbor) igang på 15000 eksempler i datasættet (så det er lettere beregneligt). Denne algoritme regner ikke efter sandsynligheder eller andre optællinger af eksempler, som de andre WEKA-algoritmer, som jeg har eksperimenteret med. IB1 kigger på alle eksemplerne i træningssættet som punkter i det euklidiuske rum. Hvert punkt har en klassificering. Testsættet undersøges ved at indsætte hvert eksempel i dette rum, og give det den klassificering, som det nærmeste punkt fra træningssættet har. Men også denne må se sig slået – kun 84.44 % klassificeredes korrekt:

```
=== Confusion Matrix ===

    a    b   <-- classified as
   66  426 |    a = yes_neonat
  352 4156 |    b = no_neonat
```

Ligeledes blev en *support-vector machine* algoritme snydt til at klassificere ingen på neonatal afdeling.

Endnu et forsøg er blevet gennemført for at tvinge algoritmerne til at tage ved lære af de underrepræsenterede eksempler (altså de børn der ender på neonatal). 5245 børn ender på neonatal afdeling, mens 47630 slipper. Jeg

prøvede at udligne forskellen ved at lave et datasæt af eksempler med ca. lige mange eksempler i hver klasse ved at indsætte alle ikke-neonat-eksemplerne 9 gange. Jeg trænede OneR algoritmen på dette udvidede datasæt, og evaluerede det derefter på hele det normale datasæt. OneR klassificerede denne gang blot 53.90 % af eksemplerne korrekt, men til gengæld havde den benyttet nogle mere interessante regler (endskønt de klassificerede dårligere):

```
bmi:
        0-18.93_bmi     -> no_neonat
        18.93-19.81_bmi -> no_neonat
        19.81-20.45_bmi -> no_neonat
        20.45-21.1_bmi  -> no_neonat
        21.1-21.79_bmi  -> no_neonat
        21.79-22.57_bmi -> no_neonat
        22.57-23.52_bmi -> yes_neonat
        23.52-24.85_bmi -> no_neonat
        24.85-27.34_bmi -> yes_neonat
        27.34->_bmi     -> yes_neonat
        ?       -> yes_neonat
(51118/94835 instances correct)
```

Disse regler antyder at der er en sammenhæng mellem moderens BMI og om barnet trænger til særlig behandling efter fødslen. Det ser ud til at barnet klarer sig bedst hvis moderens BMI ikke er for stor.

3.4 Forudsigelse af barnets overlevelse

Langt de fleste overlever heldigvis fødslen i dette datasæt. Kun 260 ud af de 52875 børn i datasættet overlever ikke den første uge. Det er derfor ret svært at komme til at sige noget generelt om dette emne. Men igen er gestationstiden en vigtig faktor. Man har en ringere chance for at overleve, hvis man er født meget for tidligt.

Ingen af de algoritmer jeg har præsenteret har været i stand til at skelne mellem liv og død ud fra de givne oplysninger – de foreslår alle at alle overlever, så jeg vil ikke præsentere nogle resultater ud over dette:

```
Naive Bayes Classifier

Class yes: Prior probability = 0

bmi:  Discrete Estimator. Counts =  24 19 23 17 31 23 32 19 30 40  (Total = 258)
beh_for_blærebet:  Discrete Estimator. Counts =  19 218  (Total = 237)
beh_for_nyre_bæk_bet:  Discrete Estimator. Counts =  1 227  (Total = 228)
arvelige_sygd:  Discrete Estimator. Counts =  36 214  (Total = 250)
medicin_før:  Discrete Estimator. Counts =  62 194  (Total = 256)
medicin_under:  Discrete Estimator. Counts =  58 197  (Total = 255)
allergier:  Discrete Estimator. Counts =  47 212  (Total = 259)
cigaretter_før:  Discrete Estimator. Counts =  152 14 31 22 42  (Total = 261)
cigaretter_under:  Discrete Estimator. Counts =  182 20 39 12 7  (Total = 260)
alkohol_under:  Discrete Estimator. Counts =  1 1 3 8 15 21 23 35 94  (Total = 201)
```

```
hash:  Discrete Estimator. Counts =  32 227  (Total = 259)
grav_lgd_uger:  Discrete Estimator. Counts =  33 52 17 16 21 19 30 30 19 8 1  (Total = 246)
barn_vægt:  Discrete Estimator. Counts =  131 25 22 9 12 9 5 5 10 10  (Total = 238)
```

NaiveBayes tæller eksempler. Herover er vist hver attribut, og for hver værdi attributten kan antage er angivet hvor mange af de døde der findes for denne værdi. Da der for de flestes attributters vedkommende er en ulige fordeling af eksemplerne er det svært at slutte noget ud fra disse tal, men ved BMI og barnets vægt (som jeg har taget med på input-siden), hvor eksemplerne jo er ligeligt fordelt, kan man lige ane at de mødre der har den højeste BMI (større end 27.34) har en større risiko for at miste barnet end de andre (der er 40 eksempler i denne kategori imod højst 32 i de andre). Ligeledes har børn der fødes i den første vægtklasse (500-2780 gram) en mindre sandsynlighed for at overleve end de der er tungere/større (131 eksempler i denne kategori imod højst 25 i de andre). Også børn der fødes i de næste to kategorier (2780-3250 gram) er mere udsatte.

Så indtil nu kan vi konkludere at gestationstiden har stor indflydelse på barnets vægt (ikke overraskende), som har indflydelse på barnets overlevelsesmuligheder.

Jeg prøvede igen at opveje forskellene i antal ved at lave et specielt datasæt hvor jeg indsatte de 260 negative eksempler 202 gange hver, og de andre blot en gang. OneR trænedes på dette sæt og evalueredes på hele det normale sæt. OneR genererede disse regler:

```
alkohol_under:
        alc_0   -> no
        alc_1   -> no
        alc_2   -> no
        alc_3   -> no
        alc_4   -> yes
        alc_5   -> yes
        alc_6   -> yes
        alc_7   -> no
        alc_8   -> no
        ?       -> yes
(61071/105135 instances correct)
```

Den branchede på alkohol. Disse regler kategoriserer kun 38,36 % af tilfældende rigtigt, men antyder altså en sammenhæng mellem alkoholindtag (kat. 4 – 6) og barnedødelighed. (Jeg ved desværre ikke på nuværende tidspunkt hvad kategorierne betyder, så jeg kan ikke vurdere med min almindelige sunde fornuft om der skulle kunne være noget om sammenhængen...)

3.5 Forudsigelse af gestationstid

Også denne attribut er det svært at se noget konkret om ud fra de givne data. ZeroR, OneR og J48 kategoriserer altid efter den største gruppe: 40 uger.

NaiveBayes kategoriserer også til de andre kategorier, men får færre rigtige. Jeg gætter på at der findes forklaring på hvert af tilfældene, som jeg ikke har oplysninger om.

Korrekt klassificeret	ZeroR	OneR	J48	NaiveBayes	NN(12 hid)
Crossvalidation	31.28 %	31.28 %	–	31.19 %	–
Percentage split – Train	31.02 %	31.02 %	31.02 %	30.97 %	31.30 %
Percentage split – Test	31.82 %	31.82 %	31.82 %	31.81 %	30.96 %

Igen er algoritmerne dog enige om hvor meget system der er i datasættet. Jeg tager det som et tegn på at der ikke er nogen åbenlyse årsager i de givne attributter der bestemmer gestationstiden.

4 Konklusion

Projektet har undersøgt fem machine learning algoritmers evner til at klassificere udvalgte attributter i den århusianske fødselskohorte. Algoritmerne har vist sig nogenlunde lige duelige til de opgaver de er blevet stillet over for.

12 egenskaber ved en moder, og disses indvirkning på den nyfødte er blevet undersøgt. Nogle brugbare informationer er dukket op. Graviditetens længde ser ud til at være langt den vigtigste faktor for barnets velbefindende. Den har både indvirkning på barnets fødselsvægt, barnets overlevelse og om barnet bliver indlagt på nyfødts-afdelingen. Det lyder jo alt sammen meget fornuftigt, men det er også ret udtalt i datasættet.

Ser man på fødselsvægten alene ser den ud til at vokse i takt med moderens BMI og aftage med hvor meget hun ryger. Desuden ser det også ud til at medicin indtaget før og under graviditeten samt en blærebetændelse oftest fører til et barn i den letteste kategori. Nogle af disse oplysninger er nok korrelerede (f. eks. medicin og en behandling for blærebetændelse), så man skal passe på med at konkludere at alle gravide der tager en eller anden form for medicin føder mindre børn. Allergier, sygdomme i familien og et (evt. lille) forbrug af hash eller stærkere sager ser ikke ud til at have den store indvirkning på fødselsvægten, men mit datasæt kender jo heller ikke hele sandheden.

Det viste sig langt sværere at få konkrete oplysninger ud af at kigge på om barnet bliver overført til neonatal afdeling, og om det overlever. Igen var gestationstiden den mest afgørende faktor. Børn der fødes for tidligt har det sværere. Det ser dog også ud til at kvinder med en høj BMI får flest børn overført. Ligeledes er det kvinderne med den højeste BMI der har de dårligste odds når det gælder barnets overlevelse – men altså kun 40 ud af de ialt 5115 kvinder i denne klasse mistede barnet. En anden faktor ser ud til at være alkohol (som jeg desværre ikke kender kategorierne på...).

Hvilke faktorer der er bestemmende for gestationstiden har jeg ikke kunnet finde frem til. Jeg gætter på at eksemplerne kan forklares med forskellige former for komplikationer, der ligger ud over mit datasæt.

Generelt har de afprøvede algoritmer været i stand til at lære omtrent de samme ting. Det neurale netværk har i ingen af tilfældende haft gavn af en længere processeringstid, der i alle tilfælde førte til en smule overfitting.

For at opnå bedre klassificeringsresultater og for at kunne udsige noget mere (og noget mere rigtigt) om virkeligheden må man nok tage flere informationer i brug. Mit datasæt har været meget begrænset, og informationerne i det forsimplede. Men alligevel synes jeg at machine learnings algoritmerne har gjort det muligt at opstille nogle hypoteser, som jeg, trods min ringe baggrundsviden om graviditeter, finder plausible. Nogle genetisk bestemte ting kan man aldrig opdage i denne slags spørgeskemaer, men der er helt sikkert mange flere oplysninger at komme efter, og machine learning kombineret med en vis portion baggrundsviden vil helt sikkert kunne finde frem til disse.

A Fordelinger

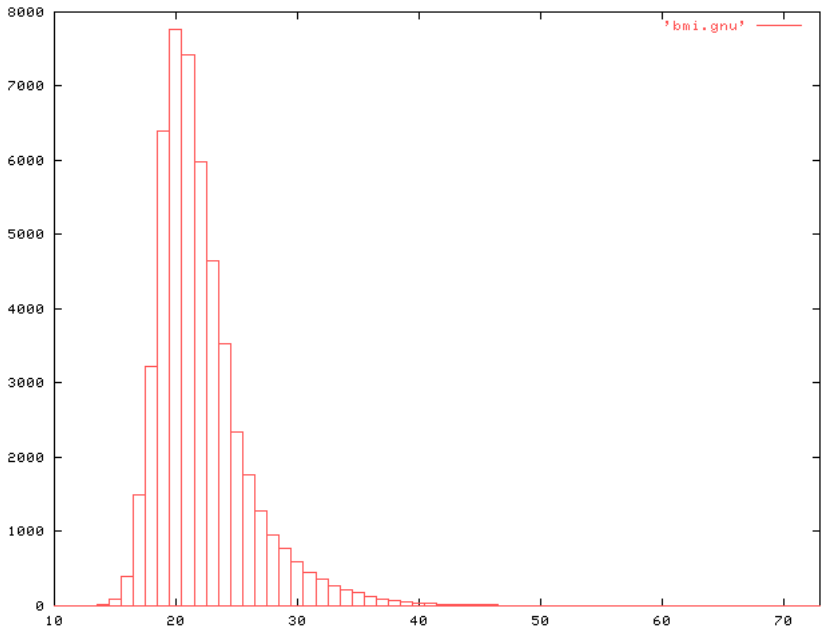
A.1 Fordelinger for ja/nej spørgsmål

Attribut	Antal ja	Antal nej	antal missings
Blærebetændelse	2446	46236	4193
Nyrebækkenbetændelse	129	46998	5748
Arvelige sygdomme	9134	41660	2081
Medicin før	12381	39305	1189
Medicin under	8913	42680	1282
Allergier	9977	41610	1288
Hash mm.	4229	47668	978
Barnet dør inden 7 dage	260	52615	0
Overført til neonatal	5245	47630	0

A.2 Fordelinger for fler-kategori spørgsmål

A.2.1 Body Mass Index

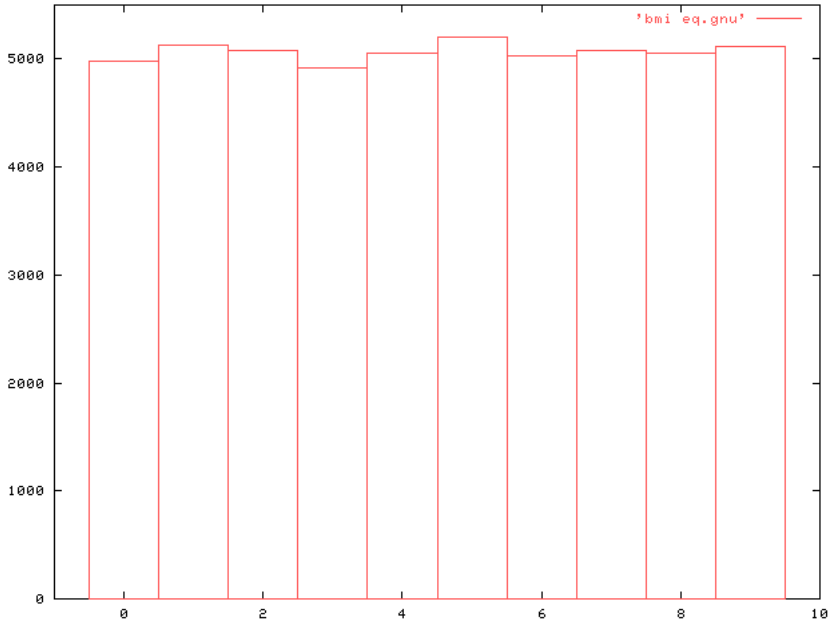
For at få et indtryk af værdierne for BMI, viser jeg her fordelingen af BMI inddelt i intervaller af 1.0 kg/m^2:



Figur 3: BMI fordelt i intervaller af 1.0 kg/m^2

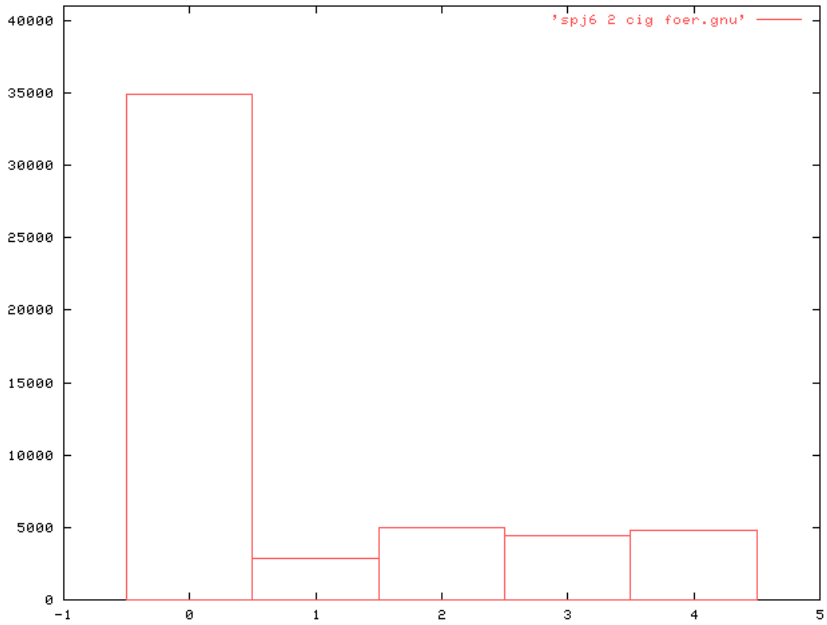
Attribut	Min	Max	Gennemsnit	Missings
BMI	10.09	73.14	22.63	2228

Kategori	BMI	Antal
0	under 18.93	4982
1	18.93-19.81	5134
2	19.81-20.45	5077
3	20.45-21.10	4913
4	21.10-21.79	5059
5	21.79-22.57	5203
6	22.57-23.52	5025
7	23.52-24.85	5085
8	24.85-27.34	5054
9	over 27.34	5115



Figur 4: BMI fordeling efter histogram equalization

A.2.2 Cigaretter før graviditeten

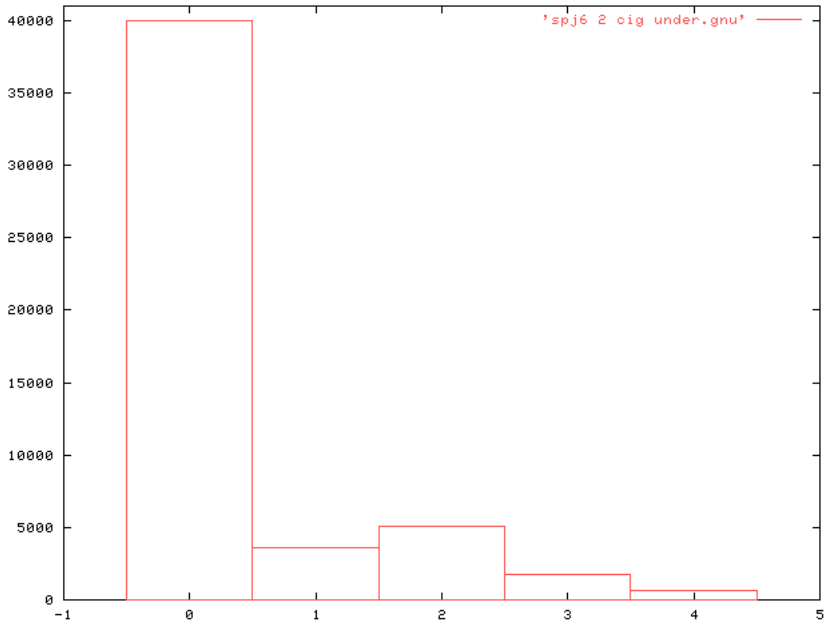


Kategori	Cigaretter pr. uge	Antal
0	0	34914
1	1-5	2881
2	6-10	5011
3	11-15	4443
4	flere end 16	4830

Figur 5: Antal cigaretter pr. uge før graviditeten inddelt i 5 kategorier

Attribut	Min	Max	Gennemsnit	Missings
Cigaretter før	0	60	4,26	796

A.2.3 Cigaretter under graviditeten

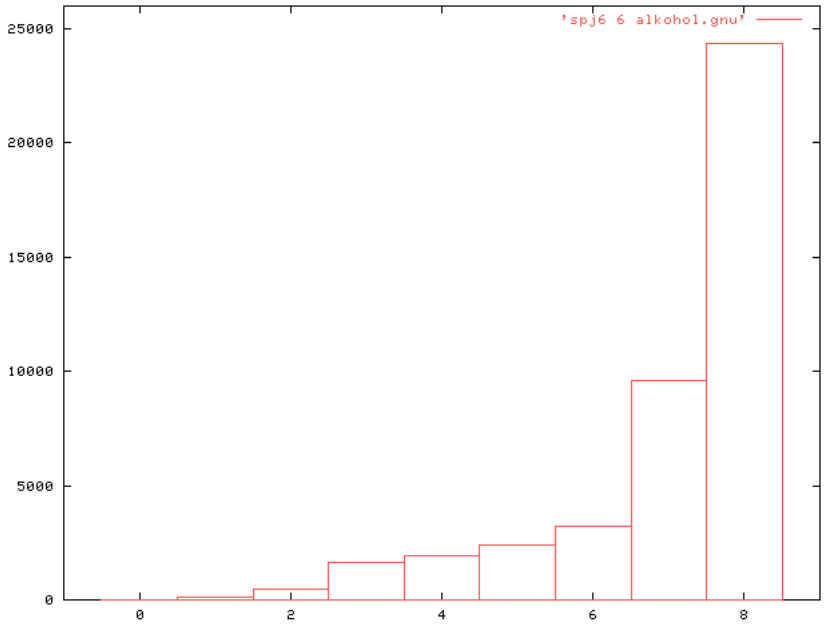


Kategori	Cigaretter pr. uge	Antal
0	0	39984
1	1-5	3579
2	6-10	5083
3	11-15	1800
4	flere end 16	660

Figur 6: Antal cigaretter pr. uge under graviditeten inddelt i 5 kategorier

Attribut	Min	Max	Gennemsnit	Missings
Cigaretter under	0	40	1,92	1769

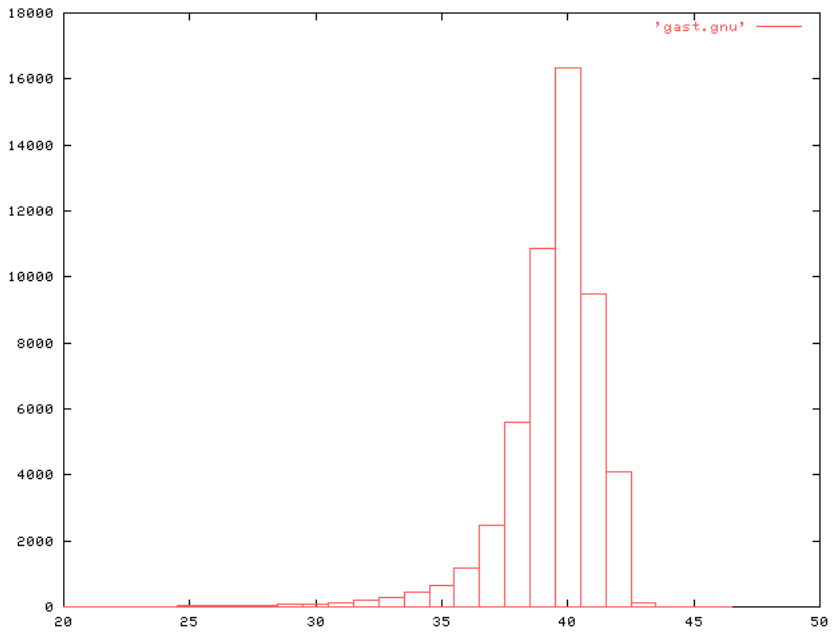
A.2.4 Alkohol under graviditeten



Kategori	Beskrivelse	Antal
0	?	10
1	?	93
2	?	496
3	?	1619
4	?	1944
5	?	2407
6	?	3250
7	?	9594
8	?	24347
missings		9115

Figur 7: Antal genstande indtaget pr. uge under graviditeten inddelt i 9 kategorier

A.2.5 Gravididitetslængde i uger



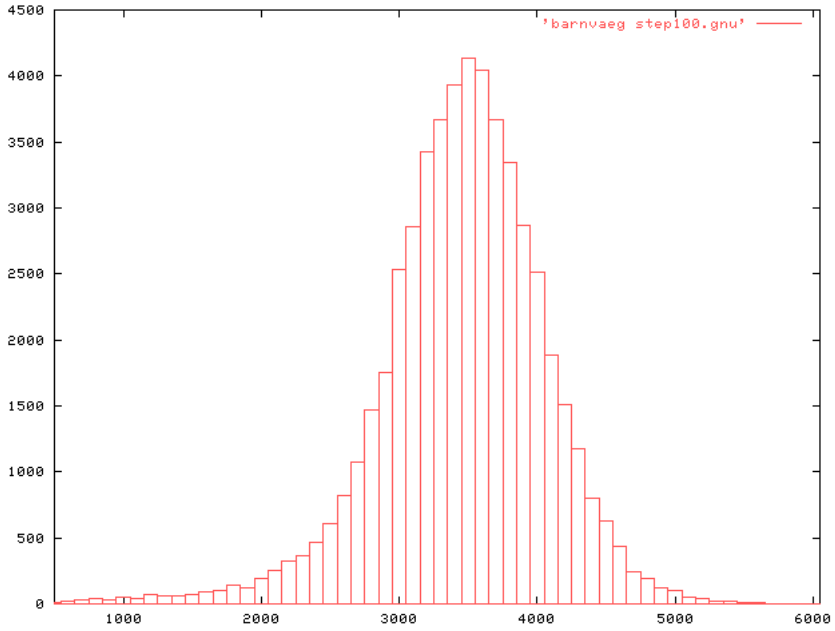
Figur 8: Gestationstid fordelt i antal pr. uge

Mine inddelinger:

Kategori	Gestationstid i uger	Antal
0	20-29	243
1	30-34	1107
2	35	660
3	36	1180
4	37	2493
5	38	5612
6	39	10883
7	40	16337
8	41	9498
9	42	4087
10	43 el. mere	121

Attribut	Min	Max	Gennemsnit	Missings
Gestationstid	21	46	39,46	654

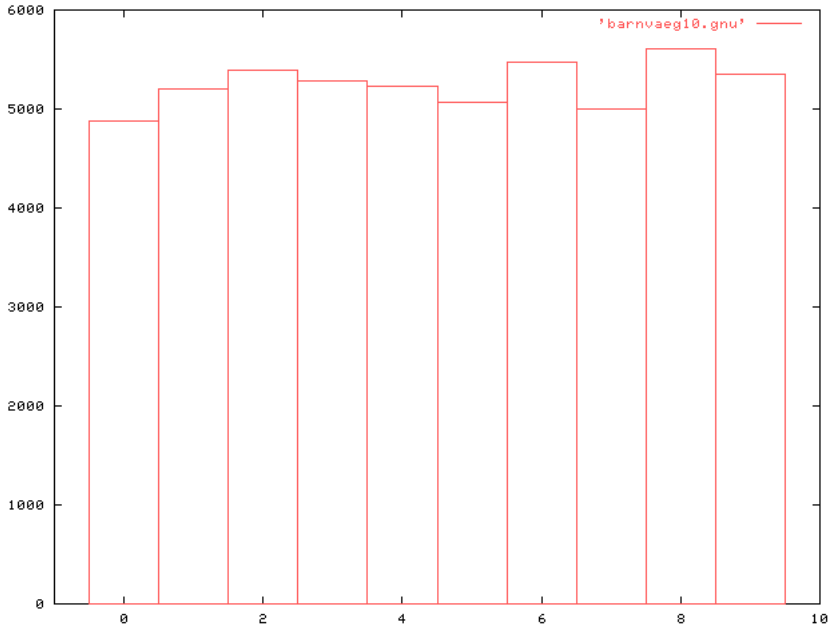
A.2.6 Fødselsvægt



Figur 9: Fordelingen af fødselsvægten inddelt i intervaller af 100 gram

Attribut	Min	Max	Gennemsnit	Missings
Fødselsvægt	500	6050	3495	380

Kategori	Fødselsvægt	Antal
0	500-2780	4875
1	2780-3060	5198
2	3060-3250	5394
3	3250-3400	5286
4	3400-3520	5226
5	3520-3650	5074
6	3650-3800	5479
7	3800-3950	4997
8	3950-4200	5612
9	over 4200	5354



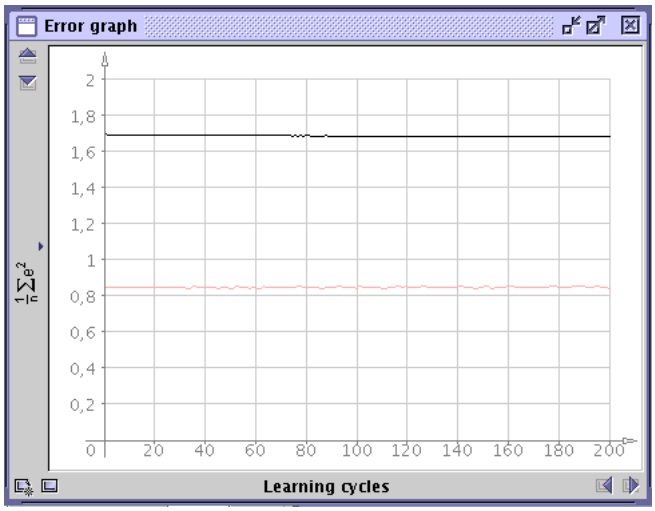
Figur 10: Fødselsvægtsfordeling efter histogram equalization

B Neuralt netværk setup

B.1 Error graphs

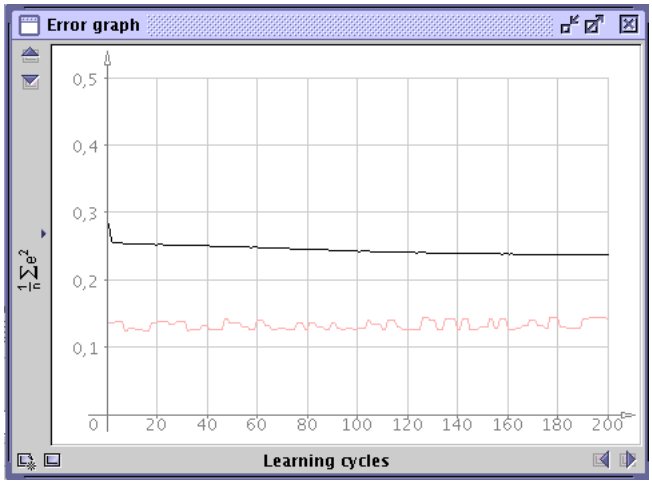
B.1.1 Klassificering af fødselsvægtattributten

Mean squared Error for trænings- (øverst) og valideringssæt (nederst) for klassificering af fødselsvægtattributten. Træningen har næsten ingen effekt.



B.1.2 Klassificering af neonatal attributten

Mean squared Error for trænings- (øverst) og valideringssæt (nederst) for klassificering af neonat-attributten. Der ses en lille tendens til overfitting.



Litteratur

[IHW00] Eibe Frank Ian H. Witten. *Data Mining.* Academic Press, 2000.

[Mit97] Tom M. Mitchell. *Machine Learning.* McGraw-Hill, 1997.