Immagine che contiene decorato, colorato, guarnizioni

Descrizione generata automaticamente

**Autori: Simone Della Porta, Antonio De Lucia, Antonio Maddaloni, Rocco Iuliano, Francesco Peluso**

1. **Business Understanding**
   1. **Introduzione al problema**
   2. **Obiettivi**
   3. **Descrizione dell’ambiente**
      1. **Specifica PEAS**
   4. **Business Success Criteria**
   5. **Tool da utilizzare**
2. **Data Understanding**
   * 1. **Scelta del dataset**
     2. **Analisi del dataset**
3. **Data Preparation**
   * 1. **Data Cleaning**
     2. **Feature Scaling**
     3. **Feature Engineering**
     4. **Data Balancing**
4. **Data Modeling**
5. **Evaluation**
6. **Business Understanding**
   1. **Introduzione al problema**

Con l’avvenire della pandemia del covid-19 iniziata nel 2019 molte organizzazioni sanitarie stanno cercando di tener traccia dei contagi e del tasso di mortalità dovuti ad essa; tutto ciò per limitare l’aumento della curva dei contagi.

* 1. **Obiettivi**

Dato il problema in esame, l’obiettivo del nostro progetto è realizzare un modello di machine learning che sia capace di predire se una persona risulta essere positiva o negativa al covid-19 sulla base di alcuni sintomi.

* 1. **Descrizione dell’ambiente**
     1. **Specifica PEAS**

|  |  |
| --- | --- |
| **PEAS** | |
| **Performance** | La misura di performance dell’agente è la sua capacità di avvicinarsi quanto più possibile ad una situazione ideale nella quale vengono predetti correttamente i veri positivi e negativi al covid-19. |
| **Environment** | L’ambiente in cui opera il nostro agente riguarda l’ambito sanitario e più in particolare la virologia. L’ambiente è:   * **Statico** in quanto l’ambiente non cambia nel tempo e non cambia alle azioni effettuate dell’agente. * **Episodico** in quanto un’azione intrapresa dall’agente in un dato istante non è influenzata dall’azione effettuata precedentemente. * **Completamente osservabile** in quanto l’agente in ogni istante può avere una visione completa dell’ambiente in cui è calato. * **Discreto** in quanto l’agente può effettuare soltanto determinate azioni e ricevere determinati impulsi. * **Non noto** in quanto l’agente non conosce il risultato delle proprie azioni. * **Singolo** in quanto l’ambiente prevede al proprio interno l’introduzione di un singolo agente.   Gli elementi dell’ambiente sono le persone che hanno determinati sintomi e che possono essere possibili positivi al covid-19. |
| **Actuators** | L’agente agisce sull’ambiente tramite lo stream di output del nostro computer nel quale andrà a predire la positività di una persona. |
| **Sensors** | L’agente riceverà gli impulsi tramite lo stream di input del nostro computer. |

* 1. **Business success criteria**

Il criterio con il quale andremo a validare il nostro sistema è il seguente:

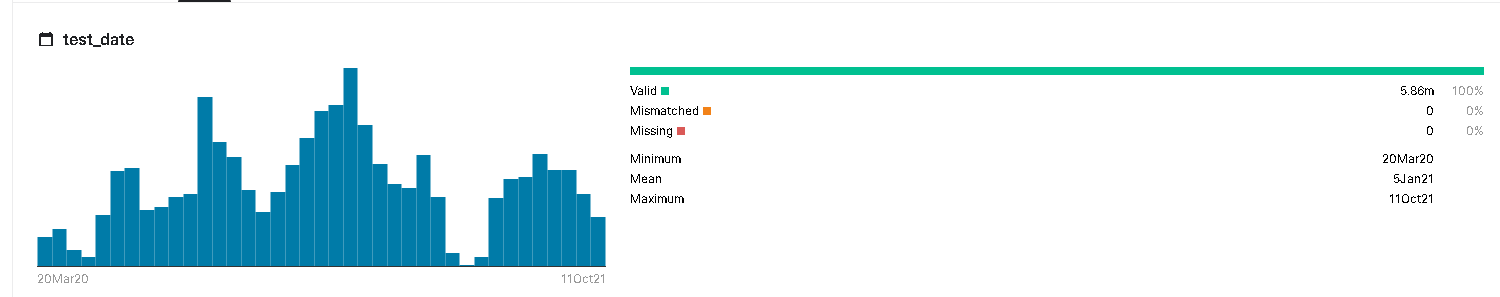
* + - * + Il modello deve avere un buon livello di accuracy, ovvero almeno il 50%. Questo perché essendo un problema di natura medica e trattiamo un virus nuovo, vuol dire che i dati a disposizione per far apprendere il modello non sono sicuramente accurati. Inoltre è da notare che il modello predice la positività o meno di una persona in base ai sintomi e non è detto che se una persona ha determinati sintomi allora ha il covid, ma potrebbe avere altre patologie che comportano sintomi simili. Un ulteriore motivo per il quale abbiamo scelto questa soglia di accuracy è che il modello potrà sbagliare molto facilmente con persone asintomatiche in quanto non hanno sintomi e risultano come persone “negative” agli “occhi” del modello.
  1. **Tool da utilizzare**

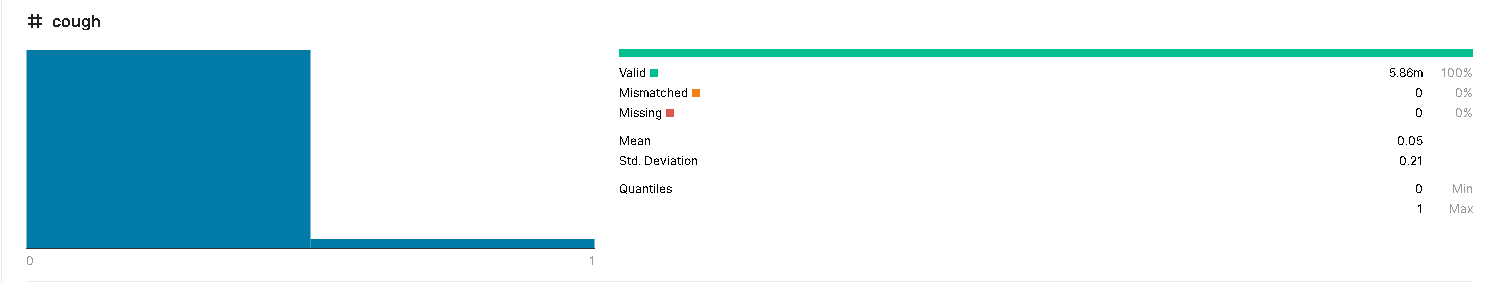
I tool da utilizzare per sviluppare il progetto sono:

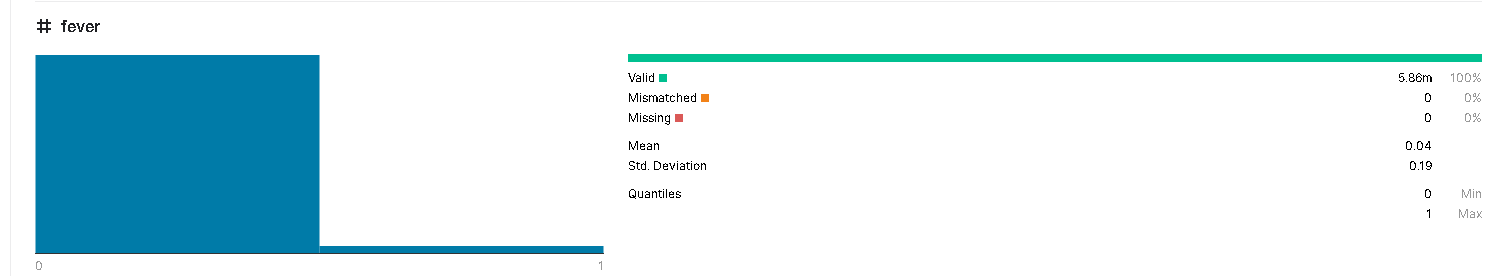
* + - * + Visual Studio Code
        + Python
        + Pandas
        + GitHub
        + Scikitlearn
        + Mathplot
        + Kaggle
        + Overleaf

1. **Data Understanding**
   1. **Scelta del dataset**

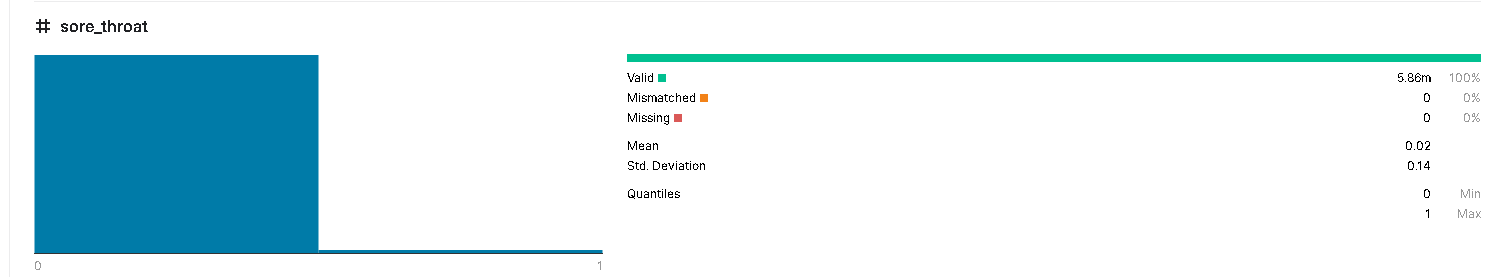
* Per la creazione del nostro machine learner, abbiamo analizzato vari dataset in rete e abbiamo deciso che quello più opportuno alle nostre esigenze e al raggiungimento dei nostri obbiettivi è il seguente: <https://www.kaggle.com/mykeysid10/covid19-dataset-for-year-2020>.
  1. **Analisi del dataset**
* Nel dataset in questione i dati sono stati collezionati nel lasso di tempo Marzo 2020-Novembre 2021, i dati sono stati reperiti dal sito del governo di Israele (<https://data.gov.il/dataset/covid-19/resource/d337959a-020a-4ed3-84f7-fca182292308>)
* Nel dataset sono presenti 10 colonne che sono:
  + **test\_data**: rappresenta la data di esecuzione del test. La data è espressa nel formato americano, quindi yyyy-mm-dd. Riportiamo l’analisi di questa caratteristica:

****

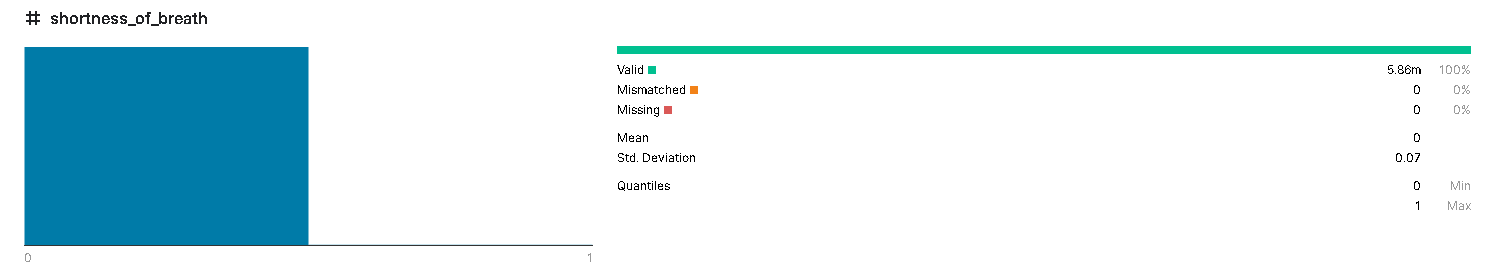
* + **cough**: indica se la persona presenta il sintomo della tosse. Viene espresso sotto forma di numero binario (1 ha la tosse e 0 non ha la tosse). Riportiamo l’analisi di questa caratteristica:
  + **fever**: indica se la persona presenta il sintomo della febbre. Viene espresso sotto forma di numero binario (1 se ha la febbre e 0 se non ha la febbre). Riportiamo l’analisi di questa caratteristica:



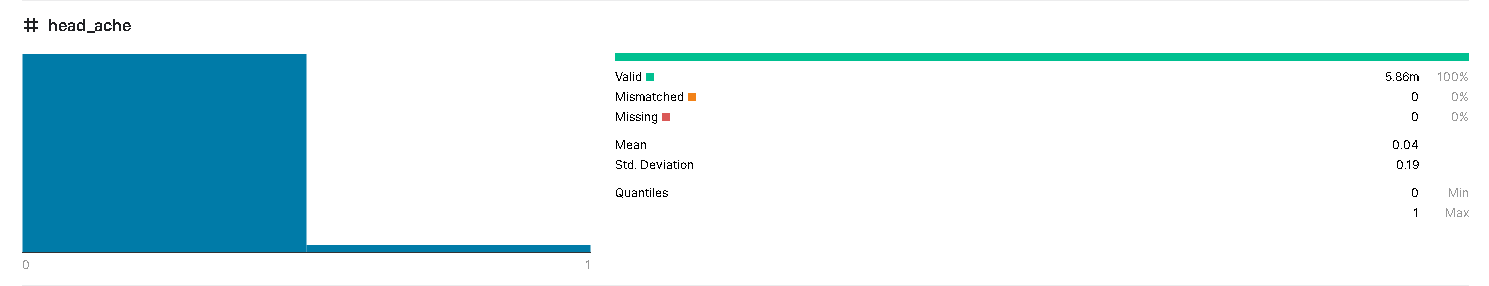
* + **sore\_throat**: indica se la persona presenta il sintomo del mal di gola. Viene espresso sotto forma di numero binario (1 se ha mal di gola e 0 se non ha mal di gola). Riportiamo l’analisi di questa caratteristica:



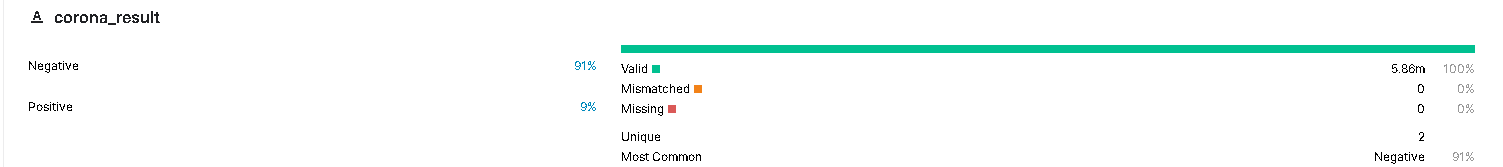
* + **shortness\_of\_breath:** indica se la persona presenta problemi respiratori ovvero fiato corto. Viene espresso sotto forma di numero binario (1 se ha il fiato corto e 0 se non ha il fiato corto). Riportiamo l’analisi di questa caratteristica:



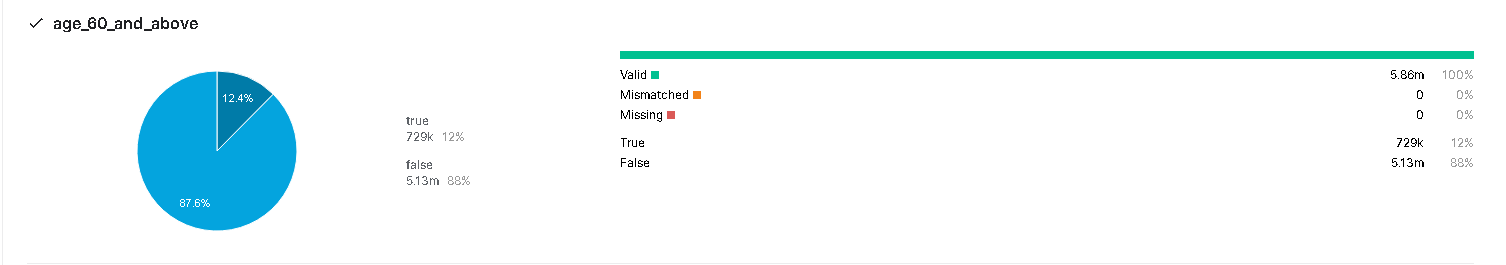
* + **head\_ache**: indica se la persona presenta mal di testa. Viene espresso sotto forma di numero binario (1 se ha mal di testa e 0 se non ha mal di testa). Riportiamo l’analisi di questa caratteristica:



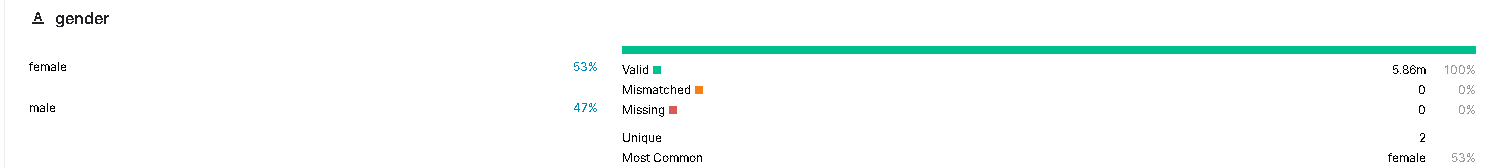
* + **corona\_result**: indica se la persona è positiva o negativa al covid. Viene espresso sotto forma di categoria (positivo o negativo). In questo dataset questa caratteristica rappresenta la **variabile dipendente.** Riportiamo l’analisi di questa caratteristica:



* + **age\_60\_and\_above**: indica se la persona ha più di 60 anni o meno di 60 anni. Viene espresso sotto forma booleana (yes o no). Riportiamo l’analisi di questa caratteristica:

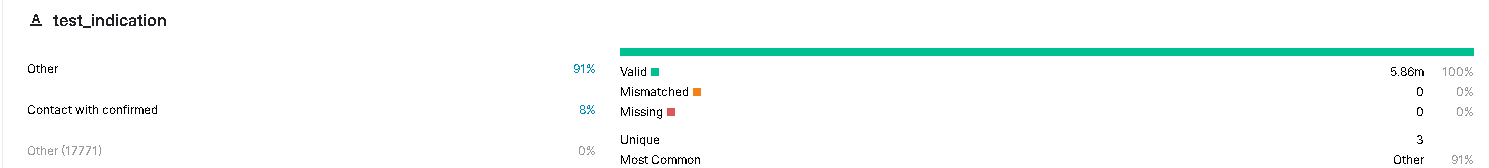


* + **gender**: indica se la persona è maschio o femmina. Viene espresso sotto forma di categoria (male o female). Riportiamo l’analisi di questa caratteristica:



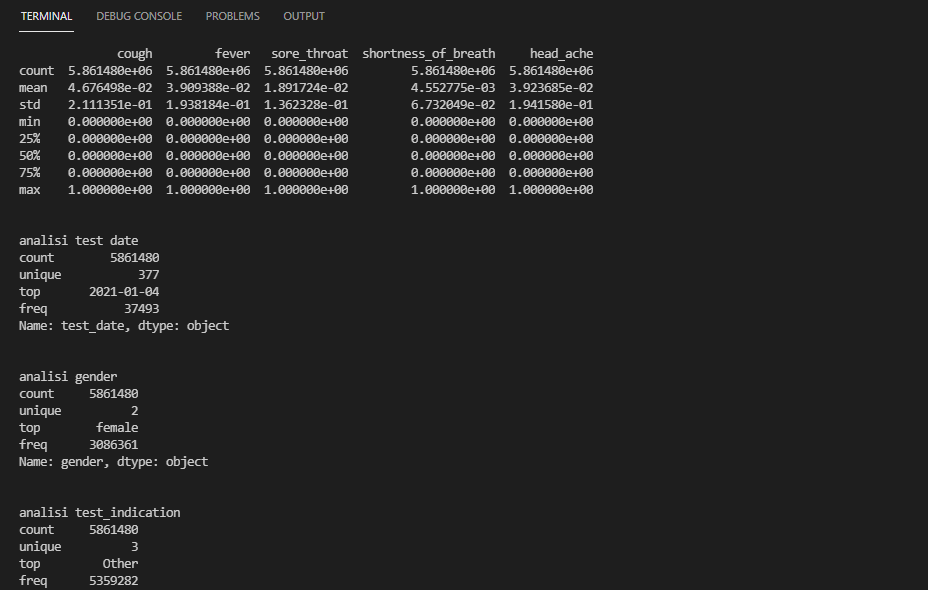
* + **test\_indication**: indica se la persona ha avuto possibili contatti con persone positive. Viene espresso sotto forma di categoria:
    - contact with confirmed indica che la persona ha avuto un contatto con una persona positiva;
    - other indica che la persona non è sicura di aver avuto contatti con positivi;
    - above indica che la persona non è sicura di aver avuto contatti con positivi ma torna da un viaggio all’estero.

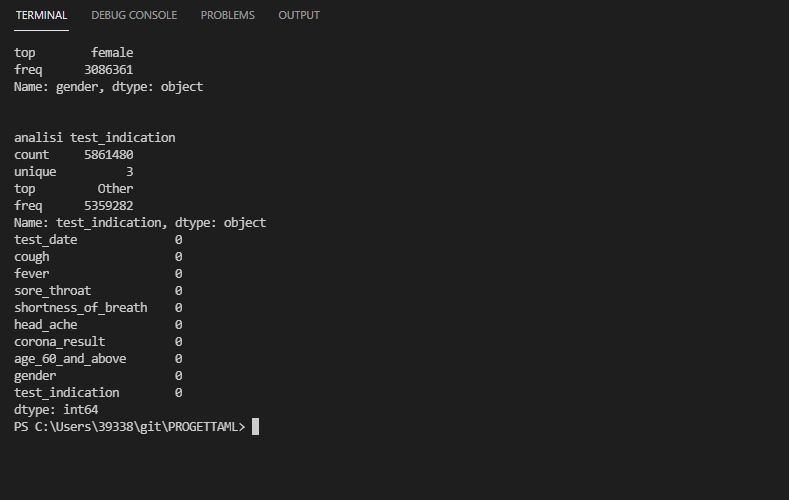
Riportiamo l’analisi di questa caratteristica:



*Nel dataset sono presenti 5.86 milioni di righe e il dataset ha la dimensione di 273mb.*

* 1. **Data quality:** Nel dataset non ci sono dati mancanti e la maggior parte dei dati sono rappresentati sulla stessa scala. C’è bisogno di prestare maggiore attenzione ad alcune caratteristiche come test\_date e test\_indication nella fase di data preparation. Questo perché test\_date e test\_indication hanno un formato e una scala diversa.
  2. **Esplorazione**: non sono state individuate correlazione tra dati.

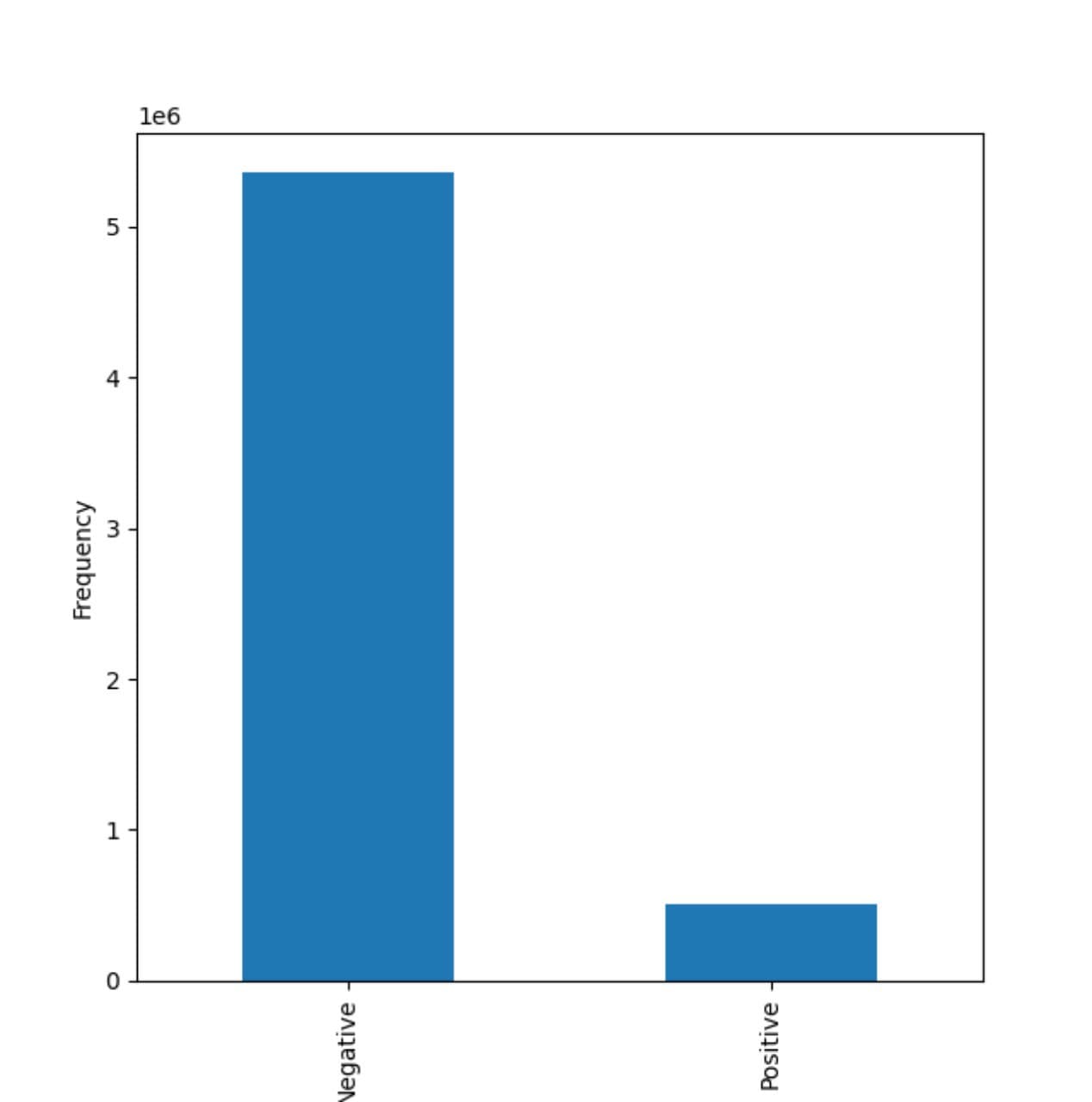


****

1. **Data Preparation**
   1. **Data Cleaning:** dalla precedente analisi dei dati abbiamo notato che non ci sono dati mancanti o rumorosi e quindi non utilizzeremo nessuna tecnica o algoritmo di data imputation.
   2. **Define data and Target Variables:** la variabile target del nostro dataset è la variabile “corona\_result” ed è la settima colonna del dataset
   3. **Data Balancing:** plottando i dati su un grafico a barre abbiamo notato che il nostro dataset è sbilanciato in quanto presenta molti casi negativi rispetto a quelli positivi. Infatti questi ultimi sono solo 505673 mentre le restanti 5355807 righe sono rappresentate dai negativi.

**Immagine che contiene testo

Descrizione generata automaticamente**

****

Per bilanciare il dataset utilizzeremo la tecnica di Undersampling perché avendo a disposizione molte righe non abbiamo il problema della sottodimensionalità dei dati. Non utilizziamo Oversampling perché la differenza tra il numero di istanze tra la classe di maggioranza e quella di minoranza è notevole e quindi facendo Oversampling aumenterebbe la probabilità di generare fake data.

Prima di effettuare l’Undersampling abbiamo diviso il dataset in test set e training set con le seguenti proporzioni: 30% e 70% rispettivamente. Questo perché non bisogna mai bilanciare il test set altrimenti andremo a falsificare il problema in analisi. Il numero di istanze ottenute per i set sono:

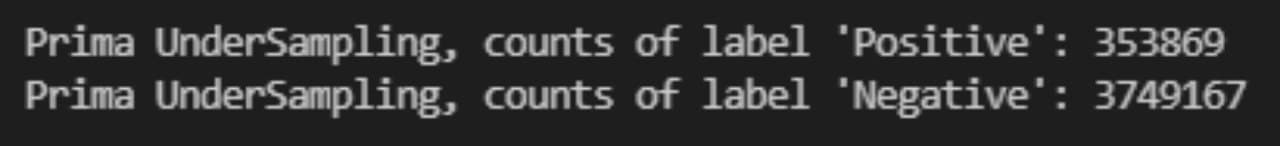
Immagine che contiene testo

Descrizione generata automaticamente

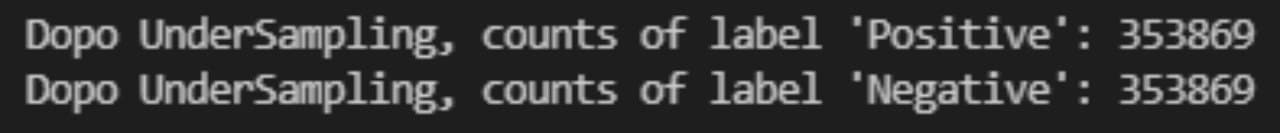
Con:

* X\_train indichiamo le variabili indipendenti del training set
* y\_train indichiamo la variabile dipendente del training set
* X\_test indichiamo le variabili indipendenti del test set
* y\_test indichiamo la variabile dipendente del test set

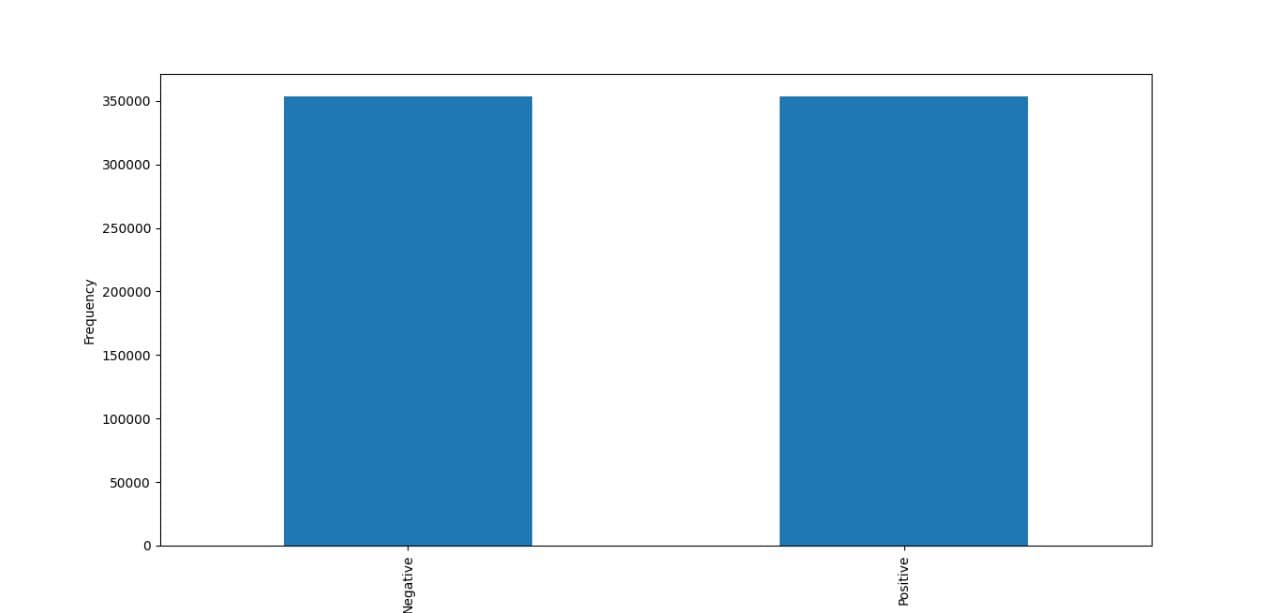
L’algoritmo utilizzato per effettuare l’Undersampling è il RandomUnderSempler. La situazione prima di applicare l’algoritmo sul training set era la seguente:



La situazione dopo aver applicato l’algoritmo sul training set è la seguente:



Plottando il dataset di train bilanciato otteniamo il seguente grafico:



Durante l’addestramento e il test del modello, verificheremo i risultati ottenuti sia applicando il data balancing e sia senza applicarlo.

* 1. **Feature Scaling:**
* All’interno del nostro dataset sono presenti feature con scala diversa, questo potrebbe portare un problema al machine learner in quanto potrebbe mal interpretare i dati e considerare una feature più importante di un’altra. Inoltre, il nostro dataset contiene il campo test\_date che potrebbe risultare ambiguo al machine learner poiché non spiega alla perfezione quello che è il nostro fenomeno che stiamo analizzando. Quindi per dare maggior interpretabilità abbiamo scalato la data in base alla curva dei contagi, infatti, abbiamo notato che la curva tende a crescere nei mesi da gennaio a maggio e da ottobre a dicembre, mentre tende a decrescere da giugno a settembre. Le feature che abbiamo normalizzato sono le seguenti:
* Gender (abbiamo sostituito male=1 e female=0)
* Test\_date (le date comprese tra gennaio e maggio e da ottobre a dicembre le abbiamo sostituite con 1 mentre le date da giugno ad agosto con 0)
* Test\_indication (abbiamo accorpato Other e Abroad con valore 0 in quanto entrambi i valori esprimono l’incertezza della persona se ha avuto contatti con un positivo, mentre Contact with confirmed con 1)
* Age\_60\_and\_above (abbiamo sostituito Yes=1 e No=0)

Abbiamo riportato tutte le feature su scala binaria proprio perché la maggior parte delle caratteristiche all’interno del nostro dataset sono sulla medesima scala.

**Immagine che contiene testo

Descrizione generata automaticamente**

**Immagine che contiene testo

Descrizione generata automaticamente**

**Immagine che contiene testo

Descrizione generata automaticamente**

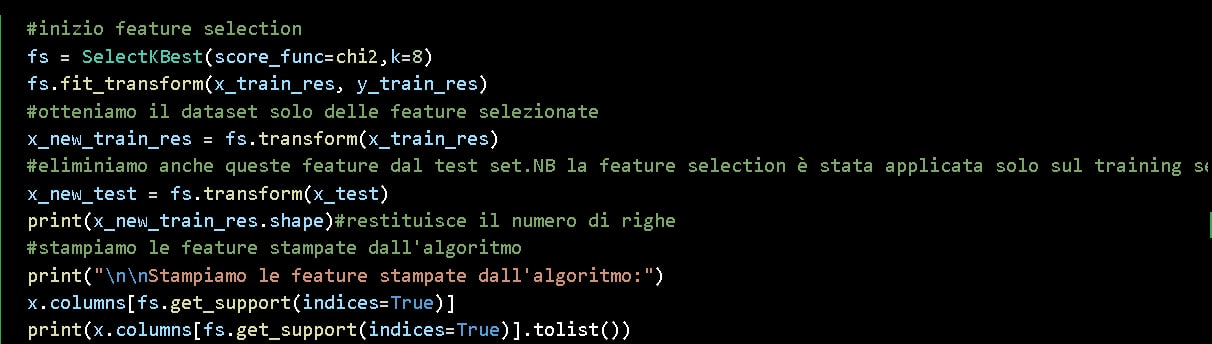
* 1. **Feature Engineering:**

Sulla base dei dati ottenuti fin ora, abbiamo analizzato le feature e attraverso la conoscenza del dominio siamo arrivati alla conclusione che le 9 variabili indipendenti sono molto rilevanti per la predizione.

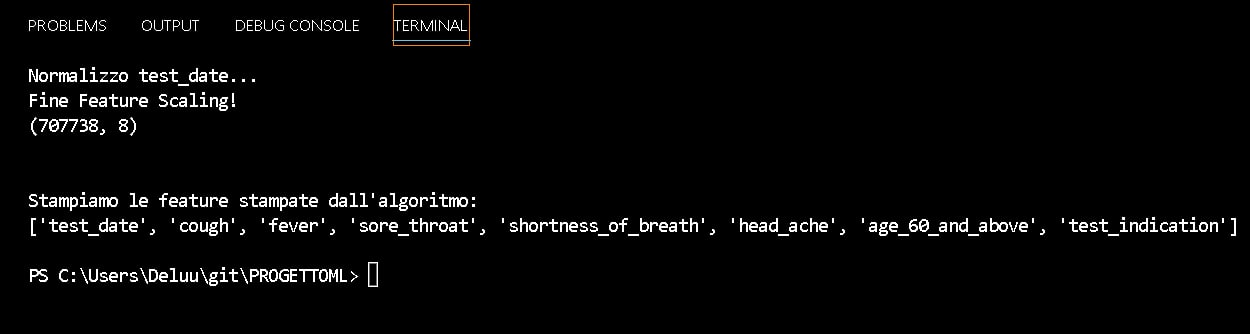
Nonostante ciò, andremo comunque ad effettuare feature selection per confrontare i risultati che otterremo dal modello di machine learning (analizziamo il modello applicando feature selection e non).

Per i dati su cui attuiamo feature selection utilizzeremo la tecnica dell’eliminazione univariata di feature attraverso l’algoritmo SelectKbest, quindi andremo a selezionare le K feature migliori correlate con la variabile dipendente.

In questo caso il K scelto è 8.



**Feature selezionate:**

****

1. **Data Modeling:**

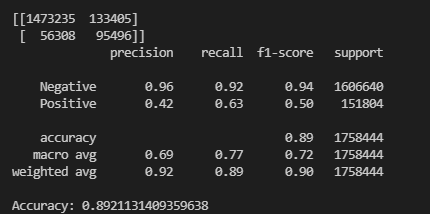
Dato che ci troviamo di fronte ad un problema di classificazione, ovvero ad un’istanza di un problema di apprendimento supervisionato; abbiamo deciso di utilizzare due tecniche differenti per realizzare il modello:

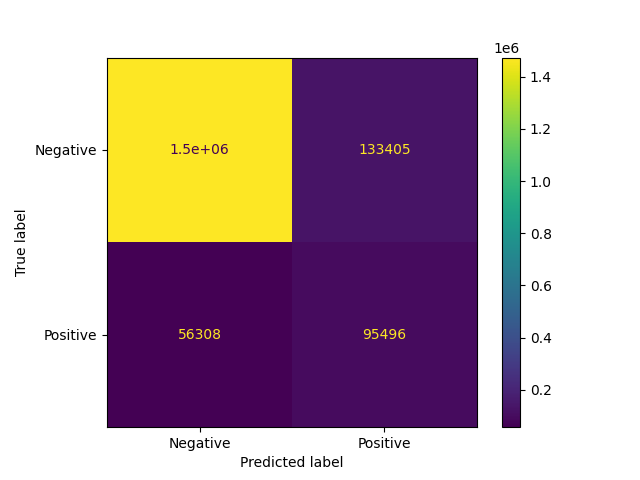
- Decision Tree (algoritmo basato su entropia).

- Naive Bayes (algoritmo probabilistico).

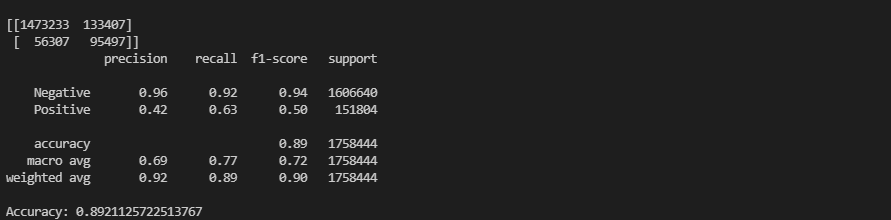
In questo modo, nella fase successiva di validazione, analizzeremo i risultati ottenuti dei due modelli in modo da poterli confrontare e determinare il modello migliore da utilizzare per il problema in questione. (Ovviamente li setteremo anche su diverse configurazioni delle fasi precedenti)

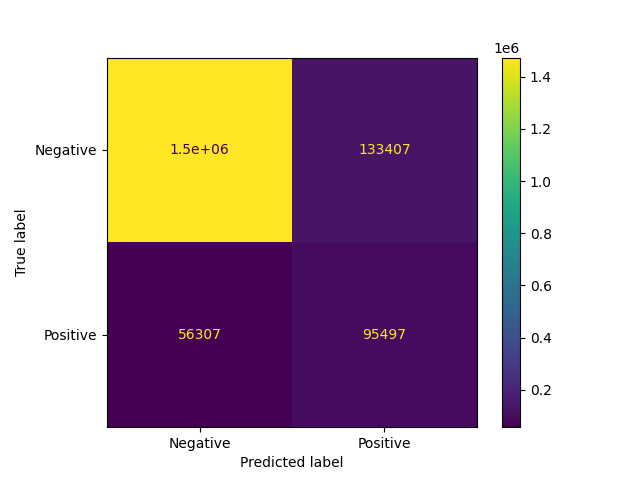
Report dei risultati del Decision Tree con data balancing e senza feature selection:



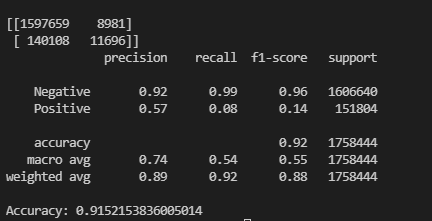
****

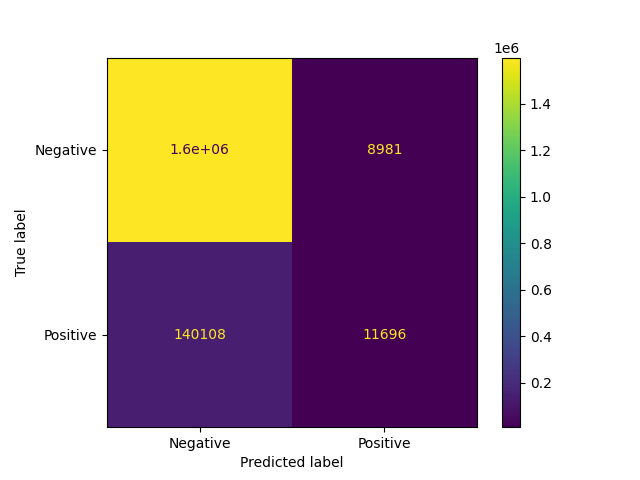
Report dei risultati del Decision Tree con data balancing e con feature selection:

****

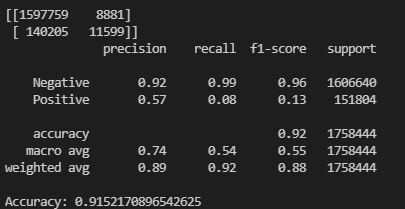
****

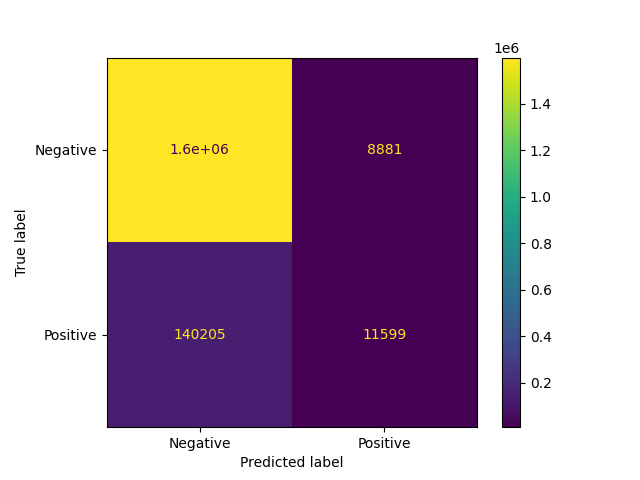
Report dei risultati del Decision Tree senza applicare data balancing e feature selection:



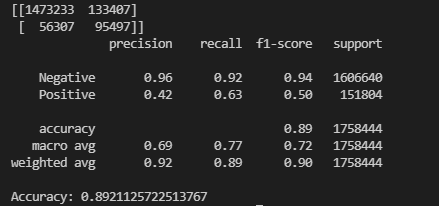


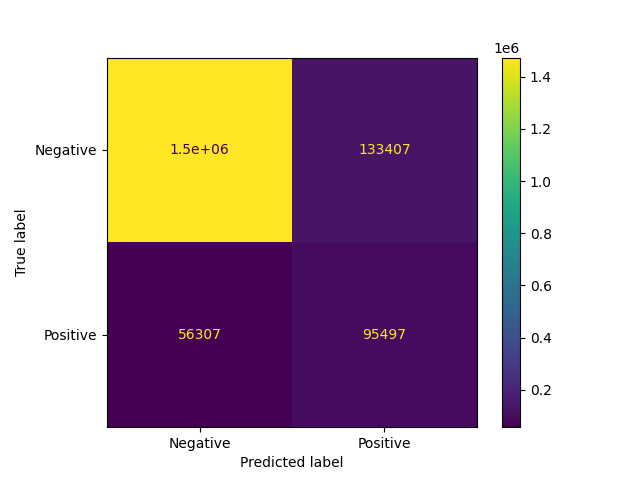
Report dei risultati del Decision Tree senza applicare data balancing ma applicando feature selection:



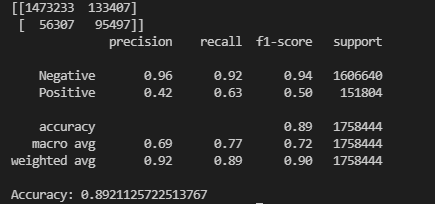


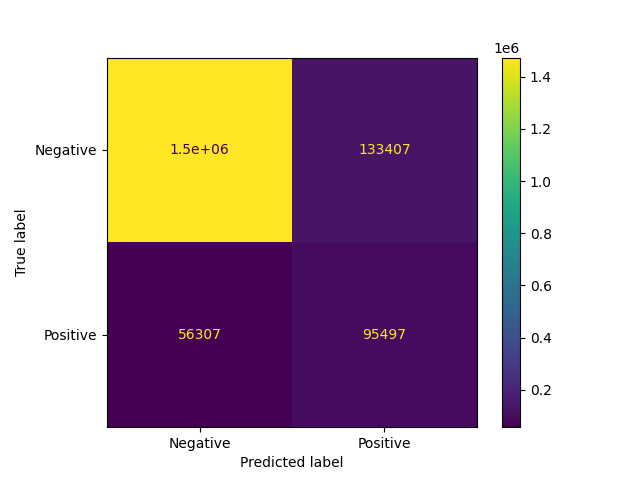
Report dei risultati del Naive Bayes con data balancing e con feature selection:





Report dei risultati del Naive Bayes con data balancing e senza feature selection:

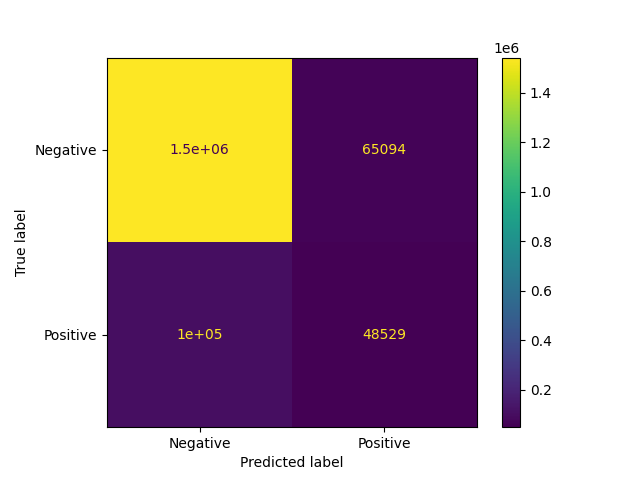




Report dei risultati del Naive Bayes senza applicare data balancing ma applicando feature selection:

Immagine che contiene testo, schermo

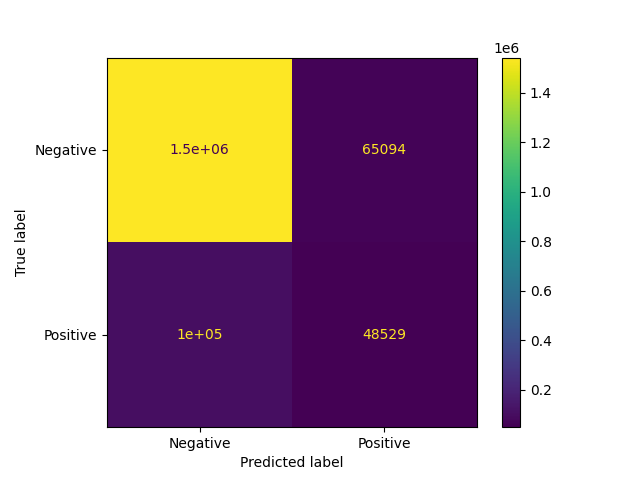
Descrizione generata automaticamente



Report dei risultati del Naive Bayes senza applicare data balancing e feature selection:

Immagine che contiene testo, schermo

Descrizione generata automaticamente



1. **Evaluation**

Osservando i risultati ottenuti applicando i due classificatori, ovvero Naive Bayes e Decision Tree, notiamo che quest’ultimo ha un’accuracy leggermente migliore di Naive Bayes, mentre per le altre metriche otteniamo gli stessi esiti. È comunque da tener in considerazione che il nostro progetto tratta un problema di natura sbilanciato e anche applicando il balancing, il classificatore sarà comunque più bravo a stabilire se una persona è negativa rispetto a stabilire se una persona è positiva. Quindi non possiamo fidarci ciecamente delle metriche, infatti nel visualizzare i risultati ottenuti senza applicare data balancing, potremmo essere tratti in inganno, dato che la precision per i positivi è aumentata così come per l’accuracy; questo perché il classificatore ha più istanze negative da cui poter apprendere e di conseguenza è più bravo a differenziare quest’ultimi dai positivi e quindi ottiene un’accuracy e una precison migliore (ma possiamo notare che la recall dei positivi è diminuita drasticamente). Possiamo affermare che il data balancing deve essere applicato altrimenti non possiamo fidarci delle metriche; ma per quanto riguarda l’operazione di feature selection, notiamo che il cambiamento degli esiti nell’applicare o meno questa tecnica non è significante poiché è minimo. Infatti, la feature eliminata (gender) non è molto rilevante per la predizione che deve effettuare il modello. Ovviamente se vogliamo avere una maggiore velocità di apprendimento del modello dobbiamo applicare la tecnica di feature selection anche se la differenza è lieve. Quindi per il problema in analisi conviene applicare come classificatore il Decision Tree con la seguente configurazione: applicando sia data balancing e sia feature selection. Inoltre il nostro progetto rispetta i business success criteria in quanto bisogna avere almeno il 50% di accuracy ma non viene richiesto che il modello abbia una certa velocità e quindi potremmo scegliere una delle due configurazioni (applicando feature selection o non).