Predicting dementia from spontaneous speech using LLM

Arianna Cipolla

22/05/2025

arianna.cipolla@studenti.unipr.it

UNIVERSITÀ
DI PARMA

Obiettivo

Utilizzare GPT-3 per prevedere la demenza analizzando il discorso spontaneo

1

2

3

UNIVERSITÀ DI PARMA

Obiettivo

Utilizzare GPT-3 per prevedere la demenza analizzando il discorso spontaneo

1

Distinguere soggetti sani da quelli affetti da Alzheimer

2

Stimare il punteggio cognitivo (MMSE)

3

Confronto dei risultati restituiti dai due metodi

UNIVERSITÀ DI PARMA

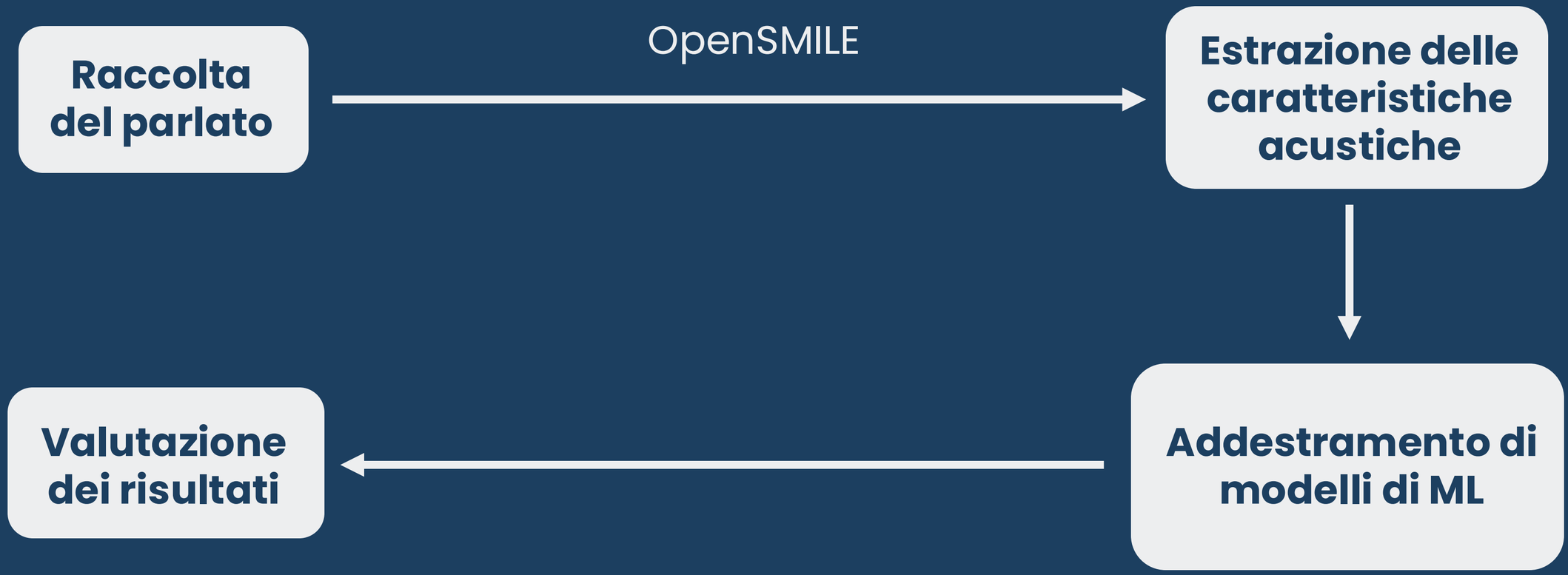
Approccio tradizionale

Test cognitivi standard
basati sul MMSE

Neuroimmagini
per osservare alterazioni cerebrali

Caratteristiche acustiche
per analizzare il parlato

Caratteristiche acustiche per analizzare il parlato



UNIVERSITÀ DI PARMA

Approccio con embedding GPT-3

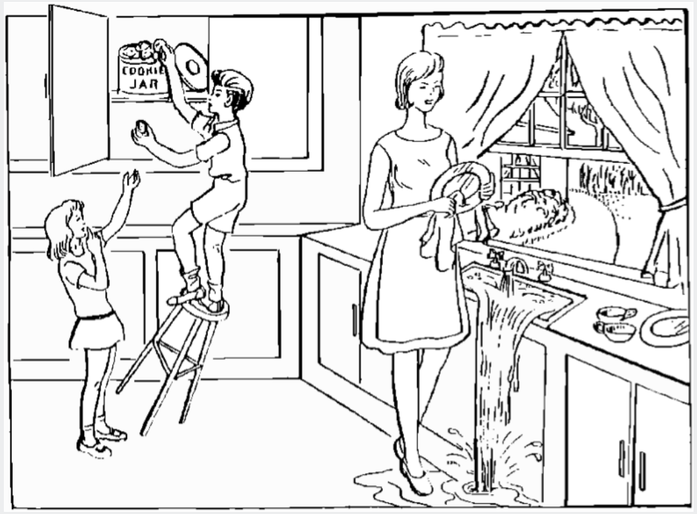
UNIVERSITÀ
DI PARMA

Strutturazione della ricerca...

UNIVERSITÀ DI PARMA

Raccolta del parlato

Il dataset utilizzato proviene dalla sfida:
"*Alzheimer's Dementia Recognition through Spontaneous Speech only*"



Da cui otteniamo **237 registrazioni vocali** che sono bilanciate secondo i dati demografici

UNIVERSITÀ DI PARMA

Estrazioni delle caratteristiche

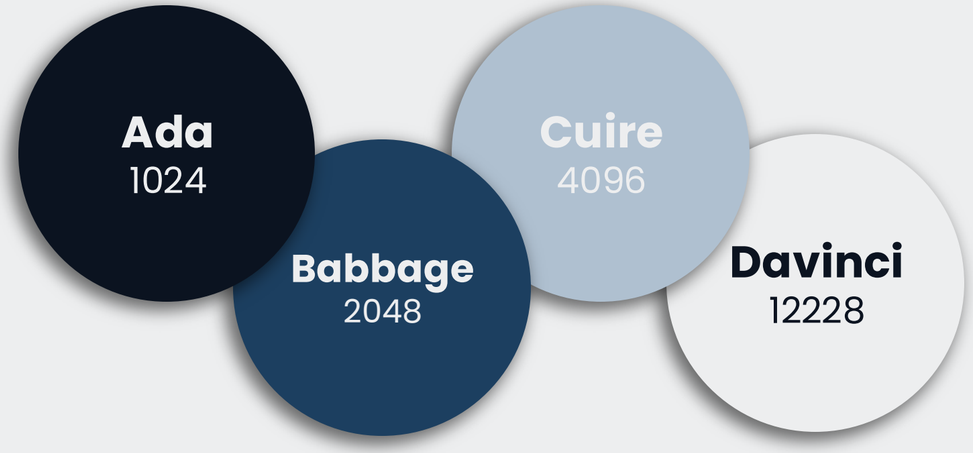
Caratteristiche acustiche:

Queste caratteristiche vengono estratte tramite **OpenSMILE**

Embedding con GPT-3:

Il **parlato** viene trasformato in **testo** tramite Wav2Vec 2.0

Dal testo ricaviamo gli **embedding** testuali utilizzando **GPT-3**, questi forniscono rappresentazioni vettoriali che catturano proprietà lessicali, sintattiche e semantiche



UNIVERSITÀ
DI PARMA

Addestramento dei modelli di ML

SVC

Support Vector Classifier

LR

Logistic Regression

RF

Random Forest

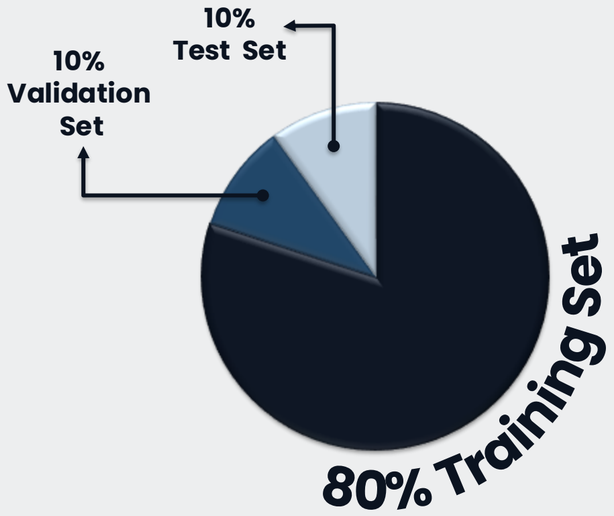
L’addestramento di questa tipologia di modelli viene utilizzata per la **classificazione dei soggetti** affetti o meno da Alzheimer

UNIVERSITÀ DI PARMA

Tipologia di addestramento

10-fold Cross Validation:

I dati vengono suddivisi in:

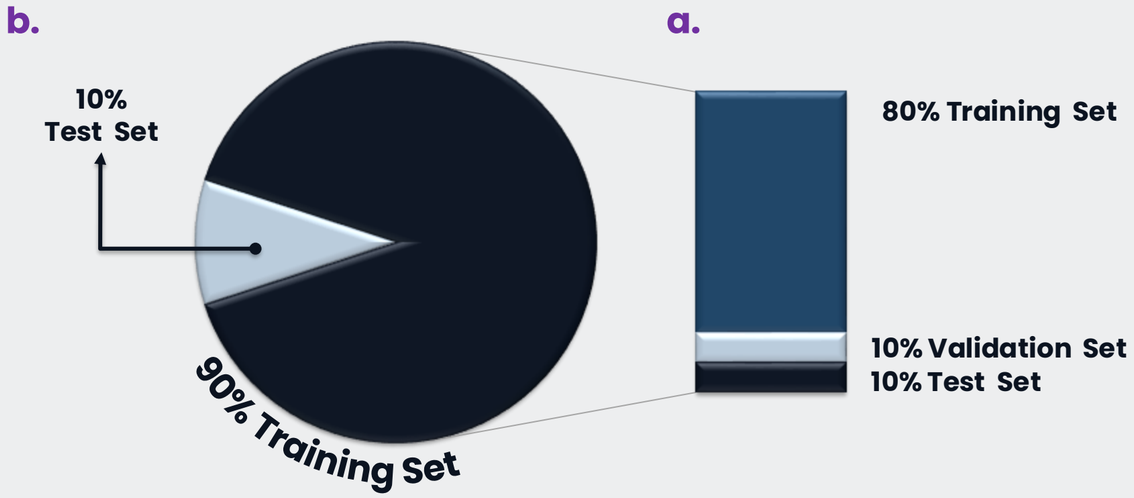


Viene applicata la modalità **10-fold cross validation**:

- Ad ogni **ciclo**: 8 fold Training set, 1 Validation set e 1 Test set
- Viene calcolata la **media** su 10 esecuzioni indipendenti

Valutazione su test set bloccato:

I dati vengono suddivisi in:



a. Si **sceglie** la combinazione di iperparametri che ha dato i risultati migliori
b. Si **riaddestra** il modello su tutto il Training test e si testa sul test set che non abbiamo mai utilizzato

UNIVERSITÀ
DI PARMA

Classificazione dei risultati ottenuti

- **Accuracy:** proporzione di previsioni corrette (sia positivi che negativi) rispetto al numero totale di esempi.
- **Precision:** proporzione dei soggetti classificati come malati che sono effettivamente pazienti affetti da Alzheimer.
- **Recall:** proporzione di pazienti con l'Alzheimer che sono stati correttamente identificati dal modello.
- **F1:** media armonica tra Precision e Recall, fornisce un equilibrio tra i due.
- **Deviazione standard:** indica quanto è stabile la performance del modello tra una suddivisione dei dati e l'altra

UNIVERSITÀ DI PARMA

Discussione dei risultati ottenuti

10-fold Cross Validation:

Caratteristiche acustiche:

Modello	Accuracy	Precision	Recall	F1
SVC	**0.697 (0.095)**	**0.722 (0.091)**	0.660 (0.120)	0.678 (0.084)
LR	0.632 (0.120)	0.645 (0.136)	0.656 (0.131)	0.647 (0.121)
RF	0.668 (0.101)	0.705 (0.156)	**0.704 (0.114)**	**0.686 (0.084)**

Valutazione:

- **Accuracy** è *migliore* con l'utilizzo di GPT-3 del **11,2%**
- **Precision** è *migliore* con l'utilizzo di GPT-3 del **12,1%**
- **Recall** è *migliore* con l'utilizzo di GPT-3 del **13,1%**
- **F1** è *migliore* con l'utilizzo di GPT-3 del **13,2%**

Il modello di ML che si è comportato meglio sembra essere **LR** con l'utilizzo di GPT-3 **Babbage**.

Embedding di GPT-3:

GPT-3	Modello	Accuracy	Precision	Recall	F1
Ada	SVC	0.788 (0.075)	0.798 (0.109)	0.819 (0.098)	0.799 (0.066)
	LR	0.796 (0.107)	0.798 (0.126)	**0.835 (0.129)**	0.808 (0.100)
	RF	0.734 (0.090)	0.738 (0.109)	0.763 (0.149)	0.743 (0.103)
Babbage	SVC	0.802 (0.054)	0.823 (0.092)	0.804 (0.103)	0.806 (0.053)
	LR	**0.809 (0.112)**	**0.843 (0.148)**	0.811 (0.091)	**0.818 (0.091)**
	RF	0.760 (0.052)	0.780 (0.102)	0.781 (0.110)	0.770 (0.047)

UNIVERSITÀ DI PARMA

Discussione dei risultati ottenuti

Valutazione su test set bloccato:

Caratteristiche acustiche

Modello	Accuracy	Precision	Recall	F1
SVC	0.634	0.657	0.622	0.639
LR	0.620	0.600	0.618	0.609
RF	**0.746**	**0.771**	**0.730**	**0.750**

Embedding di GPT-3

GPT-3	Modello	Accuracy	Precision	Recall	F1
Ada	SVC	0.788	0.708	**0.971**	0.819
	LR	0.718	0.653	0.914	0.762
	RF	0.732	0.690	0.829	0.753
Babbage	SVC	**0.803**	**0.723**	**0.971**	**0.829**
	LR	0.718	0.647	0.943	0.767
	RF	0.761	0.714	0.857	0.779

Valutazione:

- **Accuracy** è *migliore* con l'utilizzo di GPT-3 del **5,7%**
- **Precision** è *peggiore* con l'utilizzo di GPT-3 del **4,8%**
- **Recall** è *migliore* con l'utilizzo di GPT-3 del **24,1%**
- **F1** è *migliore* con l'utilizzo di GPT-3 del **7,9%**

Il modello di ML che si è comportato meglio sembra essere **SVC** con l'utilizzo di GPT-3 **Babbage**.

UNIVERSITÀ DI PARMA

Altri approcci utilizzati

Combinazione tra caratteristiche acustiche e embadding GPT-3:

Concateniamo le caratteristiche acustiche con gli embedding testuali

Tipo	Modello	Accuracy	Precision	Recall	F1
10-fold CV	SVC	**0.814 (0.115)**	**0.838 (0.133)**	0.802 (0.136)	**0.814 (0.119)**
	LR	0.800 (0.108)	0.831 (0.137)	**0.803 (0.097)**	0.809 (0.093)
	RF	0.731 (0.121)	0.741 (0.141)	0.762 (0.119)	0.745 (0.109)
Test Set	SVC	**0.802**	**0.971**	0.723	**0.829**
	LR	0.676	**0.971**	0.607	0.747
	RF	0.788	0.914	**0.727**	0.810

Embedding GPT-3:

10-fold cross validation:

Embedding	Accuracy	Precision	Recall	F1
Babbage LR	0.809 (0.112)	0.843 (0.148)	0.811 (0.091)	0.818 (0.091)

Test set:

Embedding	Accuracy	Precision	Recall	F1
Babbage SVC	0.803	0.723	0.971	0.829

Osserviamo solo un **miglioramento marginale** delle prestazioni di classificazione nella validazione incrociata a 10 fold, mentre non si osservano differenze significative nella previsione sul set di test in termini di accuratezza e F1

UNIVERSITÀ DI PARMA

Altri approcci utilizzati

Embedding con GPT-3 fine-tuning:

Viene adattato GPT-3 Babbage con il dataset posseduto

Tipo	Accuracy	Precision	Recall	F1
10-fold CV	0.797 (0.058)	0.810 (0.127)	0.809 (0.071)	0.797 (0.105)
Test Set	0.803	0.806	0.806	0.806

Embedding GPT-3:

10-fold cross validation:

Embedding	Accuracy	Precision	Recall	F1
Babbage LR	0.809 (0.112)	0.843 (0.148)	0.811 (0.091)	0.818 (0.091)

Test set:

Embedding	Accuracy	Precision	Recall	F1
Babbage SVC	0.803	0.723	0.971	0.829

Si evince che il modello ottimizzato è **inferiore** alle incorporazioni testuali basate su GPT-3

UNIVERSITÀ DI PARMA

Predizione del punteggio MMSE

Il punteggio **MMSE** (Mini Mental State Examination) è una delle misure più utilizzate per valutare la gravità dell'Alzheimer, generalmente varia da **0** (demenza grave) a **30** (normale).

Viene eseguita **un'analisi di regressione** con 3 diversi modelli:

SVR

Support Vector Regression

Ridge

Ridge Regression

RFR

Random Forest Regressor

UNIVERSITÀ DI PARMA

Risultati del punteggio MMSE

I risultati vengono riportati come errore quadratico medio (RMSE):

Caratteristiche acustiche:

Tipo	Modello	RMSE
10-fold CV	SVC	7.049 (2.355)
	LR	**6.768 (1.524)**
	RF	6.901 (1.534)
Test Set	SVC	6.285
	LR	**6.250**
	RF	6.434

Embedding GPT-3:

Tipo	Embedding	Modello	RMSE
10-fold CV	Ada	SVC	6.097 (2.057)
		LR	6.058 (1.298)
		RF	6.300 (1.129)
	Babbage	SVC	5.976 (1.173)
		LR	**5.843 (1.037)**
		RF	6.330 (1.032)
Test Set	Ada	SVC	5.6307
		LR	5.8735
		RF	6.0010
	Babbage	SVC	5.4999
		LR	**5.4645**
		RF	5.8142

UNIVERSITÀ DI PARMA

Riassumendo...

- Gli **embedding di GPT-3** (Ada e Babbage) sono efficaci per:
  - Classificare soggetti affetti o meno da Alzheimer
  - Stimare il punteggio MMSE

- Gli **embedding di GPT-3** superano le caratteristiche acustiche nei risultati di classificazione e regressione.

- **Babbage** ottiene performance migliori di **Ada** in entrambe le situazioni (Classificazione, MMSE)

- Gli embedding superano anche i modelli **fine-tuned**, in quanto hanno meno problemi di overfitting:
  - Essendo il dataset troppo piccolo

UNIVERSITÀ DI PARMA

Bibliografia

- Predicting dementia from spontaneous speech using large language models (Felix Agbavor, Hualou Liang) - December 22, 2022
- OpenAI – Vector embeddings
- Scikit-learn – Machine Learning models
- Predizione del punteggio MMSE