

BACHELORPROSJEKT vår 2022

AI-BASERT ANALYSE

av tidsseriedata fra atleter

FORZASYS
OSLOMET

SKREVET AV:

Bernadette Fanni Finheim

Hanna Bækken Nilsen

Tonje Martine Lorgen Kirkholt

Helene Birkeflet Prescott

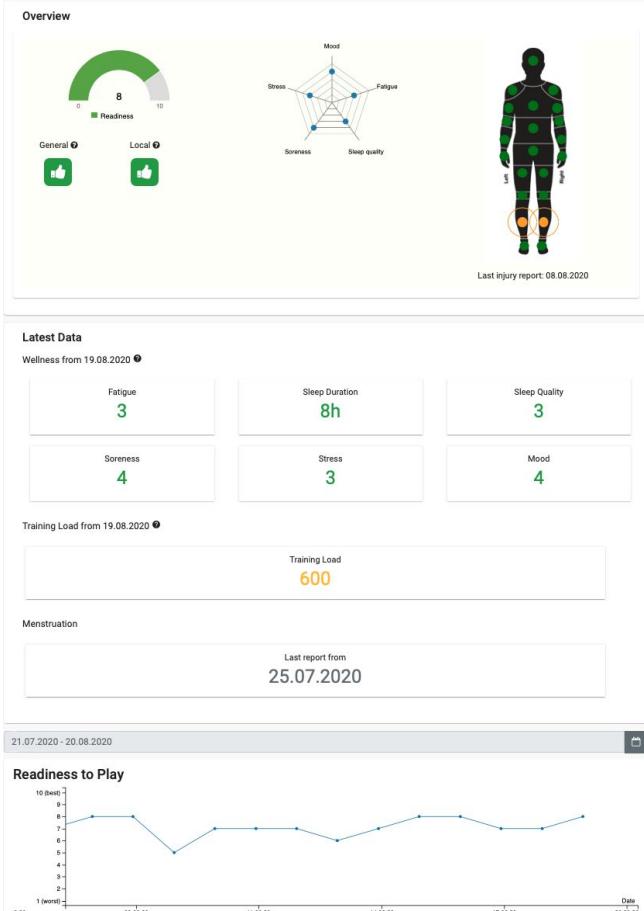


Agenda

1. Introduksjon
2. Arbeidsmetoder
3. Maskinlæringspipeline
4. Analyse og bearbeiding av data
5. Testing og trening av maskinlæringsmodellene
6. Convolutional Neural Network
7. Prediksjon av skade
8. Diskusjon og refleksjon
9. Konklusjon

Introduksjon

- ❖ Oppdragsgiver
 - *Forzasys*
- ❖ Bakgrunn og motivasjon for oppgaven
 - *Utvikling av applikasjonen pmSys*
 - *Tilby mer informasjon basert på innsamlet data*
 - *Forenkle planlegging av treningsbelastning*
 - *Bidra til forebygging av skade*



Introduksjon

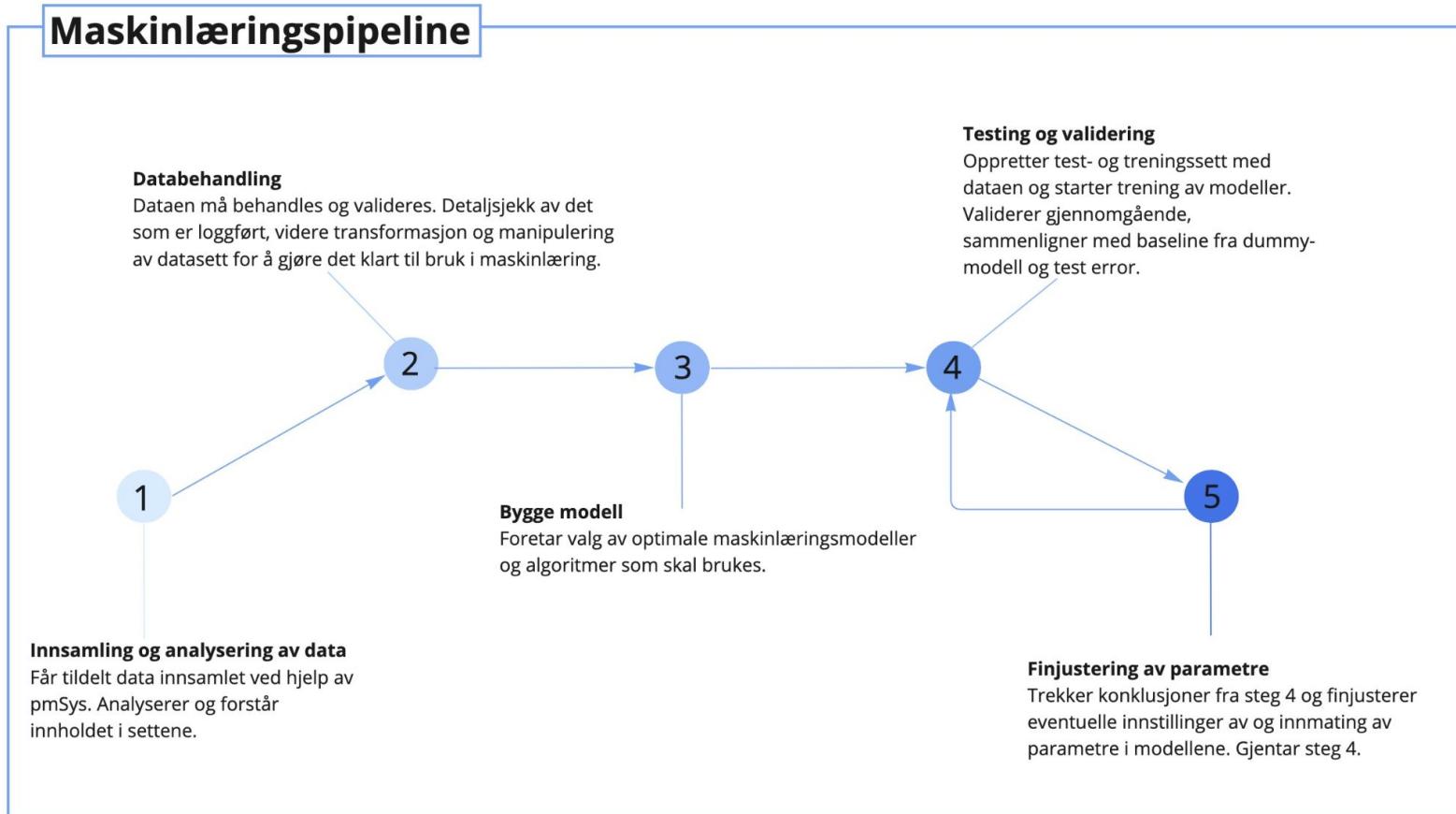
- ❖ Problemstilling
 - *Prediksjon av Readiness to Play ved hjelp av maskinlæring*
 - *Gir oss viktig informasjon om en spillers opplevde dagsform*
 - *Kan potensielt bidra til optimal treningsplanlegging*
 - *Forskningsarbeid som ikke resulterer ikke i et ferdig produkt*
- ❖ ***Målet med prosjektet er å bidra til løsninger som kan danne et grunnlag for arbeid inn i det overordnede målet; prediksjon av skade***

Arbeidsmetoder

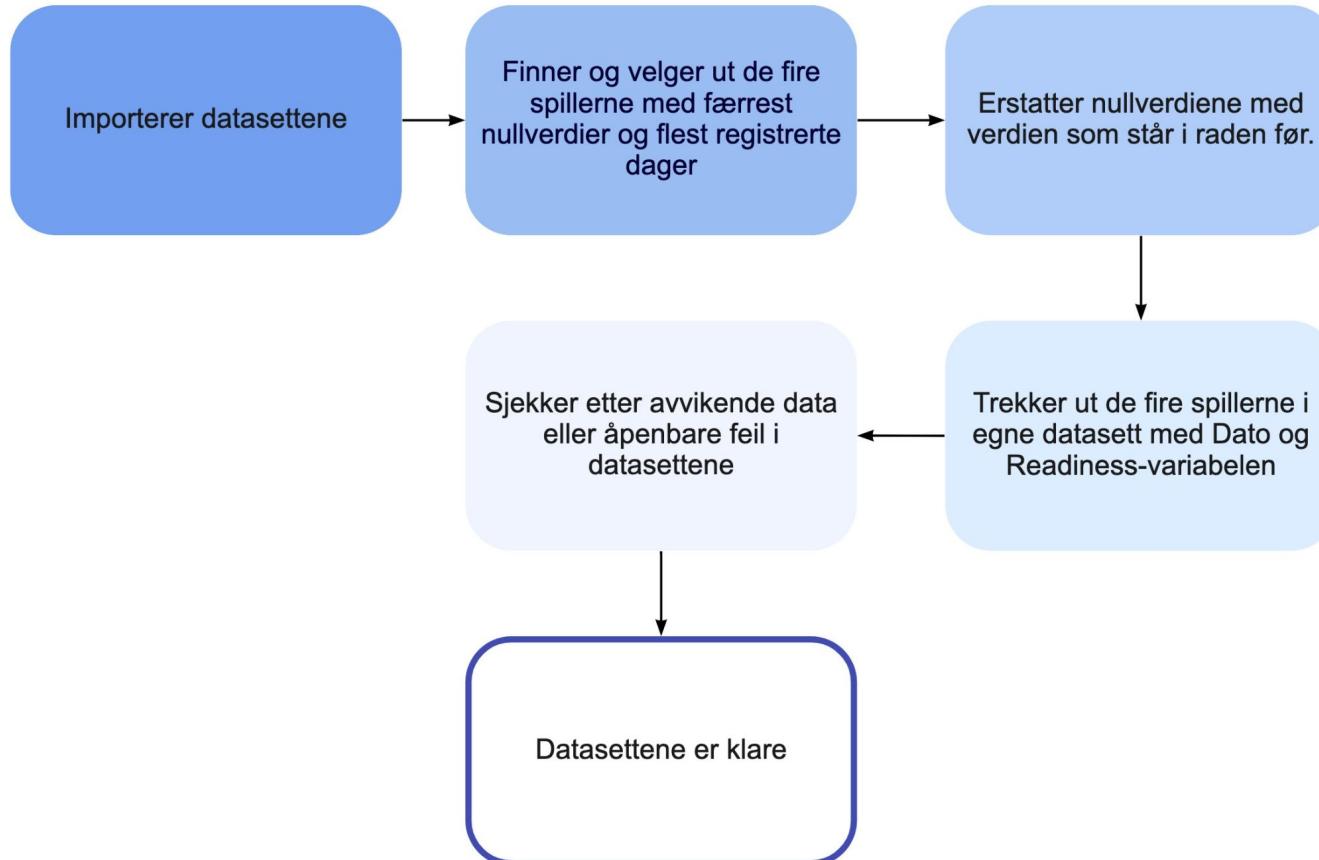
- ❖ Teamkontrakt
- ❖ Arbeidsmetodikk
 - *Scrum*
- ❖ Faglige forutsetninger
 - *Innledning fra introduksjons til AI, høsten 2021*
- ❖ Rammebetingelser
 - *Fritt valg av maskinlæringsmodeller*
 - *Datasett skal ikke publiseres med oppgaven*

Sprint 2		
Uke 6-7		
Oppgaver	Under arbeid	Utført
Oppnå forståelse for ulike rammeverk	Ser på Scikit learn-biblioteket Undersøke på ulike kodeeksempler	Forstår Pandas og numpy
Bearbeidedatasett	Behandling av nullverdier Anonymisere spillere	Lagde egne datasett for utvalgte spillere
Se etter korrelasjon i parametrerne	Finn sammenhenger mellom registrerte Readiness-verdier og resten av parametrerne	Fremstil korrelasjonsmatrise

Maskinlæringspipeline



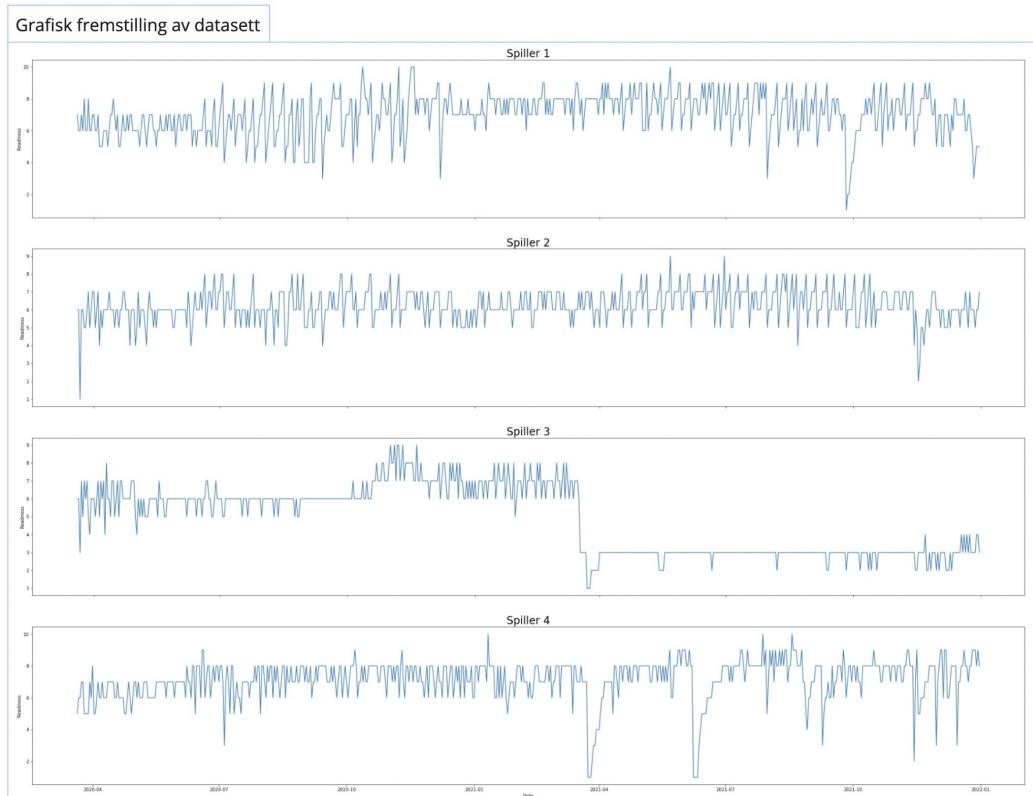
Analyse og bearbeiding av data



Innsamling og bearbeiding av data

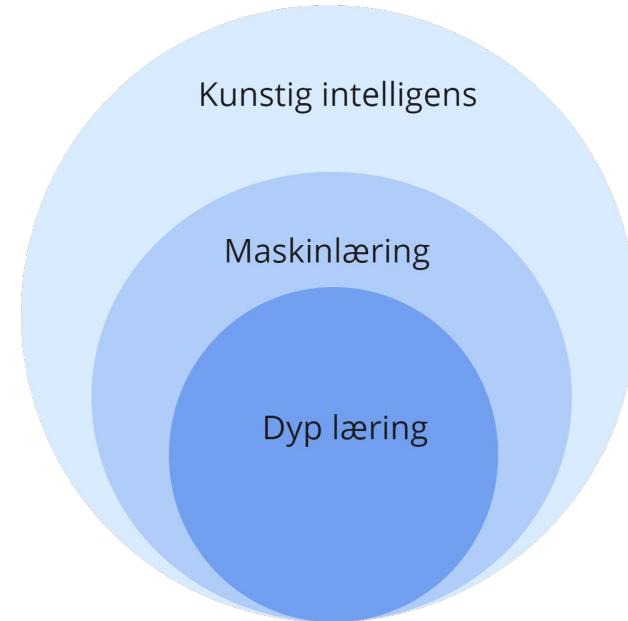
	Spiller1	Spiller2	Spiller3	Spiller4
Date				
2020-03-21	6.0	6.0	6.0	5.0
2020-03-22	6.0	1.0	3.0	6.0
2020-03-23	7.0	6.0	7.0	6.0
2020-03-24	6.0	6.0	5.0	7.0
2020-03-25	8.0	5.0	7.0	7.0
...
2021-12-27	3.0	6.0	3.0	9.0
2021-12-28	4.0	5.0	3.0	9.0
2021-12-29	5.0	6.0	4.0	8.0
2021-12-30	5.0	6.0	4.0	9.0
2021-12-31	5.0	7.0	3.0	8.0

651 rows × 4 columns



Testing og trening av maskinlæringsmodeller

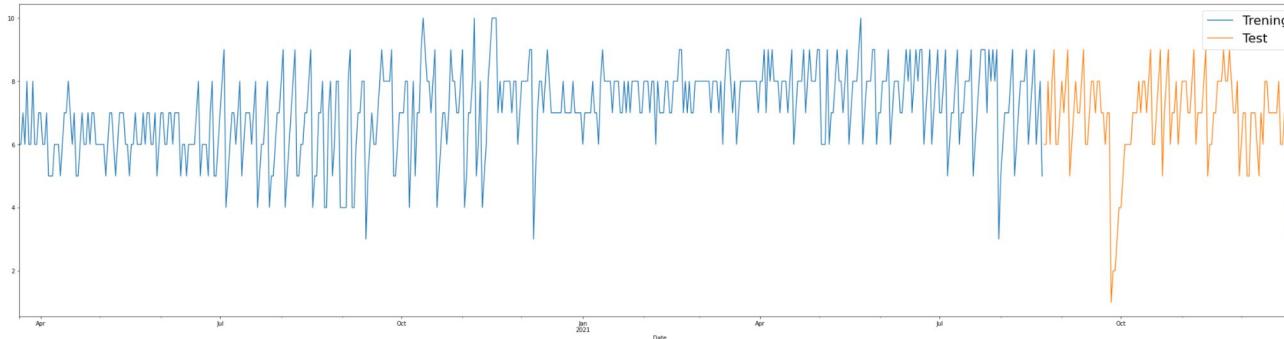
- ❖ Random Forest (RF)
 - *Ensemble modell: Decision Trees*
 - *Tradisjonell maskinlæring*
- ❖ Autoregressive Integrated Moving Average (ARIMA)
 - *Analyse av tidsserier*
 - *Tradisjonell maskinlæring*
 - *Ensemble modell: 3 ulike komponenter*
- ❖ Long Short-Term Memory (LSTM)
 - *Dyp læring*
 - *Kjent for sitt "minne"*



Testing og trening av maskinlæringsmodellene

Utgangspunkt

- **Fire datasett:** Spiller1, Spiller2, Spiller3 og Spiller4
- **651 verdier** fra tidsperioden 21.03.2020 til 31.12.2021
- Fordeling av trenings- og testsett satt til 80% og 20%
- **Treningssett: 520 verdier (21.03.2020 til 23.08.2021)**
- **Testsett: 131 verdier (23.08.2021 til 31.12.2021)**
- **Lagverdi satt til 3;** tre dager brukes til å predikere den neste



Testing og trening av maskinlæringsmodellene

Test-error

Varians / Mean squared error (MSE)

$$\frac{1}{N} \sum_{i=1}^N (y_i - \mu)^2$$

Varians / Mean absolute error (MAE)

$$\frac{1}{N} \sum_{i=1}^N |y_i - \mu|$$

R-Squared (R^2)

$$R^2 = 1 - \frac{\sum_{i=1}^N (\hat{y}_i - y_i)^2}{\sum_{i=1}^N (y_i - \bar{y})^2}$$

Baseline

Baseline for testscore - Spiller1:

MAE: 1.0992660011743982

MSE: 2.071485952391707

R²: -0.0015995274708127116

Baseline for testscore - Spiller2:

MAE: 0.7145038167938931

MSE: 1.031584082388545

R²: -0.0022086977960720233

Baseline for testscore - Spiller3:

MAE: 1.731664709336465

MSE: 3.505341789376214

R²: -0.006781095355401057

Baseline for testscore - Spiller4:

MAE: 0.841441573693482

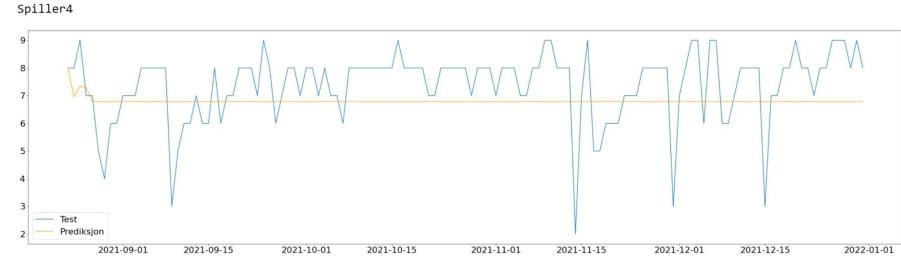
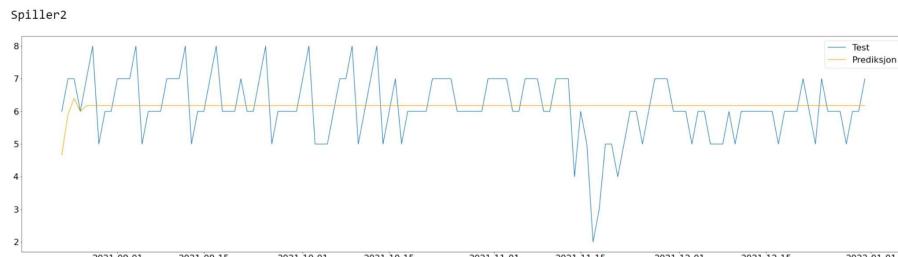
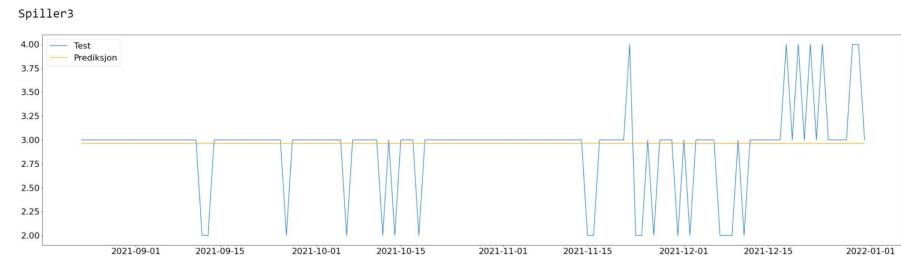
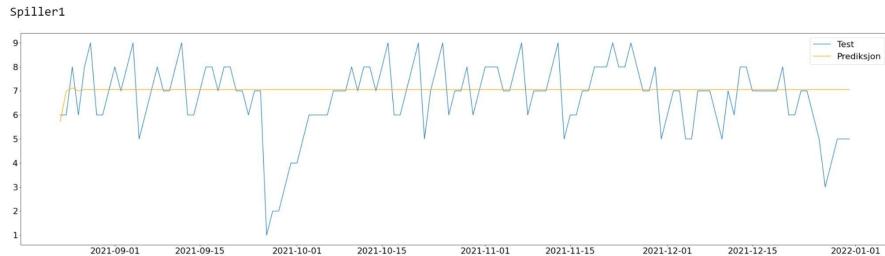
MSE: 1.1025241372690726

R²: -0.012220025662024403

Testing og trening av maskinlæringsmodellene

- ❖ Random Forest Regression
- ❖ Resultat av eksperimentene

Før hyperparameter tuning



Testing og trening av maskinlæringsmodellene

- ❖ Random Forest Regression
- ❖ Forsøk på forbedring av test score
 - *Hypertuning*

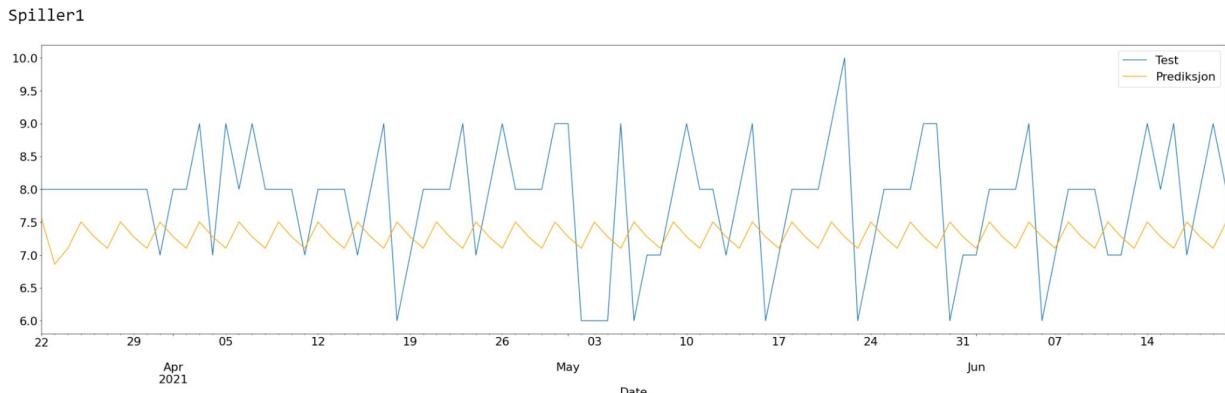
```
Spiller1:  
MSE test error for prediksjon: 2.2419845907397637  
MSE test error for prediksjon med hyper tuning: 2.418658440986441  
R2 score test error for prediksjon: -0.035379374641686834  
  
Spiller2:  
MSE test error for prediksjon: 0.9491093044203369  
MSE test error for prediksjon med hyper tuning: 0.92411795670246  
R2 score test error for prediksjon: -0.017089095363894025  
  
Spiller3:  
MSE test error for prediksjon: 0.18593324298842032  
MSE test error for prediksjon med hyper tuning: 0.18589706896368038  
R2 score test error for prediksjon: -0.011667844934775351  
  
Spiller4:  
MSE test error for prediksjon: 1.976505769843072  
MSE test error for prediksjon med hyper tuning: 1.7116073562672722  
R2 score test error for prediksjon: -0.19365201000411614
```



Testing og trening av maskinlæringsmodellene

- ❖ Random Forest Regression
- ❖ Forsøk på forbedring av test score
 - Forkortelse av datamengde i treningssett

Date	
2021-03-22	7.553259
2021-03-23	6.860631
2021-03-24	7.103187
2021-03-25	7.503742
2021-03-26	7.277012
2021-03-27	7.103187
2021-03-28	7.503742
2021-03-29	7.277012
2021-03-30	7.103187
2021-03-31	7.503742



Spiller1:

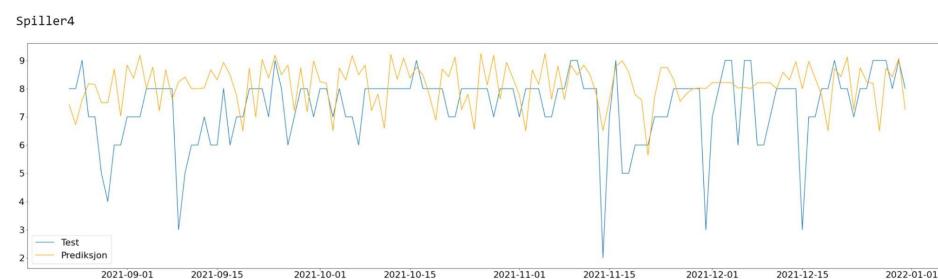
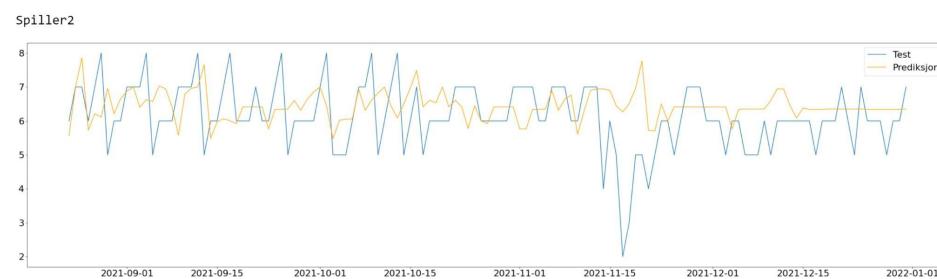
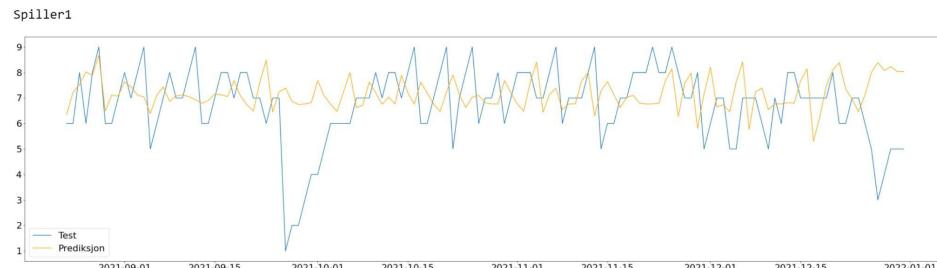
MSE test error for prediksjon: 1.4957180284464364

MSE test error for prediksjon med hyper tuning: 1.151243351260135

R2 score test error for prediksjon: -0.8766728778128698

Testing og trening av maskinlæringsmodellene

- ❖ Random Forest Regression
- ❖ Implementasjon av funksjon som forflytter lags

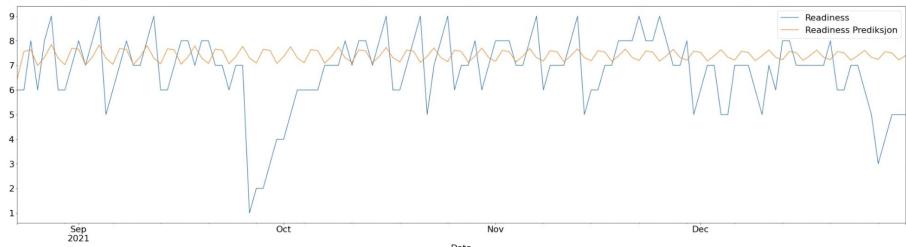


Testing og trening av maskinlæringsmodellene

❖ ARIMA

❖ Resultat av eksperimentet

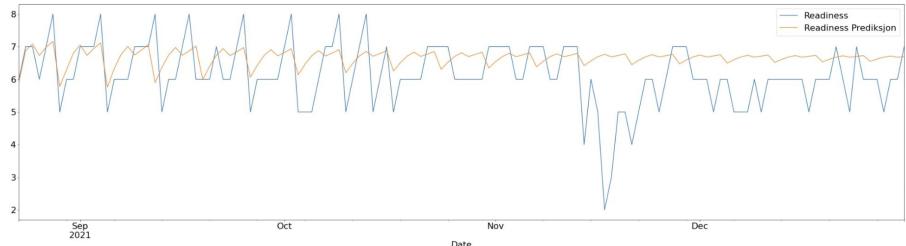
Spiller 1



---- Oversikt over testscore for Spiller1: ----

MAE: 1.153129265083629
MSE: 2.556357493374256
R²: -0.1805611125886868

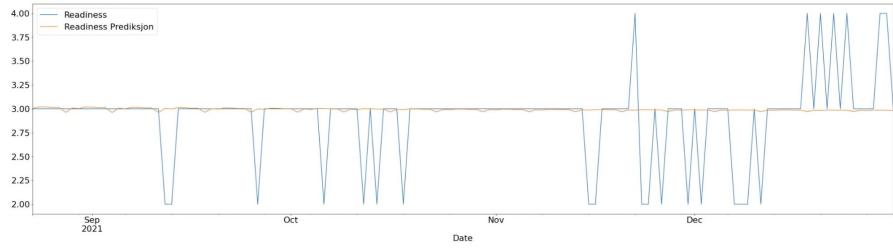
Spiller 2



---- Oversikt over testscore for Spiller2: ----

MAE: 0.8163696875522216
MSE: 1.1181858413861396
R²: -0.19827571025524793

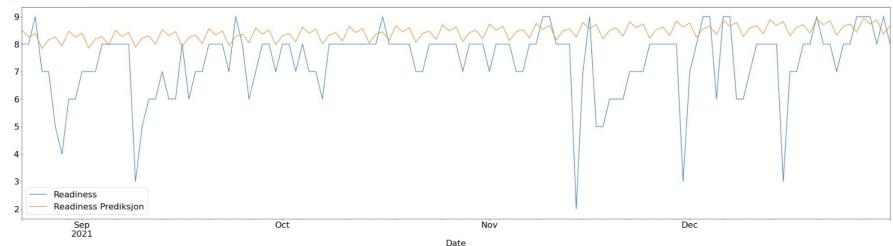
Spiller 3



---- Oversikt over testscore for Spiller3: ----

MAE: 0.20099557040035066
MSE: 0.19065927574053979
R²: -0.03738231800361569

Spiller 4

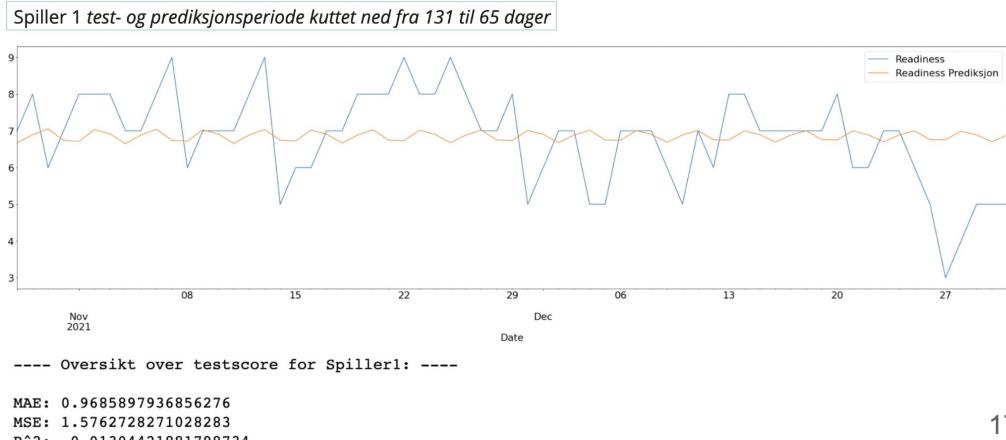
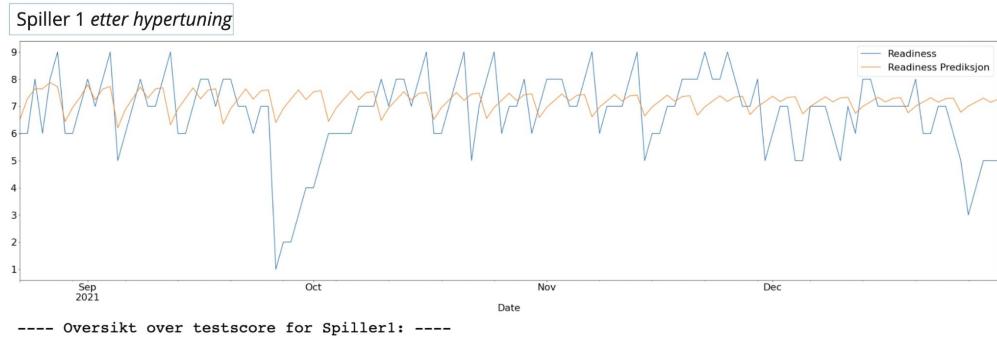


---- Oversikt over testscore for Spiller4: ----

MAE: 1.140474732394421
MSE: 2.6442074219376317
R²: -0.5968906097927824

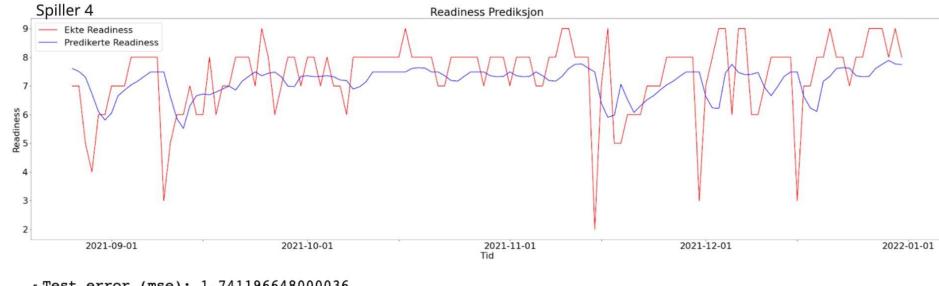
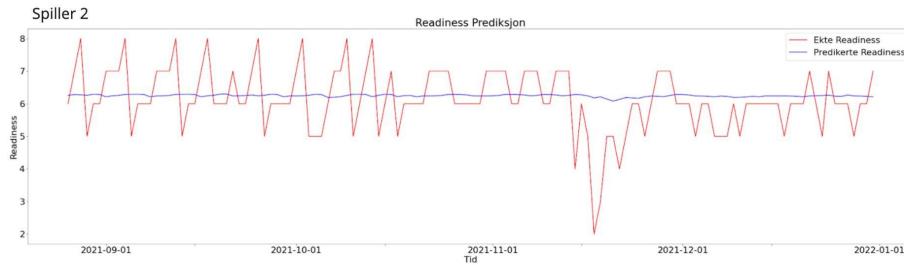
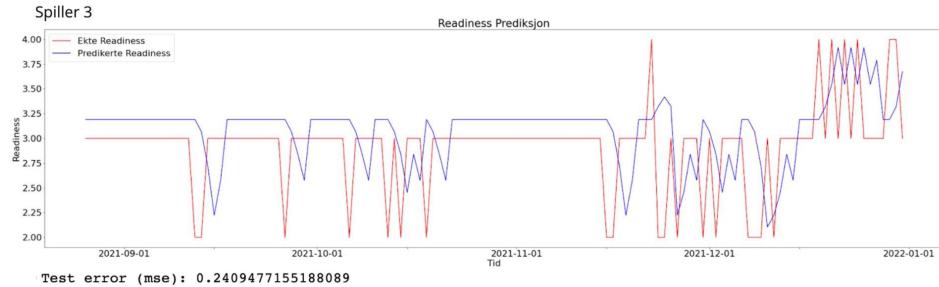
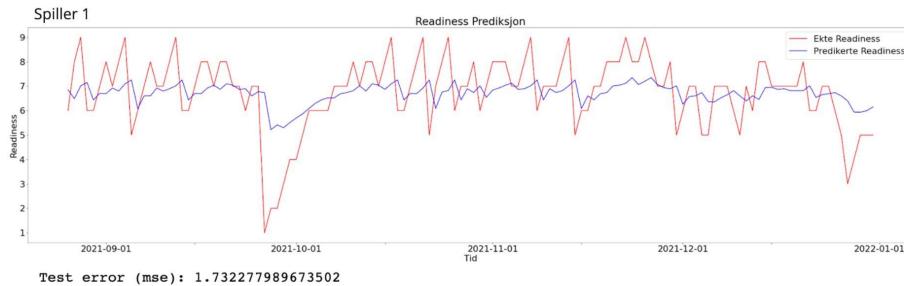
Testing og trening av maskinlæringsmodellene

- ❖ ARIMA
- ❖ Forsøk på å forbedre test score
 - *Hypertuning resulterte ikke i vesentlig bedre resultat statistisk sett*
- ❖ Eksperiment med kortere test- og prediksjonsperiode
 - *Mindre spenn i verdier gir bedre test score*
 - *Gir i flere tilfeller en gjennomsnittsprediksjon*



Testing og trening av maskinlæringsmodellene

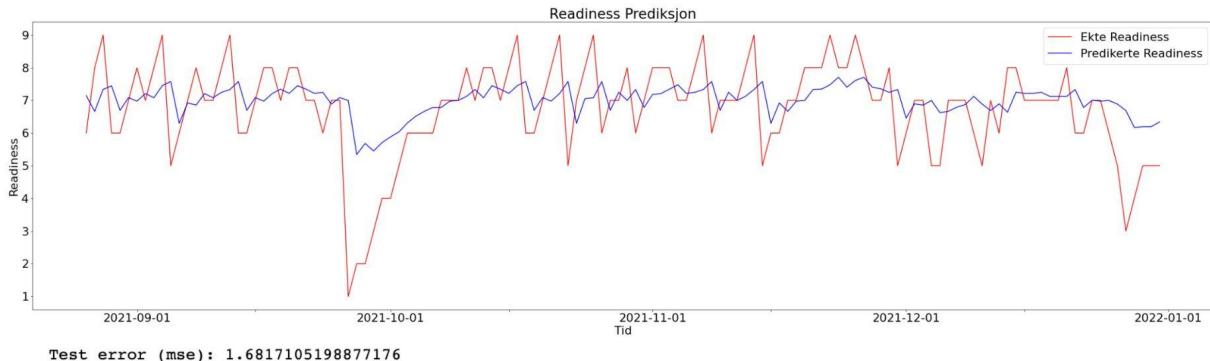
- ❖ LSTM
- ❖ Resultat av eksperimentet



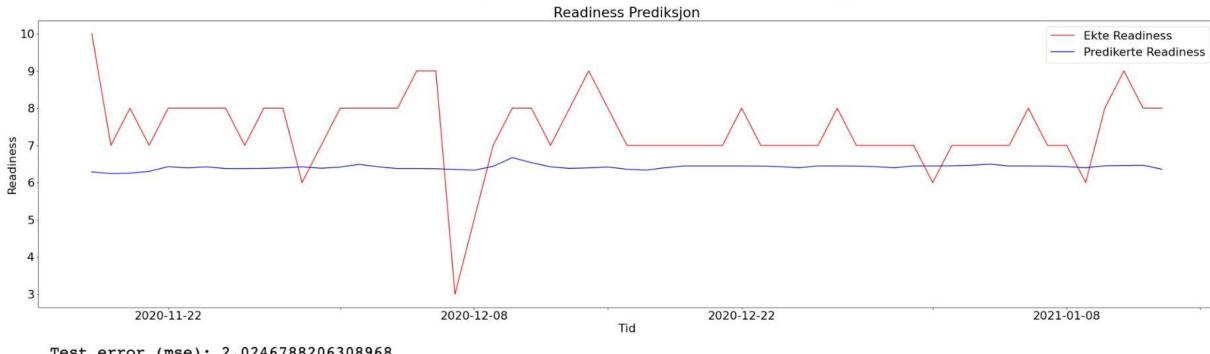
Testing og trening av maskinlæringsmodellene

- ❖ **LSTM**
- ❖ Testing av ulike datamengder
 - *LSTM trenger større datasett for bedre prediksjoner*
- ❖ Lovende potensiale

Spiller1 - periode på 652 dager:

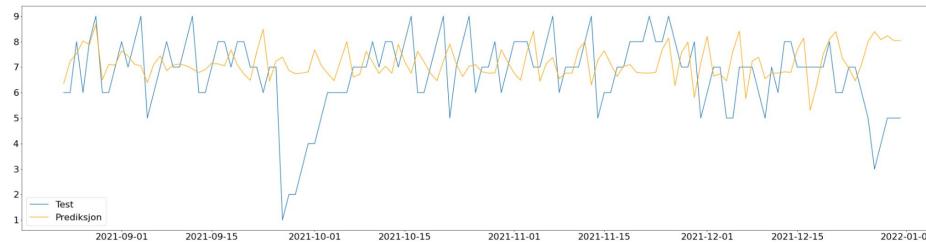


Spiller1 - periode på 300 dager:



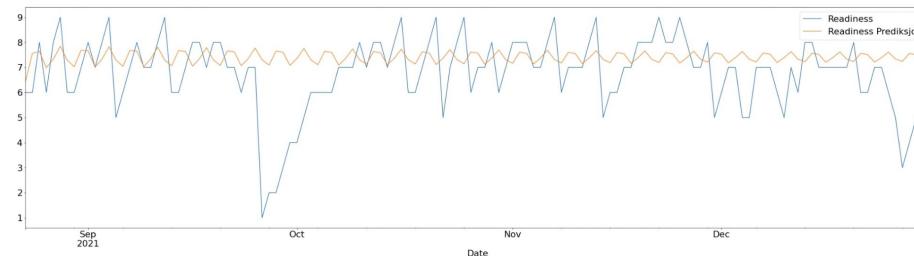
Sammenligning av testresultat

❖ Random Forest



MSE test error for prediksjon: 2.9339312699741047

❖ ARIMA

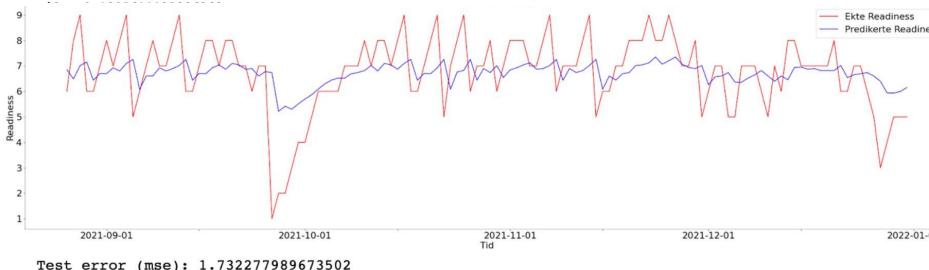


---- Oversikt over testscore for Spiller1: ----

MAE: 1.153129265083629

MSE: 2.556357493374256

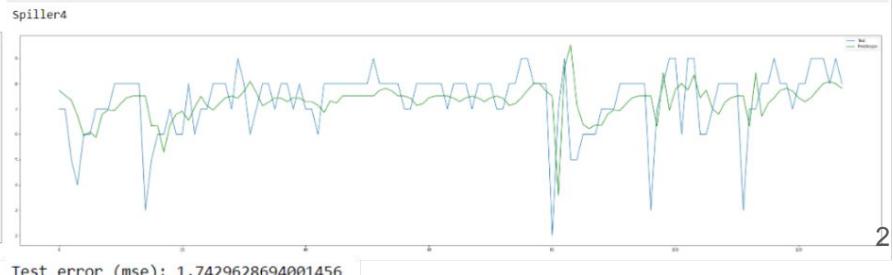
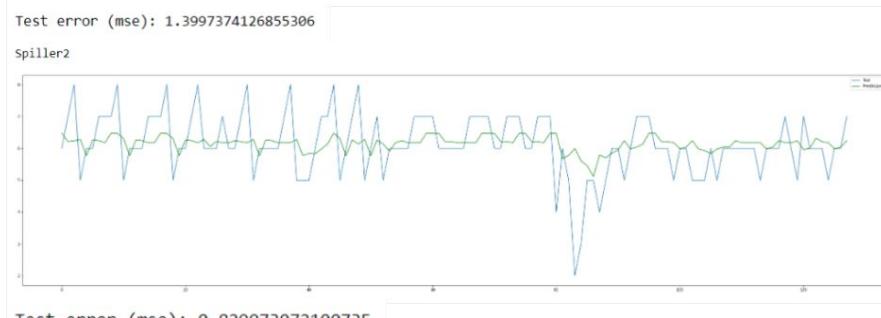
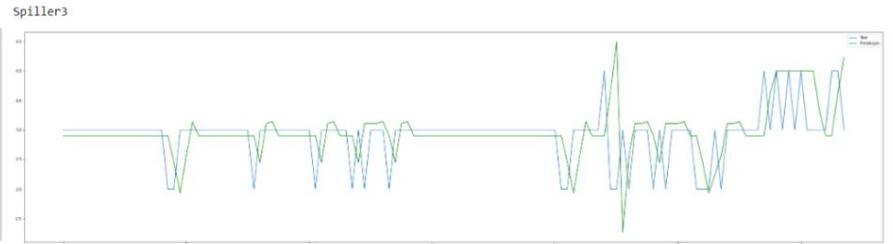
❖ LSTM



Test error (mse): 1.732277989673502

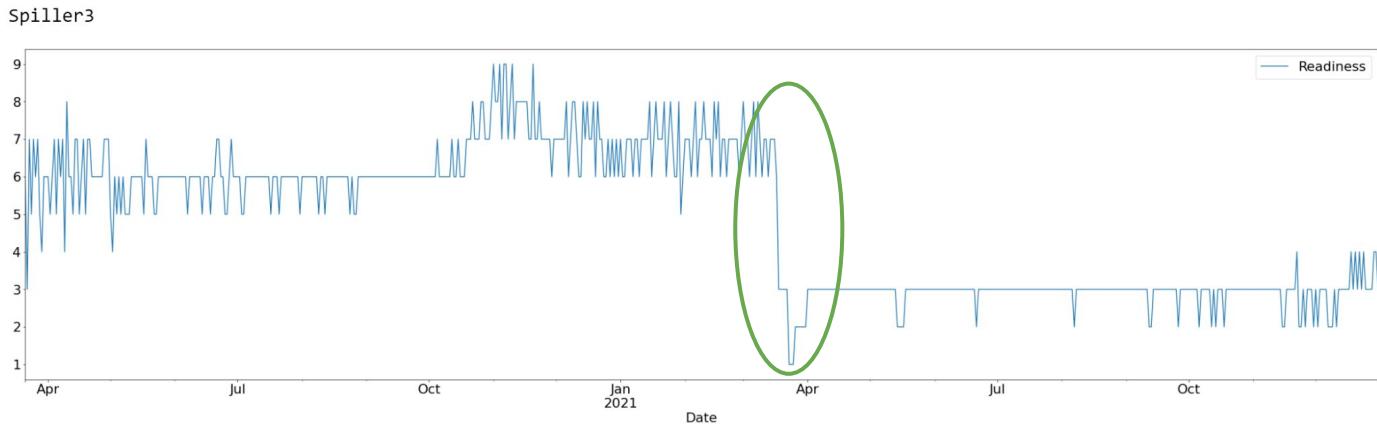
Convolutional neural network

- ❖ Convolutional Neural Network
 - *Dyp-læringsmodell*
 - *Finner topper og bunner*
 - *Gir lovende resultater*



Prediksjon av skade

- ❖ Spiller3 hadde registrert skade
- ❖ Trener modell i tidsperioden rundt skaden
- ❖ Kan ikke trekke konklusjon
- ❖ Viser potensiale



```
[[0.33553656 0.66446344]]  
Player3 risk of injury is:66%
```

Diskusjon og refleksjon

- ❖ Hva kunne ha blitt gjort annerledes
 - *Kunne brukt mer tid på testing av modeller, og ikke teorien bak*
 - *Gå i dybden på en eller to modeller, istedenfor tre*
- ❖ Læringsutbytte
 - *Større forståelse for maskinlæring*
 - *Lært om arbeidet med en utredningsrapport*
 - *Planlegging av arbeid*
- ❖ Tilbakemelding fra oppdragsiver
 - *Spennende arbeid*
 - *Steg i riktig retning*

Konklusjon

- ❖ **Overordnet mål:** Prediksjon av skade
 - **Vår problemstilling:** *Prediksjon av Readiness to Play*
- ❖ Kompleksiteten i problemstillingen krever en kompleks modell: LSTM
- ❖ Lineære algoritmer er mer krevende å sette opp for prediksjon
- ❖ Utregnet test score er ikke nok til å konkludere med om en modell er god eller dårlig
- ❖ Med resultatene fra LSTM er dyp læring den mest optimale veien videre

Takk for oss!