Metodi di Machine Learning per l'Analisi Predittiva di Prestazioni di Squadre in Ambito Sportivo

Candidato
Mirco Ceccarelli

Relatore
Prof. Marco Sciandrone

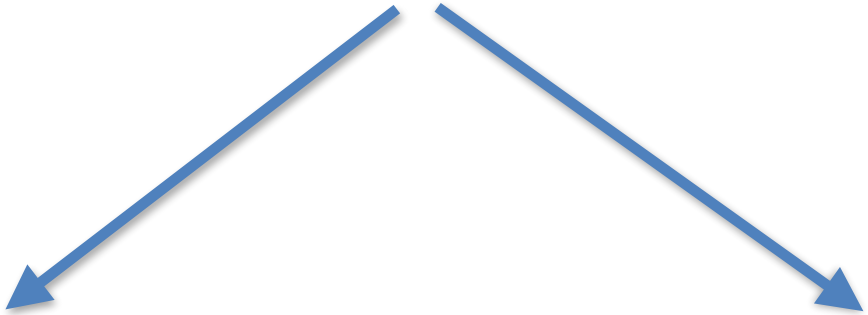
Correlatore
Dott. Matteo Lapucci

Introduzione del contesto sportivo

- In questa tesi abbiamo considerato la **Serie A di Pallavolo maschile**
  - Il **campionato** della **Lega Pallavolo Serie A** è strutturato in **2 fasi**:
    - **1ª fase**: Regular Season
    - **2ª fase**: Playoff
- Tutte le **squadre** di Serie A partecipano alla prima fase di **Regular Season.**
- Solo le **prime n squadre classificate** accedono alla fase dei **Playoff** che assegna a fine stagione il titolo di campioni.

Machine Learning per lo Sport

- Il Machine Learning nello **sport** è utilizzato per



Migliorare le **prestazioni** sportive dei **giocatori** e delle **squadre**

Ridurre il rischio di **infortuni**

Obiettivo della tesi

- **Predire** a inizio campionato, tramite l'utilizzo di **modelli** di **apprendimento automatico**, quali saranno le **squadre** che accederanno ai i **Playoff**.

Utilizzando i **dati** relativi alle **prestazioni** sportive **passate** dei singoli giocatori

Apprendimento Supervisionato

Per questa tesi è stata utilizzata questa classe di tecniche dell'Apprendimento Automatico.

Un problema risolvibile utilizzando questi metodi è:

- **Classificazione:** gli output sono costituiti da un numero finito di classi distinte (**due** nel nostro caso)

Sono stati utilizzati i seguenti modelli di Apprendimento Supervisionato:

- **Regressione Logistica**
- **Support Vector Machine (SVM)**

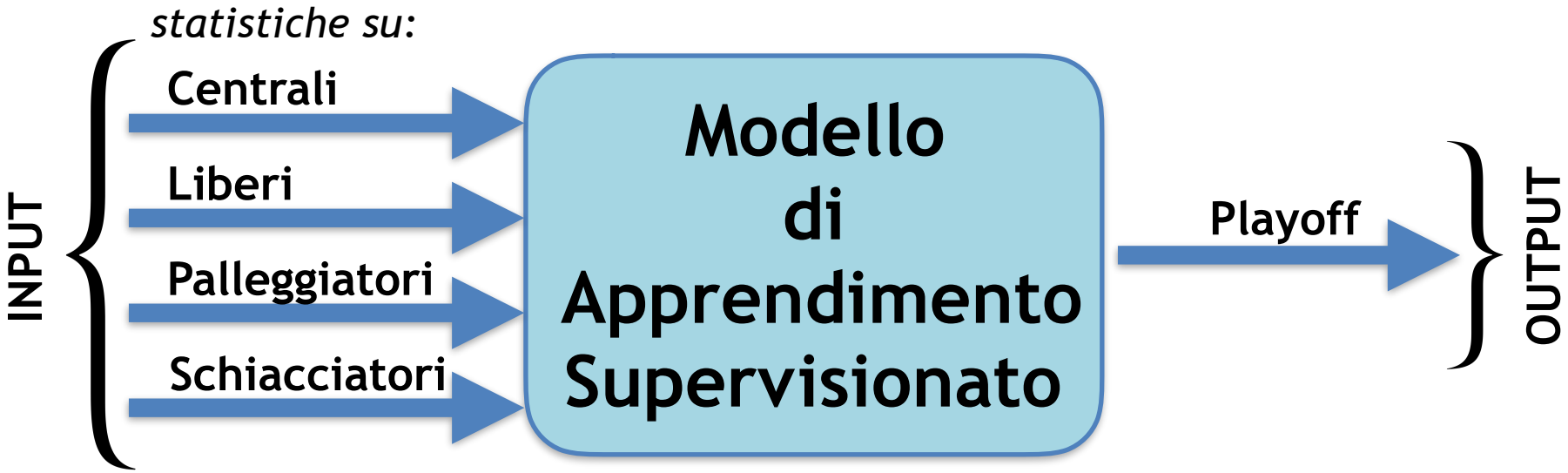
Il Dataset della Lega Pallavolo Serie A maschile

La Lega Pallavolo Serie A ci ha fornito un dataset rappresentante:

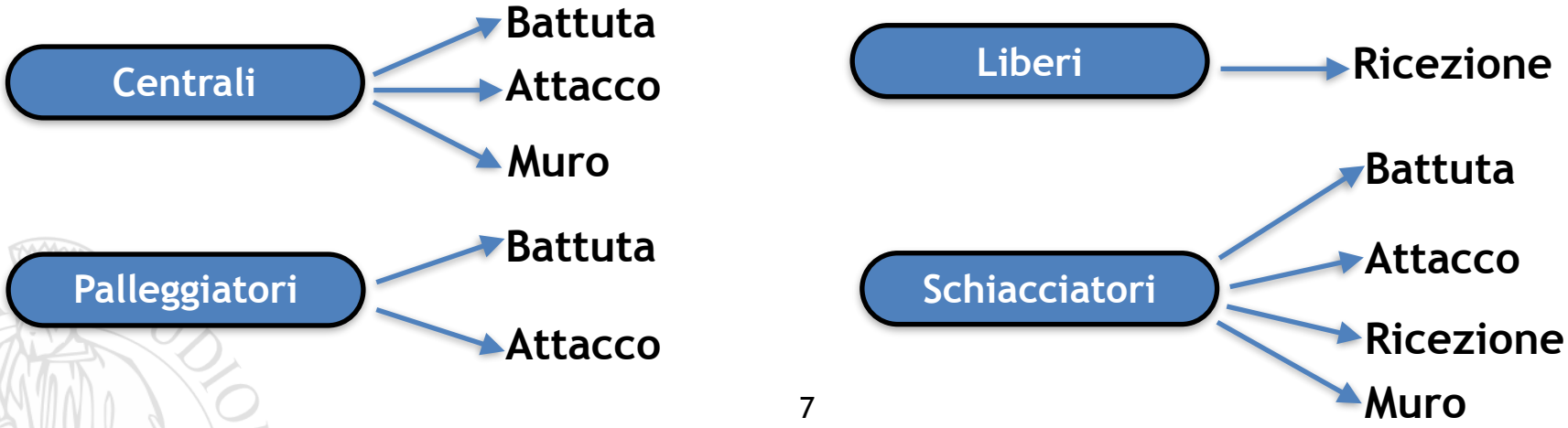
- Le statistiche (**battuta**, **ricezione**, **attacco**, **muro**), sia positive che negative, di tutti i giocatori (**Centrali**, **Liberi**, **Palleggiatori**, **Schiacciatori**) di ogni squadra per le stagioni che vanno dal 2001/2002 al 2017/2018.
- Informazioni relative ai **Playoff** $\begin{cases} \mathbf{1} & \text{se Squadra } i \text{ Sì Playoff} \\ \mathbf{0} & \text{se Squadra } i \text{ No Playoff} \end{cases}$ e **Posizione** in classifica.

Stagione	Squadra	C1_bat_pos	C1_bat_neg	C1_ric_pos	C1_ric_neg	C1_att_pos	C1_att_neg	C1_mur_pos	C1_mur_neg	C2_bat_pos		Playoff	Posizione
2001	Asystel Milano	0,305479608	0,633914331	0,563636364	0,375757576	0,898312572	0,041081368	0,832374377	0,107019563	0,301570681		1	5
2001	Borgocanale Taranto	0,256178288	0,525639894	0,706158358	0,075659824	0,718582888	0,063235294	0,675206612	0,10661157	0,223145933		0	10
2001	Bossini Sangemini Montichiari	0,257604256	0,293910896	0,472727273	0,078787879	0,503257576	0,048257576	0,456426332	0,095088819	0,159543777		1	7
2001	Casa Modena Salumi	0,429498261	0,412925981	0,727548209	0,114876033	0,789885956	0,052538286	0,778171546	0,064252696	0,18103109		1	4
2001	Icom Latina	0,1904	0,427781818	0,449586777	0,168595041	0,577448908	0,040732911	0,581385281	0,036796537	0		0	11
2001	Itas Diatec Trentino	0,206336088	0,44214876	0,459343434	0,189141414	0,61003713	0,038447718	0,613990974	0,034493875	0,155515832		0	9
2001	Lube Banca Marche Macerata	0,325219485	0,395992637	0,495833333	0,225378788	0,672541365	0,048670757	0,662735463	0,058476658	0,208605779		1	1
2001	Maxicono Parma	0,258035947	0,451054962	0,607792208	0,101298701	0,671272727	0,037818182	0,671770335	0,037320574	0,162085976		1	3
2001	Noicom Brebanca Cuneo	0,35906895	0,525779534	0,767981704	0,116866781	0,823503947	0,061344538	0,660606061	0,224242424	0,341467629	...	1	6
2001	Roma Volley	0,053620955	0,12213662	0,043939394	0,131818182	0,160340245	0,015417331	0,164772727	0,010984848	0,017112299		0	14
2001	Sempre Volley Padova	0,274263764	0,489372599	0,624793388	0,138842975	0,726686217	0,036950147	0,617408124	0,14622824	0,074618526		0	13
2001	Sira Cucine Ancona	0,195854922	0,404145078	0,458823529	0,141176471	0,573248408	0,026751592	0,534782609	0,065217391	0		0	12
2001	Sisley Treviso	0,301237374	0,613914141	0,721561772	0,193589744	0,861899164	0,053252351	0,828583129	0,086568387	0,27199123		1	2
2001	Yahoo! Italia Volley Ferrara	0,21630094	0,38369906	0,56	0,04	0,564516129	0,035483871	0,506896552	0,093103448	0,205194805		1	8
2002	Asystel Milano	0,197083224	0,354431927	0,551515152	0	0,507974482	0,04354067	0,506797707	0,044717445	0,108835187		1	4
2002	Bossini Gabeca Montichiari	0,286142164	0,435069957	0,607336523	0,113875598	0,681404959	0,039807163	0,655146889	0,066065232	0,237684538		0	10
2002	Canadiens Verona	0,097466977	0,29041181	0,387878788	0	0,358787879	0,029090909	0,349090909	0,038787879	0		0	13

Struttura Modello di Apprendimento



- **Possibili azioni** per ciascun ruolo:



Baseline

- È stata eseguita un **analisi di base** per capire quali **risultati** aspettarsi di **ottenere** e possibilmente **migliorare**.

- Si è osservato quante **squadre** entrate ai **Playoff** in una **determinata stagione**, fossero riuscite a **rientrare** ai **Playoff** anche nella **stagione successiva**.

Il risultato è che il **72%** delle squadre che partecipa in una stagione ai Playoff, è riuscita ad accedervi anche nella stagione precedente.

Analisi dei Risultati per la Regressione Logistica

- Divisione del Dataset:
  - **Training Set** → Stagioni dal 2001/02 al 2014/15
  - **Test Set** → Stagioni dal 2015/16 al 2017/18
- Viene utilizzata come **metrica: F1_score**
- **Cross-Validation** per individuare il valore ottimale dell'**iperparametro C** tra $[2^{-8}, 2^{-7}, \ldots, 2^{7}]$.
- **L'esperimento** è stato eseguito **50 volte**

Classificatore	F1_score Medio
Regressione Logistica	**75,9%**

Analisi dei Risultati per SVM Modello Lineare

- Divisione del Dataset:
  - **Training Set** → Stagioni dal 2001/02 al 2014/15
  - **Test Set** → Stagioni dal 2015/16 al 2017/18
- Viene utilizzata come **metrica: F1_score**
- **Cross-Validation** per individuare il valore ottimale dell'**iperparametro C** tra $[2^{-8}, 2^{-7}, \ldots, 2^{7}]$.
- **L'esperimento** è stato eseguito **50 volte**

Classificatore	Tipo di Kernel	F1_score Medio
Support Vector Machine (SVM)	lineare	**73,3%**

Analisi dei Risultati per SVM Modello Non Lineare

- Divisione del Dataset:
  - **Training Set** → Stagioni dal 2001/02 al 2014/15
  - **Test Set** → Stagioni dal 2015/16 al 2017/18
- Viene utilizzata come **metrica: F1_score**
- **Cross-Validation** per individuare il valore ottimale dell'iperparametro **C** e γ (**parametro** del kernel '**gaussiano**'). Dove $C \in [2^{-8}, 2^{-7}, \ldots, 2^{7}]$ e $\gamma \in [\gamma_0 \cdot 2^i \mid i = -8, \ldots, 7]$ con $\gamma_0 = \frac{1}{numeroFeatures} = \frac{1}{20}$
- **L'esperimento** è stato eseguito **50 volte**

Classificatore	Tipo di Kernel	F1_score Medio
Support Vector Machine (SVM)	gaussiano	**82,5%**

Esempio di output per il test su una determinata Stagione

TEST SU STAGIONE: 2017

Squadra	Stagione	PlayoffProbability	PredictionPlayoff	RealPlayoff	Posizione	Errore
Azimut Modena	2017	0.982443	1	1	3	0
BCC Castellana Grotte	2017	0.067559	0	0	14	0
Bios Indexa Sora	2017	0.548391	0	0	13	0
Bunge Ravenna	2017	0.626912	0	1	8	1
Calzedonia Verona	2017	0.735027	1	1	5	0
Cucine Lube Civitanova	2017	0.980922	1	1	2	0
Diatec Trentino	2017	0.888063	1	1	4	0
Gi Group Monza	2017	0.652279	0	0	10	0
Kioene Padova	2017	0.571705	0	0	9	0
Revivre Milano	2017	0.739374	1	1	6	0
Sir Safety Conad Perugia	2017	0.987725	1	1	1	0
Taiwan Excellence Latina	2017	0.519721	0	0	11	0
Tonno Callipo Calabria Vibo Valentia	2017	0.687317	1	0	12	4
Wixo LPR Piacenza	2017	0.929060	1	1	7	0

Numero Squadre Playoff= 8
Numero Squadre Stagione=14

f1 del C_max e gamma_max=0.875
accuracy del C_max e gamma_max=0.8571428571428571
balanced_accuracy del C_max e gamma_max=0.8541666666666667
precision del C_max e gamma_max=0.875
recall del C_max e gamma_max=0.875
errore medio in questo anno=0.35714285714285715

C_max e gamma_max=[(16, 0.025)]

Sviluppi futuri

- **Raffinare** i **modelli** di **apprendimento automatico** proposti, con lo scopo di renderli più robusti.
- **Rendere completo** il Dataset finale
  - **Integrando dati** relativi alle **squadre** nel loro complesso.
  - **Comprendendo** anche i **giocatori provenienti** da **altri campionati**.
- **Utilizzare** queste **tecniche** di predizione **per altri Sport**.

Metodi di Machine Learning per l'Analisi Predittiva di Prestazioni di Squadre in Ambito Sportivo

Candidato
Mirco Ceccarelli

Relatore
Prof. Marco Sciandrone

Correlatore
Dott. Matteo Lapucci