Predict Cancer

Giovanni Carbone
Matricola 0512113208

Indice

1 Introduzione

L'accurata identificazione e la previsione del tumore sono essenziali per proteggere la salute umana e ridurre il rischio di gravi conseguenze legate alla malattia. La ricerca gioca un ruolo fondamentale nel migliorare il sistema sanitario, promuovendo pratiche preventive e diagnostiche.
Questo progetto mira a colmare questa lacuna sviluppando sistemi avanzati, affidabili e intelligenti per prevedere se un tumore è benigno o maligno.

2 Definizione del problema

2.1 Obiettivi

L'obiettivo del progetto è sviluppare un modello di Machine Learning che possa anlizzare dei dati riguardanti le caratteristiche del tumore, in particolare se è un tumore beinigno o un tumore maligno, poiché la corretta identificazione di esso è essenziale per capire poi come dev'essere trattato così anche da migliorare il riconoscimento.
Tra i dati messi a disposizione, per migliorare il riconoscimento, abbiamo caratteristiche specifiche come il raggio, si intende la media delle distanze del centro dei punti sul perimetro, la struttura, ossia una deviazione standard della scala di grigio, il perimetro, l'area e tante altre caratteristiche associate.

2.2 Analisi del problema

Il modello progettato ha lo scopo di condurre un'analisi per determinare se un tumore è maligno basandosi su diverse caratteristiche. Pertanto, l'istanza di questo problema è un problema di apprendimento supervisionato, più precisamente di classificazione binaria. Per risolvere questo problema saranno confrontate diverse tecniche di Machine Learning al fine di individuare quella più efficace.

3 Analisi dei dati

3.1 Acquisizione del dataset

Il dataset utilizzato per sviluppare il seguente modello è stato preso dal seguente link:
https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic.
Inizialmente, il dataset viene scaricato dal sito in formato dataframe, ma utilizzando la libreria "Pandas" del linguaggio di programmazione "Python", viene convertito in formato CSV per semplificarne la lettura dei dati.

3.2 Struttura del datatset

Il dataset è composto da 570 righe e 32 colonne, di cui due variabili categoriche e le restanti sono varaibili continue. In particolare, ogni variabile continua assume il proprio valore. Le caratteristiche a disposizioni sono:

- **ID**: È l'identificativo di una persona.
- **Diagnosis**: Può essere M=Malignant o B=Bening, cio serve ad indicare se un tumore è maligno o benigno.
- **Radius**: È la media delle distanze di ciascun punto sul perimetro del tumore dal suo centro, quindi andrà ad indicare quant'è grande il tumore, essendo una variabile continua può assumere un numero infinito di valori, come le altre che seguono.
- **Texture**: Rappresenta la variazione dei livelli di grigio all'interno dell'immagine del tumore.
- **Perimeter**: È la lunghezza del contorno del tumore.
- **Area**: È l'area all'interno del contorno del tumore.
- **Smoothness**: Indica la variazione della lunghezza del raggio all'interno del tumore.
- **Compactness**: È una misura della "piattezza" del tumore.
- **Concavity**: Indica la gravità o la profondità delle porzioni concave all'interno del contorno del tumore.
- **Concave points**: Rappresenta il numero di punti concavi all'interno del contorno del tumore.
- **Simmetry**: Rappresenta quanto il tumore sia simmetrico rispetto al suo centro.
- **Fractal dimension**: È una misura dell'irregolarità o della complessità della forma del tumore.

4 Analisi ed Elaborazione dei dati

4.1 Variabile dipendente

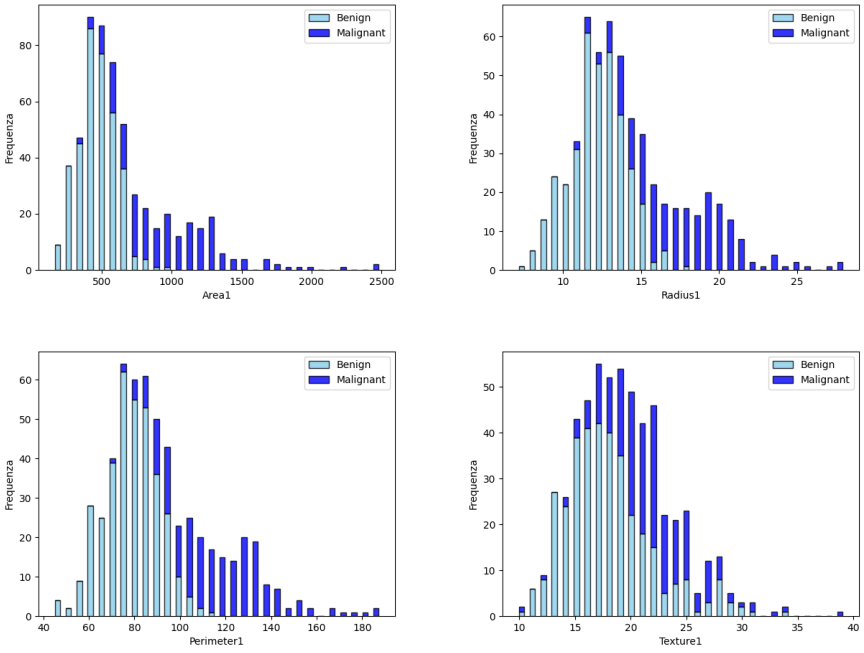
Il datatset che si sta utilizzando per risolvere il problema presenta diverse caratteristiche, tra cui troviamo una variabile dipendente. Il valore di tale variabile dovrà essere predetto dal modello, andrà a determinare se il tumore è benigno o maligno. Questa caratteristica assume due valori, assumerà "B" se il il tumore è benigno altrimenti assumerà "M" se il tumore è maligno.

4.2 Valori mancanti

Il dataset che si sta andando ad analizzare, per poi effetturare la predizione su di esso, non presenta valori nulli. Nel caso in cui il dataset avesse dei valori mancati, si potrebbero eliminare le righe contenente i valori nulli ma ciò comporta all'eliminazione di dati importanti. Ma per evitare ciò si andrebbe ad imputare o numericamente o categoricamente in base alla caratteristica che si sta andando ad imputare.

4.3 Grafici associati alla variabile dipendente

Di seguito sono riportati alcuni grafici associati alla variabile dipendente:



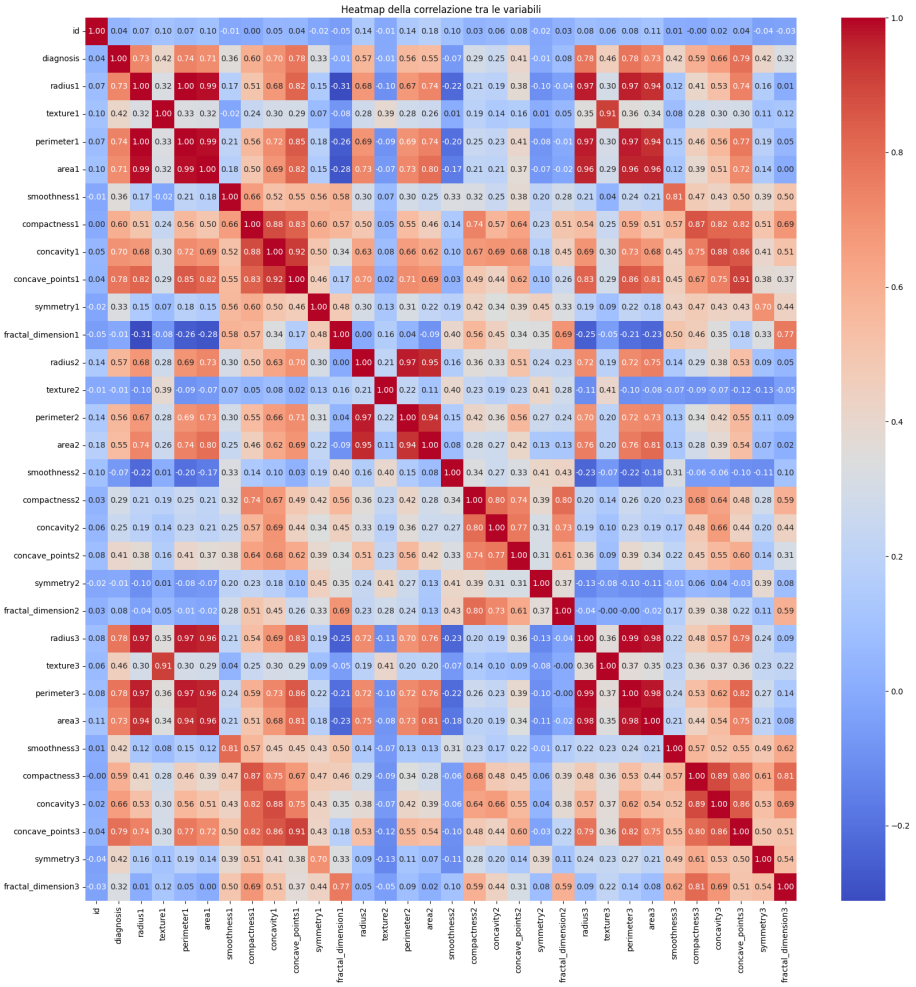
Possiamo osservare come alcune delle caratteristiche variano a seconda della variabile dipendente. Assumono valori molto vicini a loro l'una dall'altra.

4.4 Encoding delle variabili categoriche

Nel dataset sono presenti soltanto due caratteristiche che sono delle variabili categoriche, mentre tutte le restanti sono variabili continue che possono assumere infiniti valori. Poiché i modelli di Machine Learning lavorano meglio con numeri, è necessario codificare questi caratteri, che nel caso che si sta analizzando è possibile applicarlo ad un'unica colonna del dataset. In particolare per effettuare tale codifica è stato utilizzato il Label Encoding, tale tecnica serve a trasformare le variabili categoriche rin numeri interi positivi. Quindi avremmo per l'unica colonna del nostro dataset a cui è applicabile è la colonna "Diagnosis", assegnerà ad ogni tumore maligno indicato con "M" sarà 1, mentre nel caso in cui il tumore sia benigno, quindi indicato con la "B", assegnerà 0, ogni volta che incontra uno dei due tipi.

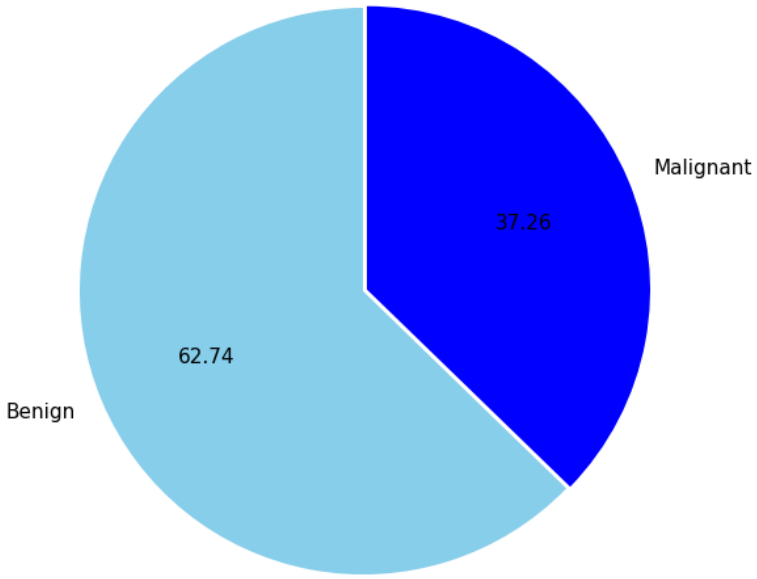
4.5 Correlazioni tra features

Di seguito sono riportate le correlazioni tra le features attraverso una heatmap:



4.6 Bilanciamento

A questo punto possiamo verificare il bilanciamento della classe da predire osservendo il grafico a torta riportato di seguito:



Come si può osservare, il dataset è sbilanciato, poiché ha una prevalenza maggiore dei tumori beningi rispetto alla minoranza dei tumori benigni. Ciò comporterà successivamente un po' di disturbo durante l'addestramento, che andremo poi a bilanciare la classe minoritaria.

5 Addestramento e Valutazione

Una volta definita la caratteristica principale che deve essere usata per la predizione del modello, aver osservato le feature e fatto si che il dataset sia pronto per l'utilizzo, l'ultimo passo da fare è quello di addestrare il modello. Poiché stiamo affrontando un problema di classificazione binaria, anche perché il dataset analizzato contiene maggiormente valori discreti, quindi si è preferito utilizzare: Naive Bayes e Logistic Regression. Le metriche utilizzate per la valutazione dei modelli sono le seguenti: Accuracy, Precision, Recall ed F1 Score. Prima di procedere, il dataset è stato suddiviso in set di training e set di testing e avremo 454 campioni per il training e 114 campioni per il testing.

5.1 Naive Bayes

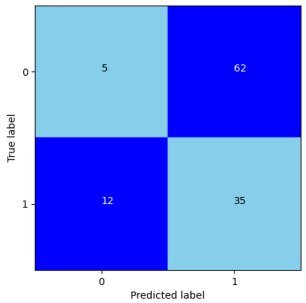
Per quanto riguarda il modello Naive Bayes, i valori ottenuti durante il training relativi ad ognuna delle metriche sono i seguenti:

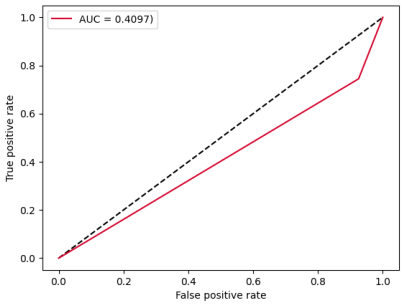
- **Accuracy**: 0.39
- **Precision**: 0.36
- **Recall**: 0.85
- **F1 Score**: 0.5

I valori ottenuti durante il testing, invece, sono i seguenti:

- **Accuracy**: 0.35
- **Precision**: 0.36
- **Recall**: 0.74
- **F1 Score**: 0.49

Di seguito sono riportate la confusion matrix e la ROC Curve:





5.2 Logistic Regression

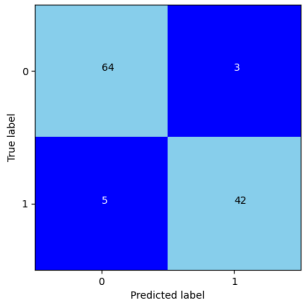
Per quanto riguarda il modello Logistic Regression, i valori ottenuti durante il training relativi ad ognuna delle metriche sono i seguenti:

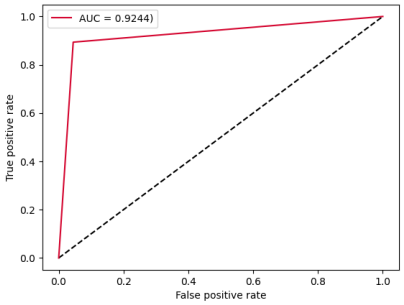
- **Accuracy**: 0.93
- **Precision**: 0.93
- **Recall**: 0.87
- **F1 Score**: 0.9

I valori ottenuti durante il testing, invece, sono i seguenti:

- **Accuracy**: 0.93
- **Precision**: 0.93
- **Recall**: 0.89
- **F1 Score**: 0.91

Di seguito sono riportate la confusion matrix e la ROC Curve:





5.3 Oversampling

Poiché il dataset analizzato è sbilanciato, presenta più dati riguardante i tumori benigni rispetto ai tumori maligni. Quindi per risolvere il problema dello sbilanciamento del dataset abbiamo applicato l'oversampling. L'oversampling comporta l'aumento del numero di campioni della classe minoritaria per bilanciare la distribuzione nel dataset. Esistono diversi modi per eseguire l'oversampling, la tecnica utilizzata per bilanciare il dataset è la tecnica SMOTE (Synthetic Minority Over-sampling Technique) che crea campioni sintetici per la classe minoritaria, piuttosto che duplicare quelli esistenti. In questo modo, il modello può apprendere da una varietà più ampia di esempi senza rischi di overfitting.

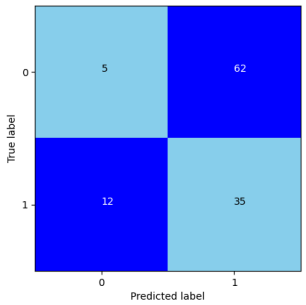
Per quanto riguarda il modello Naive Bayes, i valori ottenuti durante il training relativi ad ognuna delle metriche dopo l'oversampling sono i seguenti:

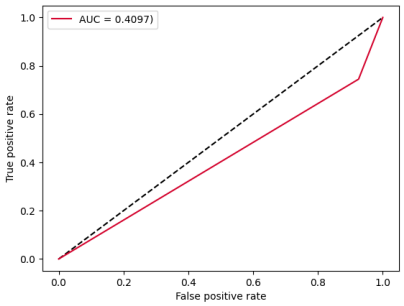
- **Accuracy**: 0.49
- **Precision**: 0.49
- **Recall**: 0.85
- **F1 Score**: 0.62

I valori ottenuti durante il testing, invece, sono i seguenti:

- **Accuracy**: 0.35
- **Precision**: 0.36
- **Recall**: 0.74
- **F1 Score**: 0.49

Di seguito sono riportate la confusion matrix e la ROC Curve:





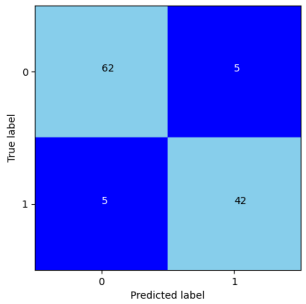
Per quanto riguarda il modello Logistic Regression, i valori ottenuti durante il training relativi ad ognuna delle metriche dopo l'oversampling sono i seguenti:

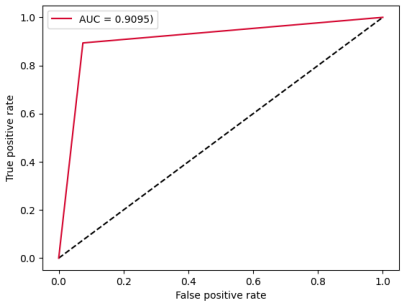
- **Accuracy**: 0.93
- **Precision**: 0.94
- **Recall**: 0.92
- **F1 Score**: 0.93

I valori ottenuti durante il testing, invece, sono i seguenti:

- **Accuracy**: 0.91
- **Precision**: 0.89
- **Recall**: 0.89
- **F1 Score**: 0.89

Di seguito sono riportate la confusion matrix e la ROC Curve:





6 Conclusioni

Osservando i risultati sopra riportati, notiamo che, sia prima di effettuare l'oversampling che dopo, il Logistic Regression ha dimostrato una performance maggiore rispetto a Naive Bayes. La sue elevate metriche durante entrambe le fasi indicano che il modello è in grado di fare previsioni accurate su entrambe le classi.
Quindi possiamo dire che il Logistic Regression è il modello più adatto per risolvere il nostro problema.