



SOCCER MATCH PREDICTION

Basso Matteo (807628)
Ferri Marco (807130)

Università degli Studi di Milano-Bicocca
Corso di Modelli Probabilistici per le Decisioni



OBIETTIVO

Una Rete Bayesiana è in grado
di predire il risultato di una
partita di calcio?



DOMINIO 01

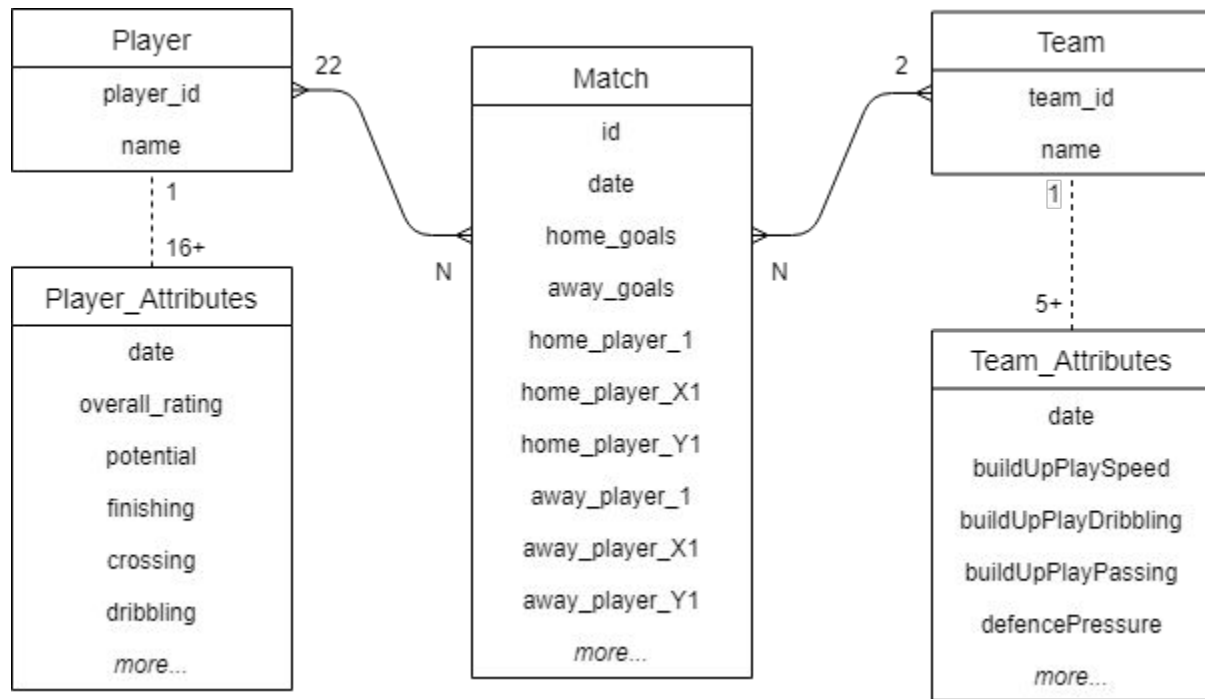
DATI

10.000 giocatori
25.000 partite



Caratteristiche dei
giocatori e delle squadre

SCHEMA





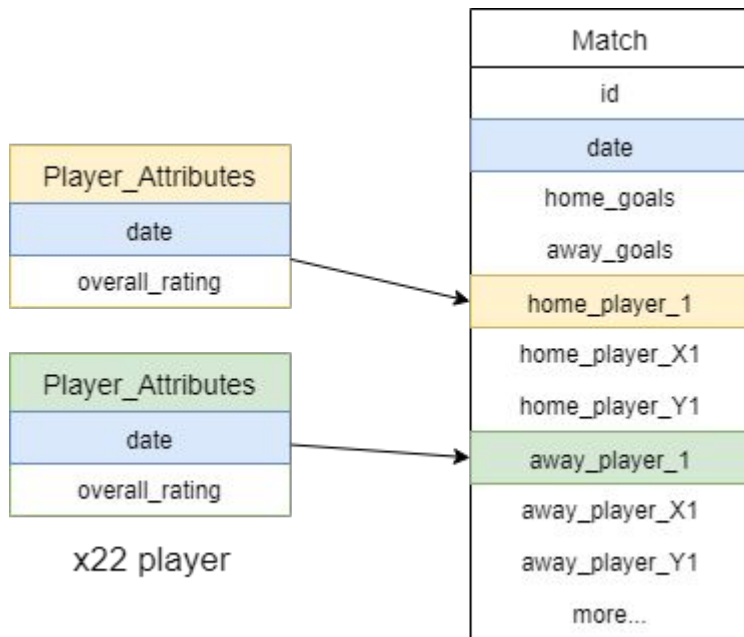
CREAZIONE DEL DATASET

02

ASSUNZIONI

- Le performance di una squadra derivano dai **giocatori**
- Il risultato della partita dipende **unicamente** dai giocatori che ne prendono parte
- Il **ruolo** del giocatore è fondamentale
- Le caratteristiche di due squadre che si scontrano sono fra loro **indipendenti**

DATA INTEGRATION



22 **join** SQL
generate via Javascript

+

eliminazione delle
righe incomplete

OVERALL RATING

Assegnazione di un punteggio a ciascun giocatore in base al **ruolo**

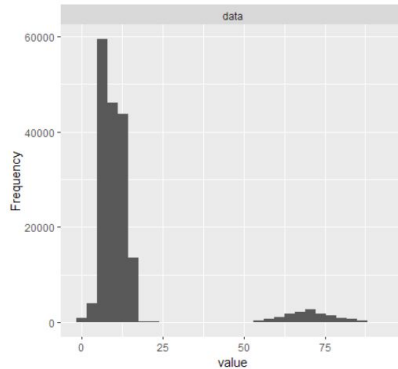


Valutazione di caratteristiche differenti per ciascun ruolo:

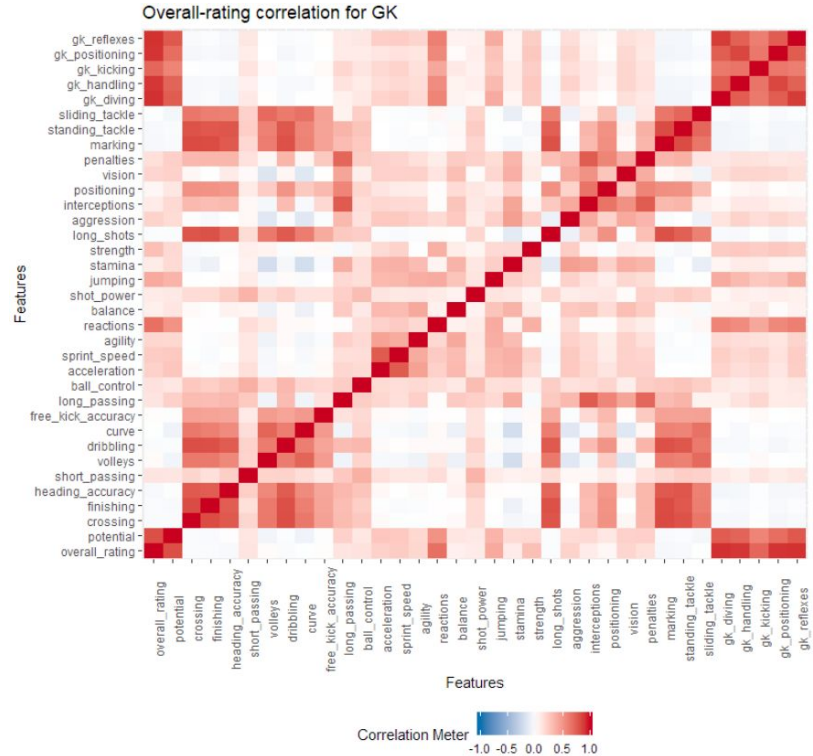
- **Portiere**: riflessi, presa, tuffo, ...
- **Attaccante**: tiro in porta, dribbling, ...
- **Difensore**: marcatura, contrasto, ...
- **Centrocampista**: cross, dribbling, ...

OVERALL RATING PER I PORTIERI

Correlazione con le
caratteristiche
tipiche dei **portieri**
(sulla destra)



distribuzione di
gk_diving

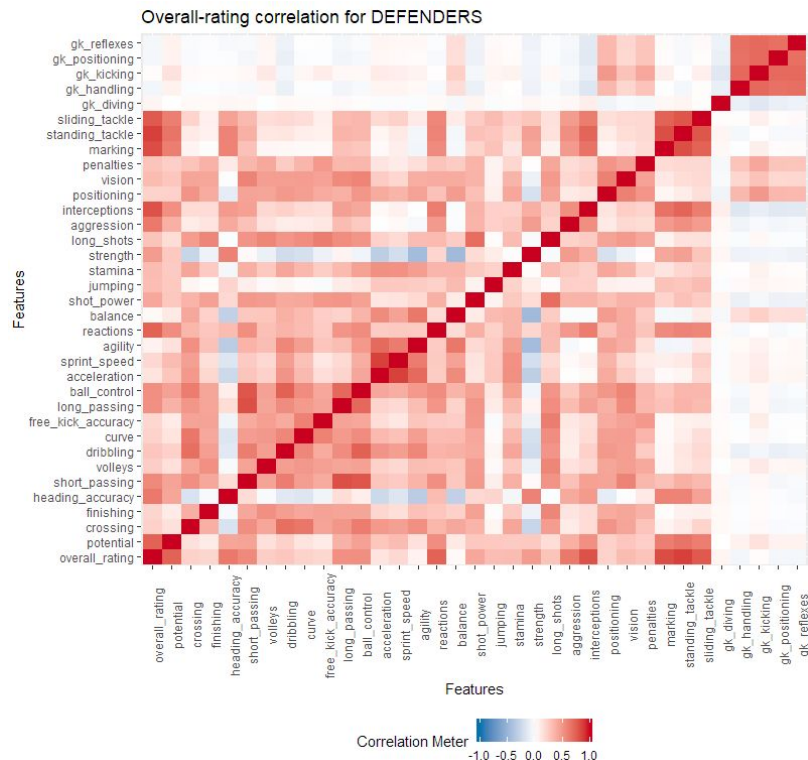


OVERALL RATING PER I DIFENSORI

Correlazione con le
caratteristiche
difensive

Distinguere un
attaccante da un difensore?

$\text{COR}(\text{marking}, \text{finishing}) = -0.6204823$

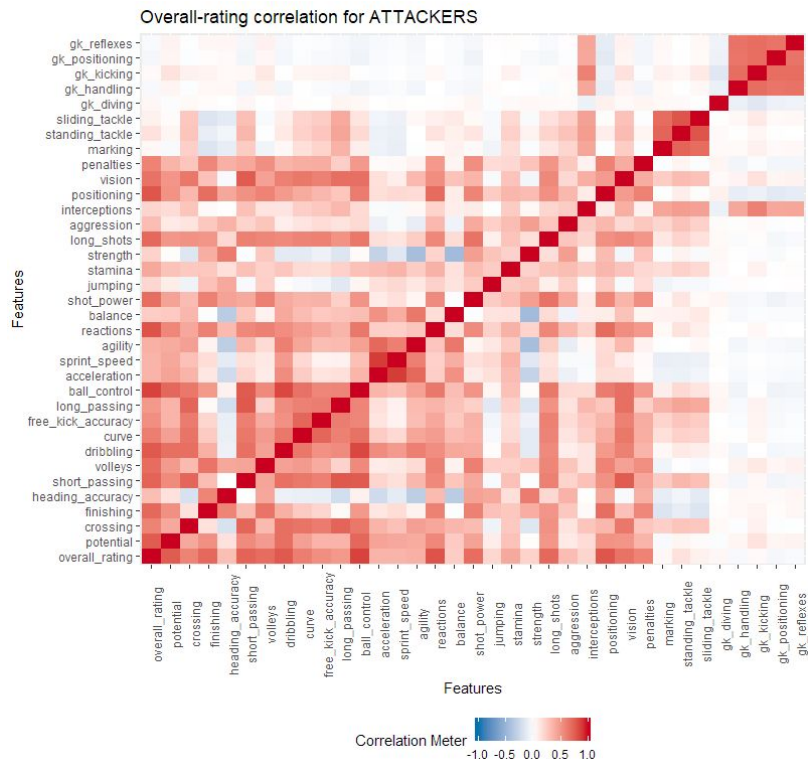


OVERALL RATING PER GLI ATTACCANTI

Correlazione con le
caratteristiche
offensive

Distinguere un
attaccante da un difensore?

$\text{COR}(\text{marking}, \text{finishing}) = -0.6204823$

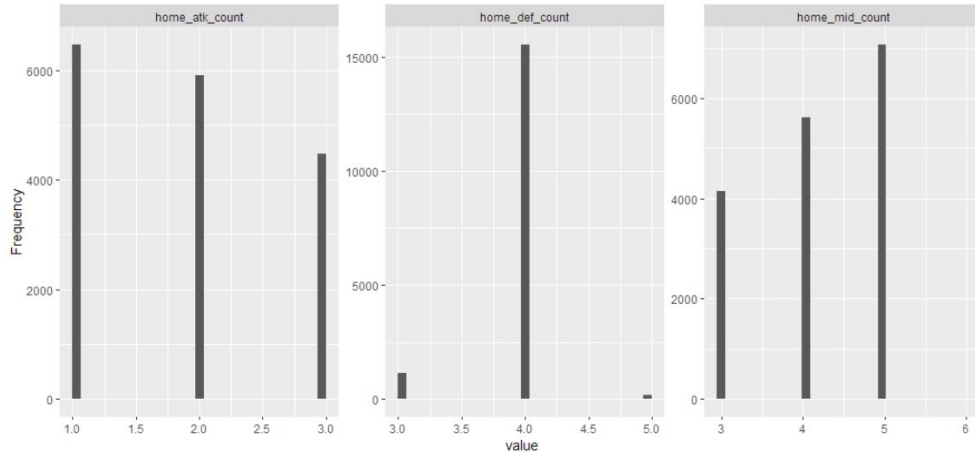


RUOLI DEI GIOCATORI

Ruoli dei giocatori **non** definiti nel dataset

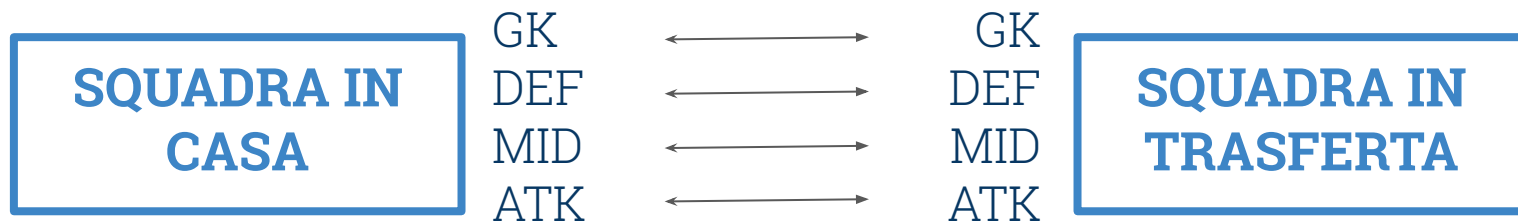


Assertiti attraverso le **posizioni** in campo



PUNTEGGI ASSEGNATI ALLA SQUADRA

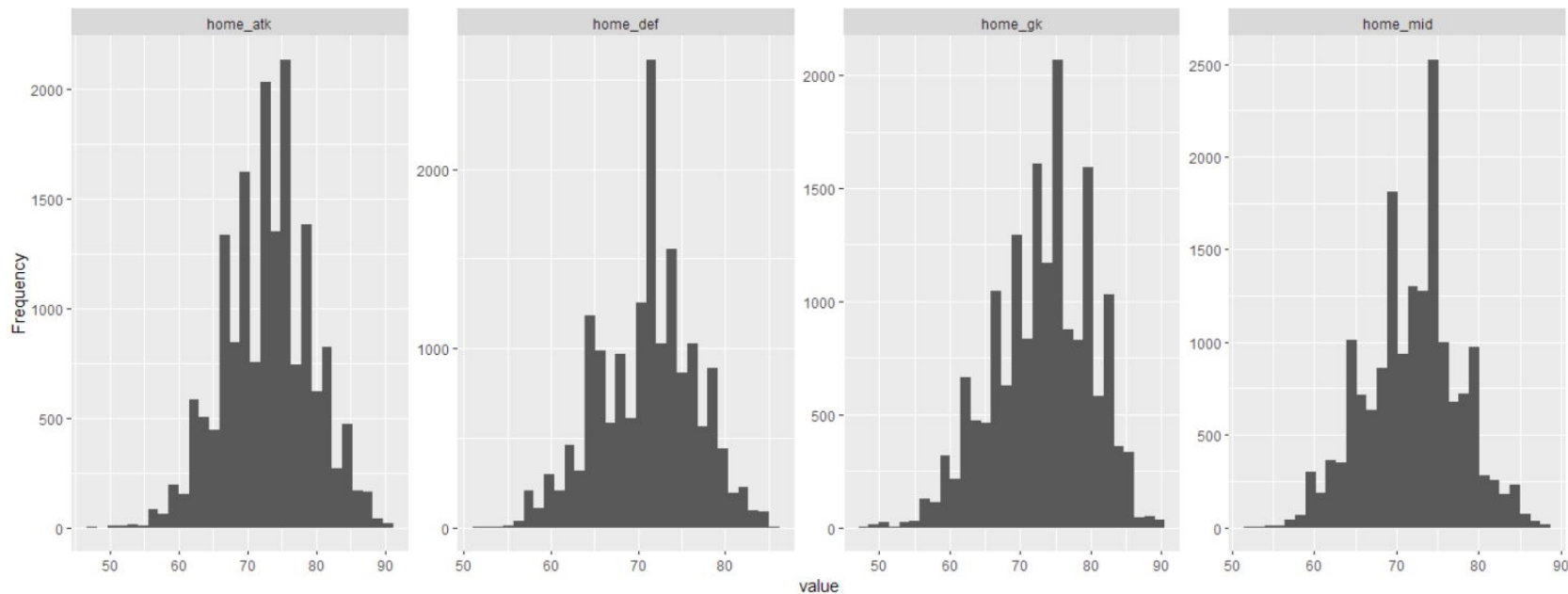
- Una valutazione per ciascun **ruolo**



vista la correlazione di **overall_rating** con il ruolo
si possono calcolare i **punteggi** attraverso la
media fra gli overall_rating dei giocatori nei diversi ruoli

FEATURE

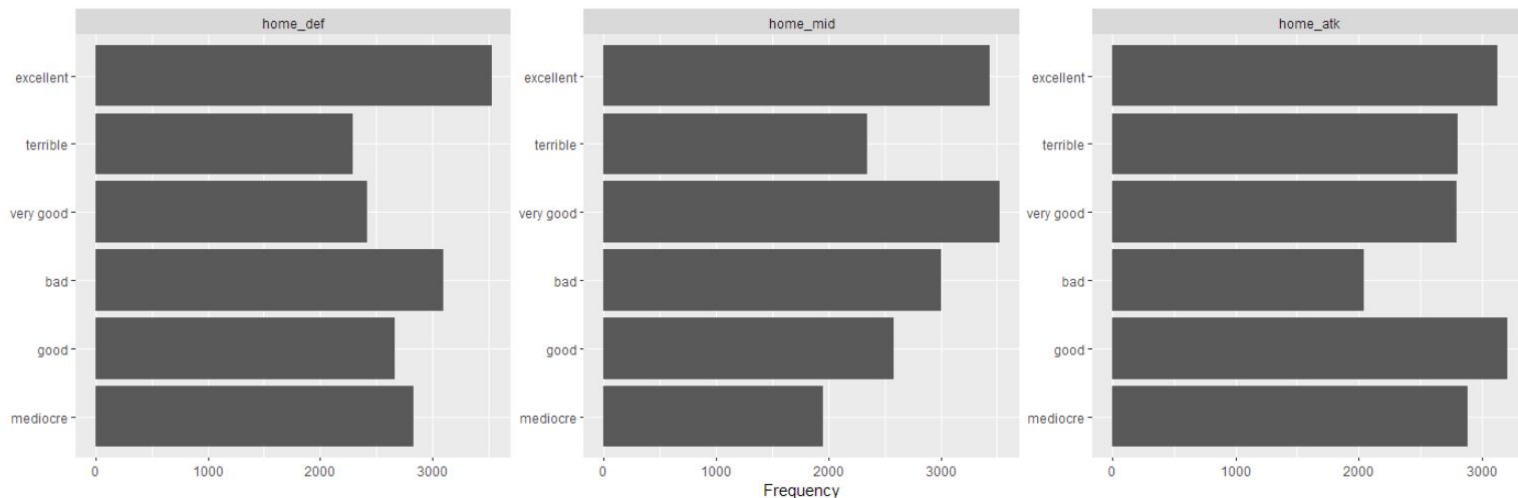
Distribuzione dei punteggi **continui** su scala da 1 a 100



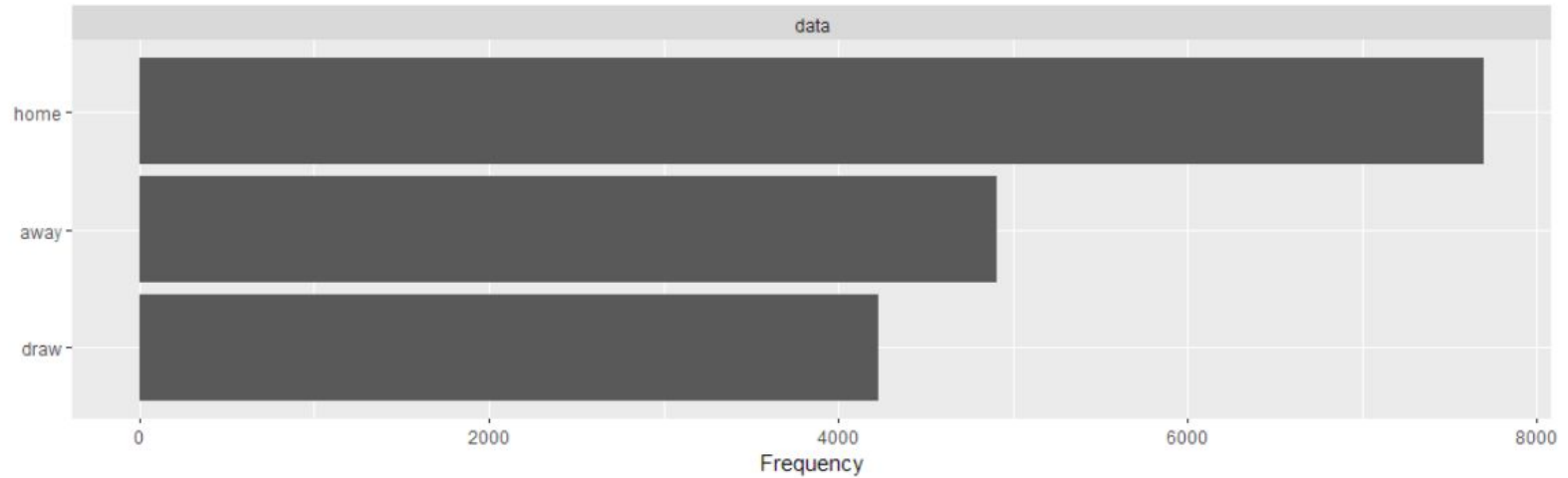
DISCRETIZZAZIONE DELLE FEATURE

Sono state effettuate degli esperimenti per la scelta del numero di intervalli e il metodo di discretizzazione.

Distribuzione post-discretizzazione per **frequenza** a **6 intervalli**



DISTRIBUZIONE DELLA VARIABILE TARGET



46.1% home

26.4% away

27.5% draw



03

RETE BAYESIANA

Network

SOFTWARE



Al fine di creare la struttura della Rete Bayesiana e stimare i parametri è stata utilizzata la libreria [bnlearn](#) del linguaggio di programmazione R

PROGETTAZIONE DI UNA RETE BAYESIANA

DEFINIZIONE DELLA STRUTTURA (DAG)



STIMA DELLE PROBABILITÀ
CONDIZIONATE

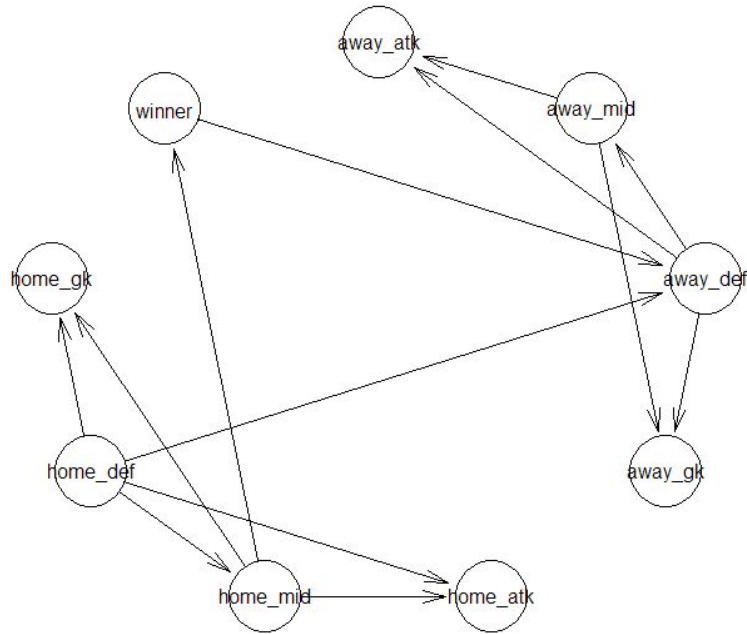


PREDIZIONE E INFERENZA

FUNZIONI DI BNLEARN

- generazione della struttura
 - *scored-based*
 - *constraint-based*
 - *hybrid*
- apprendimento delle CPT `bn.fit()`
- predizione `predict()`
- inferenza `cpquery()`
- cross fold validation `bn.cv()`

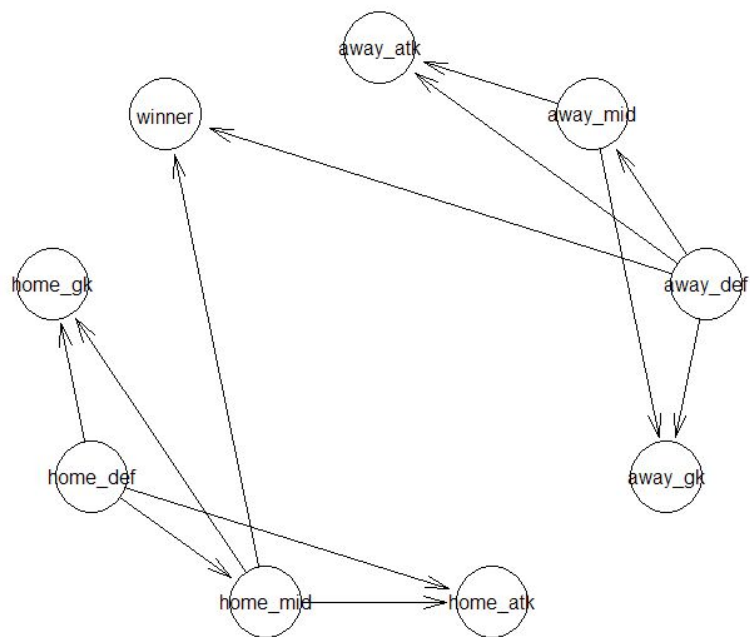
STRUTTURA RETE AUTO-GENERATA



Algoritmo
hill-climbing
greedy score-based

Viola le assunzioni per
cui i nodi delle due squadre non
devono essere raggiungibili con un
cammino orientato

STRUTTURA RETE VINCOLATA



Introduzione di una
black-list

- squadre indipendenti
- la squadra in casa punta sull'offensiva
- la squadra in trasferta preferisce la difensiva



04

PERFORMANCE

METRICHE DI VALUTAZIONE

Accuracy 0.5126

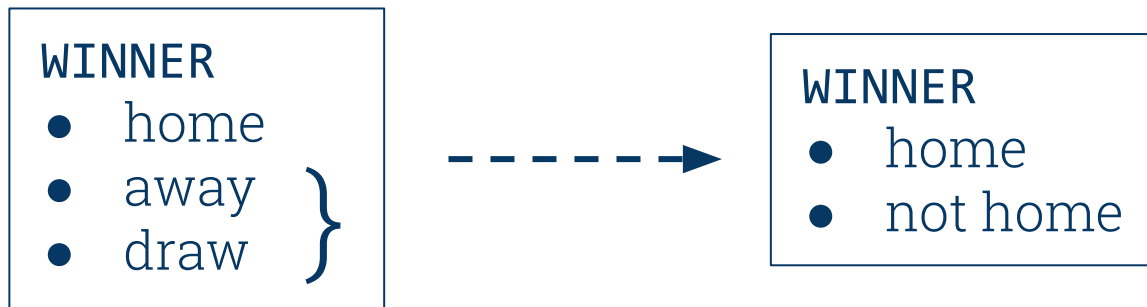
Classe	Precision	Recall	F1
HOME	0.5238946	0.849281	0.6477723
AWAY	0.482222	0.4263695	0.4516147
DRAW	0	0	0

Performance nulle per la classe DRAW (pareggio)

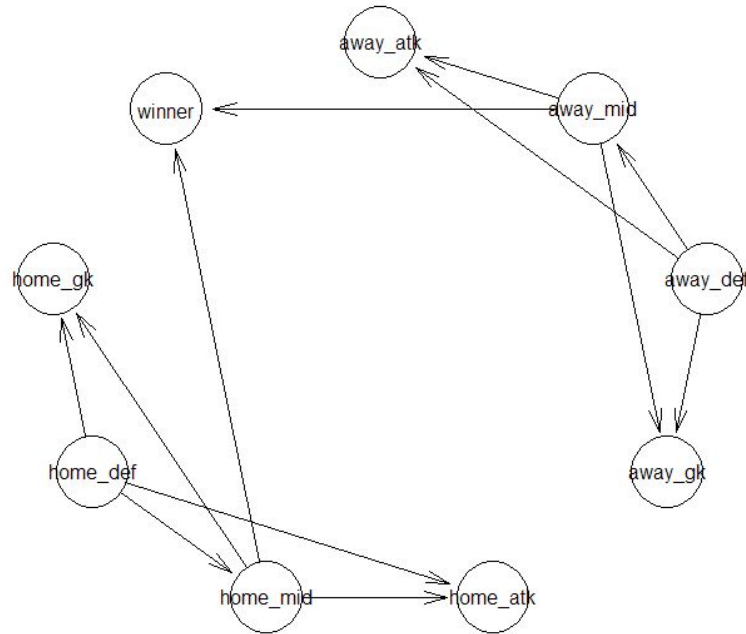
- Dataset con variabile target mal distribuita
- Difficile prevedere accuratamente un pareggio

RIDUZIONE A PROBLEMA BINARIO

A causa della cattiva distribuzione dei pareggi nella variabile target e alla conseguente impossibilità da parte della rete di classificarli, è stato deciso di ridurre il problema multiclasse a un problema di natura binaria.



STRUTTURA RETE VINCOLATA BINARIA



*struttura della rete generata
sul problema binario*

in un problema binario le
squadre giocano a pari
capacità e obiettivi

INFERENZE

Causali

$P(\text{winner} = \text{not home} \mid \text{home_mid} = \text{terrible}, \text{away_mid} = \text{excellent}) = 0.9086$

$P(\text{winner} = \text{home} \mid \text{home_mid} = \text{terrible}, \text{away_mid} = \text{excellent}) = 0.1244$

$P(\text{winner} = \text{home} \mid \text{home_mid} = \text{terrible}, \text{away_mid} = \text{terrible}) = 0.3825$

Diagnostiche

$P(\text{away_mid} = \text{excellent} \mid \text{winner} = \text{home}) = 0.1136$

$P(\text{away_mid} = \text{bad} \mid \text{winner} = \text{home}) = 0.2147$

Intercausali

$P(\text{away_def} = \text{good} \mid \text{away_mid} = \text{excellent}) = 0.0509$

$P(\text{away_def} = \text{good} \mid \text{away_mid} = \text{terrible}) = 0.0011$

Diagnostiche + causali

$P(\text{home_mid} = \text{excellent} \mid \text{winner} = \text{home}, \text{home_def} = \text{good}) = 0.1034$

$P(\text{home_mid} = \text{bad} \mid \text{winner} = \text{home}, \text{home_def} = \text{excellent}) = 0.0017$

PERFORMANCE

Accuracy 0.6257

Classe	Precision	Recall	F1
HOME	0.636251	0.4293271	0.5117034

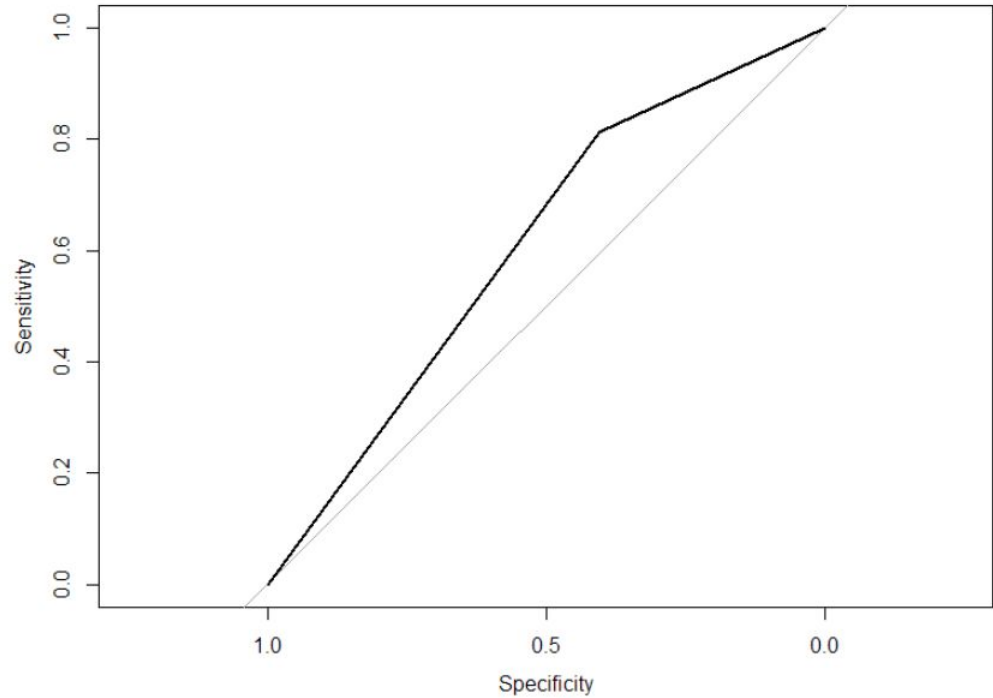
- peggioramento delle performance generali $F1$

MA

- risultato non completamente sbilanciato a sfavore di una certa classe, come nel caso binario

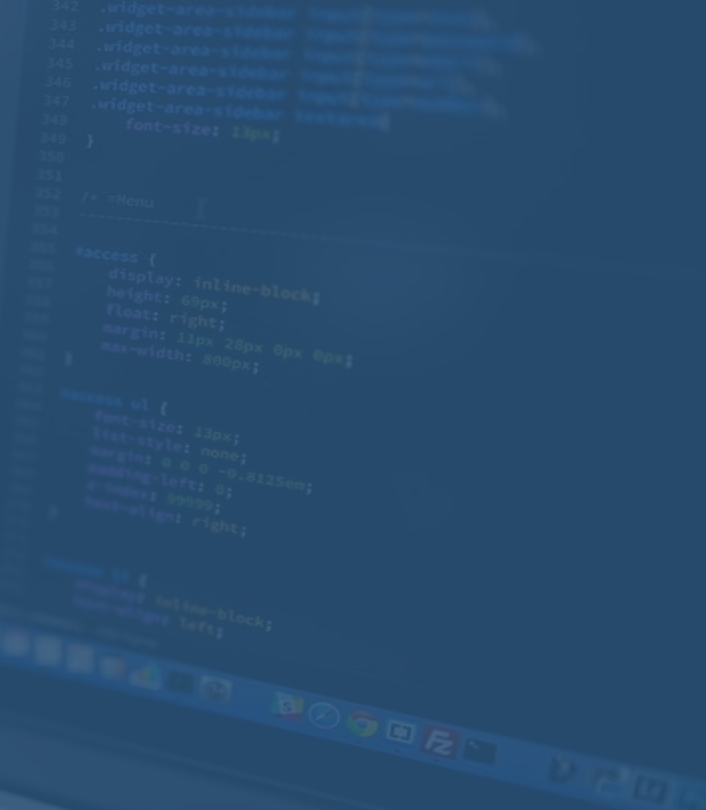
ROC CURVE

AUC 0.61



05

ESPERIMENTI



REWARD PER NUMERO DI GIOCATORI

Al fine di valutare l'impatto derivante dall'aggiunta o rimozione di un giocatore da uno dei ruoli è stata definita una nuova funzione per il calcolo del punteggio totale.

$$new_overall = overall + K * \log\left(\frac{away_atk}{avg_atk_count}\right)$$

Tuttavia, si è rivelata ininfluyente dal punto di vista delle performance.

PUNTEGGIO DI ATTACCO E DIFESA

Si è provato ad aggiungere due feature come raggruppamento del potenziale **difensivo** (portiere + difesa + mid) e **offensivo** (mid + attacco) delle squadre.

Essi sono stati impiegati in una nuova Rete Bayesiana i cui risultati si sono rivelati analoghi a quelli delle precedenti.



DEMO

06


PREDIZIONE RISULTATO

Soccer match prediction

HOME TEAM

AWAY TEAM

RESULTS



Players

Jose Dorado

Rating: 74

Yasin Karaca

Rating: 61

Arber Zeneli

Rating: 69

Eder Citadin Martins

Rating: 79

Nilson

Rating: 67

Fynn Arkenberg

Rating: 58

Abdelmalek Cherrad

Rating: 62

Jan Bednarek

Rating: 60

George Puscas

Rating: 64

Alberto Rey

Rating: 67

Gregory Mertens

Rating: 66

Nella pagina principale è possibile formare le squadre e predire il vincitore

INFERENZA

Bayesian Network

```
graph LR; hg[home_gk] --> hm[home_mid]; hd[home_def] --> hm; hm --> hw[home_winner]; ha[home_atk] --> hw; ag[away_gk] --> am[away_mid]; ad[away_def] --> am; am --> aw[away_winner]; aa[away_atk] --> aw; hw --> w[winner]; aw --> w;
```

Evidence

Home team	Away team	Winner
home_gk	away_gk	winner
home_def	away_def	
home_mid	away_mid	
very good		
home_atk	away_atk	

Query

Variable	Value
winner	home

COMPUTE RESET

$P(\text{winner} = \text{home} \mid \text{home_mid} = \text{very good}, \text{away_gk} = \text{good}) = 0.4973913$

Un dialog aggiuntivo permette la visualizzazione della Rete Bayesiana e di effettuare inferenze tramite l'inserimento di evidenze e variabile query

CONCLUSIONI

07



CONCLUSIONI

Performance non particolarmente soddisfacenti

- dataset con variabile target mal distribuita
- difficile predire un vincitore se le due squadre sfidanti non presentano caratteristiche generali molto diverse
- il risultato di una partita dipende da numerosi fattori quali il mind-set della squadra e dei singoli individui

SE FOSSE FACILE STABILIRE A PRIORI IL RISULTATO DI
UNA PARTITA CALCISTICA NON ESISTEREBBERO LO
SPORT STESSO E IL MERCATO DELLE SCOMMESSE

CONSIDERAZIONI SULLE RETI BAYESIANE

- Comode per modellare domini con dipendenza causale fra variabili stocastiche
- Facili da implementare attraverso R con **bnlearn**

Le basse performance non sono causate dalla scelta del modello.

Anche **Naive Bayes** e **Random Forest** *
hanno ottenuto risultati del tutto comparabili.

* Gli algoritmi di learning sono stati testati attraverso l'utilizzo di [Knime Analytics Platform](#) e hanno registrato la medesima accuracy dei modelli di Rete Bayesiane presentati in questo elaborato.

GRAZIE PER L'ATTENZIONE

A dynamic soccer match scene. In the foreground, a player in a yellow Brazil jersey with the number 10 is jumping towards the goal. A goalkeeper in a blue jersey is diving to the right, reaching for a soccer ball. Another player in a blue jersey with the number 2 is in the center. The background shows a large stadium filled with spectators under bright lights. The scene is overlaid with a semi-transparent green banner at the bottom.

Basso Matteo (807628)
Ferri Marco (807130)