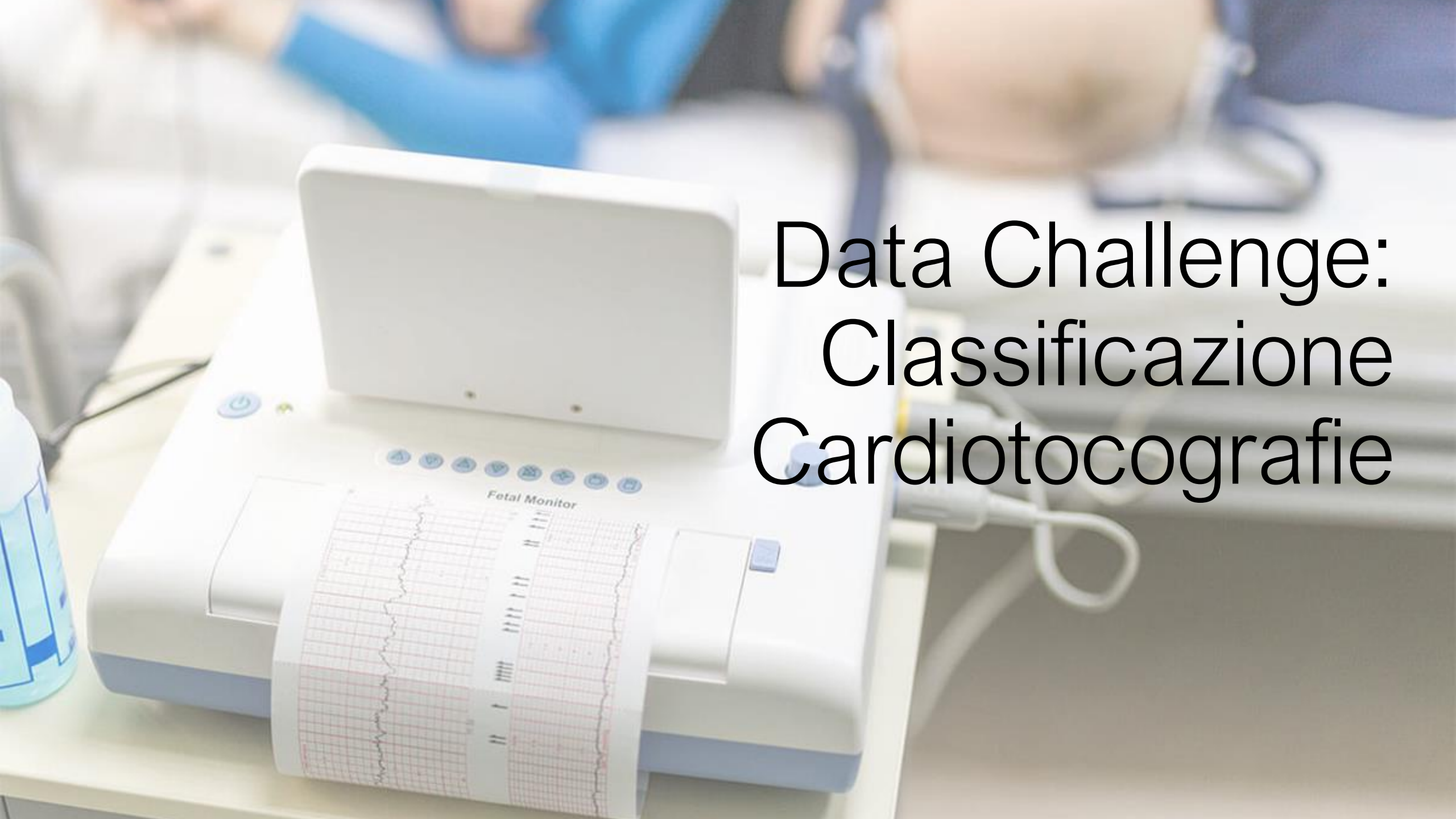


Data Challenge: Classificazione Cardiotocografie

Introduzione

Obiettivo

Sviluppare un modello di machine learning per la classificazione di condizioni di salute di un feto:

- Analisi cardiotocografie

Descrizione:

Dati provenienti da UCI repository, con circa 20 features rilevanti misurati.

Classificazione a 3 classi:

- N=normal;
- S=suspect;
- P=pathologic

Step Analisi

Data Quality:

Assessment stato di pulizia dei dati

Analisi Esplorativa:

Prima analisi e visualizzazione

Preprocessing:

Preparazione dati per fase di modellazione

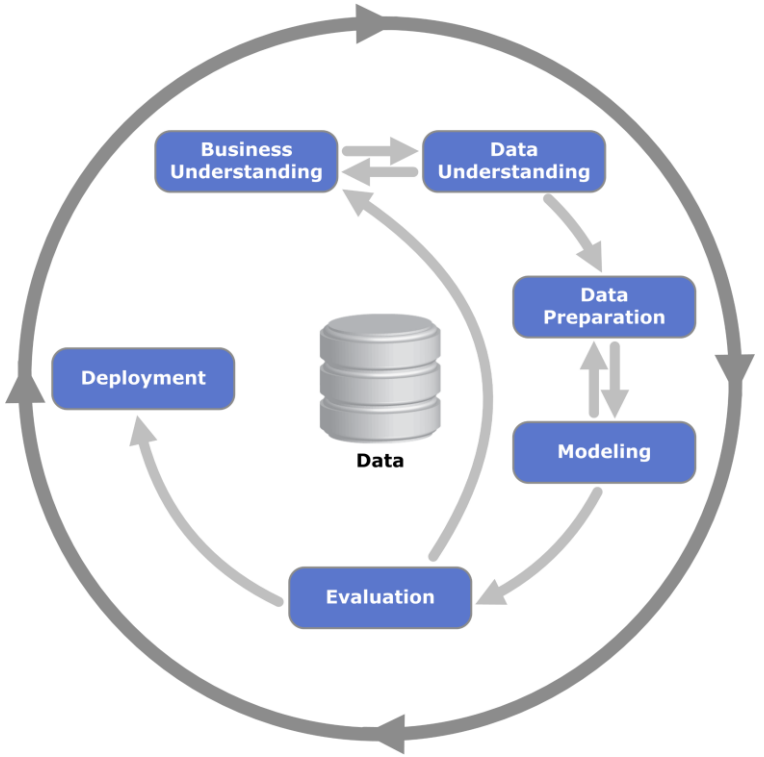
Modellazione:

Confronto modelli di classificazione (baseline)

Ottimizzazione:

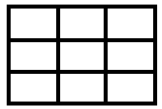
Incremento performance del modello migliore

Metodologia



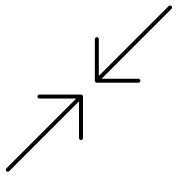
Data Quality & Esplorazione

Checks



No Missing Values &
identificati i corretti datatypes

No typos o record errati



No evidenti outlier
(i.e. valori min o max fuori scala)

Class:

- Notato sbilanciamento di classe (N~78%; S~14%; P~8%)

Correlazione:

- Notate features ad elevata correlazione (rischio collinearità) e features nulle

Distribuzione:

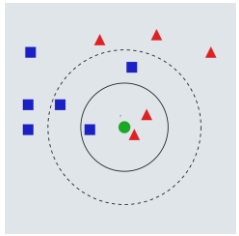
- Differenza tra distribuzioni condizionate alla classe non particolarmente informative. Notata forte similarità tra S e P
- Boxplot suggeriscono la presenza di potenziali outlier, tuttavia sono stati ignorati a causa della poca domain expertise

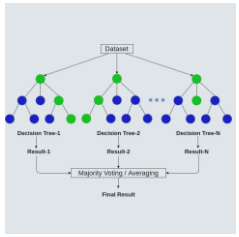
Preprocessing

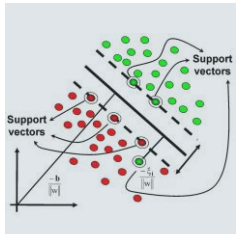
Target	*Features*	*Scala*
Ridotto da multi-class a binary: Normal - Suspect	Eliminate variabili LBE, ridondante con LB, e DR, nulla	Standardizzati i dati
Mitigazione class imbalance e maggiore rilevanza alla funzione di supporto all'utilizzatore del modello	Tralasciate altre variabili ad elevata correlazione per valutazione successiva	Standardizzazione per ridurre impatto della differenza di scala tra le diverse variabili

Modellazione (baseline)

Models *Input*







Approccio 1:
Dati non standardizzati

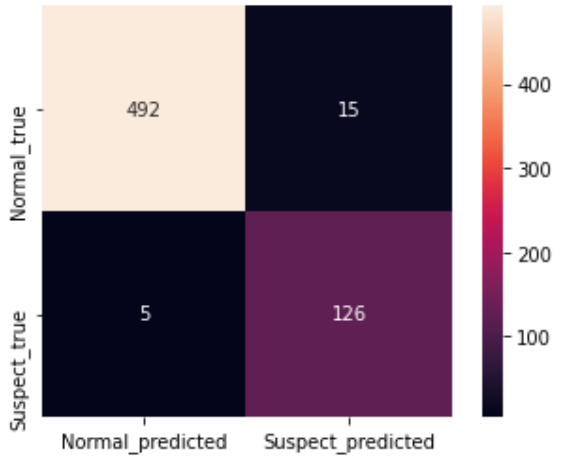
Approccio 2:
Dati standardizzati

Approccio 3:
Dati standardizzati
Classificazione multi-class

Risultati:

- Elevati livelli di accuracy in entrambi i casi, leggermente migliori con dati standardizzati (overall >90%)
- Knn e SVC hanno restituito performance più basse sia in termini di Accuracy che di Recall
- Il miglior modello baseline risulta essere il RFC

- Riduzione delle performance nella previsione multi-class



Confusion Matrix RFC:

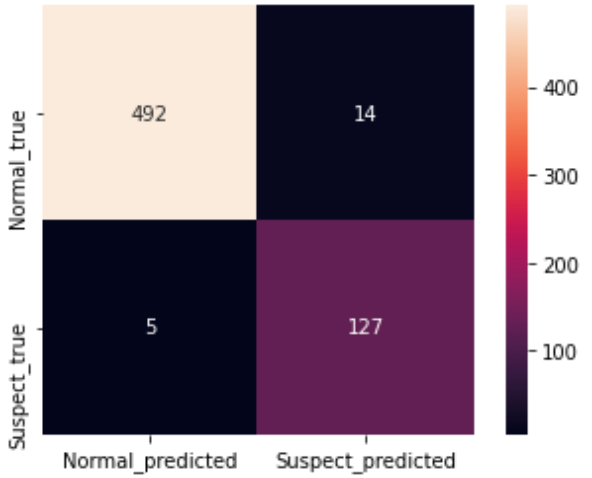
- Dati standardized
- Accuracy: 96.8 %
- Recall: 96.1 %

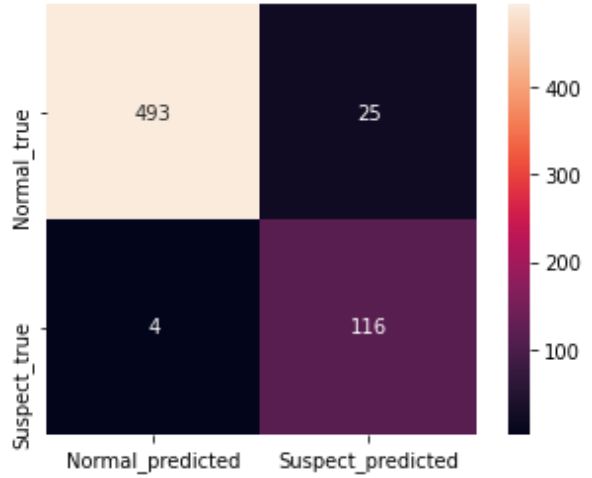
Ottimizzazione

- GridSearchCV
- Feature exclusion & Class weight

Slight increase in performance

Slight decrease in performance





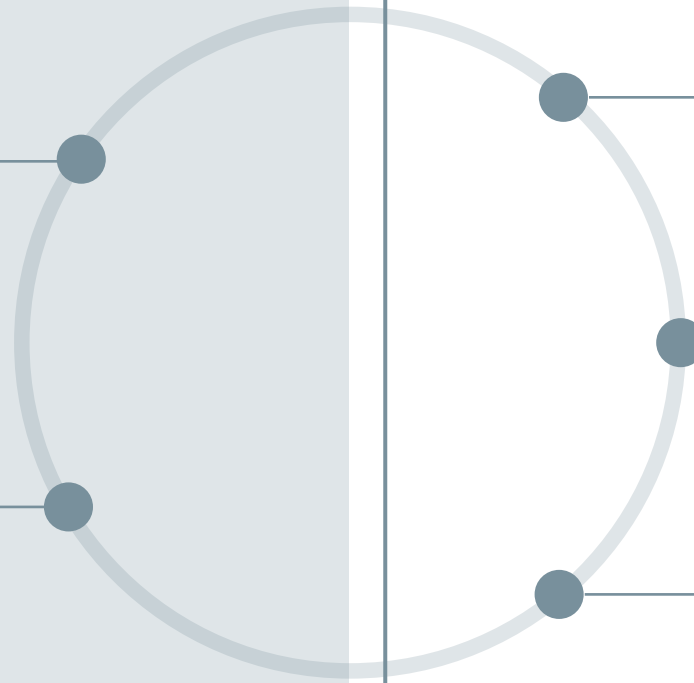
Conclusioni

Evaluation

Dei modelli utilizzati, il Random Forest Classifier è quello che ha performato meglio «overall»

Dei tre modelli RFC (baseline, ottimizzato, weighted) quello ottimizzato è risultato il migliore
F1 measure: 93%

In funzione delle necessità di utilizzo tuttavia, la scelta del modello migliore potrebbe cambiare in funzione dei tradeoff



Next Steps

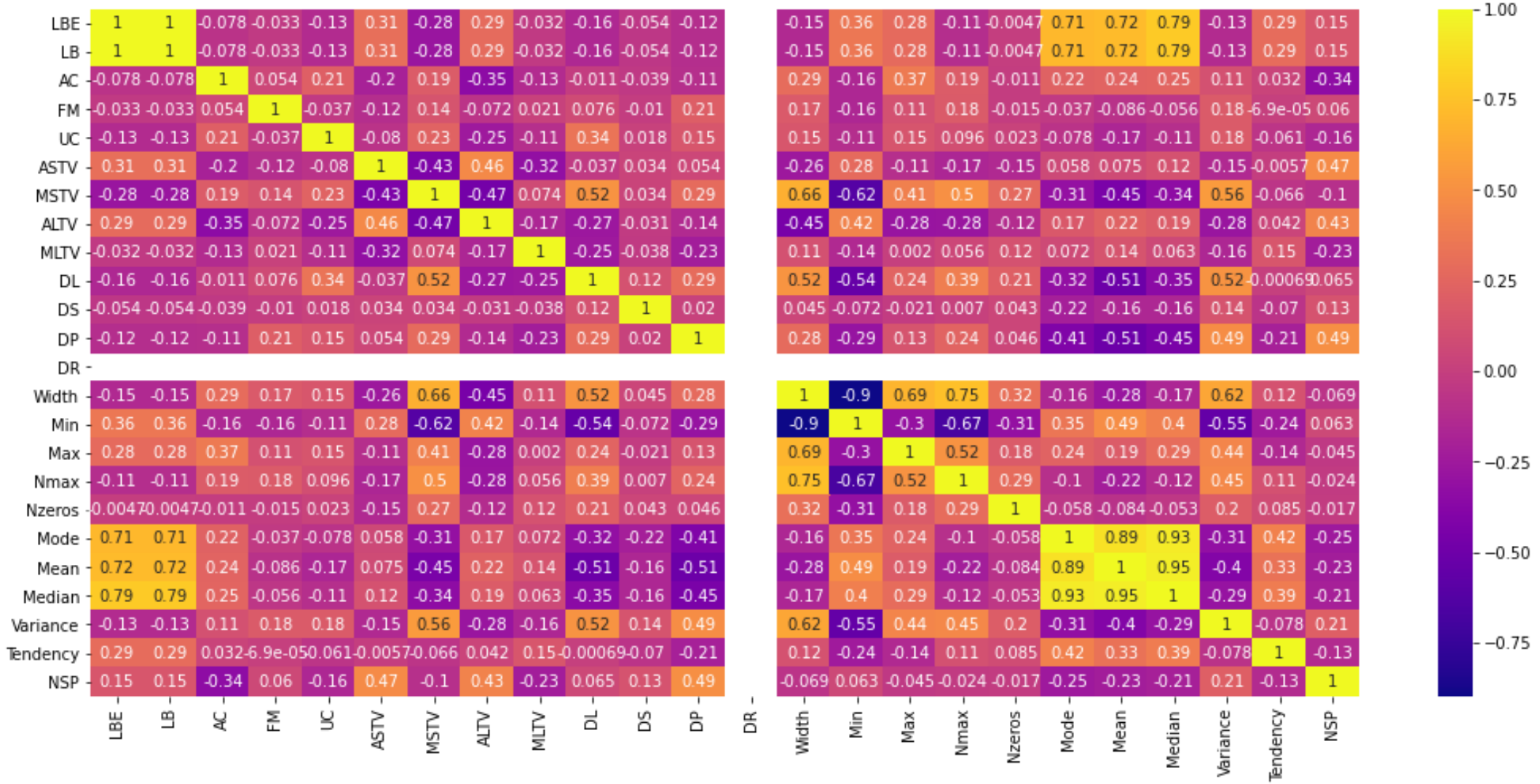
Ottenere più dati

Testare altri modelli

Continuare con il fine-tuning degli iperparametri

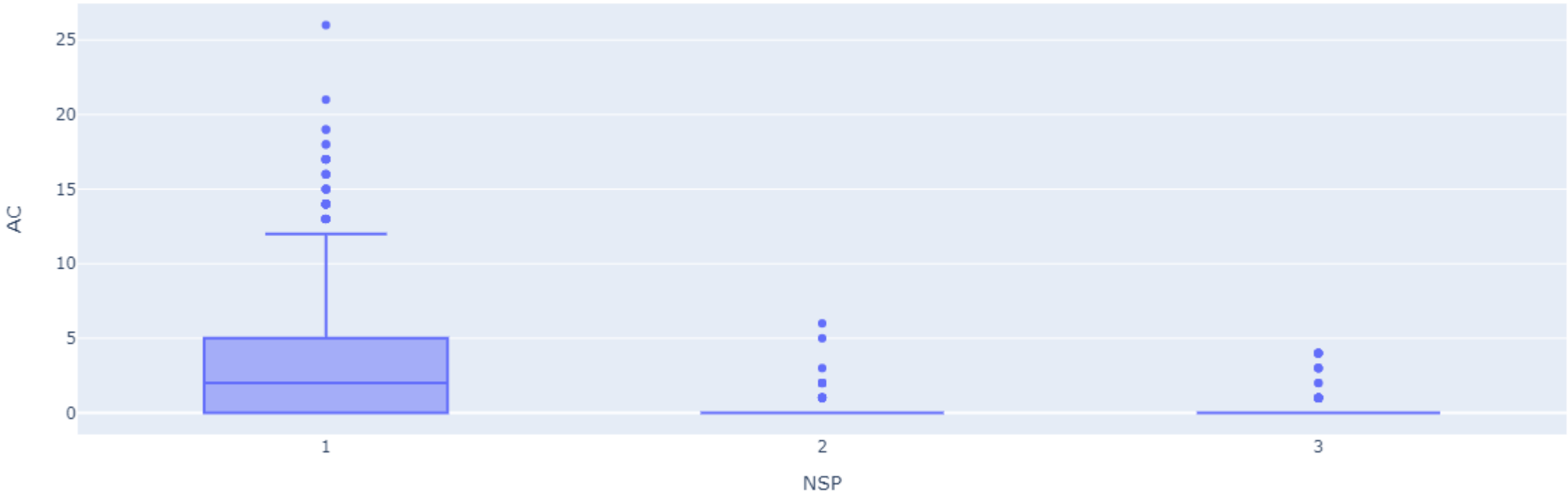
Backup

Data Understanding

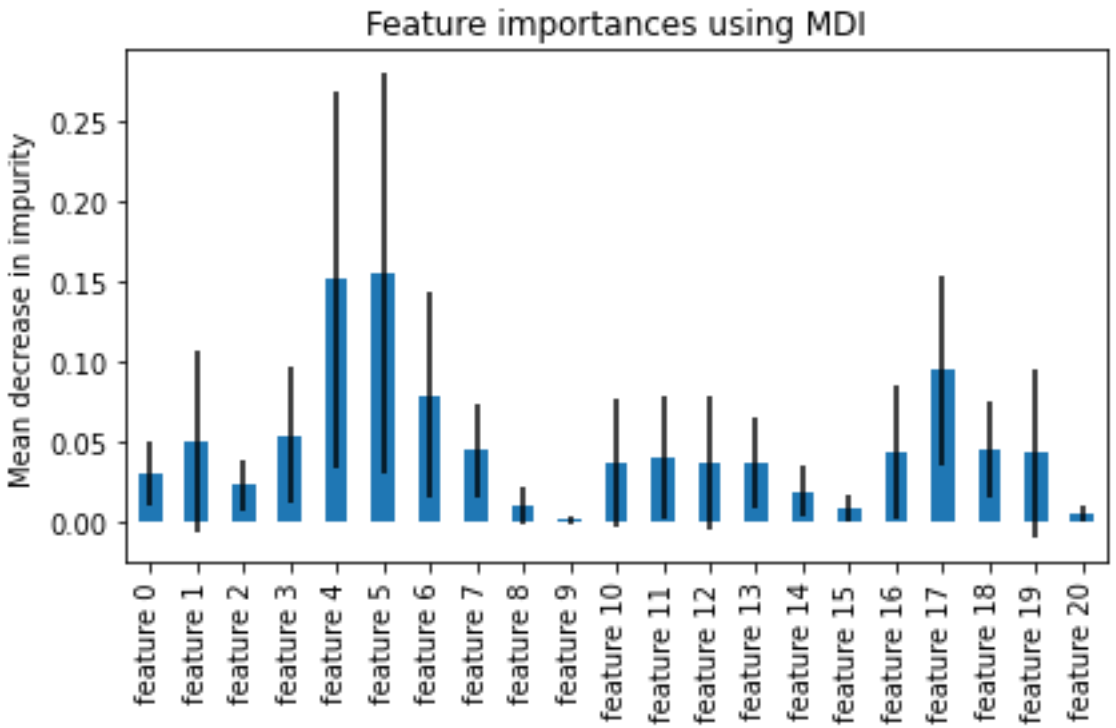


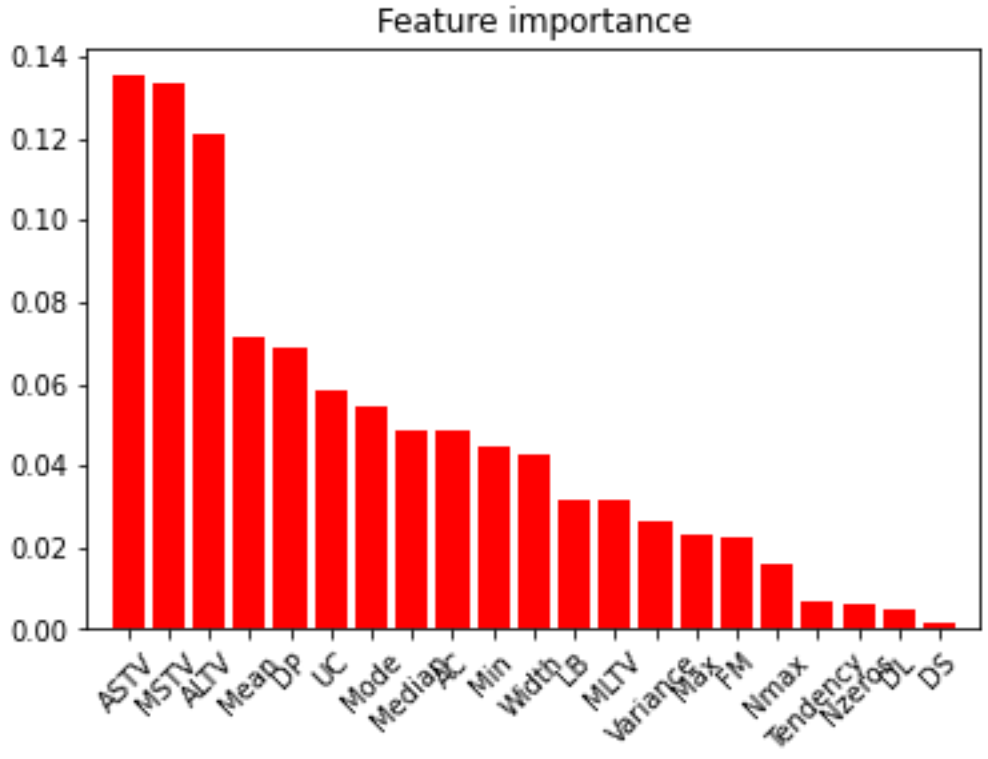
Data Understanding

Data Understanding



Data Understanding





Data Understanding

```
# confonting the 3 available models:

print("Baseline: {};".format(f1_score(rfc2_pred, y_test2)),
        "GridSearch: {};".format(f1_score(final_pred, y_test2)),
        "Tuned: {};".format(f1_score(final_pred2, y_test4)) )
```

✓ 0.1s

```
Baseline: 0.9264705882352942; GridSearch: 0.9304029304029305; Tuned: 0.9090909090909092;
```