UNIVERSITÀ DEGLI STUDI DI CAGLIARI

Dipartimento di Matematica e Informatica

Corso di laurea Magistrale

RELAZIONE PROGETTO BIG DATA

Creazione di modelli predittivi per l'analisi di emozioni da file audio

DOCENTE

Prof. Diego Angelo Gaetano
REFORGIATO RECUPERO

STUDENTI

65357 Daniele LURANI
65371 Lorenzo SUSINO

Big Data
A.A. 2024-2025

INDICE

Introduzione

Il progetto presentato si pone l'obiettivo di provare a predire l'emozione dominante in un parlato tramite l'analisi, attraverso modelli di Machine Learning (in seguito ML), di features estratte da file audio in formato .wav, registrati in tre lingue: inglese, francese e tedesco. L'approccio adottato ha previsto la creazione di un dataset multilingue, l'estrazione di feature significative e l'addestramento di modelli di ML.

In particolare, sono state esplorate due tipologie di feature: quelle acustiche e quelle testuali. Le feature acustiche sono state estratte mediante le librerie Librosa e Parselmouth, che hanno permesso di catturare informazioni spettrali e parametri utili per discriminare le emozioni. Parallelamente, è stata impiegata la tecnologia di speech-to-text basata su Speech Recognition di Google per convertire il parlato in testo, da cui sono stati ottenuti elementi testuali rappresentativi del contenuto dell'audio. Questi due insiemi di features sono stati successivamente utilizzati per addestrare modelli predittivi, sfruttando l'elaborazione distribuita offerta da Spark e integrando vari strumenti di data analysis.

Il lavoro ha infine previsto una fase di analisi e validazione dei risultati, mirata a valutare la capacità dei modelli di riconoscere correttamente le emozioni espresse nei diversi file audio.

Base di Dati

In questa sezione verranno descritte le basi di dati utilizzate per il progetto, evidenziando le fonti e le modalità di raccolta e organizzazione dei file audio in formato .wav nelle tre lingue oggetto dello studio: inglese, francese e tedesco.

2.1 Dataset Inglese

Il dataset ingelse utilizzato è il CREMA-D, un dataset comprendente 7442 clip audio originali registrate da 91 attori. Queste clip provengono da 48 uomini e 43 donne di età compresa tra i 20 e i 74 anni e da una grande varietà di etnie (Afro Americani, Asiatici, Caucasici, Ispanici, e Non specificati).

Gli attori hanno parlato ispirandosi ad una selezione di 12 frasi. Le frasi sono state presentate utilizzando una delle sei diverse emozioni (Rabbia, Disgusto, Paura, Felicità, Neutro e Tristezza) e quattro diversi livelli di emozione (Basso, Medio, Alto e Non specificato).

Per i nostri esperimenti abbiamo utilizzato una selezione di 1056 files, con proporzioni eque delle varie caratteristiche del dataset, escludendo le emozioni non compatibili con gli altri dataset.

2.2 Dataset Tedesco

Il dataset EmoDB è stato creato dall'Istituto di Scienze della Comunicazione dell'Università Tecnica di Berlino. Dieci oratori professionisti (cinque maschi e cinque femmine) hanno partecipato alla registrazione degli audio. Il dataset contiene un totale di 535 enunciati, che espriomono sette emozioni: rabbia, noia, ansia, felicità, tristezza, disgusto e neutro. I dati sono stati registrati a una frequenza di campionamento di 48 kHz e poi sottoposti a down-sampling a 16 kHz.

Per i nostri esperimenti abbiamo utilizzato una selezione di 454 files, con proporzioni eque delle varie caratteristiche del dataset, escludendo le emozioni non compatibili con gli altri dataset.

2.3 Dataset Francese

Il dataset utilizzato è stato il FESD-O, progettato per lo studio generale delle emozioni in un discorso e per l'analisi delle caratteristiche delle stesse ai fini della sintesi vocale. Contiene 79 enunciati che potrebbero essere utilizzati nella vita quotidiana in classe. Per ognuna delle 7 emozioni sono state scritte tra le 10 e le 13 frasi in lingua francese, enunciate da 32 parlanti non professionisti.

Per i nostri esperimenti abbiamo utilizzato una selezione di 323 files, con proporzioni eque delle varie caratteristiche del dataset, escludendo le emozioni non compatibili con gli altri dataset.

Sviluppo

In questa sezione verrano presentate le varie fasi di sviluppo del progetto, dal pre-processing all'addestramento dei modelli di ML.

3.1 Pre-processing

Il pre-processing ha riguardato principalmente la configurazione dell'ambiente di lavoro e l'installazione delle librerie necessarie. In particolare, è stato configurato un ambiente Python su Colab, con l'installazione delle librerie fondamentali per l'elaborazione audio e il machine learning, tra cui librosa per l'estrazione delle caratteristiche audio, parselmouth per l'analisi acustica avanzata, Speech Recognition per il riconoscimento vocale automatico, pandas per la gestione dei dati, numpy per le operazioni numeriche e scikit-learn per l'implementazione dei modelli di machine learning. Inoltre, è stato installato matplotlib per la visualizzazione grafica dei risultati e PySpark per consentire l'elaborazione distribuita dei dati, sfruttando le potenzialità di Apache Spark per parallelizzare i processi di estrazione delle features.

3.2 Feature Extraction

La fase di feature extraction ha previsto la creazione di funzioni dedicate per elaborare automaticamente tutti i file dei dataset, sfruttando la parallelizzazione tramite Apache Spark, dove possibile. Per l'estrazione delle caratteristiche audio, sono state utilizzate diverse librerie: con librosa sono stati estratti i coefficienti MFCC (Mel-Frequency Cepstral Coefficients), le feature di chroma e il spectrum contrast. Utilizzando parselmouth, invece, sono state estratte informazioni acustiche come pitch, intensity, harmonicity e formanti. Inoltre, tramite Speech Recognition, i file audio sono stati convertiti in testo, successivamente trasformato in un vettore di feature per arricchire il dataset con informazioni linguistiche. Dove possibile è stata implementata la parallelizzazione offerta da Spark.

Sono state estratte, e poi inserite nei rispettivi DataFrame, una varietà di features, così suddivise:

- 32 estratte con librosa
- 7 estratte con parselmouth
- max 1000 estratte con Speech Recognition

Il numero di features estratte con Speech Recognition varia in base al testo analizzato. Queste sono state poi usate singolarmente, o unite, in base al tipo di modello che si voleva addestrare.

3.3 Addestramento dei modelli

Sono stati utilizzati due modelli di ML, il Random Forest e il Gradient Boosting. Per entrambi i modelli è stato utilizzato un dataset etichettato, opportunamente pre-elaborato per garantire la coerenza dei dati.

- **Random Forest**: Questo modello sfrutta il meccanismo di ensemble learning, combinando un elevato numero di alberi decisionali indipendenti. La strategia di bootstrap aggregating (bagging) utilizzata consente di ridurre la varianza e di mitigare il rischio di overfitting, migliorando la robustezza del modello.
- **Gradient Boosting**: A differenza della Random Forest, il Gradient Boosting costruisce in maniera iterativa una sequenza di alberi decisionali, in cui ciascun nuovo albero cerca di correggere gli errori residui dei precedenti. Questo approccio consente di ottimizzare progressivamente la performance del modello,

Una volta addestrati, i modelli sono stati sottoposti ad una fase di validazione utilizzando le seguenti metriche:

- **Accuracy**: Misura la proporzione di predizioni corrette sul totale delle osservazioni. È utile per avere una visione complessiva delle performance, anche se in presenza di classi sbilanciate può essere meno informativa.
- **Precision**: Indica la percentuale di previsioni positive che sono effettivamente corrette, evidenziando la capacità del modello di evitare falsi positivi.
- **F1 Score**: Essendo la media armonica di precision e recall, offre un compromesso tra la capacità di identificare correttamente le classi positive e la riduzione degli errori.
- **AUC-ROC**: L'area sotto la curva ROC valuta la capacità del modello di discriminare tra le classi, analizzando il trade-off tra il tasso di veri positivi e quello di falsi positivi.

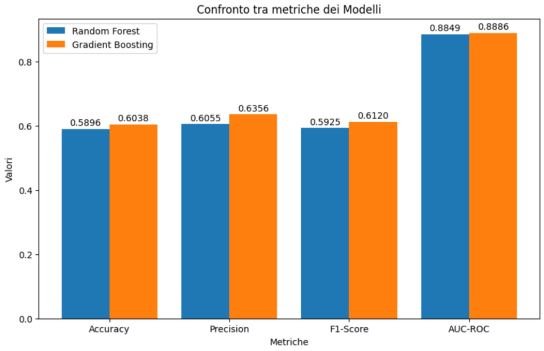
L'utilizzo di queste metriche ha permesso di confrontare le prestazioni dei due modelli e ha presentato ottimi spunti di analisi.

Risultati

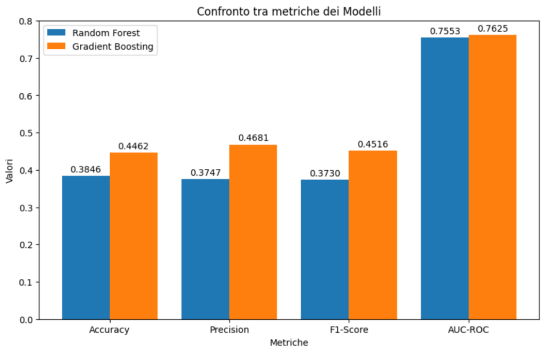
In questa sezione verranno approfonditi i risultati che abbiamo ottenuto dopo l'allenamento e l'analisi dei modelli. Per permettere di visualizzare meglio le performance del modello *RandomForest* e del modello *GradientBoosting* abbiamo deciso di utilizzare degli istogrammi che mettono in relazione le metriche analizzate.

4.1 Librosa

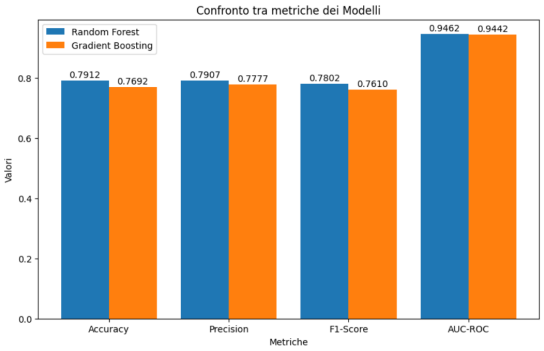
I modelli allenati con le features estratte con librosa hanno presentato risultati altalenanti a seconda del linguaggio: dopo una analisi dei dataset abbiamo ipotizzato che i motivi dovuti a queste differenze siano dati dalla natura dei dataset utilizzati. Il dataset tedesco, con il quale i modelli addestrati si sono rivelati più precisi, è composto da file audio in cui viene detta sempre la stessa frase ma con modalità diversa in modo da enfatizzare l'emozione risultante, e ciò porta il modello ad una facilità maggiore nel distinguere l'emozione. Il dataset inglese e quello francese sono formati da audio composti da una varietà di frasi più ampia, oltre ad una maggiore differenza nella distribuzione di uomini e donne. Questo porta ad una difficoltà maggiore nell'analizzare gli audio e le loro features, il che ci porta a pensare che sarebbero necessari molti più file. Questo è in parte dimostrato dal fatto che il dataset inglese, che contiene molti più audio di quello francese, ha ottenuto risultati migliori. In generale, come ci si aspetterebbe, il modello *GradientBoosting* ha performato lievemente meglio del modello *RandomForest*



(a) Dataset Inglese



(b) Dataset Francese

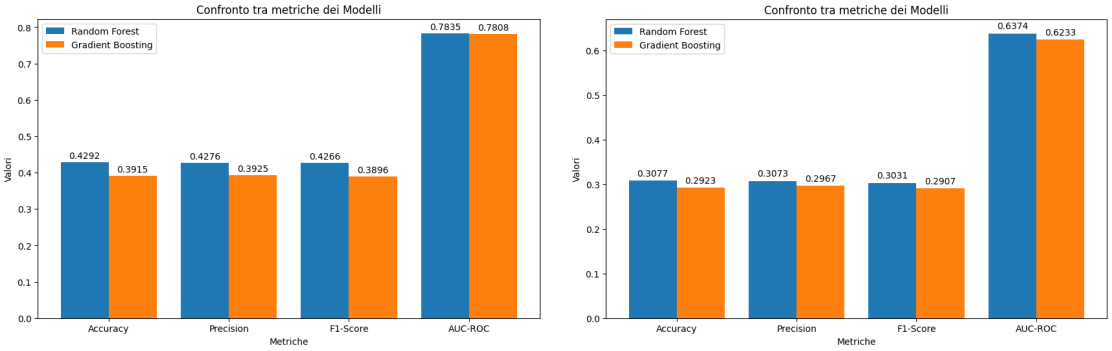


(c) Dataset Tedesco

Figure 4.1: Confronto metriche tra i vari dataset con Librosa

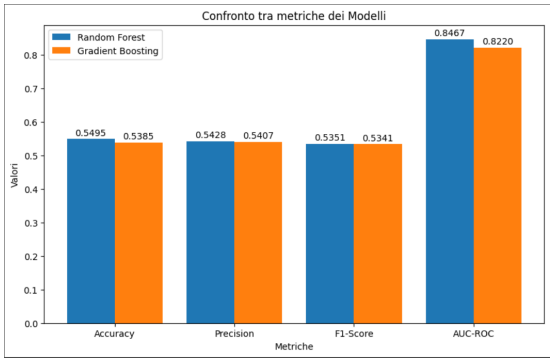
4.2 Parselmouth

I modelli addestrati con le features estratte con Parselmouth hanno presentato un calo di performance rispetto a Librosa, probabilmente dovuto al numero minore di features che sono state estratte. Inoltre i risultati ottenuti sono risultati più simili tra i vari dataset, nonostante le stesse differenze che si sono evidenziate in precedenza, e con il dataset Tedesco che riesce ancora a fornire risultati migliori.



(a) Dataset Inglese

(b) Dataset Francese

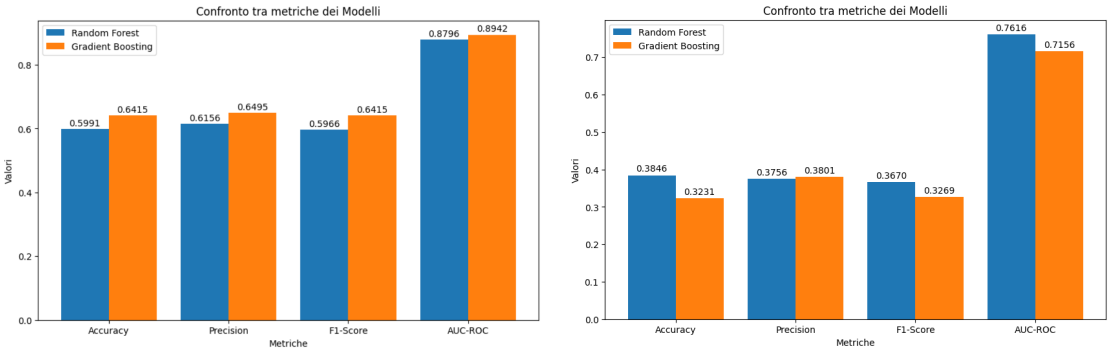


(c) Dataset Tedesco

Figure 4.2: Confronto metriche tra i vari dataset con Parselmouth

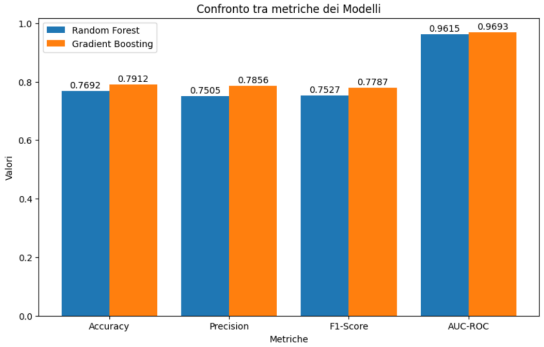
4.3 Librosa + Parselmouth

Come si poteva prevedere, la combinazione di features estratte con Librosa e Parselmouth ha portato i risultati migliori, anche se solo lievemente. Il resto risulta in linea con le osservazioni espresse in precedenza.



(a) Dataset Inglese

(b) Dataset Francese



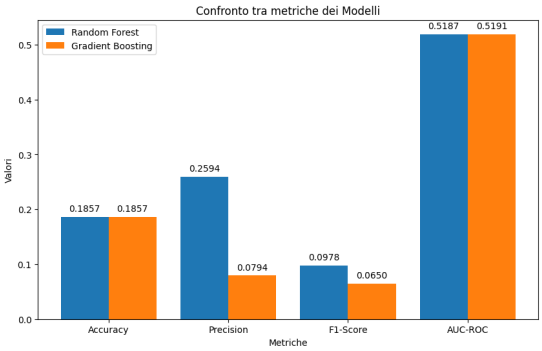
(c) Dataset Tedesco

Figure 4.3: Confronto metriche tra i vari dataset con Librosa + Parselmouth

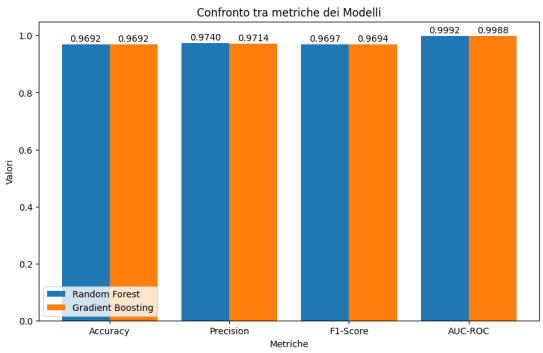
4.4 Speech Recognition

L'addestramento dei modelli utilizzando solo features estratte dai testi trascritti con Speech Recognition ha ottenuto risultati particolari: i modelli, su dataset in inglese e in tedesco hanno ottenuto performance pessime, mentre praticamente perfette su quello francese. Dopo una attenta valutazione siamo arrivati alle seguenti ipotesi:

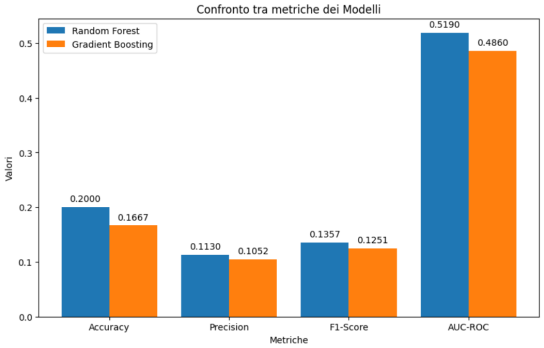
- Le lingue tedesca e inglese potrebbero aver generato trascrizioni più rumorose (con parole errate o distorte), rendendo difficile l'interpretazione da parte del modello di machine learning.
- Alcuni vettorizzatori potrebbero non gestire bene la morfologia complessa del tedesco (ad esempio parole composte), mentre il francese ha una struttura morfologica più semplice e regolare.
- I modelli potrebbero aver overfittato su lingue con distribuzioni particolari o su dati meno rumorosi (come nel caso del francese).



(a) Dataset Inglese



(b) Dataset Francese



(c) Dataset Tedesco

Figure 4.4: Confronto metriche tra i vari dataset con Speech Recognition

In generale, con i dataset utilizzati e i risultati ottenuti, non riteniamo l'estrapolazione del testo un metodo efficace per predirre le emozioni, in quanto molte frasi sono identiche e solamente lette in modo diverso per evidenziare l'emozione che si vuole esprimere.

Conclusioni

In conclusione, abbiamo visto che è possibile effettuare con una certa accuratezza una predizione sull'emozione espressa in un file audio, ma che l'accuratezza del risultato dipende da diversi fattori, principalmente riconducibili al dataset che viene utilizzato. Dai nostri esperimenti possiamo dedurre che l'utilizzo di dataset più completi e diversificati potrebbe portare a risultati soddisfacenti, ma anche che, da sola, l'estrapolazione del testo non sia una metodo efficace per questo scopo.

Bibliografia

Di seguito vengono riportate le fonti dei dataset utilizzati per il progetto:

- CREMA-D (Dataset Inglese): Cao, H., Cooper, D. G., Keutmann, M. K., Gur, R. C., Nenkova, A. and Verma, R. (2014). CREMA-D: Crowd-sourced Emotional Multimodal Actors Dataset. Disponibile su: GitHub - CREMA-D
- FESD-O (Dataset Francese): Leila Kerkeni, Catherine Cleder, Youssef Serrestou and Kosai Raoof (2020). FESD-O: French Emotional Speech Database - Oréau. Disponibile su: FESD-O
- EmoDB (Dataset Tedesco): W. Sendlmeier, Felix Burkhardt, Miriam Kienast, Benjamin Weiss and Astrid Paeschke (2005). EmoDB: A database of German emotional speech. Disponibile su: EmoDB