RELAZIONE TECNICA PROGETTO DI INGEGNERIA DELLA CONOSCENZA

Anno Accademico 2022 – 2023

Studente: Andre Alvizuri (676837)

Modello di intelligenza artificiale per la identificazione di persone con tendenze suicide

Introduzione

Il suicidio è un problema sociale che affligge la nostra società oggi, colpendo dai più i giovani agli adulti, in cui sono coinvolti molti fattori come quelli economici, sentimentali, sociali, ecc.

Motivazione

Il mio paese "Bolivia" ha il tasso piu alto di suicidi in tutta latinoamerica, questo rappresenta un grave problema per la societa boliviana, le politiche sociali, la mancanza di lavoro, una deficente educazione hanno contribuito al incremento dei suicidi negli ultimi anni, in questo progetto ho voluto analizzare cosa spinge a una persona a prendere tale decisione.

Riguardo al Dataset

I dati del dataset sono stati raccolti tramite una indagine realizzata da un medico psichiatra uruguaiano Hernán Salazar Gimenez che scrisse il paper "Valutazione del rischio di suicidio in America Latina".

Il dataset ha una quantità equa di persone di genere maschile e femminile, inoltre, vi è una vasta gamma di persone di diverse età, dai 15 ai 60 anni.

Il dataset comprende 16 domande che sono strettamente correlate ai casi di suicidio in Bolivia.

Elaborazione dei Dati

Per poter analizzare correttamente il modello, è necessario sostituire i valori "vero" e "falso" o i valori di selezione multipla con valori puramente numerici.

```
db["GENERE"].replace(["Donna", "Uomo"], [1, 0], inplace=True)
db.head()
```

[1597] ✓ 0.0s

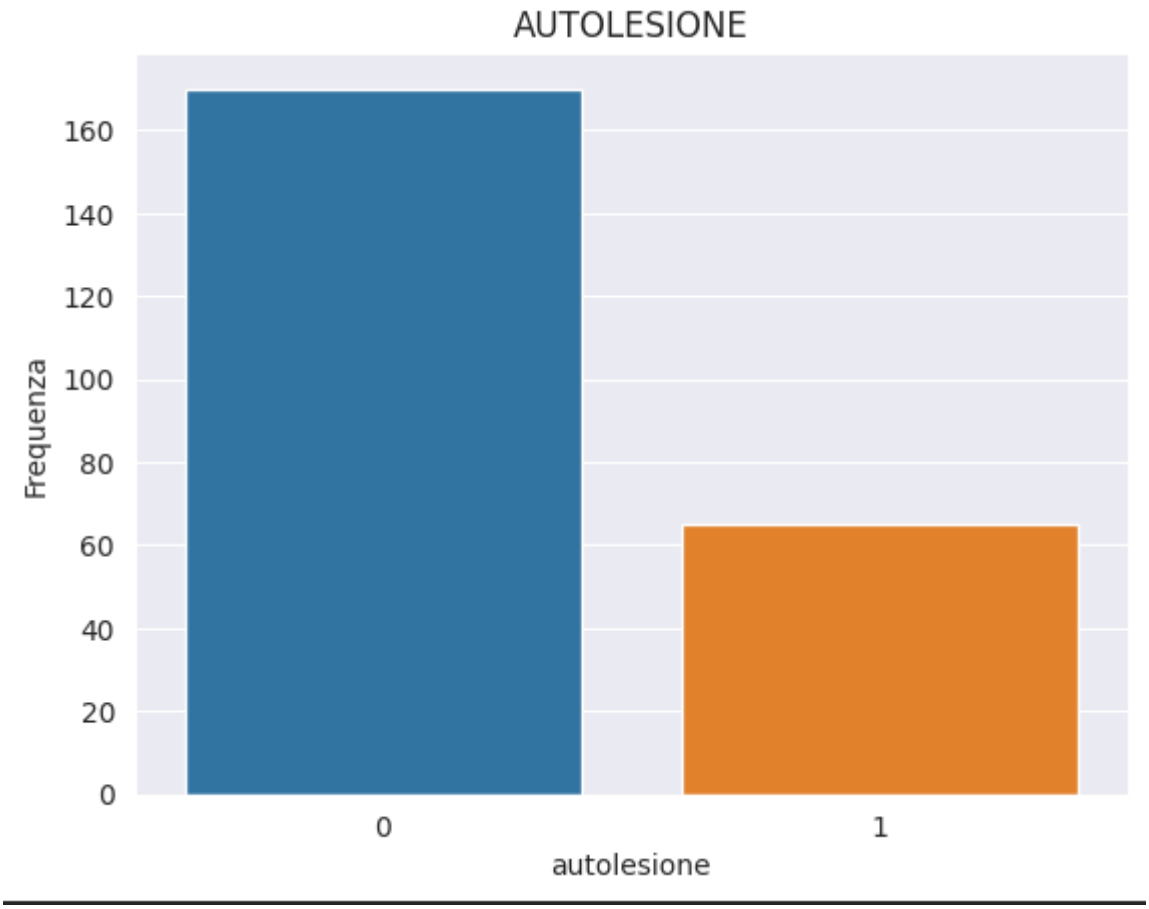
	GENERE	ETA	PERSONALITA	FUMATORE	CONSUMO_FUMATORE	ALCOL
0	1	61	Amichevole	No	Nessuno	No
1	0	24	Amichevole	No	Nessuno	No
2	1	39	Timido	No	Nessuno	No
3	0	30	Timido	No	Nessuno	Si
4	0	26	Timido	No	Nessuno	Si

```
db["CONSUMO_FUMATORE"].replace(
    ["Nessuno", "Tabacco", "Cannabis", "Sigaretta elettronica", "Hashish"],
    [0, 1, 2, 3, 4],
    inplace=True)
db.head()
```

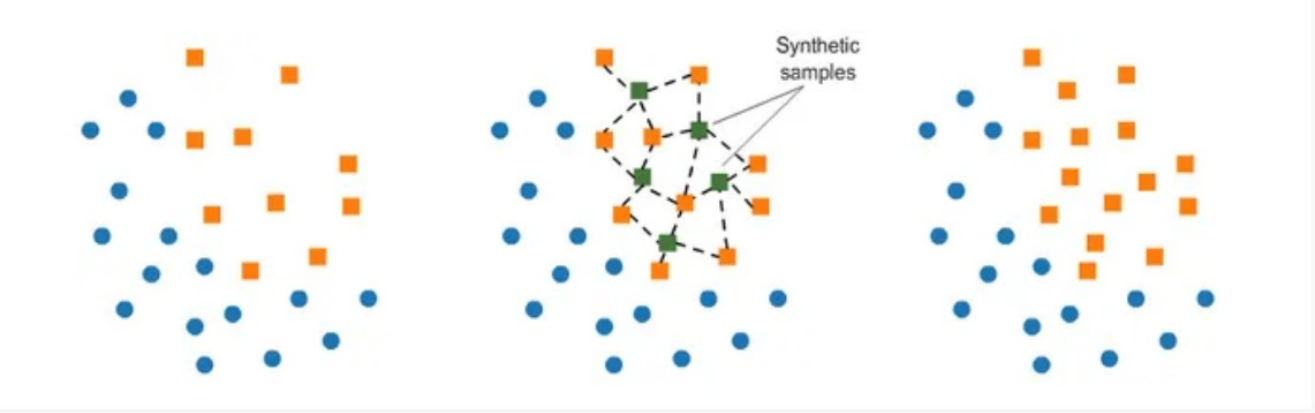
[1609] ✓ 0.0s

NALITA	FUMATORE	CONSUMO_FUMATORE	ALCOL	FREQUENZA_ALCOL	PROBLEMI	STATO_GENITORI
1	0	0	0	Mai	1	Sposati
1	0	0	0	Mai	0	Sposati
0	0	0	0	Raramente	1	Sposati
0	0	0	1	Raramente	0	Sposati
0	0	0	1	Ogni mese	0	Sposati

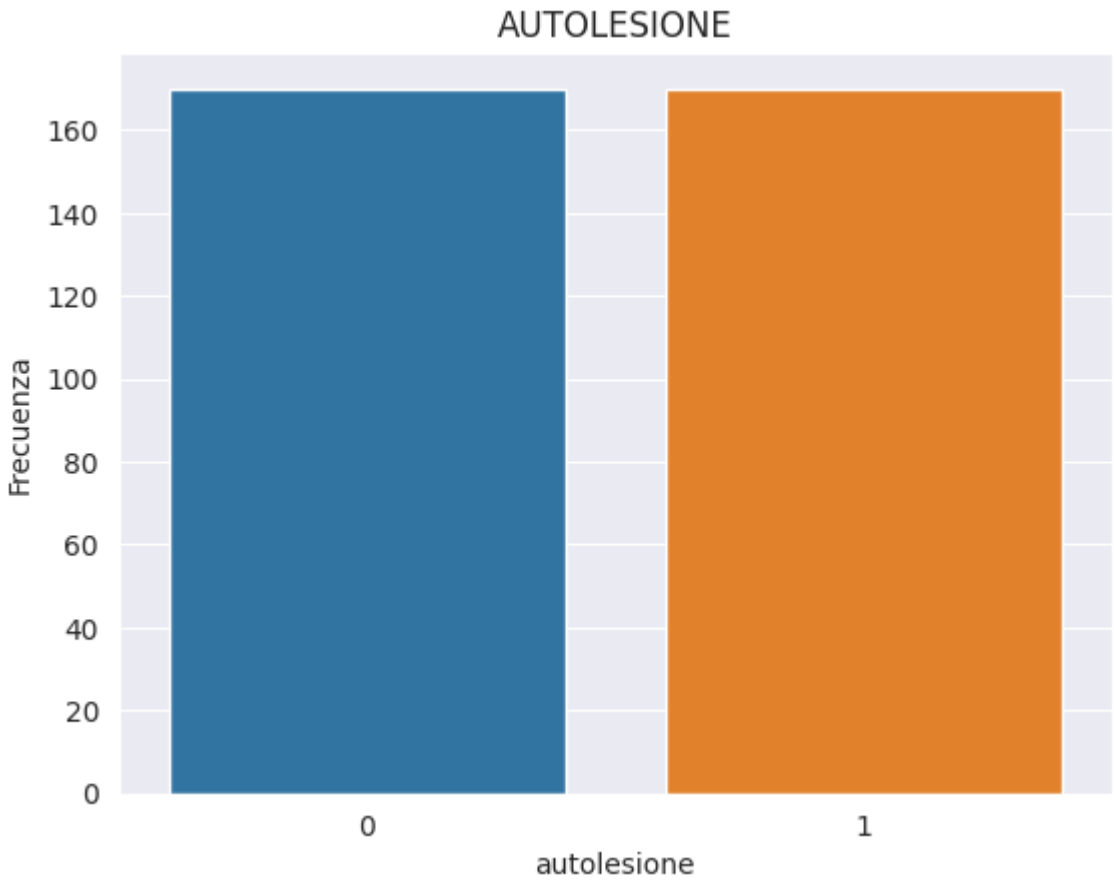
nella elaborazione abbiamo usato valori numerici e abbiamo eliminato la colonna che rappresenta l'età, facendo un analisi del dataset possiamo vedere che la variabile che fa riferimento all'autolesionismo e sbilanciata (che sara il target).



Facendo uso dell'algoritmo SMOTE proviamo a bilanciare il target, l'algoritmo SMOTE è un algoritmo di over-sampling utilizzato per gestire il problema della disuguaglianza di classe nei problemi di classificazione binaria. L'obiettivo è **generare campioni sintetici** della classe di minoranza in modo da bilanciare il dataset e migliorare le prestazioni del modello di classificazione.
L'algoritmo SMOTE funziona creando campioni sintetici per la classe di minoranza utilizzando le informazioni dei campioni già presenti nel dataset. In pratica, SMOTE seleziona casualmente un campione della classe di minoranza, individua i suoi k vicini più vicini (dove k è un parametro dell'algoritmo), e genera un nuovo campione sintetico lungo la linea che connette il campione originale con uno dei suoi k vicini scelti casualmente.
In particolare, per generare un nuovo campione sintetico, SMOTE prende in considerazione una coppia di campioni vicini (il campione originale e uno dei suoi k vicini) e crea un nuovo campione sintetico scegliendo un punto lungo la linea che connette i due campioni. Il punto viene scelto casualmente, ma deve essere all'interno della regione che connette i due campioni. Questo processo viene ripetuto per tutti i campioni della classe di minoranza, generando così un dataset sintetico bilanciato.

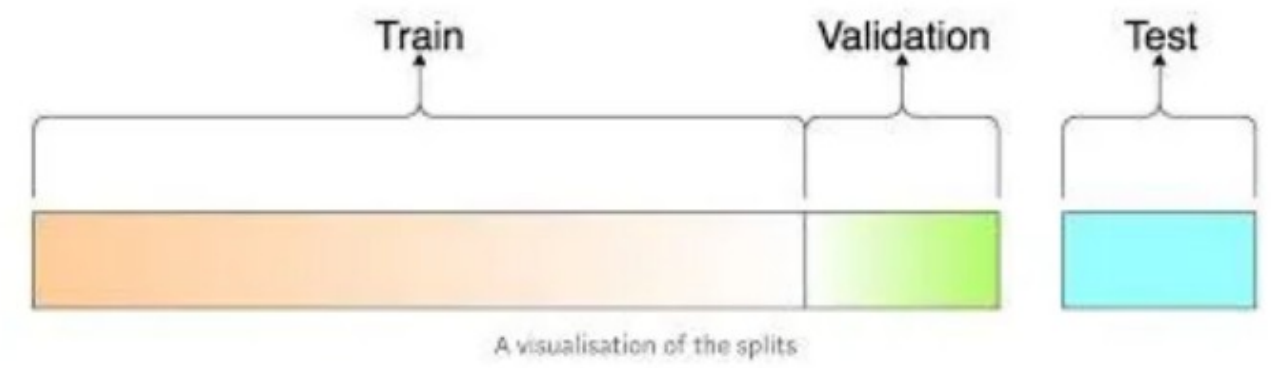


Dopo avere applicato l'algoritmo SMOTE possiamo vedere che il target e bilanciato



Ora si divide il dataset in

- 70% per il train
- 30% per il test



A visualisation of the splits

Abbiamo provato 4 diversi algoritmi per il train in modo di ottenere l'algoritmo più accurato e preciso, questi sono:

1. Random Forest
2. Alberi di Decisioni
3. Naive Bayes
4. Vettori di Supporto.

Analizzando le metriche di ogni algoritmo

- **Accuracy:** La metrica di accuratezza è una misura utilizzata per valutare quanto un modello di classificazione sia in grado di predire correttamente le classi di un insieme di dati di test. Essa rappresenta la percentuale di predizioni corrette rispetto al totale delle predizioni effettuate dal modello.
  L'accuratezza è una metrica importante per valutare le prestazioni di un modello di classificazione,
  ma può non essere sufficiente in presenza di dataset sbilanciati.

- **Precisione:** La precisione è calcolata come il rapporto tra il numero di esempi positivi correttamente classificati e il totale degli esempi classificati come positivi.
  La precisione è particolarmente importante in situazioni in cui gli errori di tipo I (false positive) sono
  costosi o indesiderati, come ad esempio nel rilevamento di tumori o nella rilevazione di frodi finanziarie.

- **Recall:** La metrica recall, nota anche come sensibilità o tasso di veri positivi, è una misura di performance utilizzata per valutare gli algoritmi di classificazione.
  Il recall è particolarmente utile quando è importante individuare tutti i casi positivi, anche a costo di
  identificare erroneamente alcuni come tali (ovvero, quando è preferibile avere falsi positivi piuttosto
  che falsi negativi).

- **F1:** La metrica F1, detta anche punteggio F1, è una metrica di valutazione utilizzata in ambito di classificazione binaria. Essa combina il valore di precision e recall, due metriche comuni nella valutazione di modelli di classificazione.
  La metrica F1 è particolarmente utile quando si desidera un'accurata valutazione del modello su

entrambe le classi, ovvero quando il dataset presenta un bilanciamento simile tra le classi positive e
negative. In questo caso, la metrica F1 rappresenta un compromesso tra precision e recall, e fornisce
un indicatore globale della capacità del modello di classificare correttamente entrambe le classi.

```
Accuracy Naive Bayes : 0.7352941176470589
recall Naive Bayes : 0.78
[[36 16]
 [11 39]]
Accuracy Decision Tree : 0.7058823529411765
recall Decision Tree: 0.76
[[34 18]
 [12 38]]
Accuracy Random Forest : 0.7647058823529411
recall Random Forest : 0.84
[[36 16]
 [ 8 42]]
Accuracy SVM : 0.7745098039215687
recall SVM : 0.88
[[35 17]
 [ 6 44]]
```

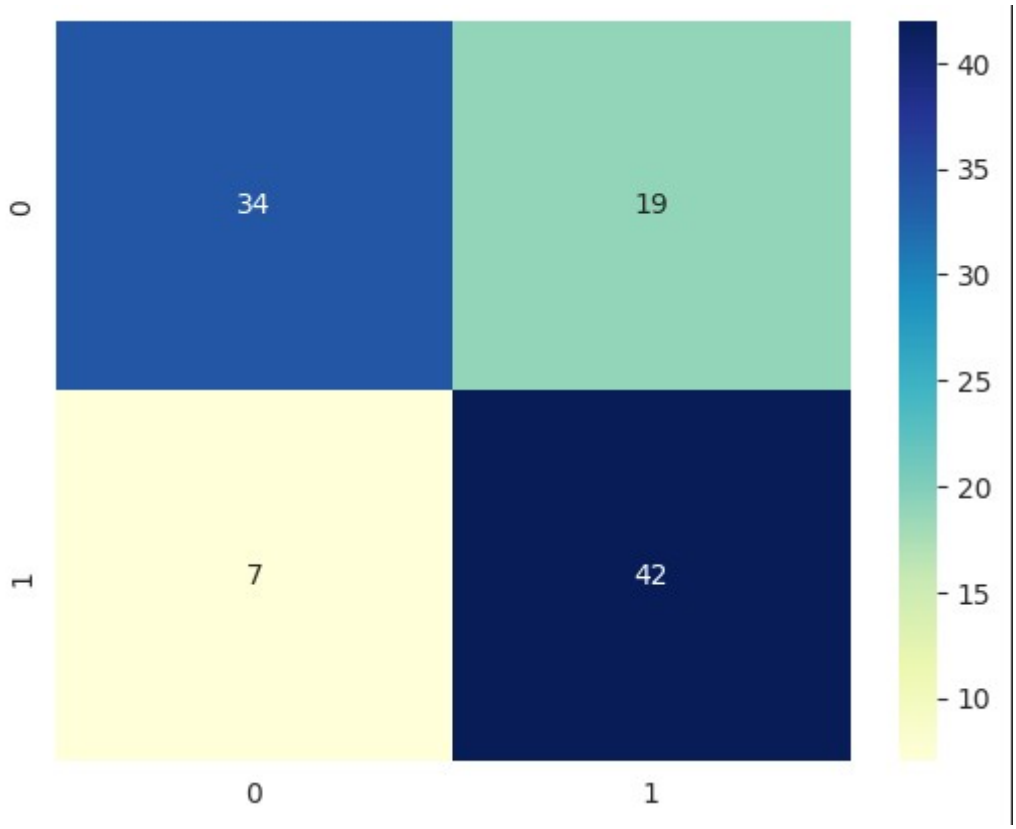
In piu si analizza anche la matrice di confusione di ogni algoritmo, la matrice di confusione è una rappresentazione tabulare che viene utilizzata per valutare le prestazioni di un modello di classificazione. È composta da quattro celle, che mostrano il numero di previsioni corrette e errate suddivise in base alle classi reali e alle classi predette.

Nella matrice di confusione, i True Positive (TP) rappresentano i casi in cui il modello ha previsto correttamente la classe positiva. I False Positive (FP) rappresentano i casi in cui il modello ha erroneamente previsto la classe positiva quando la classe reale era negativa. I False Negative (FN) rappresentano i casi in cui il modello ha erroneamente previsto la classe negativa quando la classe reale era positiva. Infine, i True Negative (TN) rappresentano i casi in cui il modello ha previsto correttamente la classe negativa.

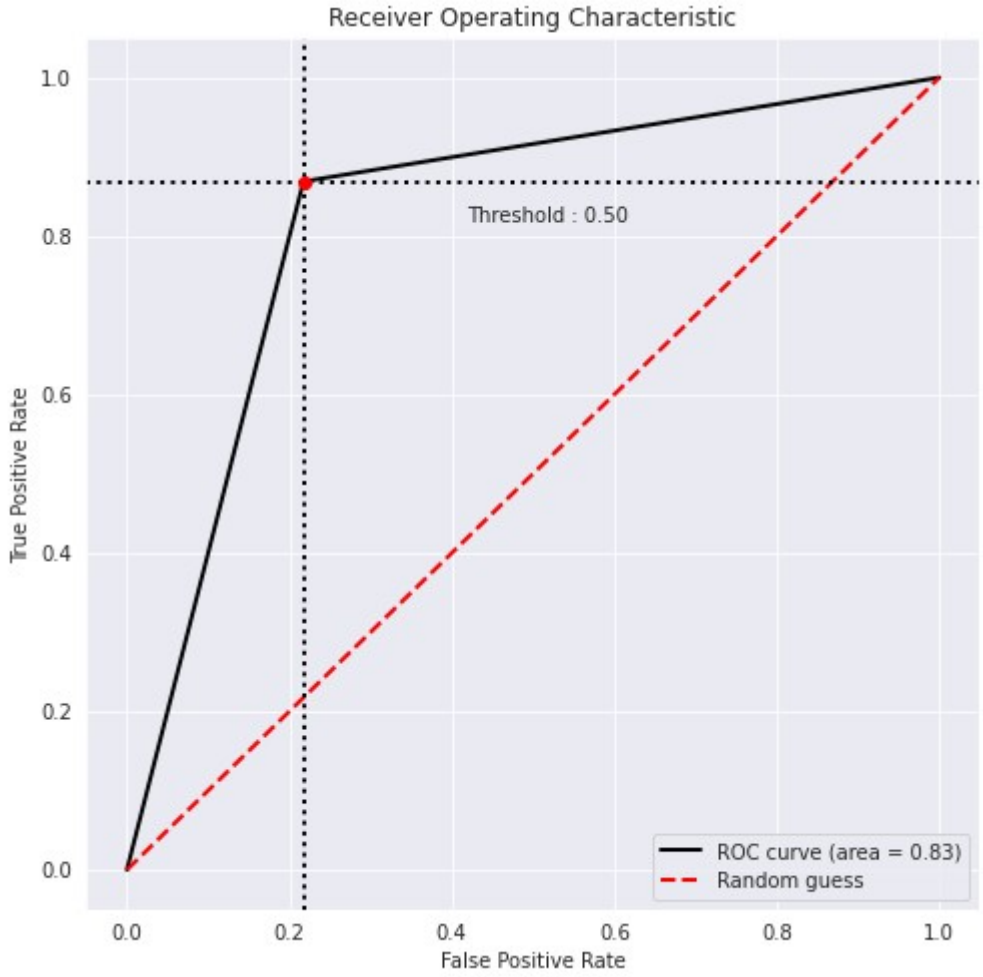
La matrice di confusione fornisce informazioni utili per valutare l'accuratezza, la sensibilità, la specificità e altre metriche di valutazione delle prestazioni di un modello di classificazione.

Predicted	0	1	All
Attuale			
0	34	19	53
1	7	42	49
All	41	61	102

Facendo un confronto tra i quattro algoritmi possiamo vedere che il modello Random Forest e quello che ottiene il risultato più accurato superando il 80% di precisione nella predizione



Si grafica la curva ROC, la curva ROC (Receiver Operating Characteristic) è una rappresentazione grafica utilizzata per valutare le prestazioni di un modello di classificazione binaria. La curva ROC mostra la relazione tra la sensibilità (True Positive Rate) e la specificità (True Negative Rate) del modello al variare della soglia di classificazione.
Un modello ideale avrebbe una curva ROC che si avvicina il più possibile all'angolo in alto a sinistra del grafico, indicando una sensibilità elevata e una specificità elevata.
In sintesi, la curva ROC fornisce una rappresentazione visiva delle prestazioni di un modello di classificazione e consente di valutarne la sensibilità e la specificità al variare della soglia di classificazione.



Il passo successivo e rispondere al test con le seguenti domande

1. Qual'è la tua identità sessuale? 0=maschio 1=femmina 2=altro
2. Qual’e la tua eta?
3. Come ti consideri? 1=amichevole 0=timido
4. Attualmente fumi? 1=sí 0=no
5. Cosa fumi? 0=niente 2=tabaco 3=marihuana 4=hashish 5=cocaína
6. Consumi bevande alcoliche? 1=si 0=no
7. Quanto spesso bevi alcolici✦
8. Senti che hai problemi in famiglia o altro? 1=sí 0=no
9. I tuoi genitori sono? 0=sposati 1=separati 2=divorziati 3=vedovo 4=altro
10. Hai una relazione amorosa? 1=si 0=no
11. Ti senti in solo anche in compagnia di altre persone 1=si 0=no
12. Soffri di insonnio? 1=si 0=no
13. Hai un disturbo alimentare? 1=si 0=no
14. Sei felice? 1=si 0=no
15. Hai mai desiderato che tu fossi morto? 1=si 0=no

Le risposte alle domande vengono salvate in un array numpy per poi essere processato per il predittore dando una risposta di Suicidio o Non Suicidio e la probabilita che il suicidio avenga