Report del progetto di Machine Learning

Ferrario Tommaso Matr. 869005 (@TommasoFerrario18)

Terzi Telemaco Matr. 865981(@Tezze2001)

Vendramini Simone Matr. 866229(@Svendra4MySelf)

6 febbraio 2024

Indice

Capitolo 1

Introduzione

Questo è un progetto per l'esame di Machine Learning del primo anno del corso di laurea magistrale in informatica dell'Università degli Studi di Milano-Bicocca.

L'intero progetto si basa sul riconoscimento della presenza di un tumore al cervello data l'immagine di una risonanza magnetica. Il dataset scelto per questo progetto è scaricabile dal seguente link.

Per il riconoscimento del tumore sono stati allenati i seguenti modelli di machine learning:

- **SVM**: è stato scelto questo modello vista la buona capacità teorica nel generalizzare.
- **Naive Bayes Gaussiano**: è stato scelto questo modello dal momento che è l'unico ad essere probabilistico.
- **Rete neurale**: è stato scelto questo modello per confrontare i primi due con una soluzione neurale.

La relazione è stata suddivisa nei seguenti capitoli:

- **Introduzione**: descrizione del dominio e presentazione dei modelli che verranno presi in considerazione per questo progetto.
- **Dataset**: descrizione di come è stato costruito il dataset a partire dalle immagini, ovvero come sono state ricavate le features, e analisi esplorativa.
- **Rete neurale**: descrizione e analisi delle performance della rete.
- **SVM**: descrizione e analisi delle performance delle SVM.
- **Naive Bayes Gaussiano**: descrizione e analisi delle performance per Naive Bayes Gaussian.
- **Analisi dei risultati**: analisi comparata dei risultati tra i tre modelli considerati.
- **Conclusioni**: conclusioni sull'elaborato.

Capitolo 2

Dataset

2.1 Struttura del dataset

Il dataset è composto da un set di 3762 immagini, ognuna delle quali è stata ottenuta dalla risonanza magnetica del cervello. Ad ogni immagine è stata associata un etichettata la quale permette di rappresentare la presenza o meno di un tumore al cervello. Nel dataset è presente una colonna, *Class*, nella quale sono presenti le etichette, ovvero:

- **Presenza del tumore**: $T = 1$
- **Assenza del tumore**: $T = 0$

Oltre alla colonna *Class*, il dataset è composto da altre 13 colonne, nelle quali sono presenti le feature calcolate sulle immagini. Tali features sono state ottenute calcolando i **momenti Hu** sulle immagini della risonanza magnetica. I momenti Hu catturano le informazioni di base sull'immagine come l'area dell'oggetto, il centroide, l'orientamento e altre proprietà.

Gli attributi di cui è composto il dataset possono essere suddivisi in 2 gruppi[1]:

1. **First Order Features**: forniscono informazioni legate alle distribuzione dei livelli di grigio dell'immagine. Queste features corrispondono alle statistiche descrittive calcolate sui valori di ciascun pixel dell'immagine. Le statistiche descrittive calcolate sono:
    - **Media**
    - **Varianza**
    - **Deviazione standard**
    - **Indice di asimmetria**
    - **Indice di kurtosis**
2. **Second Order Features**: forniscono informazioni a livello di composizione della texture dell'immagine.
    - **Contrasto**: misura la variazione locale dei livelli di grigio dei pixel. Maggiore sarà il valore allora maggiore sarà il contrasto dell'immagine.
    - **Energia**: misura l'eterogeneità o la variazione dell'intensità dell'immagine. Più piccolo è il valore allora meno variazioni di intensità e più omogenea sarà la texture, al contrario, più la texture è irregolare allora maggiore sarà il valore.
    - **ASM**: misura quanto sono distribuiti uniformemente i livelli di grigio nell'immagine. Più i livelli di grigio sono uniformemente distribuiti minore sarà il valore dell'indice.
    - **Entropia**: misura la randomicità dei livelli di grigio, quindi più piccolo è il valore, più la texture sarà uniforme.
    - **Homogeneous**: misura secondaria del contrasto. Più alto sarà l'indice allora minore sarà il contrasto dell'immagine.
    - **Dissimilarity**: misura quanto spesso differenti combinazioni dei valori di intensità dei pixel occorrono nell'immagine. Un valore alto indica che l'immagine ha una maggior variazione delle intensità dei pixel vicini, quindi più complessa sarà la texture.

- **Correlation**: misura la correlazione tra pixel nelle due diverse direzioni.
- **Coarseness**: misura quanto l'immagine è composta da regioni di intensità omogenea, ovvero quanto è granulare o fine la texture.

2.2 Analisi descrittiva

Il dataset è formato da un totale 3762 istanze, ciascuna delle quali è descritta da 15 attributi così suddivisi:

- 13 rappresentano le feature calcolate sulle immagini della risonanza magnetica. Tutti questi attributi sono di tipo *float*.
- 1, ovvero l'attributo *Image*, rappresenta l'identificativo dell'immagine a cui si riferisce l'istanza.
- 1, ovvero l'attributo *Class*, rappresenta l'etichetta associata all'immagine. Questo attributo è stato convertito da *intero* a *categorico* per poter effettuare la classificazione.

Caricato il dataset, è stato eseguito un controllo per verificare che non ci fossero valori nulli. In questo caso non sono stati trovati valori nulli, quindi non è stato necessario eseguire alcuna operazione per la gestione di tali valori.

Successivamente, si è proceduto con il calcolo delle statistiche descrittive per gli attributi numerici presenti nel dataset. Questa operazione ha permesso di ottenere le informazioni riportate nella tabella 2.1.

	Mean	Variance	Standard Deviation	Entropy	Skewness	Kurtosis	Contrast	Energy	ASM	Homogeneity	Dissimilarity	Correlation	Coarseness
count	3762	3762	3762	3762	3762	3762	3762	3762	3762	3762	3762	3762	3762
mean	9.488890	711.101063	25.182271	0.073603	4.102727	24.389071	127.961459	0.204705	0.058632	0.479252	4.698498	0.955767	7.458341e-155
std	5.728022	467.466896	8.773526	0.070269	2.560940	56.434747	109.499601	0.129352	0.058300	0.127929	1.850173	0.026157	0.000000e+00
min	0.078659	3.145628	1.773592	0.000882	1.886014	3.942402	3.194733	0.024731	0.000612	0.105490	0.681121	0.549426	7.458341e-155
25%	4.982395	363.225459	19.058475	0.006856	2.620203	7.252852	72.125208	0.069617	0.004847	0.364973	3.412363	0.947138	7.458341e-155
50%	8.477531	622.580417	24.951560	0.066628	3.422210	12.359088	106.737418	0.225496	0.050849	0.512551	4.482404	0.961610	7.458341e-155
75%	13.212723	966.954319	31.095889	0.113284	4.651737	22.640304	161.059006	0.298901	0.089342	0.575557	5.723821	0.971355	7.458341e-155
max	33.239975	2910.581879	53.949809	0.394539	36.931294	1371.640060	3382.574163	0.589682	0.347725	0.810921	27.827751	0.989972	7.458341e-155

Tabella 2.1: Statistiche descrittive degli attributi

Questa operazione ha permesso di ottenere alcune informazioni sul dataset. Innanzitutto, ha permesso di osservare che i valori associati agli attributi non sono standardizzati dal momento che nessun dato ha media 0 e deviazione standard 1.

La standardizzazione delle feature è stata fatta, utilizzando l'equazione 2.1, solo per le SVM e il modello neurale, mentre per Gaussian Bayes la standardizzazione viene fatta in automatico dalla libreria del modello.

$$F_{\mu,\sigma} = \frac{F - \mu}{\sigma} \tag{2.1}$$

In aggiunta alle statistiche descrittive, per ogni feature si è realizzato un istogramma attraverso il quale è stato possibile realizzare una prima analisi visiva sulla distribuzione dei valori. Questa operazione è stata svolta per verificare se le feature seguono una distribuzione normale.

TODO:immagine

Da questi grafici si evince che le feature *Energy*, *Homogeneity* e *Coarseness* non seguono una distribuzione normale, a differenza delle altre feature che possono essere considerate normali. In ogni caso, anche se non seguono l'ipotesi di normalità, si è deciso comunque di utilizzarli, anche se stiamo andando contro le assunzioni dei modelli che vogliamo utilizzare.

Da questa analisi, è stato possibile osservare che la feature *Coarseness* assume valori molto bassi, quindi è stato pensato di convertire questa feature ad una scala logaritmica, permettendo di aumentare la significatività dei valori.

Nonostante questa operazione, la feature *Coarseness* presenta una deviazione standard pari a circa 0.02, valore non particolarmente significativo, quindi questo ci ha permesso di escludere questa feature perché sicuramente non sarà la feature più discriminante.

Oltre all'analisi descrittiva delle feature, è stato eseguito anche un analisi sulle etichette, ovvero sulle classi. Questo è stato fatto per verificare se le classi sono bilanciate, ovvero se il numero di esempi positivi è simile al numero di esempi negativi.

In questo caso, realizzando un istogramma, riportato in figura 2.1, che rappresenta le frequenze assolute delle classi. Da questo si può notare che la distribuzione di esse è bilanciate, infatti, circa il 55% degli esempi sono negativi (assenza del tumore), contro circa il 45% degli esempi sono positivi (presenza del tumore).

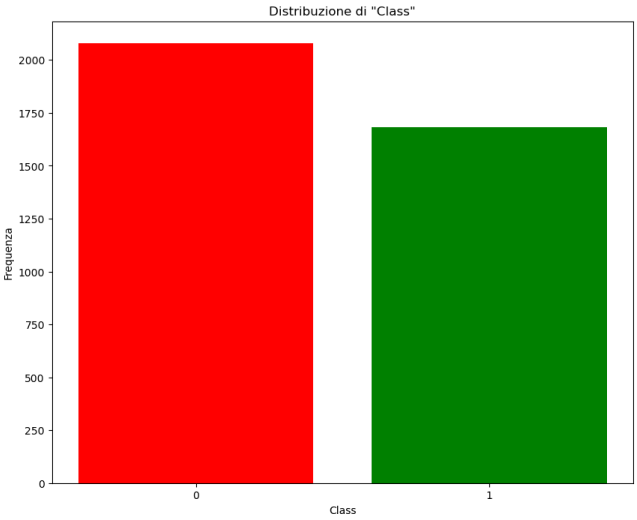


Figura 2.1: Distribuzione delle classi

Non è già stato detto prima? Successivamente risulta importante analizzare se le distribuzioni si avvicinano ad una normale standard, di conseguenza sono stati disegnati i barplot delle frequenze di ciascuna feature. Dai grafici si evince che tutte le feature eccetto *Energy* e *Homogeneity* sono distribuite normalmente.

Dal barplot delle features creato precedentemente, si possono ottenere ulteriori informazioni sulla distribuzione potenzialmente normale delle altre feature. Infatti, la distribuzione della *Standard deviation* è simmetrica, mentre tutte le altre presentano una asimmetria.

Ricapitolando, da questo primo studio descrittivo è stato modificato il dataset rimuovendo la feature *Coarseness*.

TODO:immagine

2.2.1 Analisi delle correlazioni

Il passaggio successivo è stato quello di analizzare le correlazioni tra le feature. Questo è stato fatto in modo tale da ridurre la dimensionalità dei dati, nello specifico sono state suddivise le feature in gruppi di feature altamente correlate tra loro, e in un secondo momento è stata scelta una feature per ogni gruppo.

Per fare ciò, è stata realizzata una matrice di correlazione, riportata in figura 2.2, attraverso la quale è stato possibile osservare le correlazioni tra le feature.

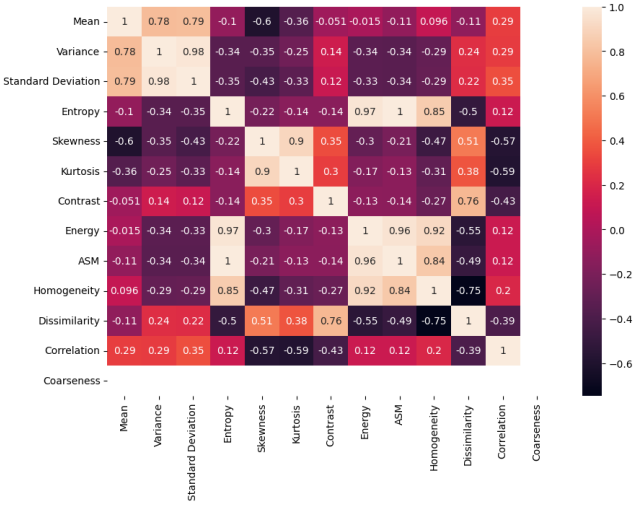


Figura 2.2: Matrice di correlazione

Dall'analisi di questa matrice, si possono osservare diverse correlazioni tra le feature. Innanzitutto, si può notare una forte correlazione positiva tra le feature *Mean*, *Variance* e *Standard deviation*. Questa correlazione

è facilmente spiegabile analizzando le immagini prodotte dalle risonanze magnetiche. Infatti, essendo in bianco e nero, se la media tende a 1 (colore bianco) allora la varianza e la deviazione standard aumentano, perché sono presenti diversi pixel bianchi. Questo comporta che le transizioni dal nero assoluto al bianco assoluto neccessitano di regioni di pixel maggiore rispetto ad una transizione tra nero assoluto e grigio (0.5).

Invece, la correlazione tra varianza è deviazione standard facilmente spiegabile perché la deviazione standard è la radice quadrata della varianza ($SD[P] = \sqrt{VAR[P]}$).

TODO:immagine

Una seconda forte correlazione positiva si può osservate tra le feature che misurano l'**uniformità dei livelli di grigio** dei pixel, più precisamente tra le feature *Entropy*, *ASM*, *Homogeneity* ed *Energy*. Queste feature quantificano delle informazioni legate alla texture dell'immagine, quindi la forte correlazione positiva può essere spiegata analizzando le texture delle immagini su cui vengono calcolate. Più precisamente se si ha un valore molto alto della feature *Entropy*, significa che la texture non è uniforme, ovvero si hanno strutture complesse e irregolari, quindi meno uniforme sarà la distribuzione dei livelli di grigio, aumentando l'indice di *ASM*, comportando di conseguenza un aumento delle variazioni di intensità dei livelli di grigio, aumentando di conseguenza anche l'indice di *Energy*, infine, (**cosa c'entra Homogeneity?**).

Al tempo stesso, la matrice di correlazione evidenza una forte correlazione positiva tra gli indici che misurano la **morfologia della distribuzione**, ovvero le feature di *Skewness* e *Kurtosis*. Questa dipendenza implica il fatto che più la distribuzione è leptokurtica (Kurtosis grande), ovvero la frequenza dei livelli di grigio dei pixel si concentrano interamente vicino alla media/mediana/ moda, allora più grande sarà la Skewness, ovvero maggiore sarà la tendenza ad avere frequenze di livelli di grigio più vicino al bianco (coda di destra più altra rispetto alla coda di sinistra).

La matrice della correlazione evidenzia anche una correlazione positiva tra le feature di *Contrast* e *Dissimilarity*, ovvero maggiore sarà il contrasto e maggiore sarà la complessità della texture.

In aggiunta dalla matrice si evidenza che le features di *Dissimilarity* e *Homogeneity* sono correlate negativo, ovvero maggiore sarà il contrasto allora minore è la complessità della texture.

Dalla correlazione delle features è possibile ridurre la dimensionalità del dataset considerando sono le seguenti features:

- **Mean**
- **Entropy**
- **Skewness**
- **Contrast**
- **Correlation**

A puro scopo didattico è stato pensato di eseguire i modelli non solo sul dataset semplificato eliminando le correlazioni, ma anche applicando l'algoritmo PCA.

2.3 PCA

Precedentemente è stato presentato un primo modo per ridurre la dimensionalità dei dati basandoci sull'analisi delle correlazioni. In seguito, è stato pensato di provare ad utilizzare un metodo di trasformazione delle feature per ridurre la loro dimensionalità e successivamente analizzare i risultati ottenuti. La scelta sul metodo da utilizzare è ricaduta su PCA.

Per prima cosa è statao necessario comprendere quante componenti delle nuove feature sono necessarie per avere la maggior parte della varianza spiegata. Per fare ciò sono state separate le features dalla classe di ciascun esempio, successivamente è stato effettuato il fit della PCA sull'istanza di ciascun dato e successivamente è stato plottato all'interno di uno scatter plot per ogni componente la sua percentuale di varianza spiegata.

TODO:immagine del plotting della PCA

Dallo scatter plot si evidenzia come con solo le prime 3 componenti, si riesce a raggiungere un totale dell'85% di varianza spiegata sui dati. In questo modo una volta eseguita la pca e applicata la trasformazione, si selezionano solo le prime 3 componenti dei nuovi dati trasformati. In aggiunta, dal momento che abbiamo ridotto la dimensionalità dei dati a 3 allora, aggiungendo anche la colonna target, si possono rappresentare all'interno di uno scatter plot a 3 dimensioni e si può notare come i dati sono separabili tramite un iperpiano.

TODO:immagine del plotting delle nuove componenti

Capitolo 3

Rete neurale

La precedente fase di analisi ha permesso di acquisire informazioni utili sulla struttura del dataset e di conseguenza permettere la selezione di un modello adatto a svolgere il compito di classificazione.

In questo capitolo verrà presentata la rete neurale utilizzata per svolgere il compito di classificazione. In particolare verrà presentata la struttura della rete neurale, il processo di addestramento e i risultati ottenuti.

Prima di presentare nel dettaglio la rete neurale, risulta necessario specificare come sono stati preparati i dati per l'addestramento della rete neurale.

3.1 Preparazione dei dati

La prima operazione svolta sui dati è stata la standardizzazione dei dati, in questo modo i dati sono stati trasformati in modo tale che la loro media sia 0 e la loro deviazione standard sia 1. Questa operazione è stata eseguita per garantire che la rete neurale non sia influenzata da valori di input con scale diverse.

La seconda operazione svolta sui dati è stata la suddivisione del dataset in training set e test set. Il training set è stato utilizzato per addestrare la rete neurale, mentre il test set è stato utilizzato per valutare le prestazioni della rete neurale. La suddivisione del dataset è stata effettuata in modo tale che il training set contenesse il 80% dei dati, mentre il test set contenesse il 20% dei dati.

La suddivisione dei dati è stata effettuata in modo da mantenere la stessa percentuale di dati positivi e negativi in entrambi i set. Questa operazione è stata effettuata per evitare che la rete neurale sia addestrata su un dataset sbilanciato.

3.2 Struttura della rete neurale

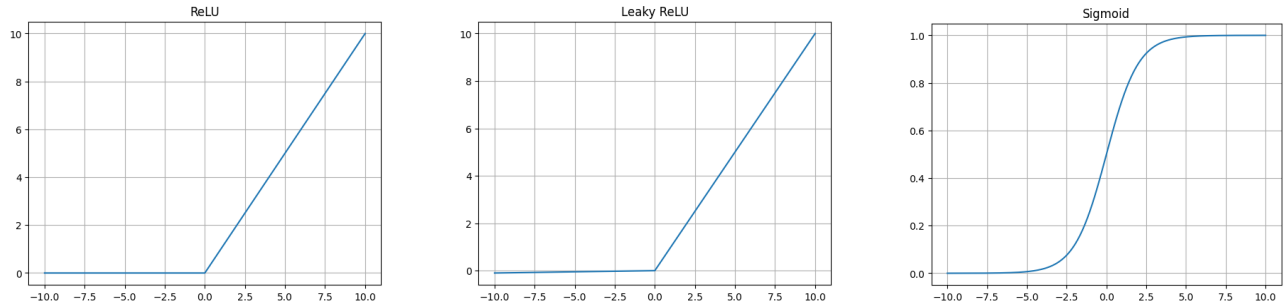
La rete neurale utilizzata per svolgere il compito di classificazione è una rete neurale feedforward. La struttura di questa rete neurale è stata definita in base ai risultati ottenuti dalla fase di analisi e attraverso un processo di grid search.

In particolare, dalla fase di analisi è stato selezionato un sottoinsieme di feature le quali sono state utilizzate come input della rete neurale. Questo sottoinsieme è composto da 5 elementi, il che ha permesso di definire la struttura del layer di input della rete neurale. Questo primo strato è composto da 5 neuroni, in cui la funzione di attivazione, come la struttura interna della rete, è stata definita attraverso un processo di grid search.

Nello specifico, il processo di grid search ha permesso di valutare le prestazioni della rete neurale al variare della funzione di attivazione, del numero di layer nascosti e del numero di neuroni per ogni layer nascosto. La ricerca della combinazione migliore di questi iperparametri è stata effettuata attraverso una cross validation a 5 fold, prendendo in considerazione solamente i dati del training set.

Visti i risultati ottenuti nella fase di analisi e la volontà di mantenere i tempi di addestramento bassi, si è scelto di mantenere una struttura di dimensioni ridotte per la rete neurale, in modo tale da evitare l'overfitting. Per questo motivo, l'operazione di grid search è stata effettuata prendendo in considerazione un numero di neuroni per layer nascosto tra 5, 10, 50, mentre il numero di layer nascosti è stato valutato tra 1 e 2. Mentre, per quanto riguarda la funzione di attivazione, sono state valutate le seguenti funzioni di attivazione: *relu*, *leaky relu* e *sigmoid*.

I risultati ottenuti da questo processo di grid search hanno permesso di definire la struttura della rete neurale. In particolare, la rete neurale è composta da:



(a) ReLU (b) Leaky ReLU (c) Sigmoide

Figura 3.1: Funzioni di attivazione utilizzate nella fase di grid search

Per concludere la descrizione della struttura della rete neurale, è necessario specificare come è composto l'ultimo layer, ovvero quello di output. Vista la natura del problema di classificazione, il layer di output è composto da un solo neurone, in cui la funzione di attivazione è la funzione sigmoide. Questa scelta è dovuta al fatto che tale funzione restituisce un valore compreso tra 0 e 1, il che permette di interpretare l'output della rete neurale come la probabilità che l'input appartenga alla classe positiva.

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{3.1}$$

Oltre alla ricerca

3.3 Addestramento della rete neurale

Definita la struttura della rete neurale, si è passati alla fase di addestramento di questa. In questa fase è stato necessario definire la funzione di perdita, l'algoritmo di ottimizzazione, il numero di epoche e la dimensione del batch.

Per quanto riguarda la funzione di perdita, è stata scelta la *binary crossentropy* in quanto è una funzione di perdita adatta a problemi di classificazione binaria. La scelta di questa funzione di perdita è dovuta alla natura del problema di classificazione che si vuole risolvere.

Gli altri iperparametri sono stati scelti attraverso un processo di grid search, nel quale sono state valutate le prestazioni della rete neurale al variare di questi iperparametri.

Questa fase di grid search è stata effettuando una cross validation a 5 fold, prendendo in considerazione solamente i dati del training set. Durante le varie iterazioni della cross validation, sono stati registrati i valori di accuratezza e tempo di addestramento della rete neurale.

Queste informazioni sono state utilizzate per scegliere i valori degli iperparametri che hanno permesso di ottenere i migliori risultati.

Per all'algoritmo di ottimizzazione il confronto è stato eseguito tra *Adam* e *SGD*, mentre per il numero di epoche e la dimensione del batch sono stati valutati i valori 100, 300 per il numero di epoche e 50, 100, 300 per la dimensione del batch.

3.4 Risultati

Il modello addestrato in precedenza è stato valutato sui dati che compongono il test set. In particolare sono state valutate le seguenti metriche: accuratezza, precisione, richiamo e F1 score. Oltre al calcolo di queste metriche, si è deciso di realizzare la curva ROC per il modello e di rappresentare la matrice di confusione.

Prima di presentare i risultati ottenuti, è necessario specificare che essendo il dataset riferito a un ambito medico, si è deciso di aggiustare il valore di threshold per la predizione del modello. In particolare, il valore di threshold è stato impostato a 0.3, in modo tale da ridurre il numero di falsi negativi.

Fatta questa precisazione, si può procedere con la presentazione dei risultati ottenuti. In particolare, nella tabella 4.1 sono presentati i risultati ottenuti dal modello addestrato.

I risultati ottenuti sono giustificati dal fatto che le due classi sono linearmente separabili.

Metrica	**Valore**
Accuratezza	??
Precisione	??
Richiamo	??
F1 score	??

Tabella 3.1: Risultati ottenuti dal modello addestrato

Capitolo 4

Bayes

La precedente fase di analisi ha permesso di acquisire informazioni utili sulla struttura del dataset e di conseguenza permettere la selezione di un modello adatto a svolgere il compito di classificazione.

In questo capitolo verranno presentati tutti i risultati ottenuti dall'apprendimento e dalle valutazioni effettuate sul modello Gaussian Naive Bayes. Da notare che si sta utilizzando Gaussian Naive Bayes pur sapendo che non tutte le features derivano da una distribuzione normale, siamo consci del fatto che non si stanno rispettando le assunzioni del modello.

4.1 Preparazione dei dati

La prima operazione svolta sui dati è stata la suddivisione del dataset in training set e test set. Il training set è stato utilizzato per addestrare la rete neurale, mentre il test set è stato utilizzato per valutare le prestazioni della rete neurale. La suddivisione del dataset è stata effettuata in modo tale che il training set contenesse il 80% dei dati, mentre il test set contenesse il 20% dei dati.

La suddivisione dei dati è stata effettuata in modo da mantenere la stessa percentuale di dati positivi e negativi in entrambi i set. Questa operazione è stata effettuata per evitare che la rete neurale sia addestrata su un dataset sbilanciato.

4.2 Addestramento di Gaussian Naive Bayes

Dal momento che il modello non ha degli iperparametri da stimare allora non è stato effettuato il processo di tuning degli iperparametri come è stato effettuato per la rete neurale. Di conseguenza è stato allenato direttamente il modello sul training set e successivamente è stata effettuata la sua valutazione calcolando le metriche di valutazione.

4.3 Risultati

Il modello addestrato in precedenza è stato valutato sui dati che compongono il test set. In particolare sono state valutate le seguenti metriche: accuratezza, precisione, richiamo e F1 score. Oltre al calcolo di queste metriche, si è deciso di realizzare la curva ROC per il modello e di rappresentare la matrice di confusione.

Prima di presentare i risultati ottenuti, è necessario specificare che essendo il dataset riferito a un ambito medico, si è deciso di aggiustare il valore di threshold per la predizione del modello. In particolare, il valore di threshold è stato impostato a 0.3, in modo tale da ridurre il numero di falsi negativi.

Fatta questa precisazione, si può procedere con la presentazione dei risultati ottenuti. In particolare, nella tabella 4.1 sono presentati i risultati ottenuti dal modello addestrato.

Metrica	**Valore**
Accuratezza	??
Precisione	??
Richiamo	??
F1 score	??

Tabella 4.1: Risultati ottenuti dal modello addestrato

I risultati ottenuti sono giustificati dal fatto che le due classi sono linearmente separabili.

Bibliografia

[1] Namita Aggarwal e RK Agrawal. “First and second order statistics features for classification of magnetic resonance brain images”. In: (2012).