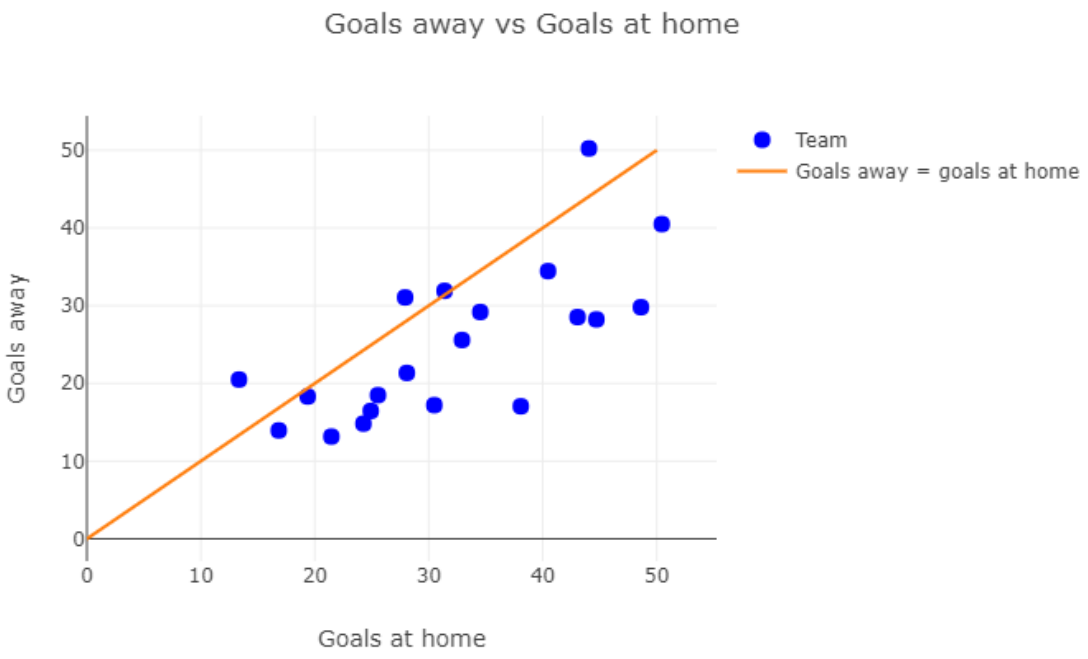
Dataset Serie A 2016-2017

Dopo i consigli ricevuti, ho modificato il mio codice così che fornisse dati più interessanti e leggibili all'analista. Riporto qui di seguito prima i nuovi plot riguardanti l'analisi esplorativa poi quelli riguardanti i classificatori.

Goal fatti in casa vs fuori casa

Solo Napoli, Sassuolo e Palermo fanno più goal fuori casa che in casa. Ho corretto piccolezze, fra cui il titolo del grafico.

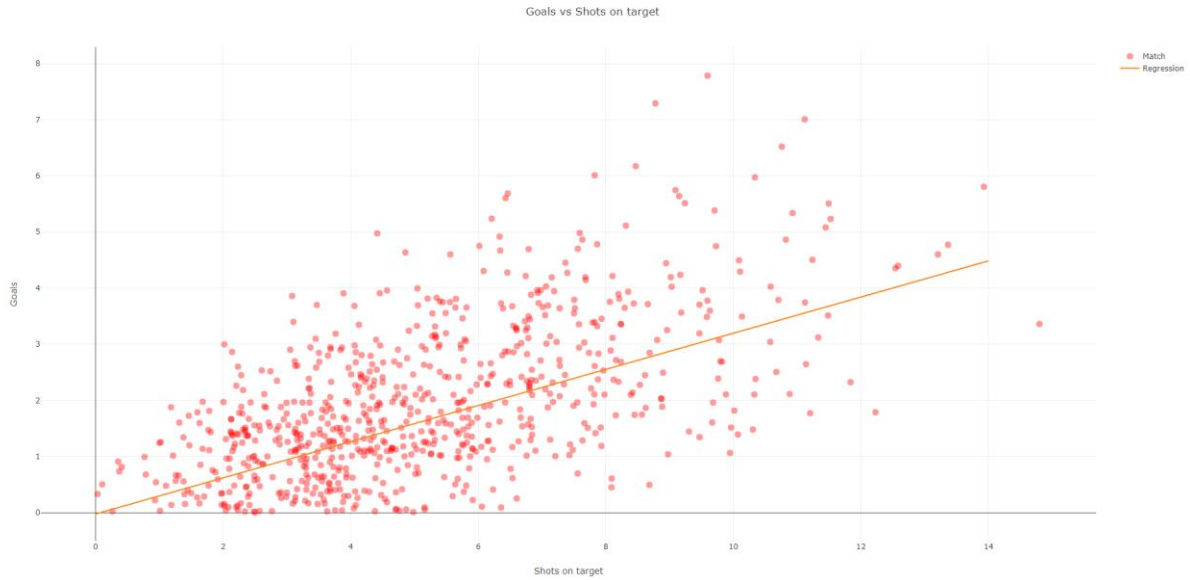


Tiri nello specchio vs goal

Trovo una buona correlazione fra i dati quindi noto che più una squadra tira in porta, più ha possibilità di fare goal.

La retta di regressione passa molto vicina all'origine, perciò decido di calcolare le statistiche dettagliate che sono riportate di seguito:

Calcolo la regressione
Intercept (b): 0.1594113684482812
Slope (w):0.2995273977980415
Correlation coefficient:
0.551564727655489

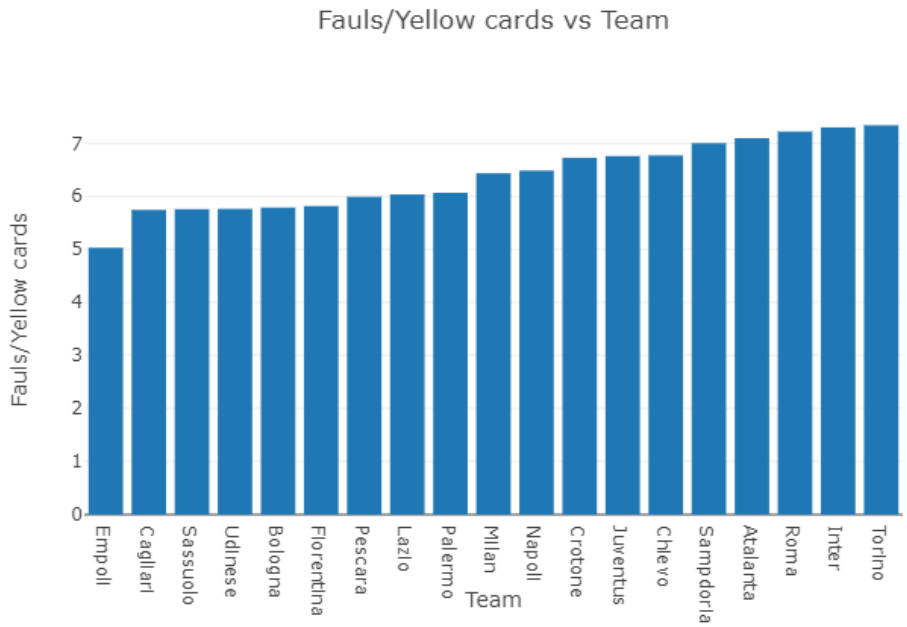


Tiri per goal

Per ogni squadra mostro i tiri che in media effettua per ottenere un goal, non tronco più i valori calcolati ma li mostro in ordine ascendente.

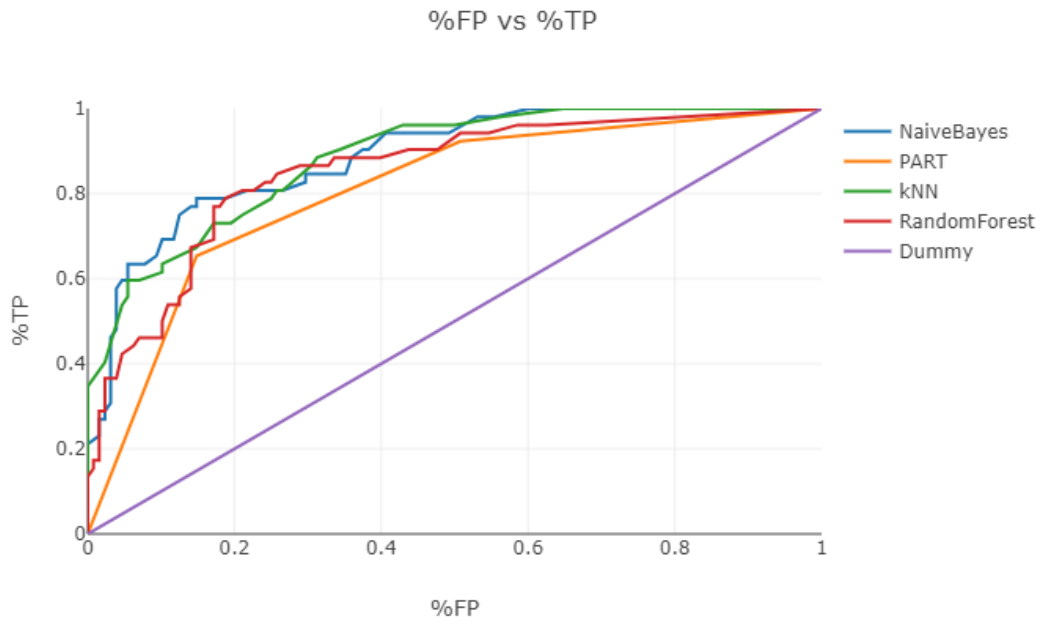
Falli per cartellino giallo

Per ogni squadra mostro i falli che fa in media per ottenere un cartellino giallo, non tronco più i valori calcolati ma li mostro in ordine ascendente.



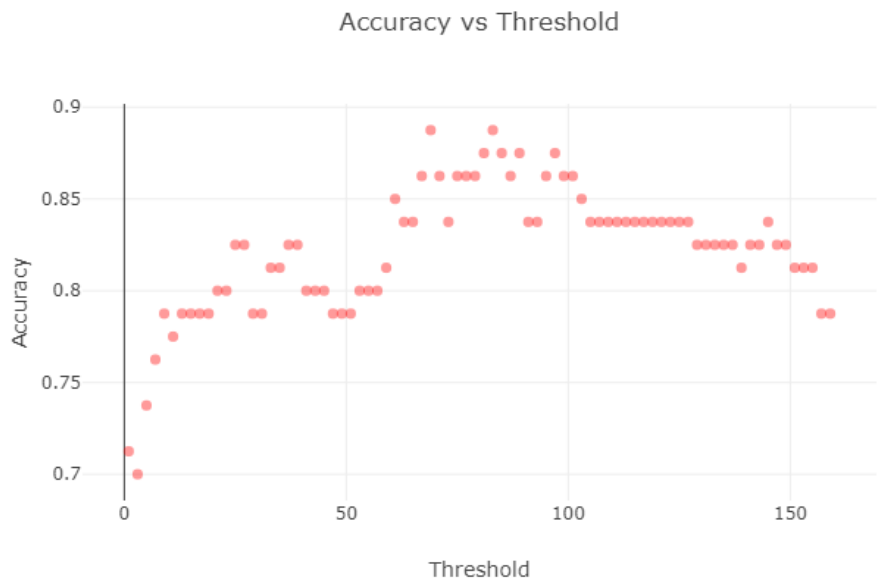
Classificatori

Rielaboro l'implementazione delle varie classificazioni in modo da riuscire a stampare sullo stesso grafico più ROC.

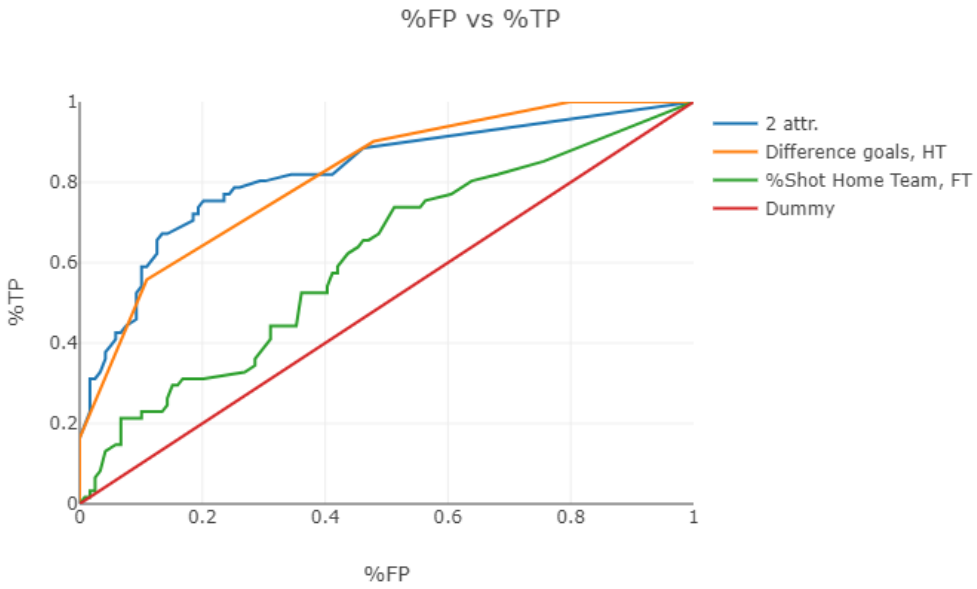


Si nota che in genere tutti i classificatori provati portano risultati soddisfacenti, ma che il kNN (con k = 91) e il NaiveBayes sono i migliori.

Per quanto riguarda il kNN ho deciso di fissare k = 91 poiché dopo diversi test mi sono accorto che, quando k si avvicina a 90, in genere ottengo buoni risultati. Questa è solo un ipotesi dato che i dati presi in considerazione vengono randomizzati ad ogni *run* ed è quindi difficile trovare il k migliore.



Per mostrare meglio l'impatto che i vari attributi hanno sulla ROC, usando il RandomForest calcolo: prima la ROC con entrambi gli attributi poi solo con l'attributo differenza reti a metà tempo e in fine solo con l'attributo che calcola la percentuale di tiri effettuata dalla squadra di casa sull'arco di tutta la partita. In seguito plotto le 3 ROC ottenute.



Si nota che già con la percentuale dei tiri si riesce ad ottenere un buon risultato, questo perché l'attributo in questione è calcolato sull'arco di tutta la partita e perciò ha un impatto notevole sul problema di classificazione che mi sono posto.