NTNU
Kunnskap for en bedre verden

Department of Computer Science

TDAT 3025 - Anvendt maskinlæring med prosjekt

Prediksjon av fotballresultater

Av:
Marcus Sevaldsen

November, 2020

Innhold

1 Introduksjon

1.1 Maskinlæringens potensiale i fotballen

Fotball har i lang tid vært en av verdens største idretter, og engasjerer mange mennesker over hele verden. For følgerne av sporten kan det av flere årsaker være interessant å forutse utfallet av enkeltkamper. Dette gjelder gjerne bookmakere og eksperter som ønsker å gi så gode spådommer som mulig, men det kan også være interessant for mannen i gata for å ta velinformerte valg innen betting og veddemål. Fotballens formål og regler er relativt enkle, men det viser seg likevel å være mange ulike faktorer som ligger bak hvert enkelt resultat. I senere tid har bruk av omfattende statistikk blitt et viktig verktøy i analysene, noe som har introdusert mange nye og spennende måter å gjøre dypdykk i resultater. Det er likevel vanskelig å vite hvordan de ulike tallene i statistikken kan tolkes og vektlegges, og mye som skjer i løpet av en fotballkamp er også rene tilfeldigheter. En løsning på dette problemet kan være å bruke maskinlæring for å dykke enda dypere i statistikken, og på den måten finne komplekse sammenhenger som kanskje ikke er åpenbare ved første øyekast.

1.2 Formål med oppgaven

Hovedmålet med dette prosjektet er å utforske om maskinlæring sammen med avansert statistikk kan være et godt virkemiddel for prediksjon av fotballkamper. Det blir spesifikt undersøkt hvor godt lineær regresjon og nevrale nettverk egner seg til problemstillingen, og hvordan mengde og kompleksitet av datapunkter påvirker nøyaktigheten i disse modellene. For å evaluere modellene vil det være interessant å undersøke hvordan de presterer sammenlignet med noen grunnleggende referansepunkter. Det ultimate målet vil være å overgå bookmakernes spådommer. Ved å betrakte resultatene er det ønskelig å utforske følgende spørsmål:

- Hvilke modeller ser ut til å fungere best?
- Hvilke datapunkter har størst innvirkning på resultatet?
- Kan noen av modellene måle seg med bookmakernes spådommer?

2 Bakgrunn

2.1 Fotballens hovedregler og formål

For å forstå hva som ligger bak et fotballresultat, må man først og fremst ha en grunnleggende forståelse for hvordan kampene utspiller seg. Her er reglene kort forklart [3]:

- Hver kamp består av to lag med 11 spillere, 10 utespillere og en målvakt.
- Begge lagene har hvert sitt mål de skal forsvare på hver sin side av banen.
- Målet er å få ballen i motstanders mål og samtidig hindre motstanderen i å gjøre det samme.
- Hver kamp består av to omganger på 45 minutter, og når kampen er over er det laget med flest scoringer som har vunnet kampen.
- For et lag har altså hver kamp 3 mulige utfall, seier, uavgjort eller tap.
- I de fleste ligasystemer gir seier 3 poeng, uavgjort 1 poeng og tap 0 poeng, hvor laget med flest poeng etter endt sesong vinner.

2.2 Viktige fotballbegreper

I denne oppgaven brukes det en del fotballfaglige begreper og forkortelser som er viktige å ha oversikt over. Det refereres til disse begrepene som et gjennomsnitt over tidligere kamper.

- **Expected goals(xG):** Forventet antall mål av et lag i en enkelt kamp, basert på kvaliteten på målsjanser som er skapt.
- **Expected goals against(xGA):** Forventet antall innslupne mål for et lag i en enkelt kamp, basert på kvaliteten på målsjanser i mot.
- **Goals per game(gpg):** Lagets gjennomsnittlige antall scoringer over tidligere kamper.
- **Conceded per game(cpg):** Lagets gjennomsnittlige antall mål innsluppet over tidligere kamper.
- **Points per game(ppg):** Lagets gjennomsnittlige antall poeng over tidligere kamper.
- **Passes allowed per defensive action(PPDA):** Antall pasninger det forsvarende laget i gjennomsitt tillater det angripende laget før det begås en forsvarende handling. Sier noe om intensiteten i forsvarsspill og press.
- **Deep progressions(deep):** Antall pasninger og driblinger inn i motstanders siste tredjedel i en enkelt kamp.
- **Deep progressions allowed(deep allowed):** Antall pasninger og driblinger mostander tillattes inn i siste tredjedel i en enkelt kamp.

2.3 Maskinlæringsmetoder

For å løse problemstillingen i oppgaven brukes det forskjellige varianter av veiledet maskinlæring [12]. Dette betyr at man kartlegger inndata til ønsket utdata, ved hjelp av kategoriserte treningsdata. Man deler gjerne inn veiledet læring i to kategorier, regresjon og klassifisering:

- **Regresjon:** I et maskinlæringsperspektiv betyr regresjon at man estimerer en funksjon (f) til å kartlegge fra inndata (x) til kontinuerlige eller nummeriske utdata (y).
- **Klassifisering:** Klassifisering betyr å estimere en funksjon (f) til å kartlegge fra inndata (x) til helt katergoriske utdata (y).

Under disse to kategoriene finnes det mange forskjellige metoder. I denne seksjonen blir metodene brukt i denne oppgaven forklart på et overordnet i nivå.

2.3.1 Lineær regresjon

Lineær regresjon [11] er en av de enkleste og mest brukte tilnærmingene til prediksjon av kontinuerlige variabler. Den antar at det er en lineær relasjon mellom inndataen og utfallet, gitt ved følgende likning:

$$y = b_0 + [b_1, b_2, .., b_i] * [x_1, x_2, .., x_i]$$

Her er b-verdiene koeffisienter for å vekte de ulike inndata-variablene, hvor b_0 beskriver bias i modellen og resten er de individuelle vektingene for hver enkelt variabel.

2.3.2 MLP - Multilayer Perceptron Network

MLP [6] er en enkel variant av kunstige nevrale nettverk, og en underkategori av det som kalles et "feedforward"-nettverk [6]. I denne typen modeller beveger informasjonen seg kun forover gjennom lagene, som vil si at det ikke er noen løkker eller sykluser nettverket. Man kaller gjerne et "feedforward"-nettverk for et multi-layer-nettverk dersom det er flere enn 2 skjulte lag i modellen. Her er en illustrasjon:

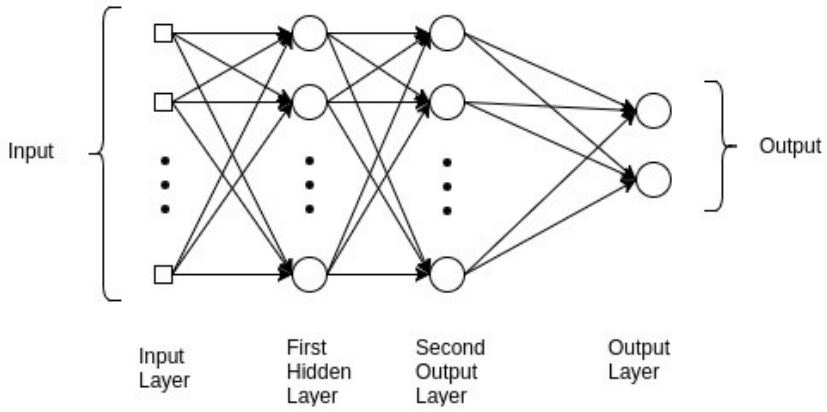

Figure 1: MLP

2.3.3 Random forest

Random forest [13] er en type beslutningstre som både kan brukes både til regresjon og klassifisering. De er satt sammen av flere individuelle trær, som alle blir trent på et gitt antalle tilfeldige deler av datasettet. I et slikt tre har vi en gitt input som sendes inn på toppen, og som deretter traverserer nedover treet mens dataen deles i mindre og mindre deler. Hvert tre vil så gi sin prediksjon, og modellen velger da prediksjonen som er mest valgt. Her er en illustrasjon:

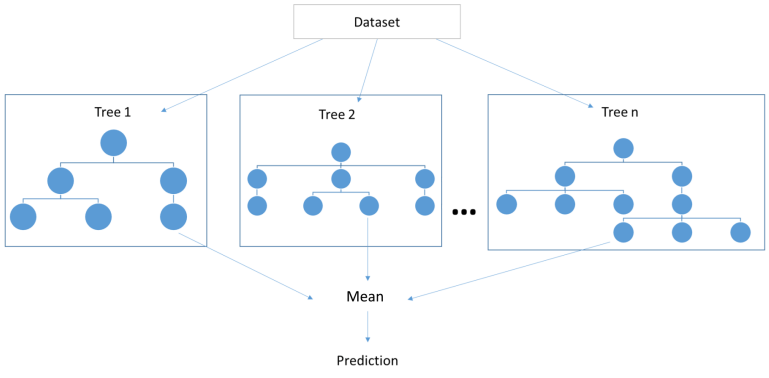

Figure 2: Random Forest

2.4 Tidligere relevant arbeid

Når man skal undersøke tidligere arbeid på dette området er det nyttig å se til flere idretter enn bare fotball, da det er gjort mye god forskning på prediksjon innenfor flere forskjellige ballidretter.

2.4.1 Sannsynlighetsmodeller (1950-2000)

De første statistiske modelleringene og forskningsprosjektene kom først rundt 1950-tallet. I denne perioden helt frem til 1974 bestod undersøkelsene stort sett av sannsynlighetsmodeller, hvor de brukte statistikk fra tidligere kamper til å modellere antall mål for hvert lag i en kamp. Her var spesielt Poisson-fordelingen og negativ binomisk fordeling mye brukt. Den kanskje mest suksessfulle modellen i denne perioden ble utviklet av Dixon og Coles [2], nettopp ved hjelp av en Poisson-fordeling.

2.4.2 Maskinlæring (2000-)

Maskinlæring har trolig vært testet til problemstillingen lenge før 2000-tallet, men det er først her man begynner å finne god forskning på området. Her begynte man å spå utfallet av kampene(seier/uavgjort/tap) direkte ved hjelp av klassifiseringsmodeller, og på den måten unngikk man feilmarginen som kommer av gjensidig avhengighet mellom hvert lags antall mål. Metodene har også blitt mer avanserte frem til idag, og i nyere tid har man spesielt sett gode resultater med regresjonsmodeller og kunstige nevrale nettverk. I en forskningsrapport fra 2015 oppnådde man en nøyaktighet på 54.7% ved hjelp av klassifisering gjennom et kunstig nevralt nettverk [7]. Logistisk regresjon [4], "Support Vector Machines" [4] og Bayesianske nettverk [9] har også vist gode resultater i nyere tid.

3 Metode

3.1 Datainnsamling

I dagens fotball blir det registrert enorme mengder statistikk fra hver kamp, både lagbaserte og spillerbaserte tall. Disse dataene er likevel utfordrende å samle inn, ettersom de største dataleverandørene ikke tilbyr åpent tilgjengelige datasett eller databaser. Det finnes likevel en del nettsider som fremstiller mye god statistikk, og dette er fullt mulig hente ved hjelp av for eksempel webskraping. I dette prosjektet ble det brukt et åpent tilgjengelig datasett fra kaggle.com, et stort websamfunn for datavitenskap og maskinlæring. Datasettet inneholder lagstatistikk fra 268 kamper i 2019/20-sesongen av den engelske toppdivisjonen Premier League, og er satt sammen av data fra flere ulike åpne kilder. Rådataen i sitt originale format er ikke spesielt nyttig i denne problemstillingen. Hver rad i datasettet inneholder tall fra en enkelt kamp, kun sett fra det ene lagets perspektiv. For å enkelt kunne bruke dataen i maskinlæringsmodellene måtte de behandles slik at en rad inneholdte all informasjon fra en kamp, i tillegg til å fjerne alle uønskede og ikke-numeriske verdier. Datastrukturen er illustrert i figur 3.

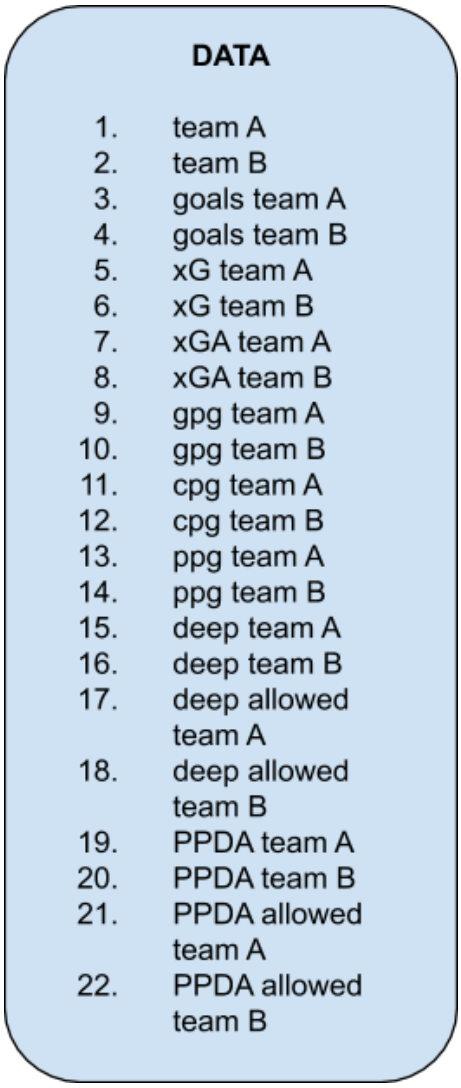

Figure 3: Fullstendig datastruktur

3.2 Valg av modeller

I denne oppgaven ble hovedfokuset på ulike varianter av regresjons- og klassifiseringsmodeller, og da spesielt lineær regresjon og MLP-nettverk. Grunnen til dette var at man har sett gode resultater med disse modellene tidligere, men også fordi de er enkle å implementere. Med lineær regresjon forsøkte man altså å predikere resultater ved å spå antall mål for hvert lag, mens man ved hjelp av MLP-nettverket forsøkte å klassifisere kamputfallet direkte. I tillegg til disse metodene ble det tatt i bruk en enkel variant av "Random Forest"-regresjon som et referansepunkt, da man på et generelt grunnlag ofte kan se bedre resultater med denne modellen enn med for eksempel lineær regresjon.

3.3 Trening av modellene

3.3.1 Datasplitting

Før trening av modellene ble datasettene splittet i trenings- og test-data. Dette er en vanlig maskinlæringsstrategi for å sikre at modellene klarer å prestere på usett data. Ved en slik splitting av datasett var det viktig å finne et passende forhold mellom mengden trenings- og test-data, slik at man kunne få trent modellene på mest mulig data og samtidig ha nok data å evaluere modellene på. Dette var spesielt en utfording når man hadde lite data i utgangspunktet, men valget falt her på 75% treningsdata og de resterende 25% til testing. Det er også vanlig å bruke en metode som kalles "cross-validation"-testing, hvor man deler dataen inn i flere forskjellige trenings- og testsett. Tidligere forskning tydet likevel på at denne metoden ikke nødvendigvis var passende for denne problemstillingen grunnet den tidsmessige strukturen i dataen, og det ble derfor ikke brukt tid på dette i denne oppgaven [7]. I tillegg var det her ønskelig å sørge for at modellene ble evaluert på samme testdata, da et lite datagrunnlag kunne føre til store variasjoner og lite sammenlignbare nøyaktigheter.

3.3.2 Optimalisering

For å optimalisere parameterne i modellene ble det brukt "Stochastic Gradient descent". Dette er en av de vanligste og mest brukte metodene for å optimalisere både regresjons- og klassifiseringsmodeller, og dessuten veldig enkel å implementere [10]. Optimaliseringen forutsetter at det beregnes avvik mellom prediksjoner og virkelige resultater for hver iterasjon, noe som kalles "loss" i modellen. Metoden forsøker å finne de parameterne som minimerer "loss", gjennom å justere dem for hver iterasjon. Det finnes flere måter å berenge "loss", og de ulike variantene egner seg gjerne til ulike problemstillinger [8]. Her er algoritmene som ble valgt i denne oppgaven:

- **Optimalisering:** Stochastic gradient descent
- **Loss:** Mean Squared Error(regresjon), Cross Entropy Loss(MLP)

3.4 Valg av datapunkter

Når man skal trene prediktive maskinlæringsmodeller er det viktig å tenke nøye gjennom hvor mange og hvilke datapunkter man bruker. Dersom man jobber med store komplekse datasett vil for mange punkter kunne påvirke både tidsbruken og nøyaktigheten ved trening negativt [5]. Ofte vil også datasett bestå av informasjon som ikke har sterke sammenhenger til resultatet man ønsker å forutse, noe som kan skape forstyrrende støy for modellene. Selv om det i dette prosjektet ble brukt et lite datasett vil det være svært relevant dersom man skulle gjort de samme forsøkene i større skala. Hva man bør ta hensyn til og hvordan man velger ut data varierer ut ifra hvilke modeller man jobber med. Som nevnt tidligere var hovedfokuset i dette prosjektet lineær regresjon og nevrale nettverk. For å tydelig undersøke påvirkningen av ulike datapunker på resultatene, ble det for begge gjort tre ulike faser av testing:

- **Avansert datasett:** I første fase ble modellene testet med all relevant data som var tilgjengelig i det behandlede datasettet. Dette var totalt 18 datapunkter. Se figur 4.

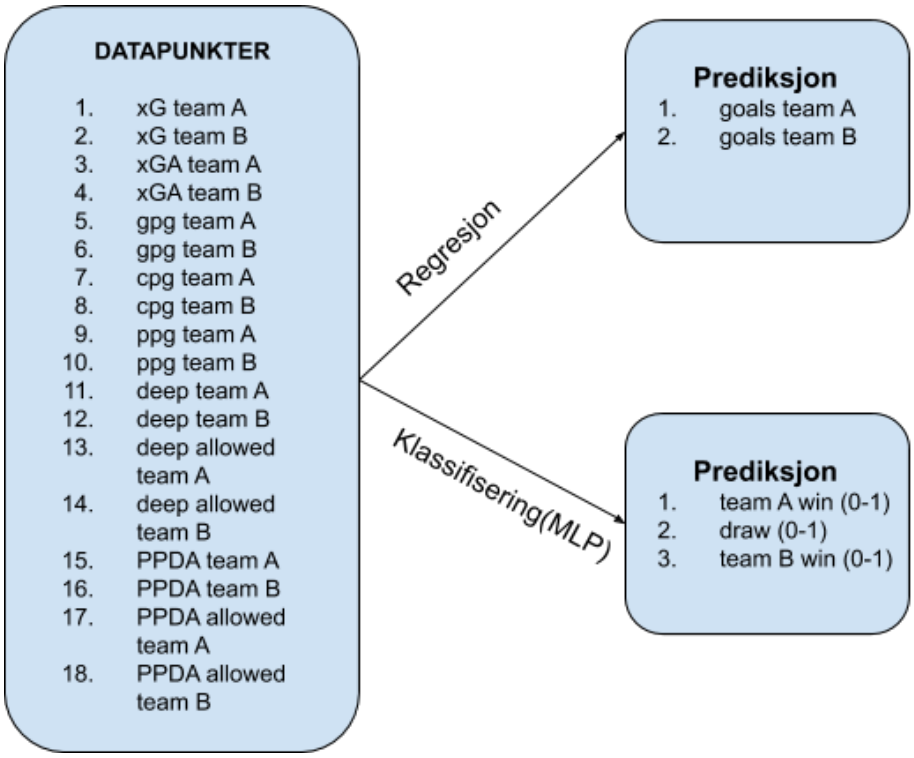

Figure 4: Modell med avansert datasett

- **Simpelt datasett:** I neste fase ble modellene testet med et enkelt utvalg av data beskrivende for lagenes offensive og defensive slagkraft. Disse ble ikke valgt helt tilfeldig, men det ble ikke gjort noen grundig analyse på forhånd. Disse modellene bestod av 10 datapunkter. Se figur 5.

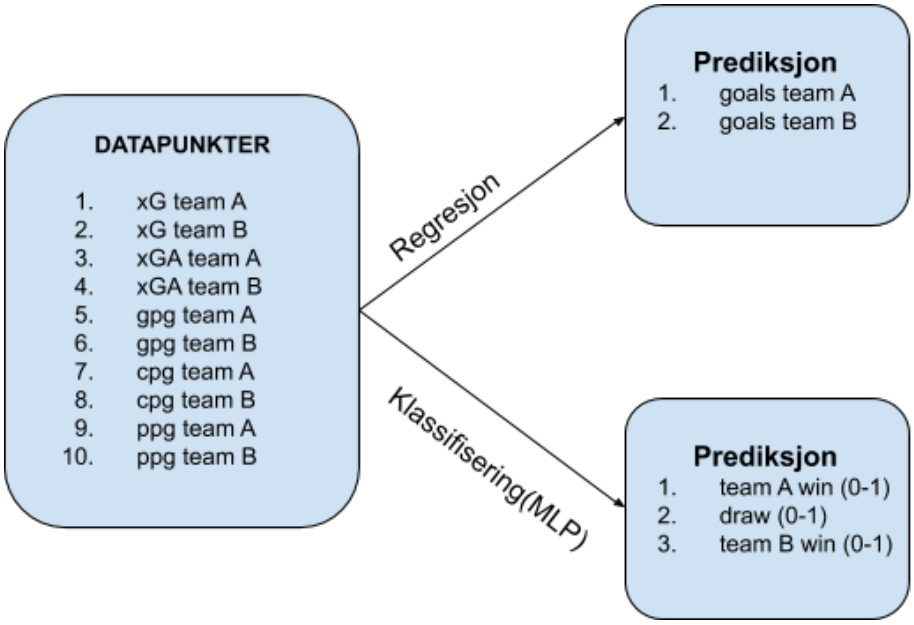

Figure 5: Modell med simpelt datasett

- **Optimalisert datasett:** I siste fase ble det gjort grundigere forarbeid. Ved å først analysere viktigheten av datapunktene, kunne man velge ut kun de punktene som så ut til å styrke resultatene. For lineær regresjon ble dette gjort ved å studere hvordan modellen vektet hvert enkelt punkt og fjerne de som hadde negativ eller lav innvirkning. Fremgangsmåten var lignende for klassifisering, men her ble vektingene undersøkt ved hjelp av "permutation importance". Resultatene av disse analysene kan ses i kapittel 4.

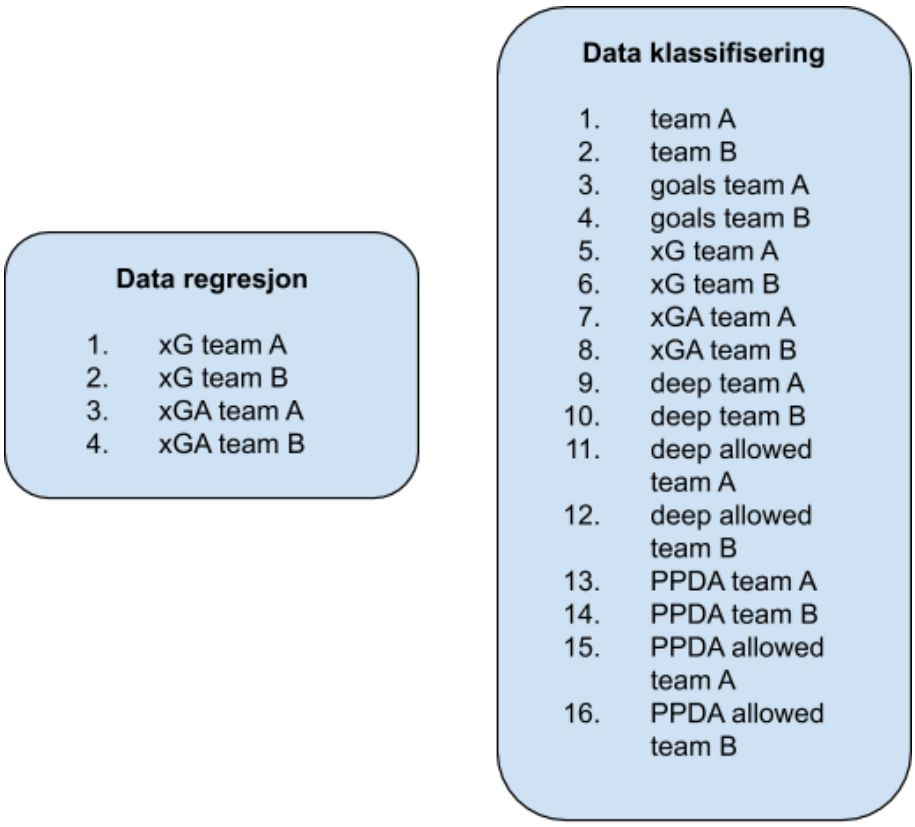

Figure 6: Modeller med optimaliserte datasett

3.5 Evaluering

Modellene ble hovedsaklig evaluert ved nøyaktighet i prediksjonene på den usette testdataen. Altså i hvor stor prosentandel av disse kampene ble det spådd riktig utfall. For regresjonsmodellene var det ikke et krav at modellen måtte gjette riktig antall scoringer, men prediksjonen ble betraktet som riktig dersom kamputfallet var likt(seier/uavgjort/tap). På grunn av relativt lite testdata, ble nøyaktigheten betraktet sammen med loss for å gi en sikrere evaluering. Dersom en modell skulle oppnå høy nøyaktighet på testdataen, men likevel ha stigende loss, ble evalueringen betraktet som usikker. Dette også dersom det skulle oppstå stort avvik mellom nøyaktighet på treningsdata og testdata, noe som kan indikere at man enten overtilpasser eller undertilpasser modellen. Som tidligere nevnt var det også ønskelig å sammenligne modellenes prestasjon med noen enkle referansepunkter, altså Random Forest og odds-prediksjoner.

4 Resultater

I denne seksjonen oppsummeres resultatene som ble observert i oppgaven. Observasjonene fra hver enkelt modell blir presentert, og alle resultater blir også fremstilt samlet med referansepunktene.

4.1 Lineær regresjon

4.1.1 Avansert modell

Nøyaktighet: 57,25%

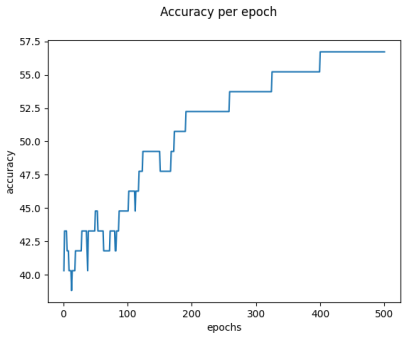

Figure 7: Nøyaktighet under trening

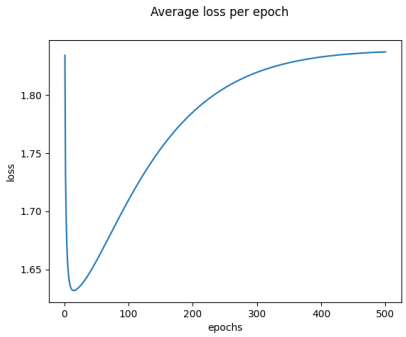

Figure 8: Loss under trening

Den avanserte modellen ble altså trent med alle datapunktene som var tilgjengelig. Dette ga nokså interessant en nøyaktighet på hele 57,25%, samditig som nøyaktigheten steg i takt med loss. Et lovende resultat, men også veldig vanskelig å vite hvor god modellen faktisk var ut ifra dette.

4.1.2 Enkel modell

Nøyaktighet: 56,72%

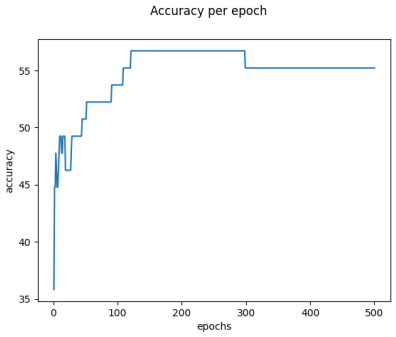

Figure 9: Nøyaktighet under trening

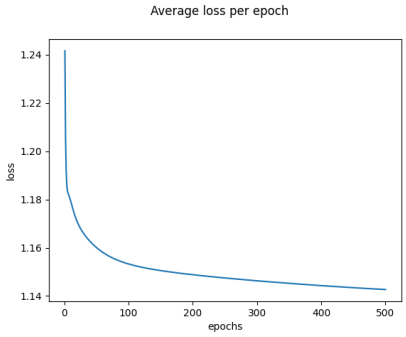

Figure 10: Loss under trening

I den enkle modellen ble det som nevnt valgt ut noen datapunkter som man antok burde ha en sammenheng med faktisk resultat. Denne modellen oppnådde også høy nøyaktighet på 56,72%, men i dette tilfellet mye mer lovende siden vi ser at loss i tillegg konvergerte mot minimum. Nøyaktigheten så også ut til å stabilisere seg underveis i treningen.

4.1.3 Optimalisert modell

Nøyaktighet: 52,25%

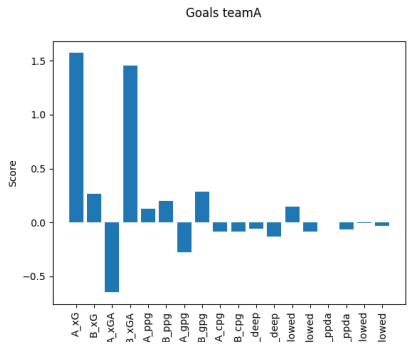

Figure 11: Vekting av data lag A

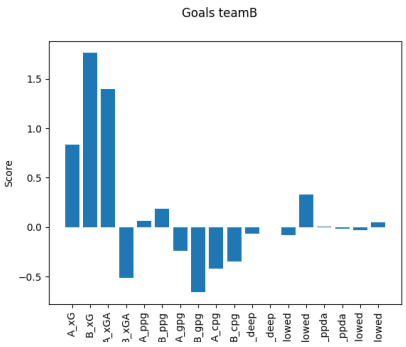

Figure 12: Vekting av data lag B

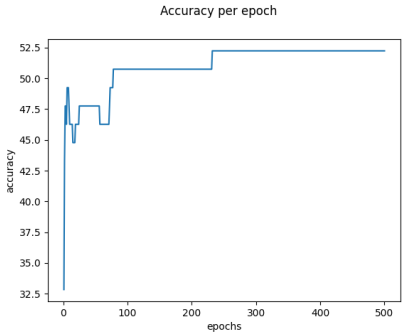

Figure 13: Nøyaktighet under trening

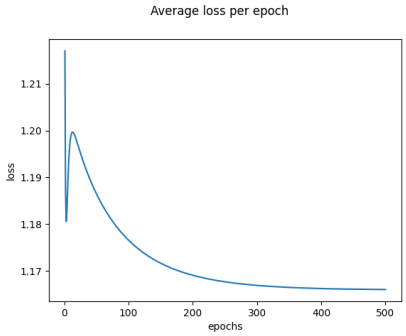

Figure 14: Loss under trening

Den optimaliserte modellen bestod kun av datapunkter som virket lovende ut ifra analysene av datapunktene i figur 11 og 12. Likevel så vi en ganske betydelig lavere nøyaktighet enn hos de to andre modellene, selv om 52,25% også er et lovende resultat. Det at loss konvergerte mot et minimum tyder i hvert fall på at modellen kan være troverdig.

4.2 MLP-klassifisering

4.2.1 Avansert modell

Nøyaktighet: 52,25%

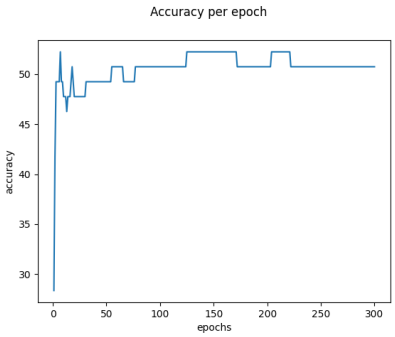

Figure 15: Nøyaktighet under trening

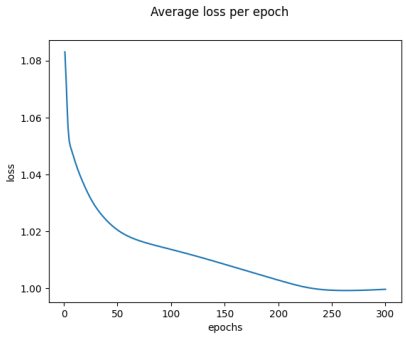

Figure 16: Loss under trening

Den avanserte modellen oppnådde med mlp-klassifisering en nøyaktighet på 52,25%. Man kan se at loss konvergerte mot et minimum, mens nøyaktigheten stabiliserte seg etterhvert.

4.2.2 Enkel modell

Nøyaktighet: 52,25%

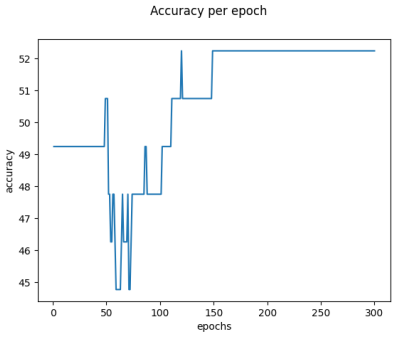

Figure 17: Nøyaktighet under trening

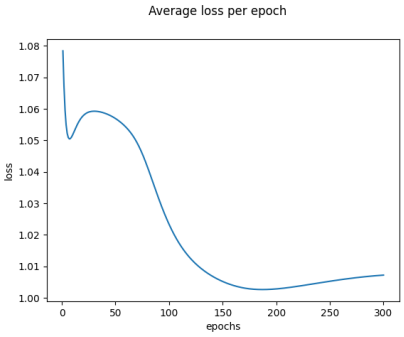

Figure 18: Loss under trening

Den enkle modellen klarte samme nøyaktighet som den avanserte modellen, altså 52,25%. Nøyaktigheten stabiliserte seg helt flatt halvveis i treningen, og vi ser også at det er omtrent på dette punktet loss nådde sitt minimum.

4.2.3 Optimalisert modell

Nøyaktighet: 55,22%

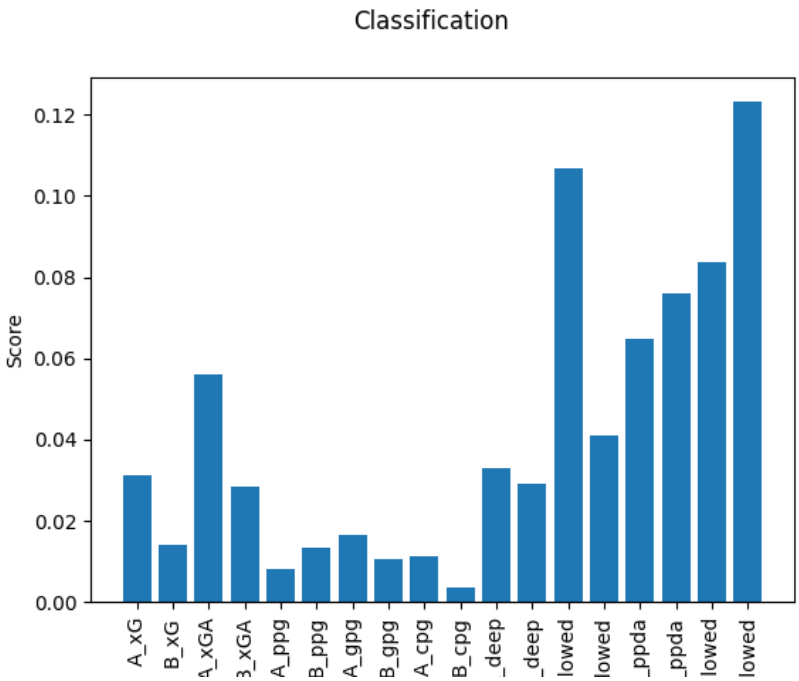

Figure 19: Vekting av data

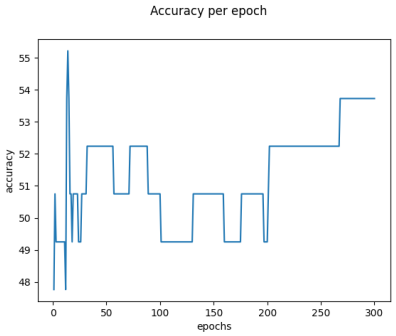

Figure 20: Nøyaktighet under trening

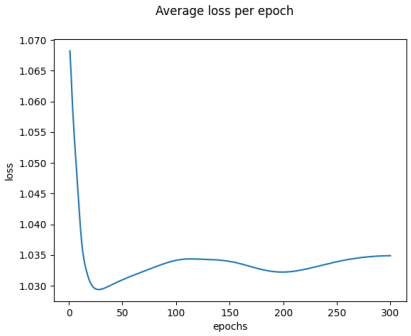

Figure 21: Loss under trening

Den optimaliserte modellen for MLP-klassifiseringen hadde veldig ulike datapunkter enn for regresjon. Analysene i figur 19 indikerte at de fleste datapunktene kunne fungere bra i modellen, så få av dem ble faktisk valgt bort. Vi ser at loss var på sitt laveste og nøyaktigheten på sitt desidert høyeste omtrent samtidig. Den høyeste nøyaktigheten som ble observert var 55,25%, altså ble det den MLP-modellen med best resultat.

4.3 Sammenligning med referansepunkter

Nedenfor gis en full oversikt over alle modellenes prestasjoner.

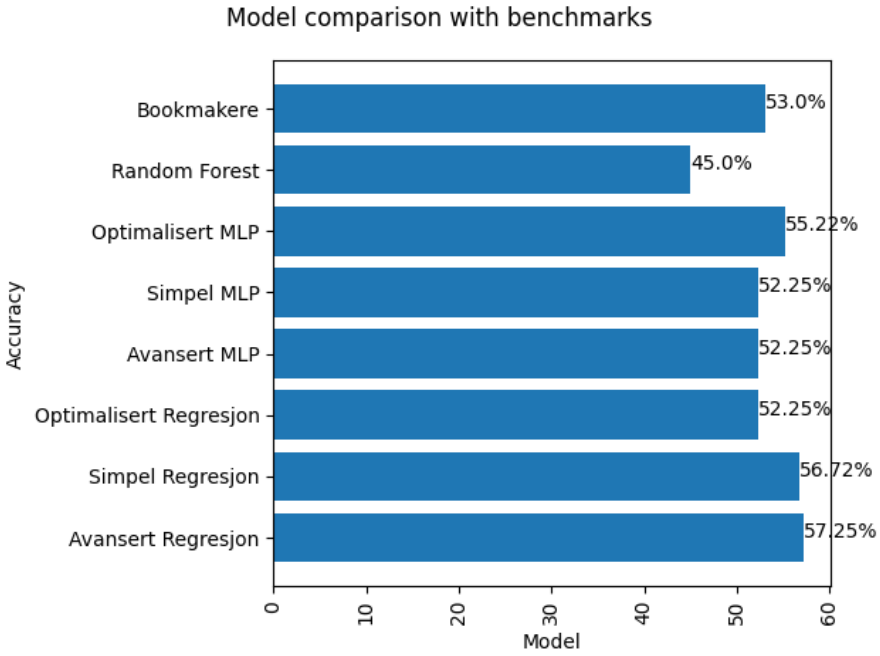

Figure 22: Sammenligninger med referansepunkter

Resultatet ble altså at tre av modellene oppnådde høyere nøyaktighet enn odds-prediksjonene fra bookmakerne, mens samtlige presterte bedre enn Random Forest som kom klart dårligst ut. Aller best nøyaktighet klarte regresjonsmodellen hvor alle datapunkter ble benyttet, dette til tross for at loss ikke konvergerte. De lineære regresjonsmodellene klarte seg også generelt bedre en MLP-modellene.

5 Diskusjon

5.1 Oppsummering

Målet med oppgaven var å undersøke hvor godt maskinlæring egnet seg til å forutse fotballresultater basert på tidligere kampdata, og da spesielt ved regresjon og klassifisering. Ved hjelp av "stochastic gradient descent" lyktes det å trene tre ulike modeller av både lineær regresjon og MLP-klassifisering, hvor alle modellene oppnådde over 50% nøyaktighet. I tillegg ble det altså hentet inn odds-prediksjoner og en enkel variant av Random Forest som referansepunkter for modellene. Selv med meget lite datagrunnlag leverte alle modellene lovende resultater, og hele tre av dem klarte leverte bedre enn odds-spådommene fra bookmakerne. Tar man i betraktning begrensninger på tid og ressurser er dette en god indikasjon på at maskinlæring kan være et meget nyttig verktøy for problemstillingen.

5.2 Utfordringer

Selv om resultatene var lovende støtte man likevel på noen utfordringer underveis, hvor noen av disse kan skape noe usikkerhet rundt resultatene.

5.2.1 Innhenting av data

Som tidligere nevnt var det meget utfordrende å samle inn store mengder data av god kvalitet, noe som vanskeliggjør trening og evaluering av maskinlæringsmodeller [1]. Et datasett på kun 268 kamper er altså langt i fra optimalt. Spesielt nevrale nettverk har vist seg å prestere betydelig bedre med større mengder data, noe som kan være en forklaring på hvorfor de presterte svakere enn regresjonsmodellene i dette tilfellet.

5.2.2 Analyse av datapunkter

Det ble forsøkt å analysere viktigheten av hvert enkelt datapunkt i datasettet, for å se om man kunne optimalisere modellene med å kun ta med sterke sammenhenger. For lineær regresjon ble det sett på vektingene av hvert enkelt datapunkt med en veldig simpel implementasjon av metoden. Problemet er at vi utfører regresjon med to selvstendige utfall, nemlig antall mål for de to lagene. De datapunktene som har en sterk sammenheng til antall mål for det ene laget, vil ofte ha en svak sammenheng til det andre laget. Dette kunne man også se av analysen, noe som gjør det vanskelig å velge ut hvilke datapunkter som faktisk vil styrke det samlede resultatet. Man ser også at den modellen som burde vært optimalisert faktisk presterte svakere enn de to andre. Her ville en mulig løsning vært å kjøre to ulike regresjoner for å spå hvert lags antall mål, helt uanvhengig av hverandre.

Dette problemet ser man blir unngått for klassifiseringen, da alle datapunktene blir vektet etter deres innvirkning på det endelige resultatet. Denne modellen ble også den beste av klassifiseringsmodellene.

5.2.3 Evaluering

Evalueringen av modellene ble gjort med enkle målinger som nøyaktighet og loss. Når man hadde såpass lite datagrunnlag som i dette tilfellet, så skapte det en del problemer hva gjelder sammenligningsgrunnlag og troverdighet. I et datasett med såpass få kamper var det relativt sannsynlig at to ulike modeller kunne oppnå samme antall riktige prediksjoner, og da kunne det blitt vanskelig å avgjøre hvilken modell som faktisk er best. Fotball er også en idrett som består av mye tilfeldigheter, og det vil alltid være tilfeller hvor utfallet av en kamp ikke har sterk sammenheng til den underliggende statistikken. Når man da hadde få kamper å både trene og teste på, ble det vanskelig for modellene å avgjøre hvilke kamper som er av mindre betydning. Om datasettet også hadde stor andel av resultater som kunne regnes som usannsynlige, så er det ikke sikkert modellene ville prestert like bra på mer balansert usett data. Med litt mer tid ville det nok her vært ønskelig å implementere minst en ekstra måling, som for eksempel "F1-score" eller "Precision". Disse tar også hensyn til falske positive og falske negative prediksjoner, som kunne gitt en enda mer troverdig evaluering av modellene.

5.3 Fremtidig arbeid

Det er liten tvil om at det er rom for utvidelser og forbedringer, selv med lovende resultater på veldig kort tid.

5.3.1 Finne mer data

Den store begrensingen i denne oppgaven var uten tvil mangelen på data. Her hadde det nok vært mye å hente dersom man enten hadde kjøpt tilgang til databaser fra store aktører, eller tatt seg tid til å hente ut mer data manuelt. Dette vil trolig kunne bedret både prestasjonen til modellene, i tillegg til at det som det som nevnt ville styrket troverdigheten til resultatene betraktelig.

5.3.2 Optimalisere trening

Treningen av modellene ble gjort med relativt enkle metoder, og selv om de burde vært godt egnet til problemstillingen kunne det vært nyttig å utforske enda flere. "Cross-validation"-testing kunne som nevnt tidligere vært en mulig implementasjon for å styrke modellenes prestasjon på usett data. Forutsett at man også hadde samlet inn mer data kunne det vært aktuelt å splitte den i flere "buckets" isteden for å behandle hele datasettet samtidig.

Interessant kunne det også vært å trene modellene med flere forskjellige optimaliseringsalgoritmer. I denne oppgaven ble det kun brukt Stochastic Gradient Descent, men det finnes en rekke flere algoritmer som hadde vært interessant å teste. Loss-funksjonene ble valgt ut i fra tidligere erfart "best practice", men det kunne likevel også her vært nyttig å teste flere varianter. På den måten kunne man selv tatt en avgjørelse på hva som var best egnet for akkurat denne problemstillingen.

5.3.3 Andre metoder

Det finnes flere maskinlæringsmetoder som burde vært meget godt egnet til oppgaven. Det er som nevnt i kapittel 2.4 ikke bare med lineær regresjon og nevrale nettverk det tidligere var sett gode resultater, og det ville vært meget interessant å implementere flere modeller i fremtiden.

6 Conclusion

Basert på resultatene i denne oppgaven ser maskinlæring ut til å kunne ha stort potensiale innen prediksjon av fotballkamper, og med lite data og relativt enkle metoder oppnådde man på kort tid gode resultater. Flere av modellene klarte også målet om bedre nøyaktighet enn bookmakerne, selv om det var de lineære regresjonsmodellene kom aller best ut.

Noe usikkerhet var det likevel, og det ville vært aktuelt å utvide dette prosjektet ytterligere for å kunne styrke resultatet og konklusjonene. Det er var for eksempel vanskelig å slå fast hvilke datapunkter som hadde sterkest sammenhenger med resultatet, selv om xG-målingene så ut til å ha ganske sterk vekting i begge metodene. I MLP-modellene ble det også indikert at lagenes "PPDA" og "deep progressions" kan være av stor betydning, men ytterligere undersøkelser ville vært nødvendig for å kunne dra endelige konklusjoner.

Referanser

[1] Jason Brownlee. *How Much Training Data is Required for Machine Learning?* URL: https://machinelearningmastery.com/much-training-data-required-machine-learning/.

[2] Mark J. Dixon Stuart G. Coles. *Modelling Association Football Scores and Inefficiencies in the Football Betting Market.* URL: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/1467-9876.00065.

[3] Richard C. Giulianotti. *Football.* URL: https://www.britannica.com/sports/football-soccer.

[4] Babak Hamadani. *Predicting the outcome of NFL games using machine learning.* URL: http://cs229.stanford.edu/proj2006/BabakHamadani-PredictingNFLGames.pdf.

[5] SAURAV KAUSHIK. *Introduction to Feature Selection methods with an example (or how to select the right variables?)* URL: https://www.analyticsvidhya.com/blog/2016/12/introduction-to-feature-selection-methods-with-an-example-or-how-to-select-the-right-variables/#:~:text=Top%20reasons%20to%20use%20feature,the%20right%20subset%20is%20chosen..

[6] John McGonagle. *Feedforward Neural Networks.* URL: https://brilliant.org/wiki/feedforward-neural-networks/.

[7] Y. Joustra N. Tax. *Predicting The Dutch Football Competition Using Public Data: A Machine Learning Approach.* URL: https://www.researchgate.net/profile/Niek_Tax/publication/282026611_Predicting_The_Dutch_Football_Competition_Using_Public_Data_A_Machine_Learning_Approach/links/5601a25108aeb30ba7355371/Predicting-The-Dutch-Football-Competition-Using-Public-Data-A-Machine-Learning-Approach.pdf.

[8] Ravindra Parmar. *Common Loss functions in machine learning.* URL: https://towardsdatascience.com/common-loss-functions-in-machine-learning-46af0ffc4d23.

[9] Nazim Razali. *Predicting Football Matches Results using Bayesian Networks for English Premier League (EPL).* URL: https://iopscience.iop.org/article/10.1088/1757-899X/226/1/012099/pdf.

[10] Sebastian Ruder. *An overview of gradient descent optimization algorithms.* URL: https://arxiv.org/pdf/1609.04747.pdf.

[11] Dennis J. Sweeney. *Statistics.* URL: https://www.britannica.com/science/statistics/Experimental-design#ref367488.

[12] Mark Ryan M. Talabis. *Supervised Learning.* URL: https://www.sciencedirect.com/topics/computer-science/supervised-learning.

[13] Evaldas Vaiciukynas. *Architecture of the Random Forest model.* URL: https://www.researchgate.net/%20figure/Architecture-of-the-random-forest-model_fig1_3016386435.

Appendix

A Eksempelresultater fra lineær regresjon

	teamA	teamB	predictedScore	actualScore	correct
0	Aston Villa	Brighton	0-1	1-1	X
1	Crystal Palace	Man City	0-2	2-2	X
2	Bournemouth	Norwich	1-1	0-1	X
3	Sheffield United	Arsenal	1-1	1-1	V
4	Wolves	Southampton	1-0	3-2	V
5	Chelsea	Newcastle Unit	2-0	0-1	X
6	Leicester	Burnley	1-0	1-2	X
7	Man Utd	Liverpool	1-1	0-2	X
8	Brighton	Bournemouth	1-1	1-3	X
9	Newcastle United	Everton	0-1	2-2	X
10	Man City	Sheffield Unitec	2-0	1-0	V
11	Southampton	Crystal Palace	1-0	2-0	V
12	Watford	Aston Villa	1-1	1-2	X
13	Arsenal	Chelsea	0-1	2-2	X
14	West Ham	Leicester	0-1	1-4	V
15	Norwich	Tottenham	0-1	1-2	V
16	Burnley	Man Utd	0-1	2-0	X
17	Liverpool	Wolves	1-0	2-1	V
18	Liverpool	West Ham	2-0	2-0	V
19	Chelsea	Leicester	1-1	2-2	V
20	Southampton	Liverpool	0-1	0-4	V
21	Brighton	West Ham	1-1	3-3	V
22	Aston Villa	Bournemouth	0-1	1-2	V
23	Sheffield United	Crystal Palace	1-0	1-0	V
24	Norwich	Newcastle Unit	0-0	0-0	V
25	Everton	Watford	1-0	3-2	V
26	Wolves	Man Utd	1-1	0-0	V
27	Arsenal	Burnley	1-1	0-0	V
28	Man City	Tottenham	2-0	0-2	X
29	Crystal Palace	Everton	0-1	1-3	V
30	Watford	Brighton	1-1	1-1	V
31	Bournemouth	Sheffield Unitec	0-1	1-2	V
32	Leicester	Wolves	1-1	0-0	V
33	Burnley	Southampton	1-0	2-1	V
34	Liverpool	Norwich	2-0	1-0	V
35	Tottenham	Aston Villa	1-0	3-2	V

Figure 23: Utvalg av resultater