Match predictor Serie A

Steffen Ulvestad, August Mareno Hansen, 02.11.2025

1: BESKRIV PROBLEMET

OMFANG / *SCOPE*

Målet med prosjektet er å lage et web-basert verktøy som kan gi sannsynlig utfall av kamper i den italienske fotballigaen Serie A.

Systemet bruker maskinlæring for å vurdere hvert lag er basert på tidligere sesongdata, og estimerer sannsynligheter av utfall.

Maskinlæring er en god løsning fordi resultatet av fotballkamper påvirkes av mange faktorer som form, forsvar, tidligere kamper osv. Til sammen utgjør dette et mønster som er vanskelig å fange manuellelt. Ved å trene en modell på faktisk data kan systemet lære disse mønstrene og sette de sammen til en sansynlighet. Løsningen presenteres som et enkelt nettsted der brukeren velger to lag fra dropdown meny. Når brukeren trykker Predict, hentes statistikk og modellen returnerer sannsynligheter og en sannsynlig vinner.

I dag finnes mange prediksjonssider, men de fleste krever omfattende registreringer eller bruker proprietære data. Dette prosjektet fokuserer på åpen data og en transparent, gjennomsiktig modell.
Brukeren trenger bare en nettleser,ingen installasjon og får umiddelbart prediksjonen.

Business objective

Formålet er først og fremst læring og demonstrasjon av et komplett maskinlæring produkt, men prinsippet kan utvides for:

- betting-analyse,
- sportsjournalistikk,
- eller innholds-apper som viser forventet kampbilde.

Målgruppe

Fotballinteresserte, sportsanalytikere og studenter som ønsker å forstå hvordan maskinlæring kan brukes på sportsdata.

Ressurser

Prosjektet kjøres lokalt på PC med Python, scikit-learn og FastAPI. Ingen spesiel maskinvare er nødvendig siden datamengden er moderat.

METRIKKER

Modellens ytelse måles med nøyaktighet på testsettet etter treningen.
Logistisk regresjon gir typisk ca. 0.70–0.80 nøyaktighet for å klassifisere lag som over/under gjennomsnittet. Dette anses som akseptabelt siden kampresultater har høg grad av tilfeldigheit.
I tillegg brukes software-metrikker:

- Respons-tid
- Oppetid under test
- Brukeropplevelse

En bra prototype defineres som:

- Modellen kan lastes og gi prediksjon uten feil
- Nøyaktighet ≥ 0.70
- Frontend viser resultat for valgte lag på under 1 sekund

2: DATA

Datasettet består av lagstatistikk fra Serie A – hentet fra åpne kilder (fbref.com / Kaggle).
Hver rad representerer et lag i en gitt sesong og inneholder:

- Kamper spilt
- Seire, uavgjort, tap
- Mål for/mot og målforskjell
- Poeng
- Noen tilleggsfelt som “top scorer” og “goalkeeper” for visning i UI.

Modellen trenes i “supervised learning” modus, der label y = 1 hvis laget hadde over gjennomsnittli poengsum den sesongen, ellers 0.

Data krever enkel rensing:

- konvertering av kolonnenavn til liten bokstav,
- håndtering av NaN,
- beregning av features som win ratio og goals per game.

Feature-engineering og skalering utføres i klassen FeatureBuilder som inngår i ML-pipen.
Datasettet inneholder kun åpne sportsdata og ingen personopplysninger, så ingen personvernhensyn er aktuelle.

3: MODELLERING

Modellen som brukes er en Logistic Regression-modell implementert i scikit-learn.
Dette er en enkel baseline som egner seg godt når målet er å klassifisere lag som sterke eller svake.
Pipelinen:

1. FeatureBuilder lager beregnede features:
   wn_ratio, goals_per_game, goals_against_per_game, gd
2. StandardScaler normaliserer data
3. LogisticRegression treningsfase
4. Modellen lagres med joblib til models/model.joblib

Baseline-ytelsen estimeres ved train/test-split (80/20) og accuracy_score.
Feilprediksjoner undersøkes ved å se på lag som ble feil klassifisert (lag med uvanlig høy xG eller mange uavgjort).
Mulige forbedringer:

- inkludere xG / xGA for bedre offensiv/defensiv balanse
- teste RandomForest eller XGBoost
- bruke poeng per kamp som kontinuerlig target

4: DEPLOYMENT

Systemet er satt opp som et komplett web-produkt:

- Backend: FastAPI med endepunkter for
  /api/teams, /api/team/{team}, /api/compare
  ML-modell lastes med joblib og prediksjoner beregnes via predict_proba.

- Frontend: HTML + CSS + JavaScript
  Nettstedet kjøres lokalt via port 5500.
  Når brukeren velger to lag og trykker Predict, sendes en POST til API-et og resultatet vises med prosentfordeling.

Plan for vedlikehold og forbedring:

- Oppdatere datasettet ved slutt av hver sesong
- Retraining av modell automatisk ved nye data
- Mulighet for å utvide til andre ligaer eller legge til spiller-features
- Legge til logging og monitorering av prediksjon over tid

5: REFERANSER

- *Kaggle datasets – Italian Serie A Season Statistics (2023/2024)*
- *https://fbref.com/en/comps/11/Serie-A-Stats#all_results2025-2026111',*
  - *Football Reference stats for Serie A*

- *scikit-learn documentation (https://scikit-learn.org/)*
- *FastAPI documentation (https://fastapi.tiangolo.com/)*
- *Uvicorn / Joblib / Pandas docs*
- Anas Riad : https://www.youtube.com/watch?v=Lngf-g369A4&t=275s

6: Redegjørelse for bruk av kunstig intelligens

I arbeidet med denne oppgaven har vi benyttet kunstig intelligens (KI) som verktøy på følgende måte: Vi har brukt Github copilot og ChatGPT under utviklingen av nettsiden. Vi har brukt den for feilsøking, struktur, idemyldring,hjelp til valg og oppsett av prosjektet. Vi har også brukt den til å finne forbedringer på koden for å lage et bedre produkt.