UNIVERSITÀ DEGLI STUDI DI SALERNO

DIPARTIMENTO DI INFORMATICA

Corso di Laurea Triennale in Informatica

Football Predictor
Prof. Fabio Palomba

Alfonso Anzelmo Mat. 0512113154

ANNO ACCADEMICO 2024/2025

Indice

Capitolo 1

Introduzione

Il calcio è lo sport più popolare al mondo, con miliardi di appassionati che lo seguono e milioni di persone coinvolte nelle scommesse sui risultati delle partite. Tuttavia, come ogni altro sport, è per sua natura imprevedibile. Nonostante l'analisi dettagliata delle squadre e delle prestazioni, il risultato di una partita è influenzato da molteplici fattori come la condizione fisica dei giocatori, le strategie, gli imprevisti e perfino il clima.

Fare predizioni accurate sulle partite di calcio è di grande interesse per tifosi, allenatori, media e scommettitori, alimentando così un ampio business legato alle scommesse sportive. Attualmente, numerosi siti e modelli predittivi, come Forebet, offrono pronostici sui risultati. Tuttavia, questi strumenti spesso risultano poco affidabili: nel campionato di Premier League 2017/18, ad esempio, Forebet ha correttamente predetto solo 2 risultati su 10 per ogni giornata, sottolineando il rischio elevato di affidarsi esclusivamente a tali previsioni.

Con questo progetto, ci proponiamo di sviluppare un modello predittivo che riduca i margini di errore, offrendo un'accuratezza e precisione superiori. Utilizzando tecniche avanzate di machine learning, il nostro modello classificatore utilizzerà dati del campionato Premier Legue 20/21 e 21/22 per predire con maggiore sicurezza se una squadra vincerà o perderà una partita.

Capitolo 2

Descrizione dell'agente

2.1 Specifica PEAS

L'acronimo PEAS sta per Perfomance, Enviroment, Actuators, Sensors e viene utilizzato per descrivere un ambiente.

Perfomance: è la misura di prestazione adottata per valutare l'operato di un agente. Nel caso di FootballPredictor è predire se una squadra vince, perde o pareggia una partita.

Environment: l'ambiente comprende l'insieme delle squadre di calcio, i risultati passati, e altri dati rilevanti. Include anche variabili esterne come le condizioni atmosferiche e gli infortuni che potrebbero influire sui risultati.

Actuators: servono all'agente per poter compiere azioni. Nel nostro caso sarà lo schermo per mostrare le predizioni.

Sensors: permettono all'agente di ricevere input dall'esterno. Nel caso di FootballPredictor saranno il dataset da cui prendere informazioni.

2.2 Proprietà dell'ambiente

Completamente osservabile: i sensori forniscono all'agente l'accesso allo stato completo dell'ambiente in qualsiasi istante.

Deterministico: lo stato successivo dell'ambiente è determinato interamente dall'azione eseguita dall'agente.

Episodico: Ogni previsione sarà indipendente da un'altra.

Statico: l'ambiente non cambia mentre l'agente elabora.

Discreto: l'ambiente ha un numero di stati discreto, chiaramente definiti.

Singolo Agente: Vi è un 'unica entità che prende decisioni nell'ambiente.

Capitolo 3

CRISP-DM

Il CRISP-DM rappresenta il ciclo di vita di progetti di intelligenza artificiale e data science. Si può paragonare ad un modello di ciclo di vita waterfall con feedback ed è composto : Business understanding, Data understanding, Data Preparation, Data modelling, Evalutation e deployment.

3.1 Business Understanding

Questa fase consiste nel chiarire gli obiettivi di progetto e i requisiti. Si tratta di identificare il problema da risolvere e stabilire criteri di successo che guideranno le fasi successive.

3.1.1 Obiettivi di business

In linea con il rinnovamento e il crescente interesse, abbiamo pensato di creare un modello che possa stimolare ulteriormente l'interesse verso questo sport. L'obiettivo di business è prevedere il risultato di una partita con un modello preciso, così da essere significativo ai a tifosi, scommettitori e professionisti del settore.

1. **Accuratezza delle Predizioni**: Migliorare la precisione delle previsioni sui risultati delle partite, fornendo un modello più affidabile rispetto alle soluzioni esistenti.
2. **Incremento del Coinvolgimento**: Aumentare l'interesse e l'engagement dei tifosi attraverso predizioni interessanti e coinvolgenti che arricchiscono l'esperienza sportiva.

3.1.2 Analisi del problema

Possiamo notare che l'agente dovrà essere in grado di predire una variabile categorica, quindi modelleremo un classificatore.

3.1.3 Tecnologie e Tool necessari

Useremo il linguaggio Python e diverse sue librerie. Pandas: Libreria per la manipolazione e analisi dei dati. Sklearn:Libreria per il machine learning composta di algoritimi, metriche.

3.2 Data Understanding

Questa fase riguarda l'analisi dei dataset necessari per raggiungere gli obiettivi del progetto.
Si esplorano i dati disponibili, verificandone la qualità, la completezza e la rilevanza per il problema da risolvere.

3.2.1 Descrizione del dataset

Il dataset è composto da due stagioni del campionato di Premier Legue 20/21 e 21/22. Come descritto nella documentazione, ci si aspetterebbero 38(giornate) * 20(squadre) * 2 (stagioni) = 1520 righe, invece ci sono solo 1389.Questo poiché lo scraping è stato fatto mentre la stagione 2022 era ancora in corso.

- time orario in cui si è tenuta la partita
- comp competizione per cui si è tenuta la partita
- round giornata della competizione
- day nome del giorno in cui si è tenuta la partita
- venue partita in casa o fuori casa
- result L lose, W win, D draw
- gf goal che la squadra ha eseguito
- ga goal che la squadra ha subito
- opponent nome della squadra opponente

- xg #goal attesi che la squadra esegua
- xga #goal attesi che la squadra subisca
- captain capitano della squadra
- formation formazione della squadra
- referee nome dell'arbitro
- sh tiri che la squadra ha eseguito
- sot tiri che la squadra ha eseguito alla porta
- fk calcio di punizione eseguiti
- pk calci di rigore eseguiti
- penalty kicks attemps tentativi di calci di rigore
- season anno della competizione
- team squadra di casa

3.3 Data Preparation

Questa fase consiste nel trasformare, pulire e organizzare i dati per renderli pronti per la fase di modellazione.
Ci concentriamo su :

- **Pulizia dei Dati**: Gestire valori mancanti, duplicati o inconsistenti.
- **Trasformazione delle Variabili**: Creare o modificare le variabili per renderle più utili al modello. Ad esempio, convertiremo la variabile `target` (risultato della partita) in una variabile categoriale numerica (es. vittoria = 1, pareggio = 0, sconfitta = -1).
- **Riduzione della Dimensionalità** (se necessario): Eliminare variabili poco rilevanti o ridondanti, come le colonne `notes` e `match report` se non contribuiscono significativamente al modello.
- **Standardizzazione e Normalizzazione**: Portare le variabili numeriche, come `gf` (gol fatti) e `ga` (gol subiti), su una scala comune per facilitare l'apprendimento del modello.
- **Suddivisione del Dataset**: Dividere il dataset in set di addestramento e test .

3.3.1 Feature selection

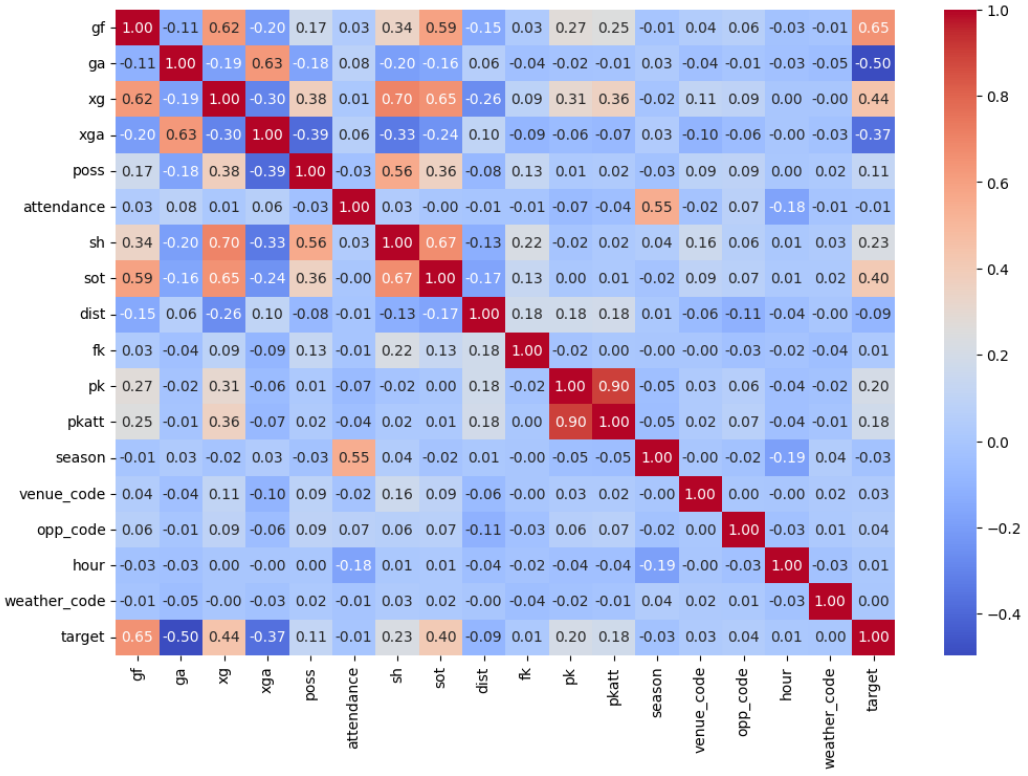
Questa è una delle fasi più importanti del processo poichè è in questa fase che si selezionano le variabili indipendenti più rilevanti dal dataset, col fine di migliorare le prestazioni del modello.
Maggiore è la cura in questa fase maggiore saranno i vantaggi derivanti, tra cui:

- riduce la dimensionalità del dataset
- contribuisce a migliorare la velocità del modello
- aumenta l'accuratezza del modello
- riduce rischio di overfitting.

Abbiamo deciso di eseguire la fase di feature selection articolandola in 3 fasi:

- **Rimozione di Variabili a Bassa Varianza**: Calcoliamo la varianza delle feature numeriche e rimuoviamo quelle con varianza inferiore ad una soglia (abbiamo scelto 0.1).
  Le variabili con bassa varianza forniscono poca informazione poichè i valori di cui sono composte non variano molto.
- **Calcolo della Correlazione**: Crea una matrice di correlazione e visualizza la correlazione tra tutte le feature e la variabile target.
  In sostanza, la correlazione tra due variabili è una misura di relazione tra esse. Se la correlazione tra la variabile target è forte con una variabile indipendente allora al variare di una anche l'altra cambierà. Quindi rimuoviamo le feature che hanno una correlazione bassa (abbiamo scelto valori minori di 0.1) con la variabile target.
- **Test Chi-Quadrato per Variabili Categoriali**: Eseguiamo il test Chi-quadro tra la variabile target e ciascuna variabile indipendenti. Se il valore ottenuto dal test è maggiore di 0.05, significa che la variabile indipendente è scarsamente correlata con la variabile target, quindi viene rimossa. In sostanza questo test misura l'indipendenza della distribuzione di due variabili, Maggiore è l'indipendenza minore è importante l'informazione data dalla variabile indipendente.

'result', 'gf', 'ga', 'opponent', 'xg', 'xga', 'poss', 'captain', 'formation', 'sh', 'sot', 'pk', 'pkatt', 'team', 'target'



3.3.2 Data cleaning

Per pulire il dataset abbiamo utilizzato una serie di tecniche per renderli più coerenti con le fasi successive, tra queste l'eliminazione di righe con valori Nan all'interno.

3.3.3 Divisione del dataset in train e test set

Ora dividiamo il dataset in due parti:

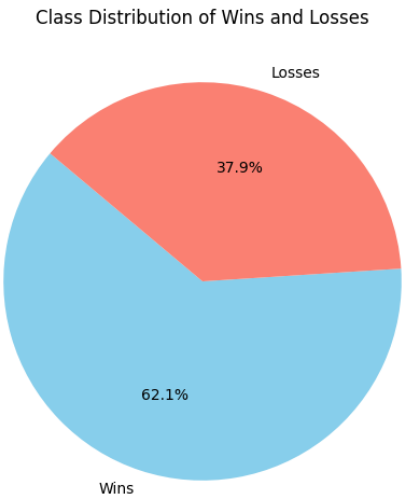
- training set: dati di addestramento
- test set: dati per valutare l'accuratezza del modello sulle predizioni che eseguirà

La scelta è stata di dividere il dataset per data ovvero per il train set di tutti i dati che hanno data inferiore al primo gennaio 2022, viceversa per il test set.

In questo modo si hanno 1100 dati per train set e circa 300 per il test set.

3.3.4 Data balancing

I nostri dati si presentano in una forma già bilanciata. Quindi non sono state applicate tecniche di over/under sampling.



3.4 Data Modelling

In questa fase l' obiettivo è costruire, addestrare il modello di machine learning che possa prevedere accuratamente la variabile target sulla base delle feature selezionate.

3.4.1 Scelta dell'algoritmo da utilizzare

Ricordiamo che l'obiettivo è di predire il numero di Goal segnati da una squadra in un match per questo siamo di fronte ad un regressore.

Esistono centinaia di algoritmi di regressione ma siamo interessati solo ad alcuni di essi :

- Decision Tree
- Random Forest Classifier
- Naive Bayes

Abbiamo scelto questi algorimtmi perchè ognuno apporta una vista differente alla classificazione. Il decision Tree fornisce una interpretabilità dai dati

tramite la costruzione di regole. Naive Bayes anche se più semplice è comunque efficente su dataset di piccole dimensioni. Random Forset è il più potente, è un algoritmo di ensemble learning che utilizza come base learner una serie di decision tree poi aggrega le predizioni di ognuno per migliorare la predizione finale. Per questo motivo è il più robusto nonchè con la maggiore accuratezza e precisione.

3.4.2 Addestramento

Ecco l'algoritmo per addrestramento:

```
train = grouped_matchesLast3Perfomed[grouped_matchesLast3Perfomed["date"] < '2022-01-01']

test = grouped_matchesLast3Perfomed[grouped_matchesLast3Perfomed["date"] > '2022-01-01']

predictors += ["gf_last3Performed","ga_last3Performed","sh_last3Performed","sot_last3Performed","pk_last3Performed","pkatt_last3Performed"]

rf.fit(train[predictors], train["target"])

preds = rf.predict(test[predictors])

print(precision_score(test["target"],preds))
print(accuracy_score(test["target"],preds))

combined = pd.DataFrame(dict(actual = test["target"], prediction = preds), index = test.index)
combined
```

3.5 Valutazione

3.5.1 DecisionTreeClassifier

```
accuratezza ->  0.644927536231884
precisione ->  0.5294117647058824
```

3.5.2 Naive Bayes

```
accuratezza ->  0.677536231884058
precisione ->  0.5949367088607594
```

3.5.3 Random Forest Classifier

```
accuratezza ->  0.7137681159420289
precisione ->  0.654320987654321
```

Capitolo 4

Conclusioni

L'approccio proposto combina modelli semplici e complessi, sfruttando i punti di forza di ciascuno per migliorare le previsioni sui risultati delle partite di calcio. La struttura dell'approccio segue un metodo sistematico, in mode che ogni fase sia ben definita da scelte metodologiche consapevoli. Come in tutte le cose anche quest'approccio è migliorabileper esempio, avere dati aggiuntivi, come le condizioni meteorologiche o informazioni sugli infortuni dei giocatori
Oppure eseguire l'esplorazione di integrazione di modelli alternativi o approcci ensemble, come il Gradient Boosting, potrebbe offrire nuove opportunità per affinare le previsioni.

4.1 Glossario

- CRISP-DM → Cross-Industry Standard Process for Data Mining
- CSV → Comma-separated values
- API → Application programming interface

4.2 Riferimenti

- GitHub: Football Predictor
- Wikipedia Premier Legue 2020/21 , 2021/22

Grazie per l'attenzione