Egen projekt del 1

Projektbeskrivning:

Med hjälp av ML (Machine Learning) så skall vi försöka förutse fotbollsmatcher vilken lag kommer att vinna mellan två lag. Med hjälp av historisk statistik och olika former så skall jag försöka ta fram en modell som är någorlunda träffsäker.

Mål:

Målet är att kunna ställa två lag mot varandra och förutse resultatet. Typ av resultat kan vara H-lag (hemma lag) vinst, B_lag (borta lag) vinst eller att det blir oavgjort.
Godtyckligt resultat är +60% träffsäkerhet, men om det skulle innebära att resultate är 100%, så kommer Nils att avsluta sitt dagliga arbete och sysslar med day-betting 🙂

Datakälla:

Vi kommer primärt att använda att datakälla som finns gratis tillgängliga på nätet. I första hand. så kommer data att hämta från siten “http:///fbref.com”. Ev om data inte räcker så kan det hända att data måste hämta från mer kommersiella sidor eller apier

Primärt så kommer data att vara fokuserad på den Engelska ligorna (model utgår fråm Premier League). Anledning till det är att ligan är väl känd och det finns gått om detaljdata att inhämta.

Datavärde och typer:
Historisk data kommer inte att vara linjärt, dvs att för varje säsong i ligan så kommer vissa lag i tabellen att åka ut och nya lag kommer in. De nya lag som kommer in i ligan kommer ev inte ha fullständig historisk data för den nya ligan de tillhör. (Om man vill så kan vi hämta historisk data för den lag som flyttas upp. Man kan ta data från den tidigare ligan. Oklart om det kommer att göra). Just nu så är antal data rader på ca 1052, som består av matcher under perioden 2020-2022. Det finns möjlighet att ta ut data längre tillbaka i tiden.

Typa av datavärde är faställde efter dem regler i ligan, datakälla kommer att bestå både text och siffror. Man kommer att behöva konverter vissa textfält till numeriska värde. Och i vissa fall så kommer vissa fällt att innehålla “null”-värde (troligen de fall då ett lag åker ur ligan rep ny lag som kommer in).

Hantering av datakälla:

- Datafält som inte har något värde (null).
- Data konvertering till nummerik. (klass lag namn, spelform, bort/hemma-lag etc…)
- Lagra data historisk data i database (ingen måste för projekten, men kan vara bra längre fram om man skal titta på reaktids data.)

Källa:
Web Scraping url: http:///fbref.com. Gratis men med begränsning på anslutning.
Kommersiella url: https://footystats.org. Har tillgång till api, men kostar.

Problem Type:

Troligen så är fallet en klassificeringsproblem av typen supervised learning. Vi kommer att känna till historisk resultat, vi kommer till och börja med att grupper data i följande typer.

- Laget form: Ta fram laget 5 senaste matcherna och klassificera resultat som lagters form. Ex (summering av vins = 1, förlust = -1 & oavgjort = 0)
- Laget procentuellt resultat vid hemmamatcher rep bortamatcher.
- Ta fram statestik på de spel form som har generert mest vinst i historisk data. (spel form, dvs lag uppställning och dess positioner. Ex 4-4-2 där svenska follbollandlaget oftas ställer upp med. 4st backar, 4 mittfälltare och 2st forwarder)
- Ta fram den spelform som har bra statistik vid spel på hemmaplan.
- Ta fram den spelform som har bra statistik vid spel på bortaplan.
- Vickning mellan spelform, dvs spelform för hemmalag mot spelform för bortalag. (H_lag kör 4-4-2 och B_lag kör 4-3-3)
- Vad innebär lag med stor bollinnehav, genererar det vinst för laget ..? (lag med x% bollinhav = vinst eller ..?)
- Har spelform någon inverkan på bollinnehav ..? (lag med spel form X har X% bolinnehav).
- H_lag förväntad antal gjorda mål.
- B_lag förvänta antal gjorda mål.

Resultat och lösning/problem:

Jag har inte fastställt hur jag skall definiera resultatet. Allt är lite beroende på vilket resultat som klassificering presenterar. (kan behöva bolla lite)

Att ta tänka på.

- Ett sätt är att poängsätta resultat på klassificering grupperna och ta ut ett medelvärde. Detta innebär att resultate måste vara av samma datatyper. (just nu är det blandad mellan numerisk och procent värde)
- Sätta/ta fram ett gränsvärde för resultatet. Dvs vad innebär den resultat värde. (Om x-värde är över/under z-värde, är det en vinst och vad skall z-värde vara ..?)

Om det finns tid:

Önskade funktioner som ev kan implementer eller testa.

- Testa klassificering modellan med unsupervised learning.
- Utöka modellen till Reinforcement Learning och deep neural networks med realtids data. (Om det går att förutse resultat i levi match).