Analisi dell'abbandono sportivo

Giorgio Pilotti 755626 Matteo Provasi 782922 Nicolò Monti 769709

INTRODUZIONE

Il dataset analizzato raccoglieva le risposte di 3500 ragazzi riguardo le loro abitudini sportive presenti e passate.

Gli obiettivi principali sono stati:

- *Effettuare un'analisi di classificazione per individuare quali fossero gli aspetti più rilevanti che determinassero l'abbandono sportivo dei ragazzi prima dei 17 anni d'età.*

- *Su richiesta dei fornitori dei dati*
  - *Età in cui i ragazzi hanno abbandonato o pensano di abbandonare il proprio sport*
  - *Cause di abbandono e fattori di protezione*
  - *Eventuali relazioni con le variabili sociodemografiche rilevate*
  - *Differenze fra ragazzi e ragazze*
  - *Eventuali miglioramenti per un futuro questionario*

INTRODUZIONE

Il questionario era composto da 210 domande sia a livello sociodemografico che sulle abitudini sportivi dei ragazzi.

In base all'abitudine sportiva il rispondente veniva classificato in uno dei seguenti casi:

- *Ragazzo che ha praticato uno sport ed ha pensato di abbandonarlo*
- *Ragazzo che ha praticato uno sport e non ha pensato di abbandonarlo*
- *Ragazzo che ha praticato più sport, continua a praticare il più significativo ed ha pensato di abbandonarlo*
- *Ragazzo che ha praticato più sport, continua a praticare il più significativo e non ha mai pensato di abbandonarlo*
- *Ragazzo che ha abbandonato il suo sport preferito prima dei 17 anni*
- ~~*Ragazzo che non ha mai praticato sport*~~

PRE-PROCESSING

È stata necessaria una procedura di pulizia dei dati sul dataset originale prima di procedere con le analisi

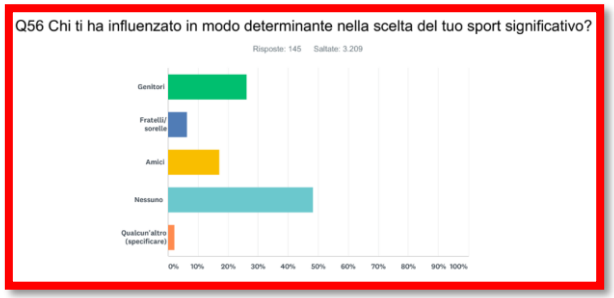
- Ad ogni domanda non era associata una colonna, ma era presente una variabile per ogni possibile risposta
- Due righe di intestazione
- A fronte delle 210 domande nel questionario risultavano 2170 variabili

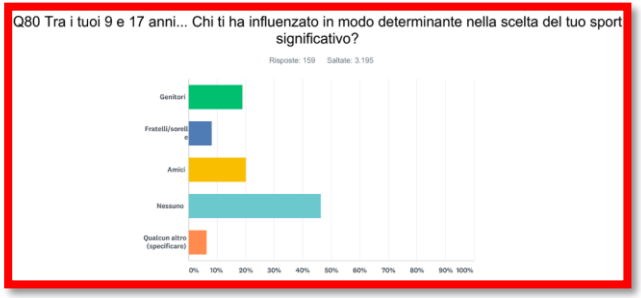
Sei.maschio.o.femmina.	X
Maschio	Femmina
	Femmina
Maschio	
	Femmina
	Femmina
Maschio	
Maschio	
Maschio	
	Femmina

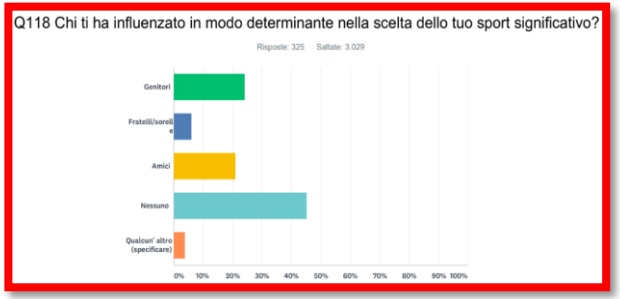
PRE-PROCESSING

A seconda di come un rispondente veniva classificato, rispondeva solo a delle determinate domande nel questionario

Tuttavia analizzando il dataset si è notato che la maggior parte delle domande erano in realtà identiche fra i diversi casi



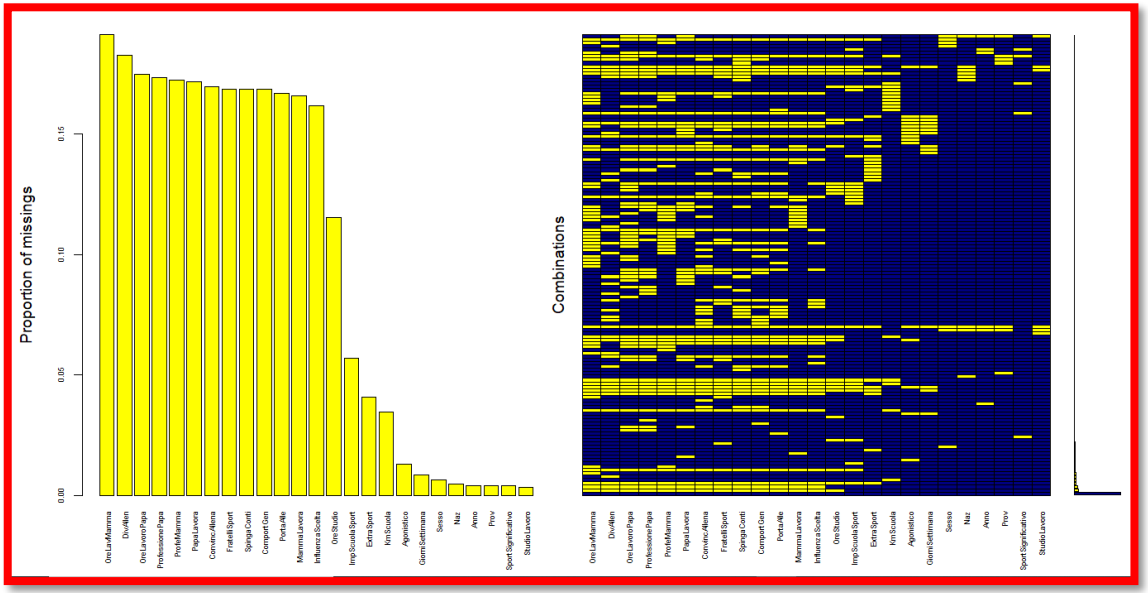




Nei casi in cui era possibile si sono raggruppate le osservazioni all’interno di un’unica variabile

PRE-PROCESSING

Successivamente si sono valutate le proporzioni di dati mancanti a livello di variabili e di osservazioni



- Alcune persone rispondevano solo alla prima parte del questionario
- Diverse persone hanno svolto il questionario più di una volta.
- Alcune domande erano lasciate vuote in quanto non c’era l’obbligo di risposta forzata
- Sono stati rimossi i record con un elevato numero di missing
- Imputazione dei valori mancanti con moda o mediana

DATASET SISTEMATO

Per alcune variabili è stato necessario modificare i valori per uniformarli a livello di formato (anno di nascita, risposte a domande aperte)

Sono state rimosse le variabili a varianza nulla o che erano classificate come near zero variance

Alla fine del pre processing il dataset finale è stato ridotto a 2225 osservazioni e 65 variabili.

MODELLI CLASSIFICATIVI

Per svolgere il compito di classificazione sono stati implementati i seguenti modelli

- Alberi classificativi (valutati per accuracy e con matrice di costi)
- Random forest
- Modello logistico
- Naive Bayes

ALBERI CLASSIFICATIVI

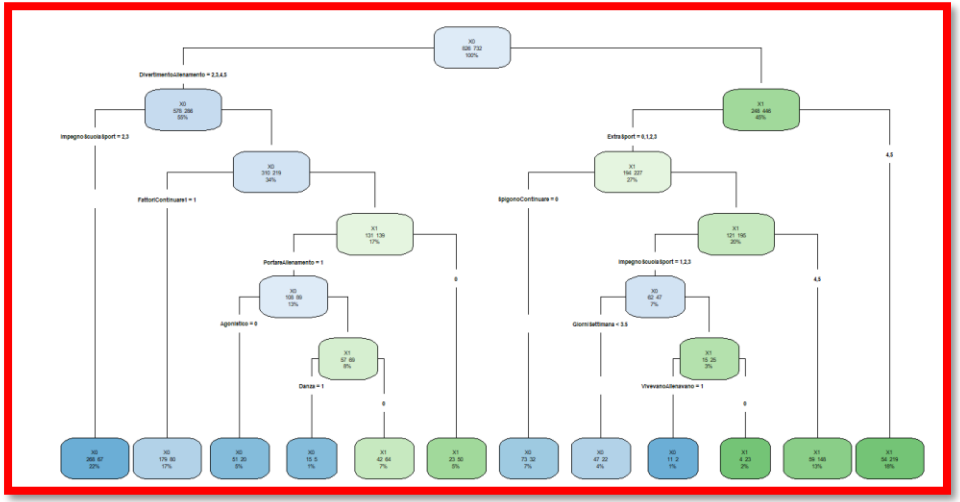
Albero con tuning sul cp per massimizzare l'accuracy

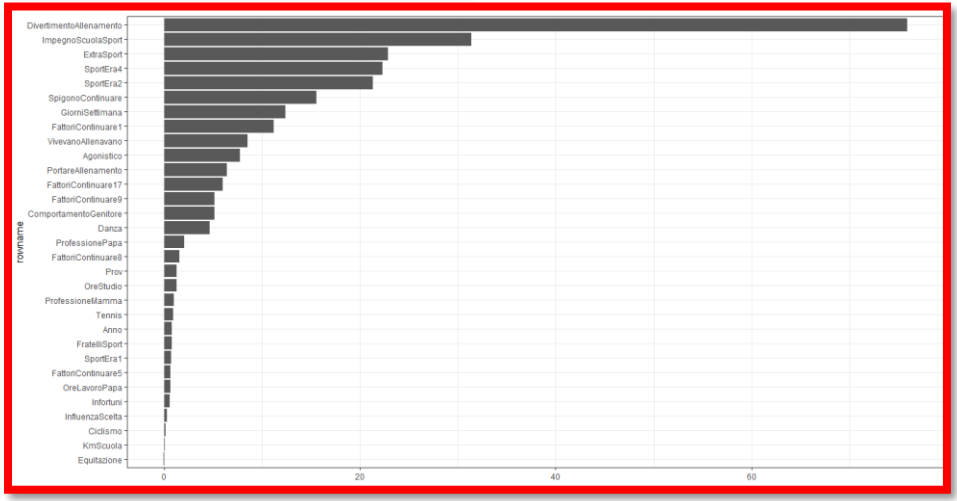
Albero Post-pruning : 12 foglie

- Train: Accuracy 0.7368, Sensitivity 0.7797

```
Confusion Matrix and Statistics

          Reference
Prediction  X0  X1
        X0 275 112
        X1  79 201

               Accuracy : 0.7136
                 95% CI : (0.6777, 0.7477)
    No Information Rate : 0.5307
    P-Value [Acc > NIR] : < 0.0000000000000002

                  Kappa : 0.4216
 Mcnemar's Test P-Value : 0.02059

            Sensitivity : 0.7768
            Specificity : 0.6422
```

Non ci limitiamo a vedere l'accuracy, ma anche il NIR

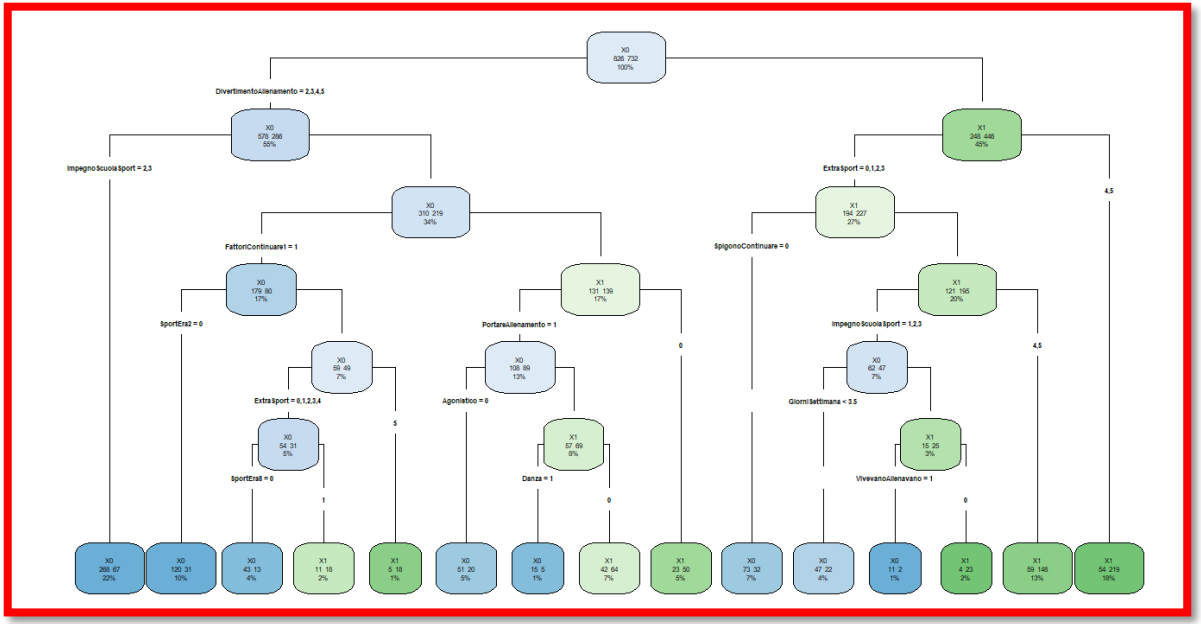




ALBERI CLASSIFICATIVI

Albero con matrice dei costi

- -5 per abbandono classificato correttamente
- +2 per abbandono classificato erroneamente




```
Confusion Matrix and Statistics

          Reference
Prediction  X0  X1
        X0 263 101
        X1  91 212

               Accuracy : 0.7121
                 95% CI : (0.6761, 0.7463)
    No Information Rate : 0.5307
    P-Value [Acc > NIR] : <0.0000000000000002

                  Kappa : 0.421
 Mcnemar's Test P-Value : 0.516

            Sensitivity : 0.7429
            Specificity : 0.6773
```

Post- pruning : 15 foglie

- Accuracy sul train 0.7497
- Sensitivity sul train 0.7603

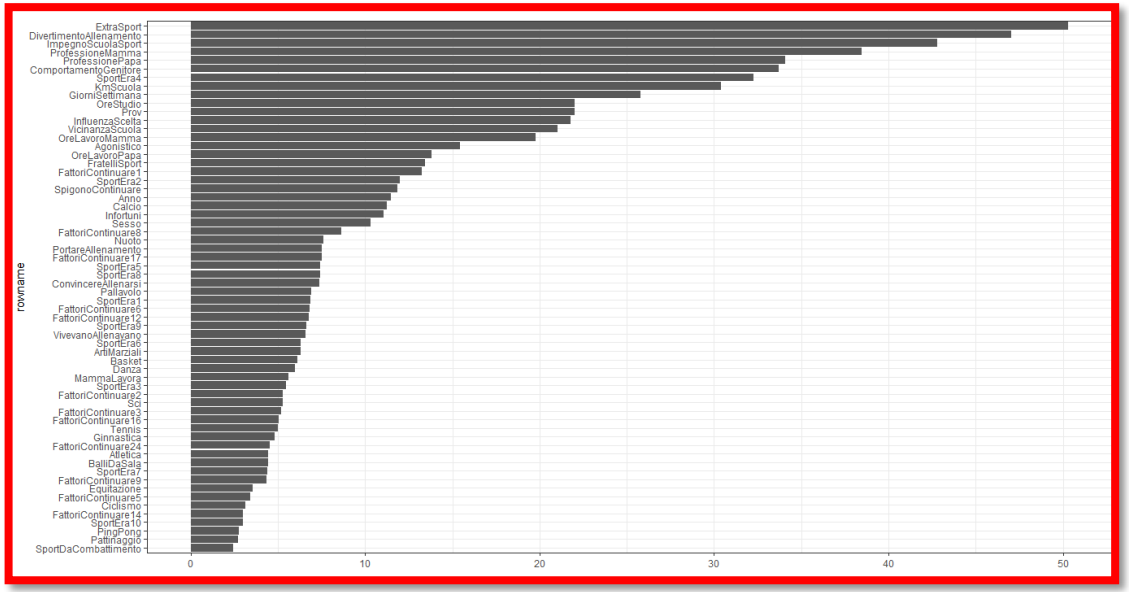
RANDOM FOREST

Random forest creata utilizzando 500 alberi

```
Confusion Matrix and Statistics

          Reference
Prediction  X0  X1
        X0 297 117
        X1  57 196

               Accuracy : 0.7391
                 95% CI : (0.704, 0.7721)
    No Information Rate : 0.5307
    P-Value [Acc > NIR] : < 0.00000000000000022

                  Kappa : 0.4704
 Mcnemar's Test P-Value : 0.000007721

            Sensitivity : 0.8390
            Specificity : 0.6262
```

Sensitivity molto buona

Da notare che la variabile più importante è cambiata rispetto a quella degli alberi di classificazione

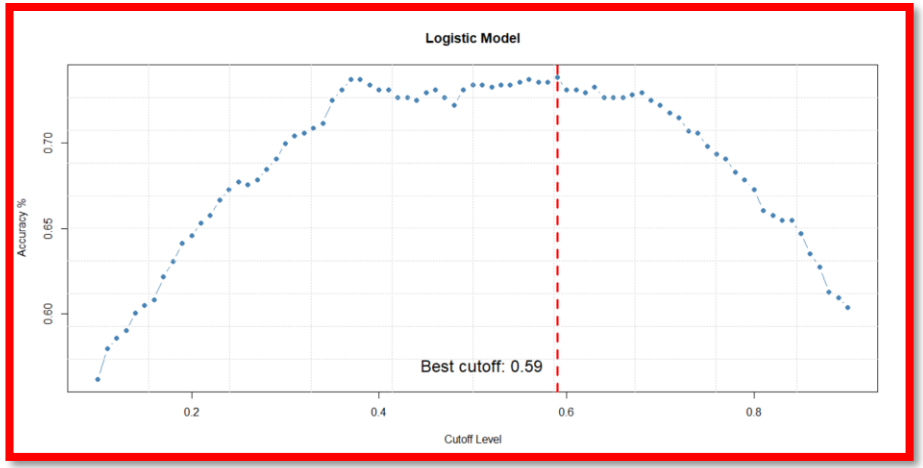


MODELLO LOGISTICO

Variabili con VIF più elevato, situazione borderline per la prima

	Name	GVIF
1	ProfessioneMamma	6.194581
2	ComportamentoGenitore	3.077717
3	MammaLavora	2.903735
4	ProfessionePapa	2.592226
5	Sesso	2.571540

Scelta del cutoff migliore per l'accuracy

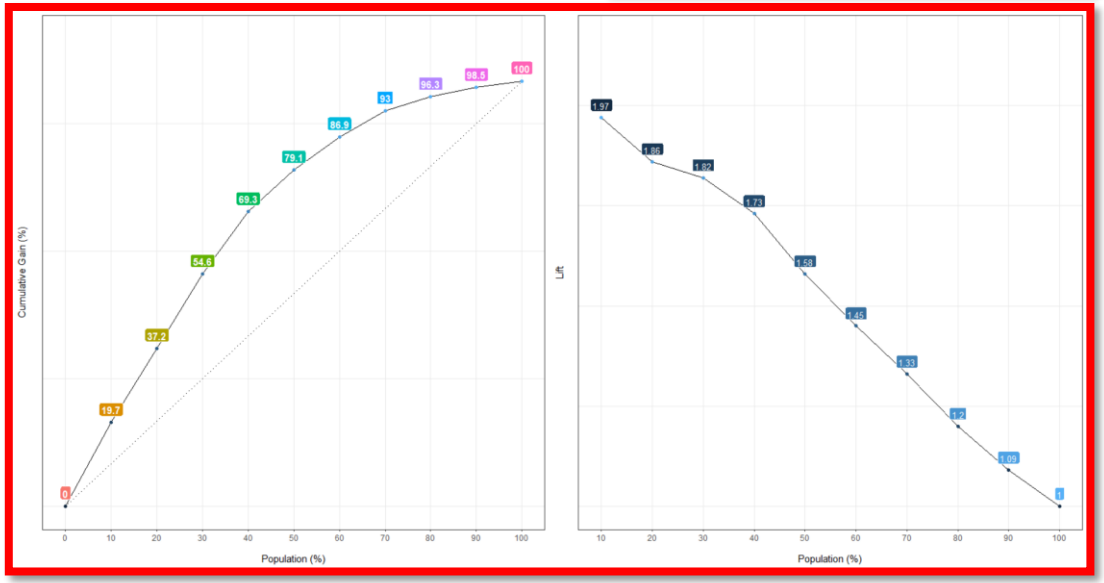



```
Confusion Matrix and Statistics

          Reference
Prediction  X0  X1
        X0 276  98
        X1  78 215

               Accuracy : 0.7361
                 95% CI : (0.7009, 0.7692)
    No Information Rate : 0.5307
    P-Value [Acc > NIR] : <0.0000000000000002

                  Kappa : 0.4683
 Mcnemar's Test P-Value : 0.1521

            Sensitivity : 0.7797
            Specificity : 0.6869
```

MODELLO LOGISTICO

Odds ratio più elevati

Variable	Odds
DivertimentoAllenamento2	5.09900177
FratelliSport	3.29355729
FattoriContinuare1	2.27124930
ImpegnoScuolaSport2	2.21334174



Informazioni simili a quelle ricavate con la random forest

Cumulative gain e Lift non eccezionali, ma in linea con i livelli di accuratezza ottenuti

NAIVE BAYES

Proviamo a vedere quali sono le performance del Naive Bayes supponendo indipendenza fra le variabili

Dopo aver considerato solo le variabili factor ed eliminato quelle con più livelli, si procede con il Naive Bayes tree con metodo di ricampionamento Cross - Validation.

Discreta l'accuratezza
pessima la sensitivity

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 173  72
         1  88 223

               Accuracy : 0.7122
                 95% CI : (0.6726, 0.7495)
    No Information Rate : 0.5306
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.4203
 Mcnemar's Test P-Value : 0.2357

            Sensitivity : 0.6628
            Specificity : 0.7559
```

CONFRONTO

Confronto dei modelli in termini di accuracy e sensitivity

	Tree Acc	Tree Cost	RF	Logistic	N. Bayes
Accuracy	0.7136	0.7121	0.7391	0.7361	0.7122
Sensitivity	0.7768	0.7429	0.8390	0.7797	0.6628

ANALISI DELL'ABBANDONO

Passando al secondo obiettivo dell'analisi, si è cercato di dare delle risposte ai quesiti posti dai fornitori di dati.

Età in cui i ragazzi hanno abbandonato o pensano di abbandonare il proprio sport

Domanda presente nel questionario ma differenziata per casi. Raggruppando i risultati si ottiene che l'età media è di 14.58 anni.
La differenza fra ragazze e ragazzi è di circa un anno, 14.14 contro 14.93 anni.

Ragazzi che non praticano sport

Il 4.34% dei rispondenti non ha mai praticato sport, in proporzione le femmine sono 1.5 volte più dei maschi.

ANALISI DELL'ABBANDONO

Cause di abbandono e fattori di protezione

Anche in questo caso le domande inerenti all'abbandono e ai fattori di protezioni erano divise fra i casi

Fattori di protezione

- Allenatore
- Famiglia

Compagni di squadra

Cause di abbandono

- Scuola
- Orari scomodi

ANALISI DELL'ABBANDONO

Eventuali relazioni con le variabili sociodemografiche rilevate

Non sono emerse grandi relazioni fra le variabili sociodemografiche e l'abbandono

I fenomeni più rilevanti sono stati:

	Centro	Nord-Est	Nord-Ovest	Sud e Isole
Abbandona	202	204	222	552
Non abbandona	174	196	246	429

Forte abbandono al Sud e Isole

	Femmina	Maschio
Abbandona	708	472
Non abbandona	449	596

Notevole differenza fra maschi e femmine

MIGLIORAMENTI QUESTIONARIO

I seguenti miglioramenti sono stati i ndividuati per un futuro questionario:

- *Metodo migliore per scaricare i dati*
- *Obbligo di risposta ad alcune domande*
- *Standardizzare alcune risposte (data, provincia di residenza…)*
- *Evitare domande ridondanti dopo la suddivisione dei casi*
- *Campionamento?*

CONCLUSIONI

- Non essendo un caso didattico i risultati ottenuti con i modelli sono da considerarsi accettabili

- Sono stati individuati alcuni elementi importanti che portano all’abbandono sportivo o al contrari o sono da considerarsi fattori di protezione

- I miglioramenti proposti sono più incentrati sulla qualità della raccolta dati

Grazie dell’attenzione!