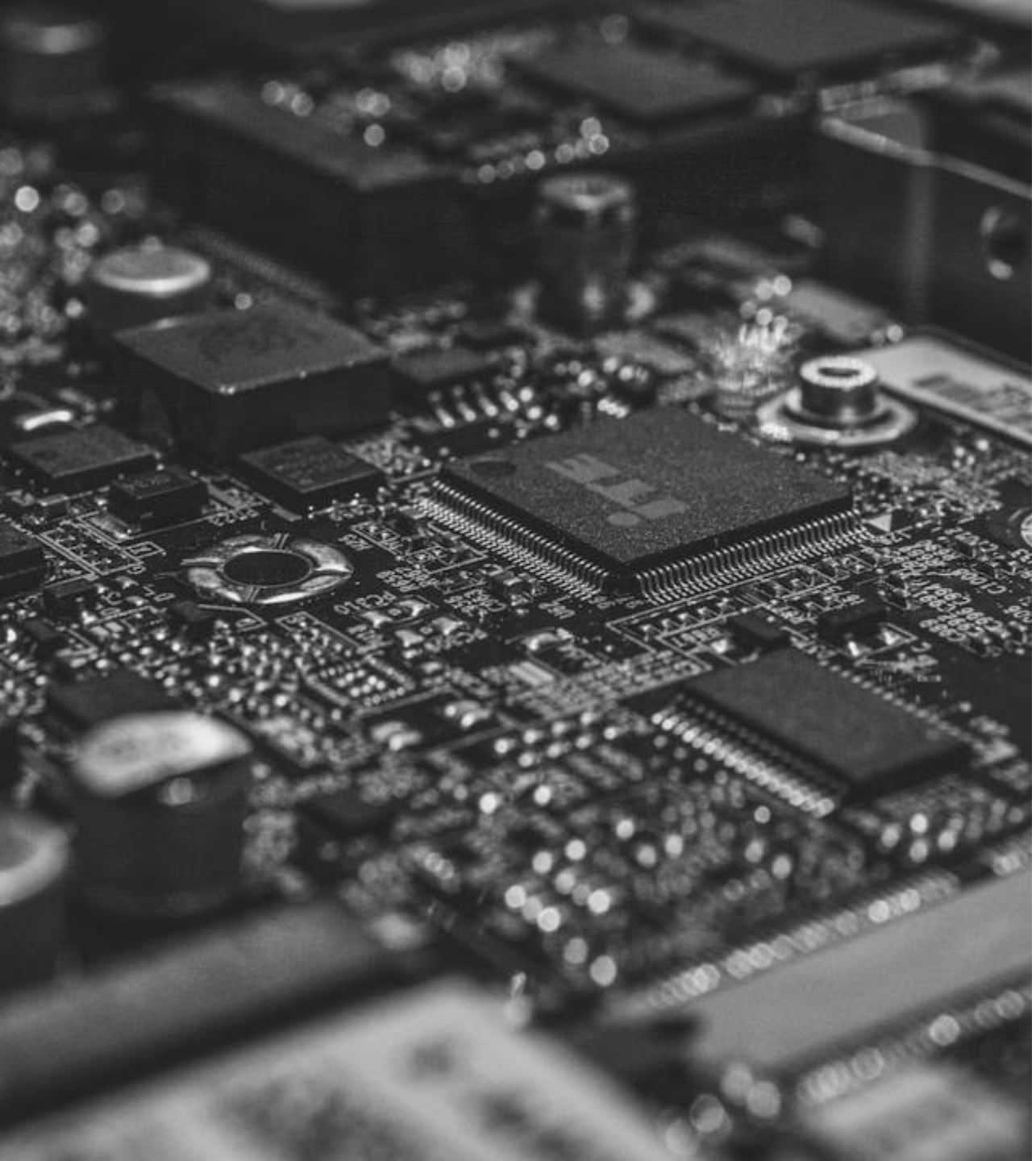


Coronavirus Diagnosis-AI

Emmanuel De Luca
05121 13925

NEXT →

Machine Learning in Medicina

Utilizzo medico

Il Machine Learning in medicina si riferisce all'applicazione di tecnologie di intelligenza artificiale di supporto ai medici nel diagnosticare ai pazienti malattie e condizioni cliniche.

Precisione nella diagnosi

L'utilizzo del Machine Learning consente una maggiore precisione nelle diagnosi mediche rapidamente.

Vantaggi nell’utilizzo

- Precisione.
- Rapidità.
- Accuratezza.

Covid-19: Overview

Panoramica sul Coronavirus e sul Covid-19

Il Coronavirus è una famiglia di virus respiratori che possono causare malattie come il SARS o la MERS. Della famiglia dei coronavirus fa parte il SARS Covid-19. Il Covid-19 è stata una pandemia globale causata dal virus SARS-CoV 2, diffusassi inizialmente nel 2019 principalmente nella cittadina di Whuan in Cina, successivamente anche in tutto il mondo.

NEXT

Applicazioni del Machine Learning al Coronavirus

- Prevenzione.
- Monitoraggio.
- Diagnosi.

Diagnosi di precisione

Diagnosi precise

L'utilizzo di modelli di Machine Learning per la medicina di precisione ha permesso lo sviluppo di modelli capaci di classificare e distinguere una malattia simile al Covid-19 dal Covid-19.

Analisi Genomica

Il Machine Learning applicato al sequenziamento del genoma del Coronavirus ha contribuito a identificare rapidamente le varianti del virus e a comprendere la loro diffusione.

NEXT →

I problemi degli approcci di Machine Learning in ambito clinico

- Dati: mancanti, artificiali, sporchi.
- Explainability.
- Interpretability.

NEXT →

Scopo del lavoro progettuale

Partendo da uno studio definito in letteratura che si propone di classificare tra influenza H1N1 e Covid-19:

- Individuare eventuali problemi.
- Implementare dei modelli che svolgono lo stesso task e che propongono delle soluzioni ai problemi individuati.
- Migliorare la spiegabilità dei modelli.

NEXT →

Gli Approcci proposti

Modelli

- Decision Tree.
- Random Forest.

Metriche utilizzate

- Verifica del bilanciamento delle classi.
- Accuracy; Precision; Recall; F1-Score.
- Feature Importance.
- SHAP.

NEXT →

Le tecnologie utilizzate

NEXT

Struttura iniziale del dataset utilizzato

Column	Non-Null Count	Null Count	Dtype
Diagnosis	1485	0	object
D	1485	0	int64
Age	1457	28	float64
Sex	1409	76	object
neutrophil	103	1382	float64
neutrophilCategorical	148	1337	object
serumLevelsOfWhiteBloodCell	151	1334	float64
serumLevelsOfWhiteBloodCellCategorical	191	1294	object
lymphocytes	156	1329	float64
lymphocytesCategorical	197	1288	object
CTscanResults	161	1324	object
XrayResults	47	1438	object
Diarrhea	450	1035	object
Fever	926	559	object
Coughing	862	623	object
SoreThroat	670	815	object
NauseaVomitting	422	1063	object
Temperature	629	856	float64
Fatigue	531	954	object
RenalDisease	226	1259	object
diabetes	226	1259	object

NEXT →

Struttura iniziale del dataset utilizzato

I problemi principali del dataset:

- Dati Discriminatori.
- Valori nulli.
- Duplicati.
- Bilanciamento delle classi.

Column	Non-Null Count	Null Count	Dtype
Diagnosis	1485	0	object
D	1485	0	int64
Age	1457	28	float64
Sex	1409	76	object
neutrophil	103	1382	float64
neutrophilCategorical	148	1337	object
serumLevelsOfWhiteBloodCell	151	1334	float64
serumLevelsOfWhiteBloodCellCategorical	191	1294	object
lymphocytes	156	1329	float64
lymphocytesCategorical	197	1288	object
CTscanResults	161	1324	object
XrayResults	47	1438	object
Diarrhea	450	1035	object
Fever	926	559	object
Coughing	862	623	object
SoreThroat	670	815	object
NauseaVomitting	422	1063	object
Temperature	629	856	float64
Fatigue	531	954	object
RenalDisease	226	1259	object
diabetes	226	1259	object

NEXT →

Struttura iniziale del dataset utilizzato

Gestione dei duplicati e dei valori nulli

Nel dataset originale erano presenti 302 records duplicati, essi sono stati eliminati per evitare potenziali data leakage e bias nelle predizioni.

La maggior parte delle feature presentava un alto tasso di valori nulli, eliminarle tutte avrebbe portato a perdita di informazioni importanti, per questo sono stati applicati i metodi di:

- Imputing.
- Eliminazione delle colonne con altissimo tasso di valori nulli.
- Variabili dummy (binarie).

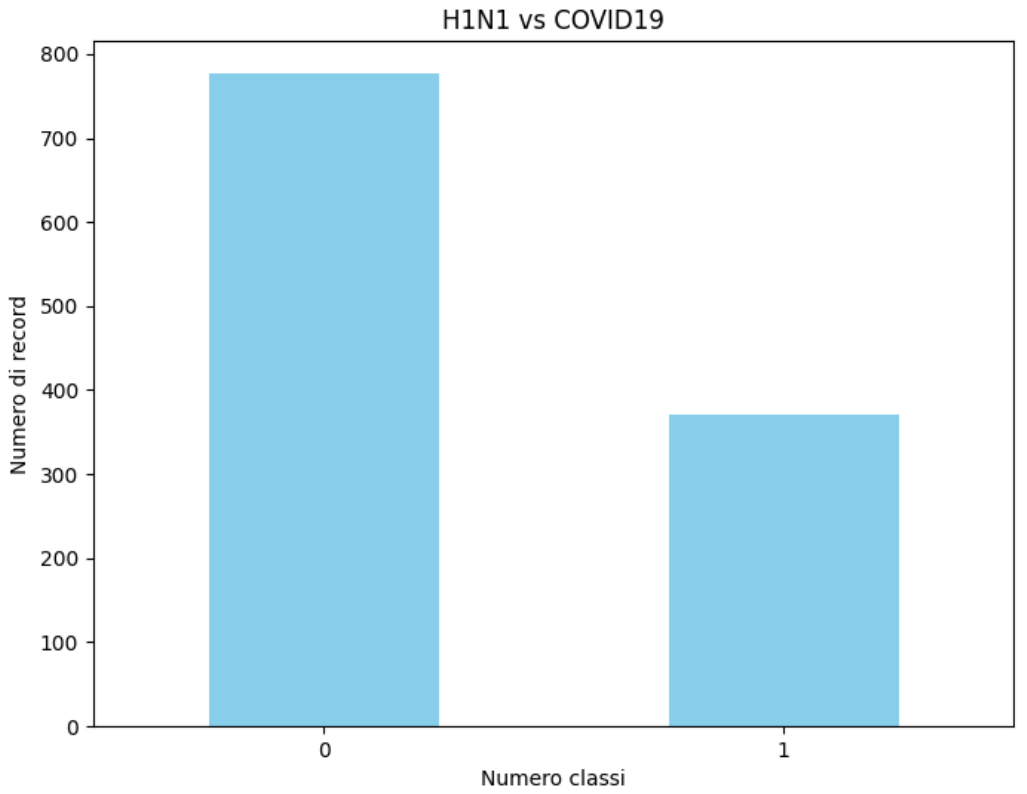
Struttura iniziale del dataset utilizzato

Il bilanciamento delle classi

Il dataset originale presentava uno sbilanciamento ampio tra le due classi, favorendo la classe «Influenza H1N1» rispetto la classe «Covid-19», risolto tramite l'utilizzo di:

- Undersampling Randomico.
- Oversampling SMOTE.

Tali metriche sono state applicate unicamente al set di training.



NEXT

Undersampling VS. Oversampling

Il problema legato all'Undersampling

L'undersampling permette di ridurre la dimensionalità di una classe in modo tale da ribilanciare la distribuzione stessa delle classi. Esso tuttavia può portare:

- Dati viziati.
- Dati poco rappresentativi.

Il problema legato all'Oversampling

L'oversampling consiste nella creazione di nuove righe della classe minoritaria partendo dalle righe appartenenti alla medesima classe già presenti nel datase. Esso tuttavia può portare:

- Presenza di troppi dati sintetici.
- Dati poco rappresentativi.

NEXT →

Undersampling VS. Oversampling

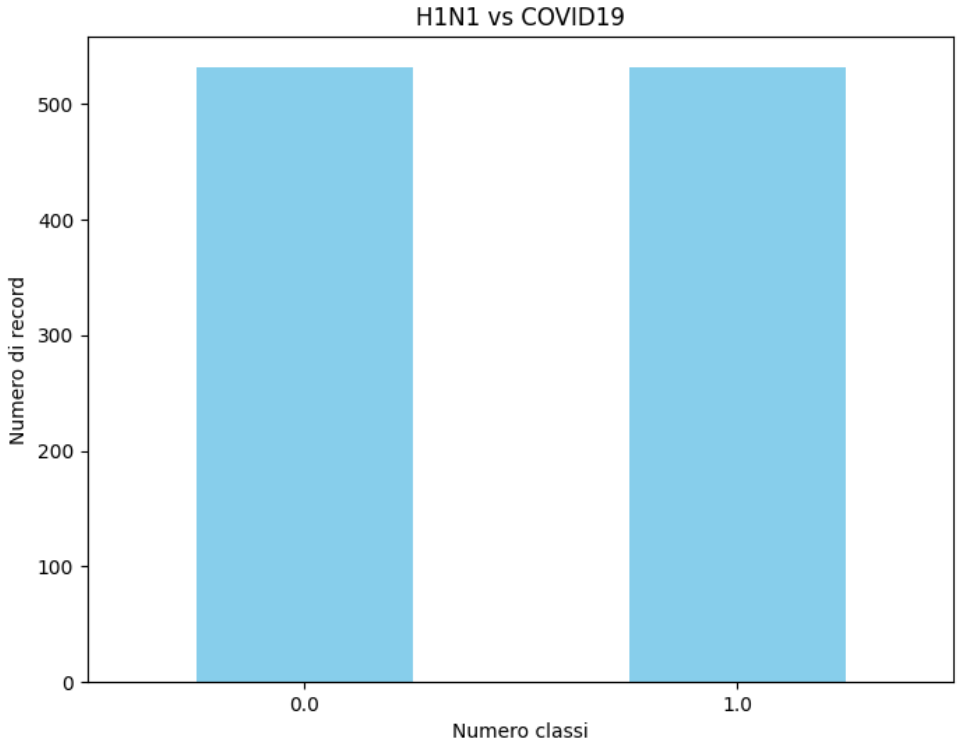
La soluzione proposta

Per evitare di perdere troppi record della classe maggioritaria e allo stesso tempo per evitare di perdere rappresentatività è stato scelto di usare:

- L'undersampling sulla classe maggioritaria, selezionando una popolazione del 60% dal set di training.

- L'oversampling del 40% rimanente sulla classe minoritaria, per evitare di introdurre troppi dati sintetici.

Di seguito il grafico rappresentate il bilanciamento dopo aver applicato le tecniche.



NEXT

Il DataSet Finale

Column	Non-Null Count	Null Count	Dtype
D	1147	0	int64
Age	1147	0	float64
Temperature	1147	0	float64
Sex_F	1147	0	int64
Sex_M	1147	0	int64
neutrophilCategorical_high	1147	0	int64
neutrophilCategorical_low	1147	0	int64
neutrophilCategorical_normal	1147	0	int64
serumLevelsOfWhiteBloodCellCategorical_Low	1147	0	int64
serumLevelsOfWhiteBloodCellCategorical_Normal	1147	0	int64
serumLevelsOfWhiteBloodCellCategorical_high	1147	0	int64
serumLevelsOfWhiteBloodCellCategorical_low	1147	0	int64
serumLevelsOfWhiteBloodCellCategorical_normal	1147	0	int64
lymphocytesCategorical_High	1147	0	int64
lymphocytesCategorical_Low	1147	0	int64
lymphocytesCategorical_Normal	1147	0	int64
CTscanResults_Inconclusive	1147	0	int64
CTscanResults_Neg	1147	0	int64
CTscanResults_Non_Presente	1147	0	int64

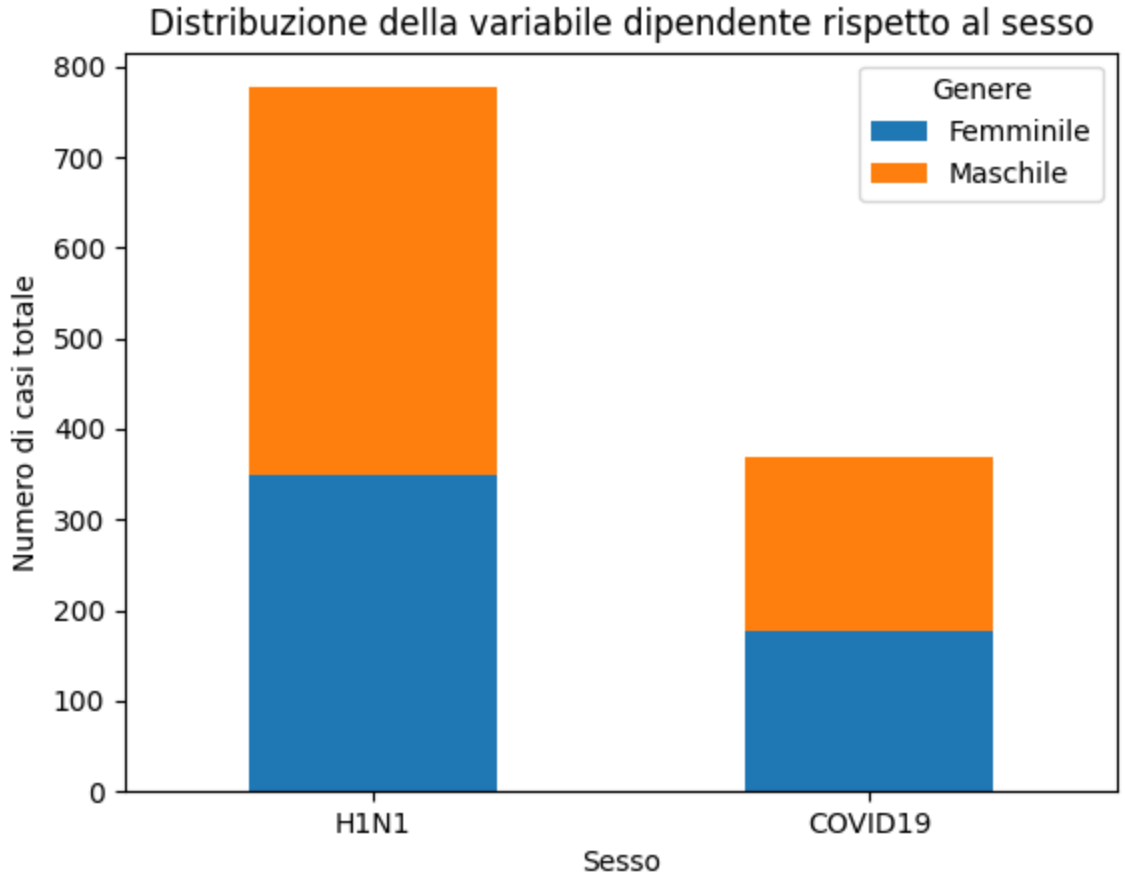
Column	Non-Null Count	Null Count	Dtype
CTscanResults_Pos	1147	0	int64
CTscanResults_neg	1147	0	int64
XrayResults_Neg	1147	0	int64
XrayResults_Non_Presente	1147	0	int64
XrayResults_Pos	1147	0	int64
Diarrhea_No	1147	0	int64
Diarrhea_Yes	1147	0	int64
Fever_No	1147	0	int64
Fever_Yes	1147	0	int64
Coughing_No	1147	0	int64
Coughing_Yes	1147	0	int64
SoreThroat_No	1147	0	int64
SoreThroat_Yes	1147	0	int64
NauseaVomitting_No	1147	0	int64
NauseaVomitting_Yes	1147	0	int64
Fatigue_No	1147	0	int64
Fatigue_Yes	1147	0	int64
RenalDisease_No	1147	0	int64
RenalDisease_Yes	1147	0	int64
diabetes_No	1147	0	int64
diabetes_Yes	1147	0	int64

Analisi della variabile dipendente

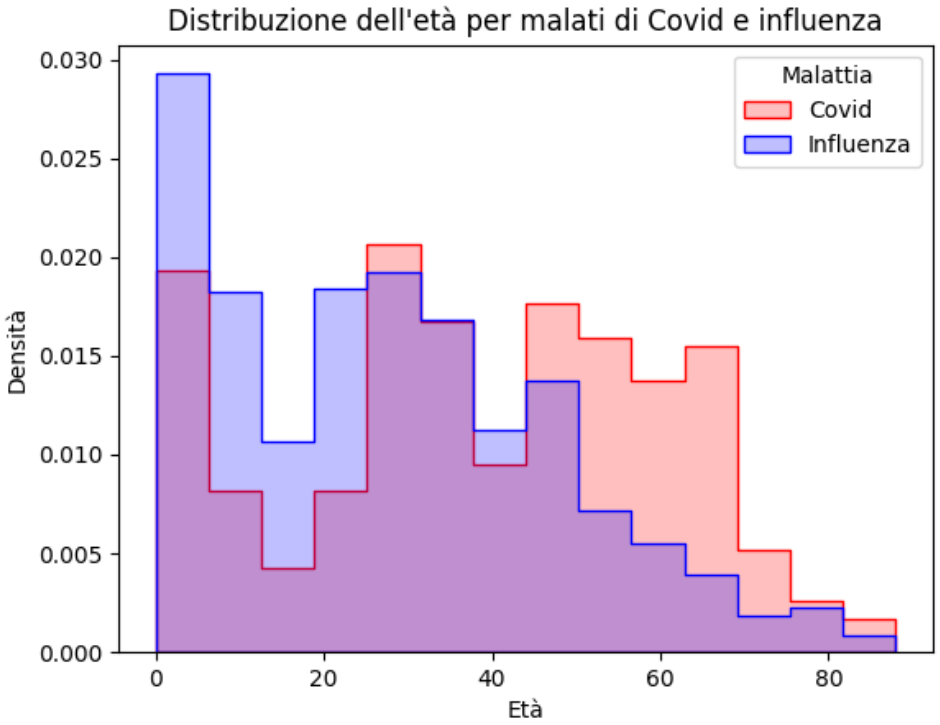
La variabile dipendente

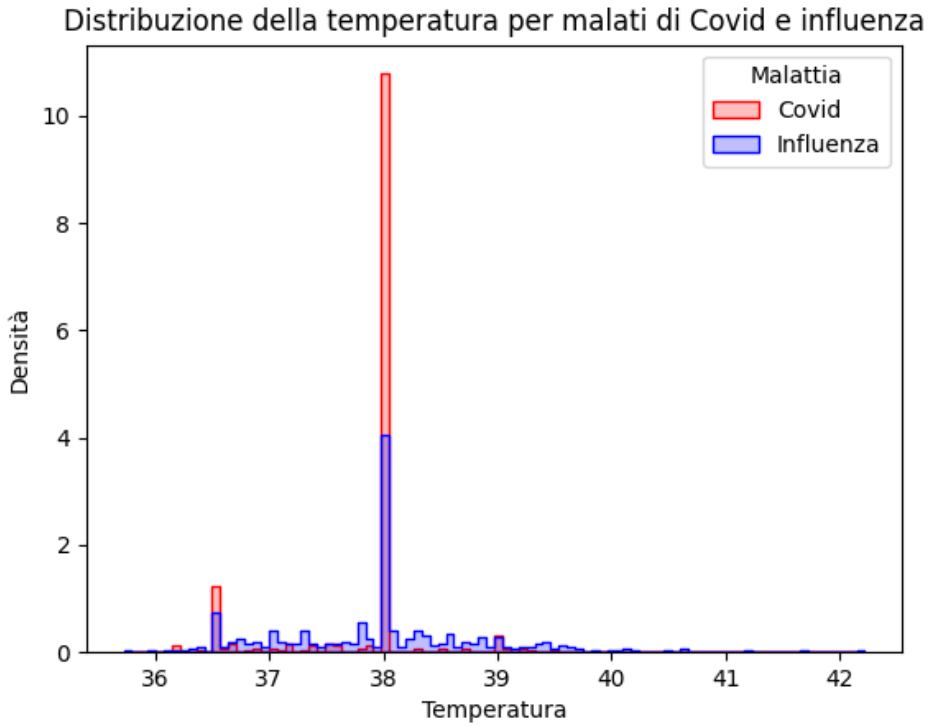
La variabile dipendente del dataset, ossia la variabile da predire, `e riportata all'interno del dataset stesso con la label "D". Il valore assunto dalla variabile `e "0" nel caso pazienti malati di "H1N1", "1" nel caso di pazienti malati di "COVID-19".

Di seguito alcuni grafici che descrivono la correlazione tra essa e le feature presenti.



NEXT


Distribuzione dell'età per malati di Covid e influenza
Malattia
Covid
Influenza
Densità
0.030
0.025
0.020
0.015
0.010
0.005
0.000
0
20
40
60
80
Età


Distribuzione della temperatura per malati di Covid e influenza
Malattia
Covid
Influenza
Densità
10
8
6
4
2
0
36
37
38
39
40
41
42
Temperatura

NEXT

Normalizzazione ed Encoding

Perché farlo?

Per evitare potenziali bias e rendere anche le variabili confrontabili con le altre variabili è necessario, in fase di preprocessing dei dati, normalizzare i valori che rappresentati dalle variabili continue in modo tale da ridurli alla stessa ”scala” e codificare le variabili categoriche che non sono accettate dai modelli di Machine Learning.

Gli approcci scelti

Per la fase di pre-processing dei dati sono stati scelti gli approcci di:

- Encoding delle variabili categoriche utilizzando le variabili dummy.
- Normalizzazione Min-Max per le variabili continue.

NEXT →

Data Splitting

Perché farlo?

Fare la divisione del dataset in set di training e set di valutazione permette di evitare qualsiasi distorsione sistematica che potrebbe essere presente nel set di dati, come l'ordine dei campioni o la loro posizione all'interno del set di dati

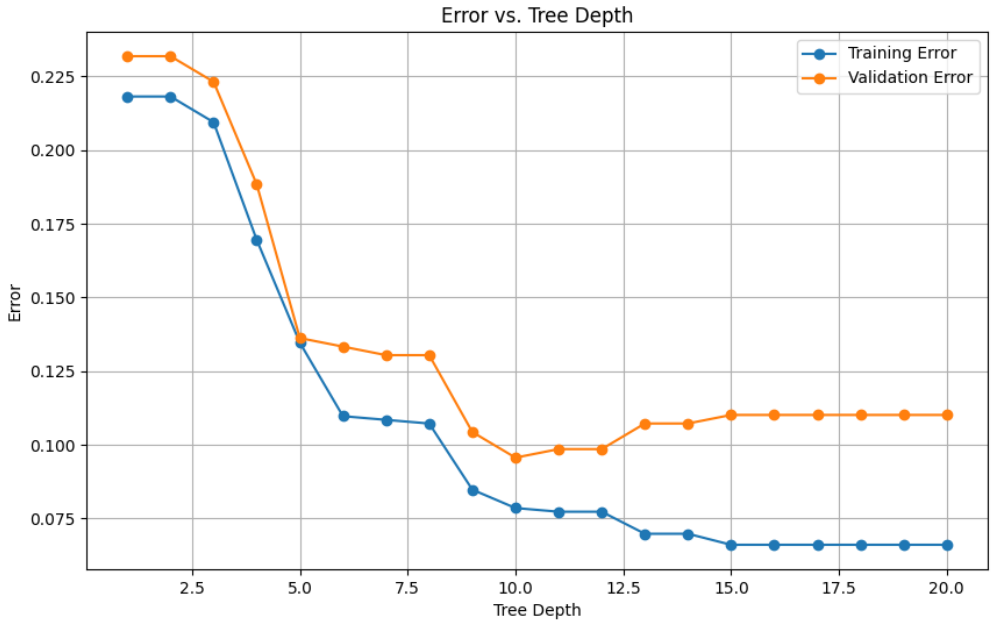
Metodo di splitting

Per la fase di splitting del set in set di training e validazione è stata utilizzata la metrica del 67/33, il 67% del dataset è stato utilizzato per l'addestramento del modello, mentre il restante 33% del dataset per la fase di testing del modello.

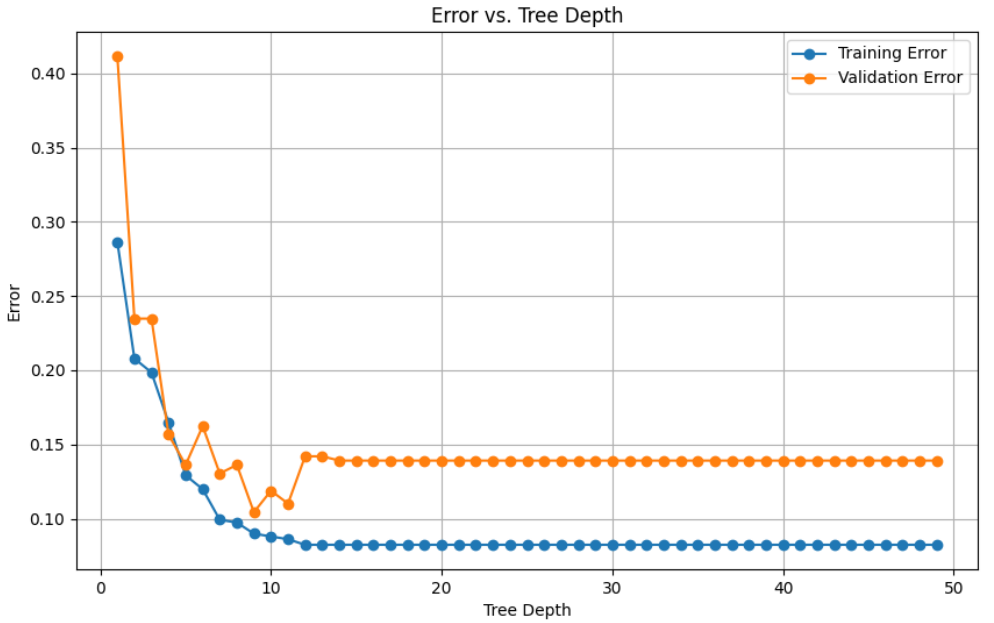
NEXT →

Addestramento e valutazione

Decision Tree: profondità dell'albero decisionale



Dataset non bilanciato



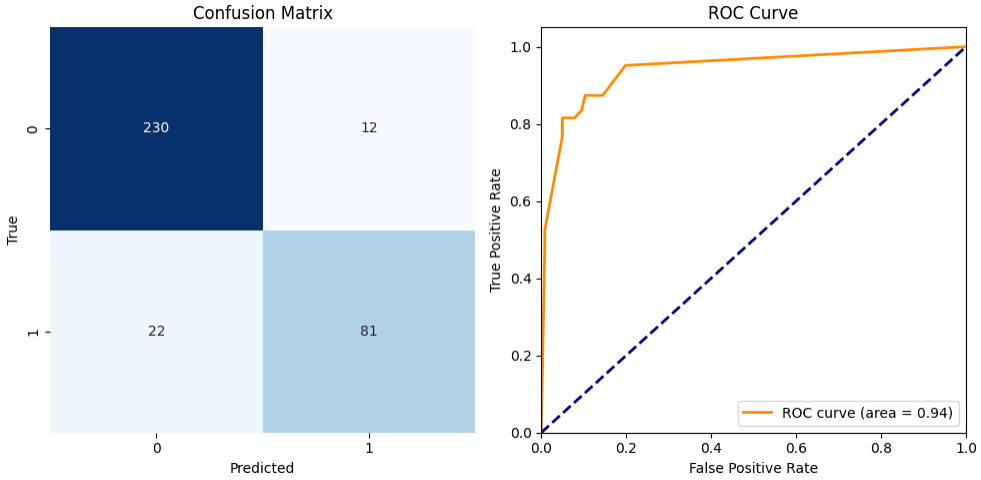
Dataset bilanciato

NEXT →

Addestramento e valutazione

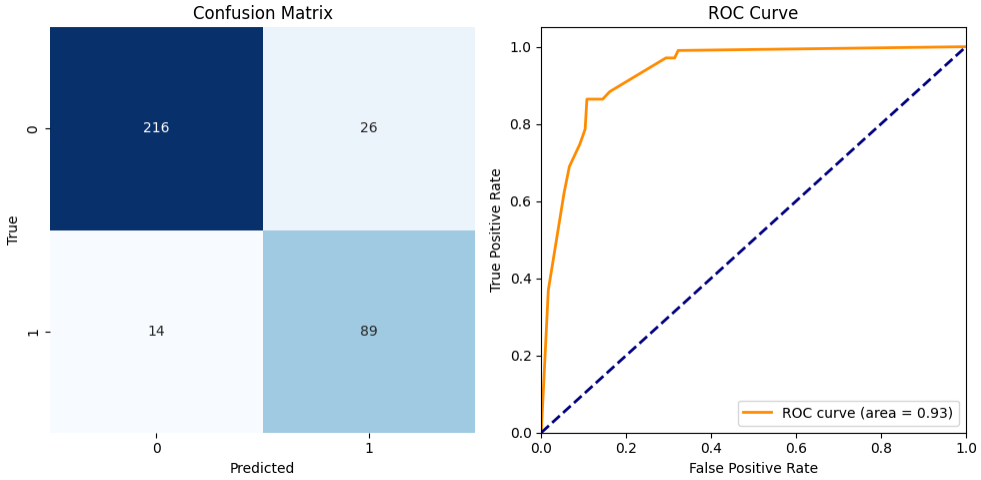
Decision Tree con Dataset non bilanciato

Accuracy: 0.90
Precision: 0.87
Recall: 0.79
F1 Score: 0.83



Decision Tree con Dataset bilanciato

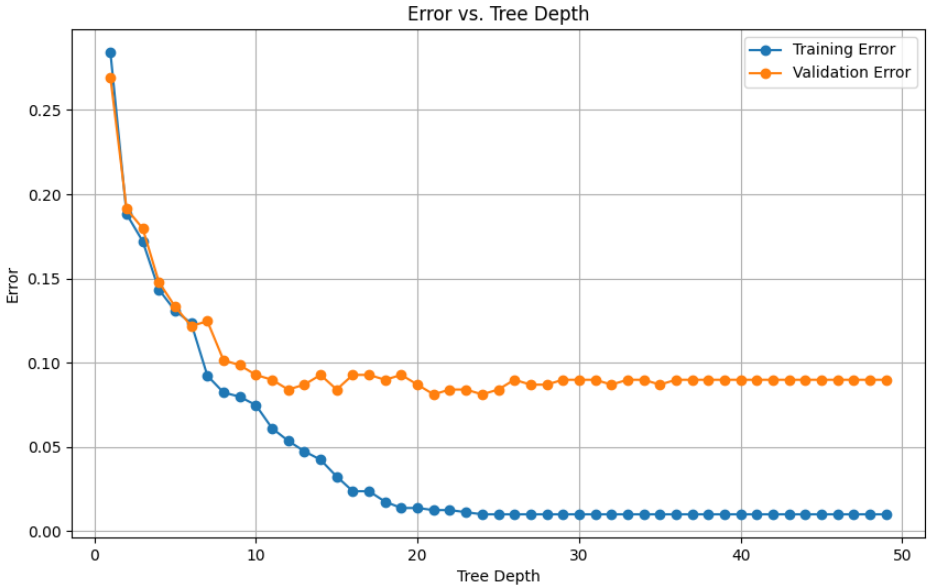
Accuracy: 0.88
Precision: 0.77
Recall: 0.86
F1 Score: 0.82



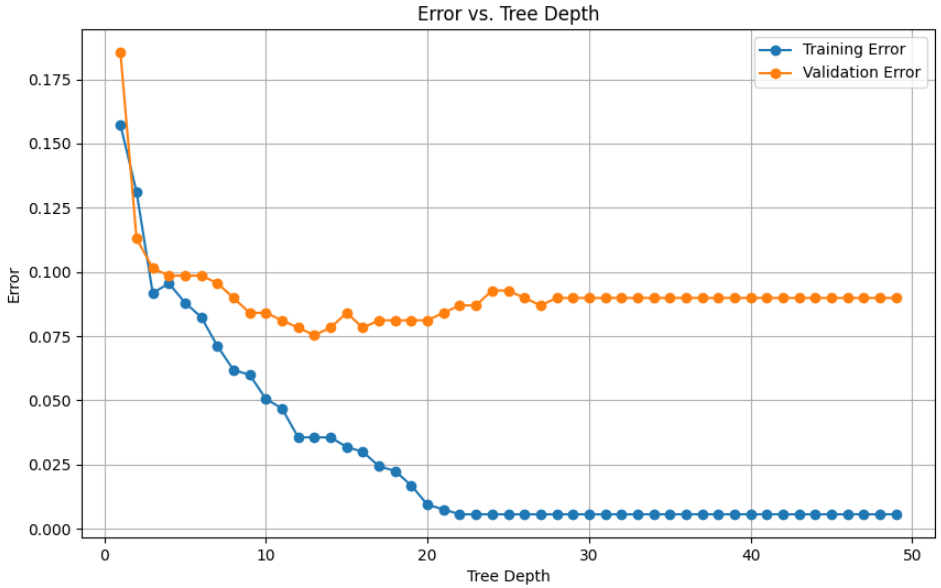
NEXT

Addestramento e valutazione

Random Forest: profondità massima degli alberi decisionali



Dataset non bilanciato



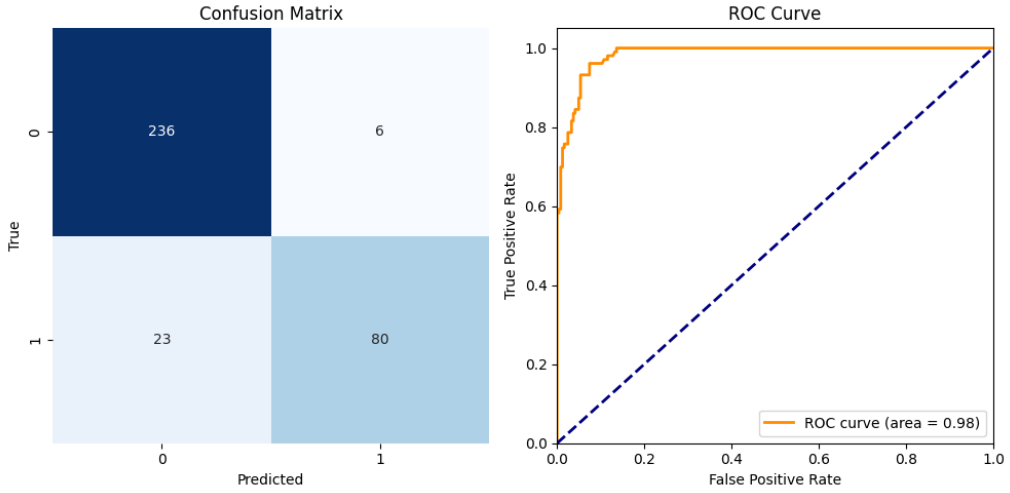
Dataset bilanciato

NEXT

Addestramento e valutazione

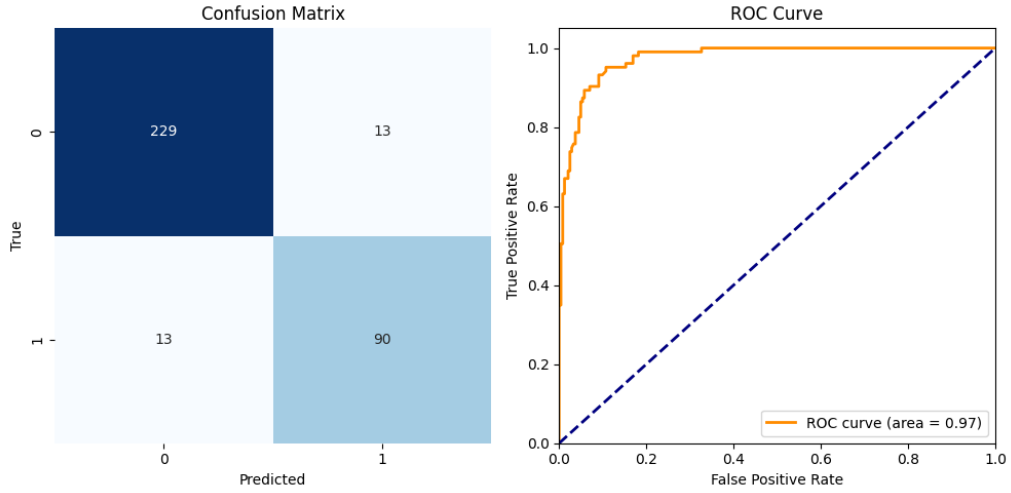
Random Forest con Dataset non bilanciato

Accuracy: 0.92
Precision: 0.93
Recall: 0.78
F1 Score: 0.85



Random Forest con Dataset bilanciato

Accuracy: 0.92
Precision: 0.87
Recall: 0.87
F1 Score: 0.87



Explainability dei modelli

Che cos'è?

L'explainability è la metrica che misura quanto si riesce a spiegare il funzionamento di un algoritmo di IA, in particolare l'obiettivo del miglioramento dell'explainability è quello di rendere chiari e intuitivi i motivi per la quale un modello di IA ha dato un certo output e come mai ha preso una data decisione.

Gli approcci scelti

Per la fase di miglioramento dell'explainability del modello sono stati scelti i seguenti approcci:

- Feature Importance.
- SHAP.
- Creazione di un prototipo di interfaccia usabile.

NEXT →

Explainability dei modelli: Feature Importance

Che cos'è?

La feature importance è una metrica che valuta quanto una feature ha impattato sui risultati di una predizione, un punteggio più alto di importance indica che una feature ha più peso sulla predizione finale rispetto ad un'altra feature con importance più bassa.

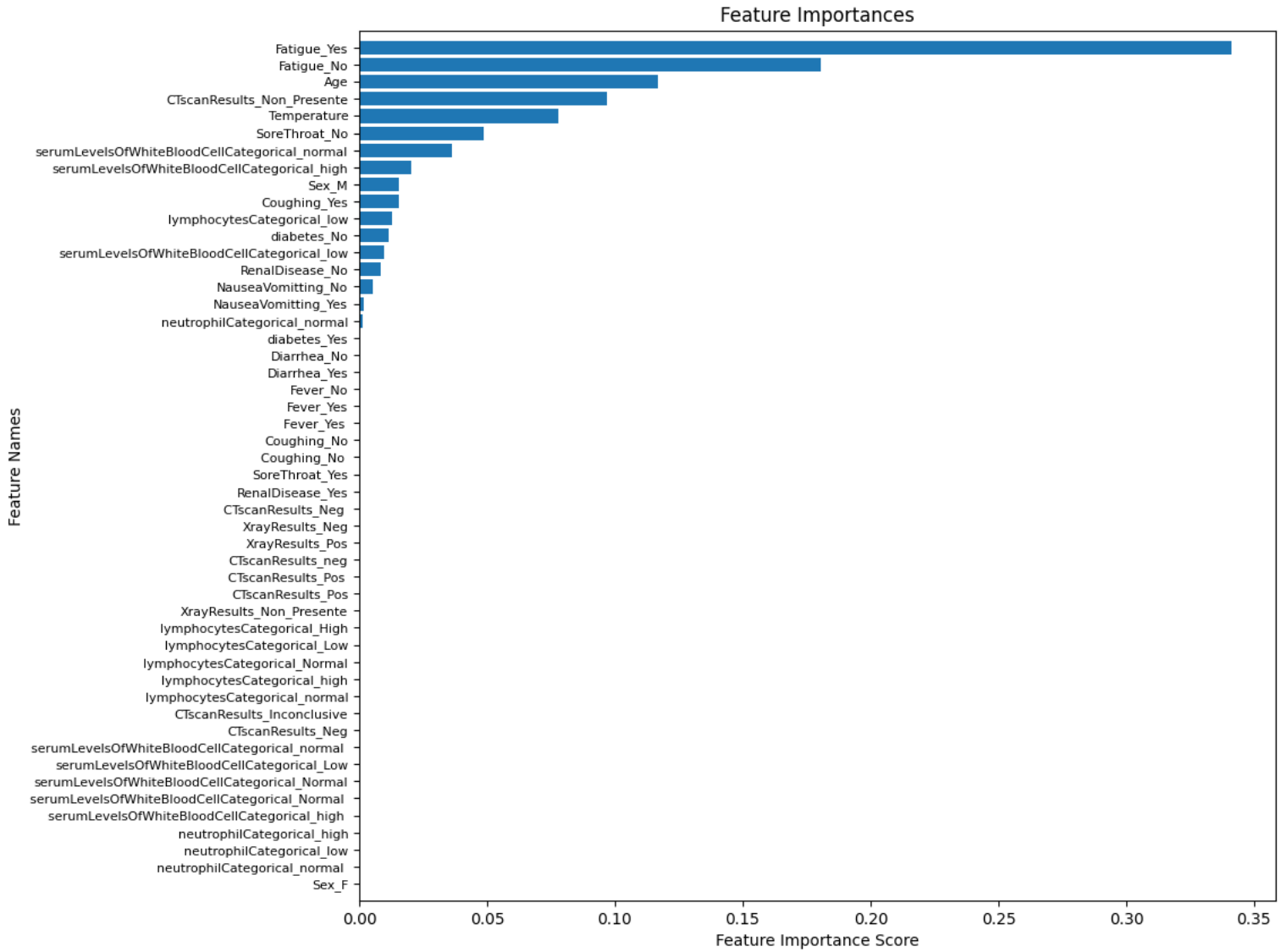
Quale dataset valutare?

Per valutare l'importanza delle feature è stato utilizzato il training sul set di addestramento bilanciato.

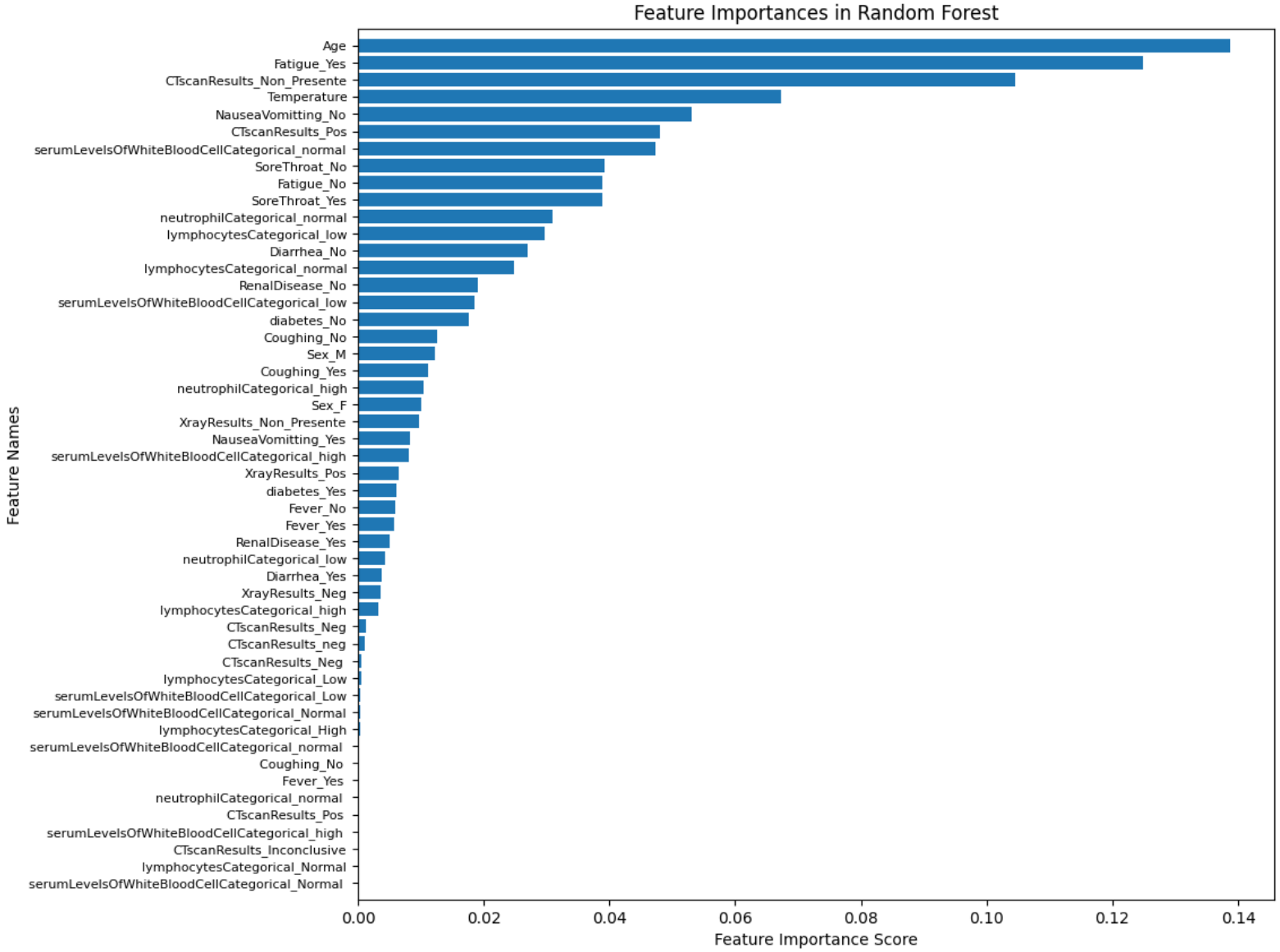
NEXT →

Comparazione Feature Importance

Feature Importance del modello Decision Tree



Comparazione Feature Importance



Feature Importance del modello Decision Tree

Explainability dei modelli: SHAP

Che cos'è?

SHAP è una metrica, che si basa sulla teoria dei giochi, utilizzata per spiegare la predizione di un modello di Machine Learning. Essa viene calcolata come somma dell'impatto di ogni feature sulla predizione finale, lo scopo è riuscire a ricostruire e a spiegare in che modo il modello ha prodotto una data predizione.

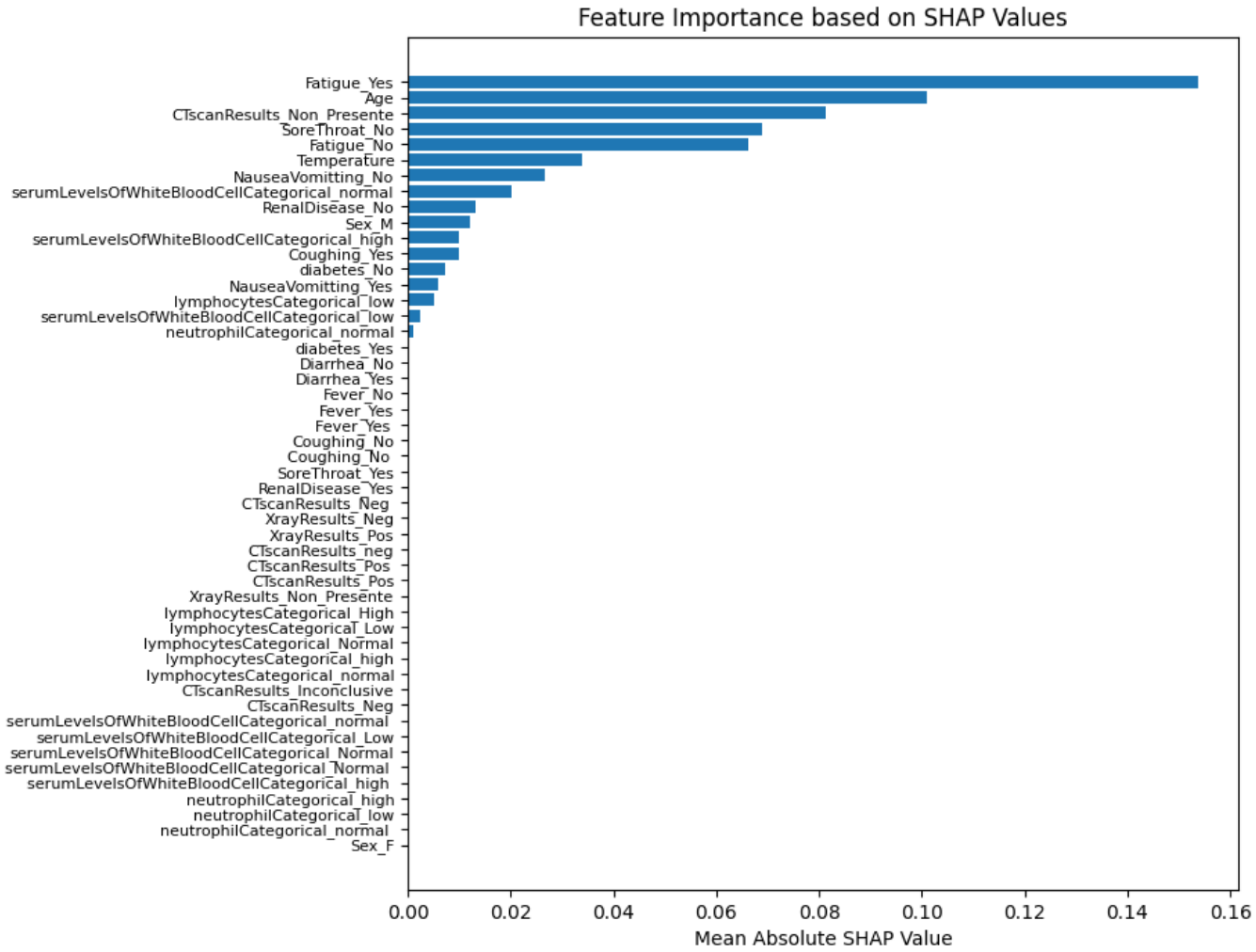
Quale dataset valutare?

Per valutare l'importanza delle feature tramite SHAP è stato utilizzato il training sul set di addestramento bilanciato.

NEXT →

Explainability dei modelli: SHAP

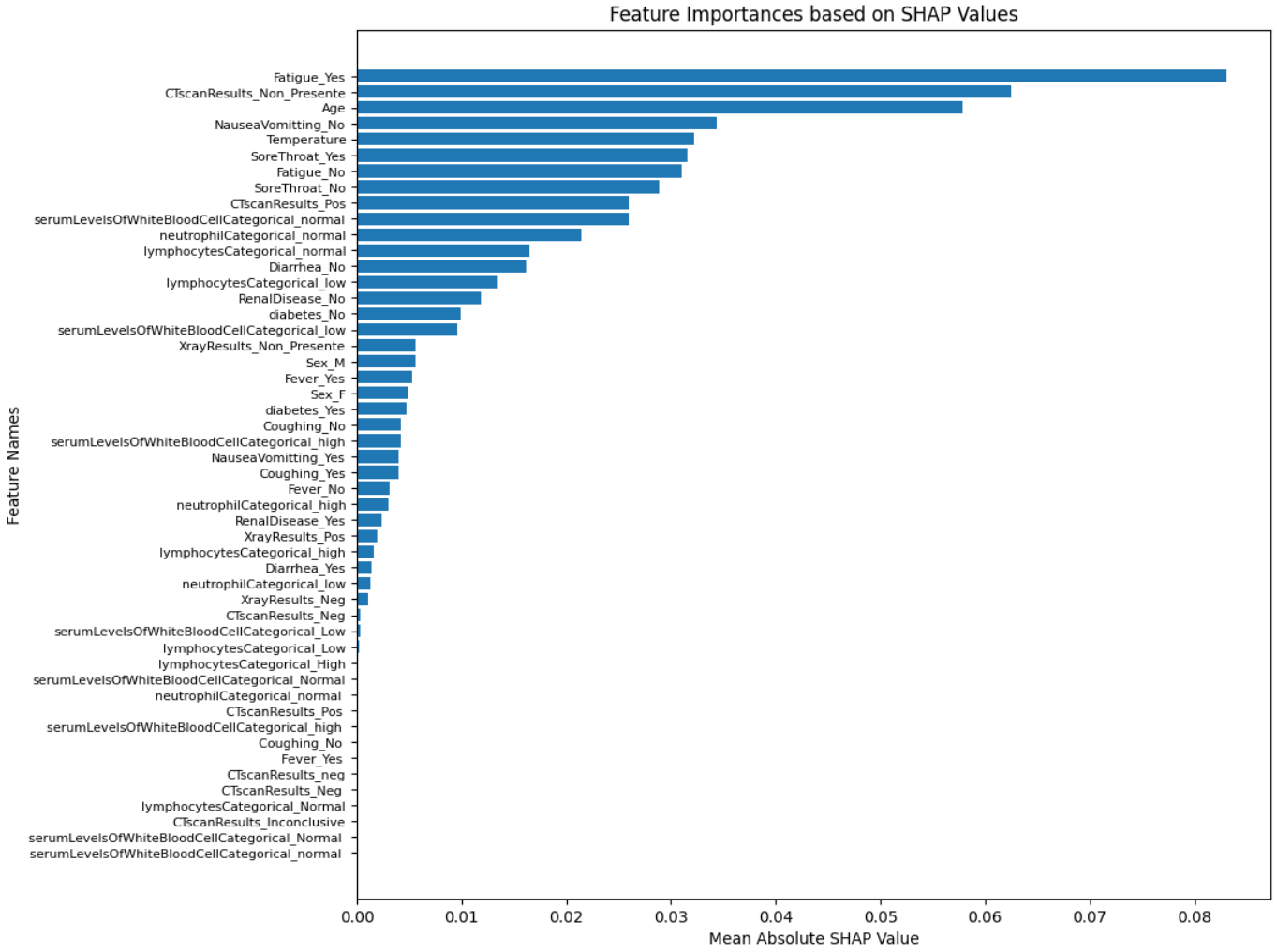
SHAP Importance del modello Decision Tree



NEXT →

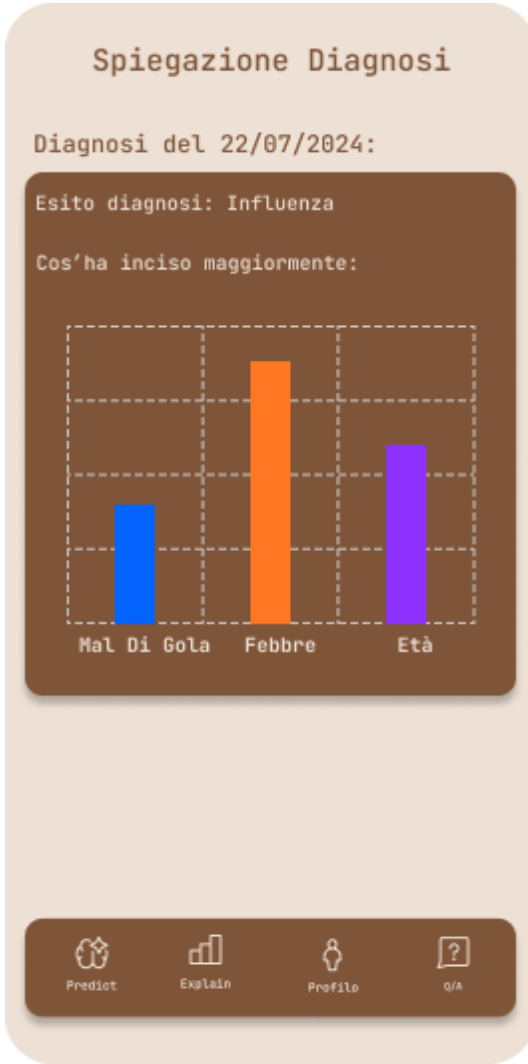
Explainability dei modelli: SHAP

SHAP Importance del modello Random Forest



NEXT

Explainability dei modelli: Prototipo dell’interfaccia



Conclusioni

Problemi individuati

- Legati ai dati presenti nel dataset.
- Legati al periodo di sviluppo dello studio scelto.
- Legati alla spiegabilità del modello.

Soluzioni implementate

- Data Cleaning.
- Poco utilizzo dei dati sintetici.
- Utilizzo delle metriche per migliorare l’explainability.

Le mie conclusioni

I problemi principali dello studio di partenza è che la qualità dei dati raccolti è profondamente influenzata dal periodo di svolgimento dello studio, i tempi precoci e i pochi dati raccolti non permettevano e non hanno permesso l ‘implementazione di un modello privo di bias e con accuratezza alta. Tuttavia il lavoro svolto sui dati presenti ha permesso di migliorarne la qualità, le metriche applicate hanno permesso di migliorare la spiegabilità delle predizioni, portando ottimi risultati in termini di obiettivi del progetto.

NEXT →