Documentazione progetto Ingegneria

della Conoscenza 2020-2021

**Indice**

1. Introduzione

2. Requisiti Funzionali

3. Manuale utente

4. Scelte progettuali

5. Architettura del sistema

6. Processo e sviluppo dell’organizzazione del lavoro

7. Conclusioni

**Introduzione**

Questo progetto è stato realizzato per l’esame di Ingegneria della Conoscenza dagli

studenti Massaro Claudio (matricola 705108), Trotti Francesco (matricola 703010) e Marino Angela (matricola 683710).

Abbiamo preso in considerazione un dominio legato ad uno dei disturbi più pericolosi per l’uomo: quello cardiaco.

Il nostro sistema è in grado di prevedere se si potrebbe essere affetti da una malattia cardiaca a

seconda delle risposte fornite in input, grazie alle informazioni contenute all’interno del dataset di riferimento.

**Requisiti Funzionali**

Avendo progettato il programma in Python, si richiede un ambiente in grado di

eseguire codice, nel nostro caso abbiamo deciso di utilizzare PyCharm poiché ci permetteva inoltre di installare le librerie esterne direttamente dall’IDE.

Relativamente alle librerie esterne importate, vediamo la necessità di installare sulla

macchina:

* **pandas**, usato per la manipolazione e l’analisi dei dati;
* **scikit-learn**, per applicare i concetti del Machine Learning;

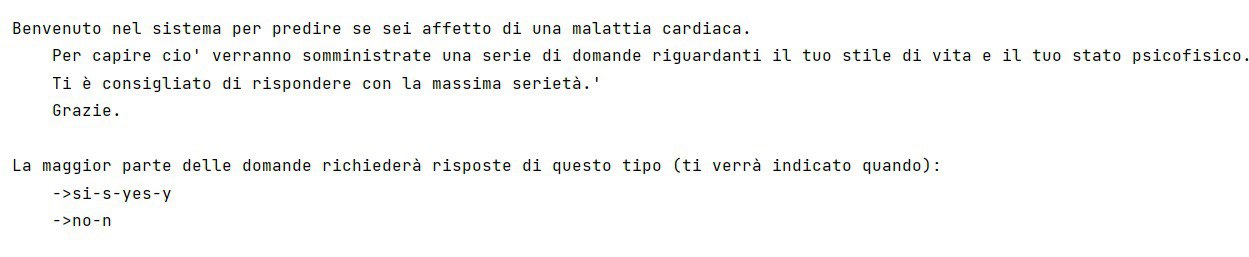
Nel caso in cui non si utilizzi Pycharm, potrebbe essere necessario l’utilizzo del terminale. Ad esempio, su Windows ritroviamo il comando:

*“python -m pip install nome\_libreria”*

**Manuale utente**

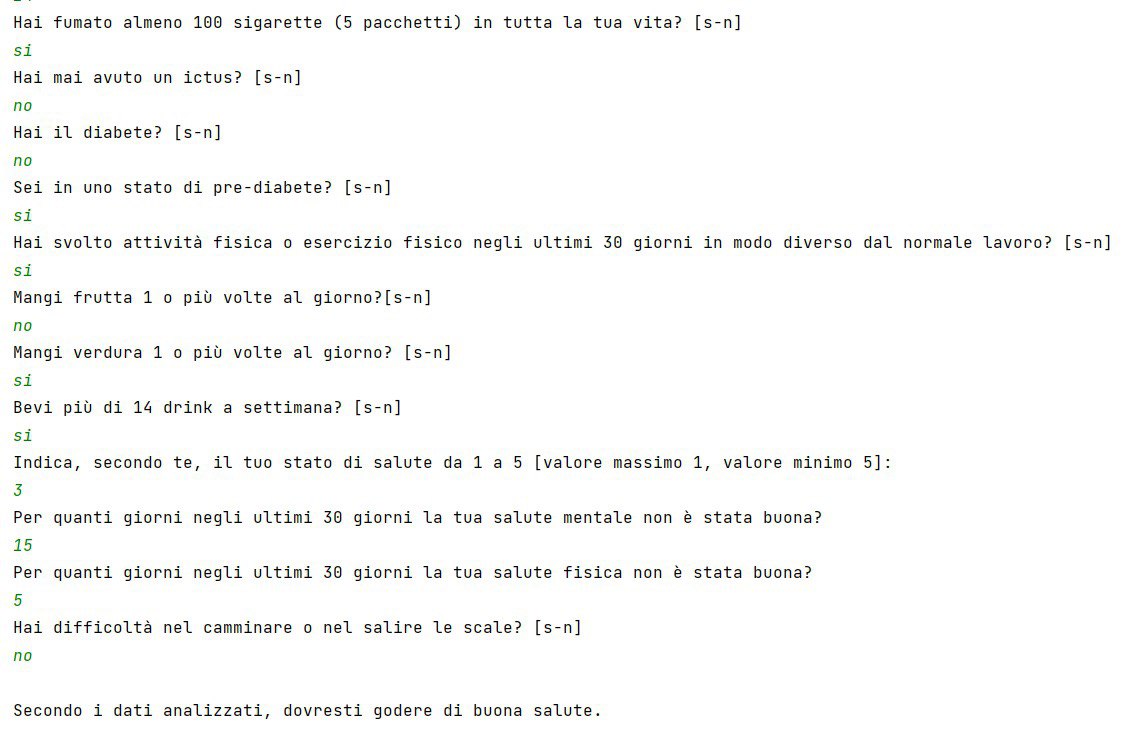
E’ possibile scaricare il progetto dal seguente link <https://github.com/clah865/icon_project.git>

Dopo averlo scaricato si dovrà eseguire il file “***system***” su un IDE.



Questo screenshot rappresenta il messaggio iniziale che vi si presenterà una volta avviato correttamente il programma.

È presente una breve descrizione in cui è esplicitato all’utente di cosa il sistema effettivamente si occupa.



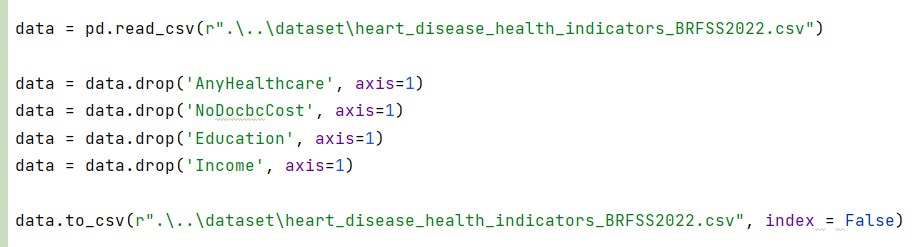
Subito dopo, inizieranno una serie di domande, che verranno richieste sequenzialmente. Al termine, il sistema restituirà la predizione.

**Scelte progettuali**

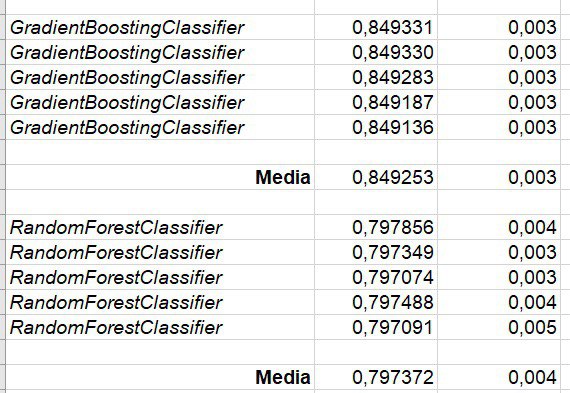
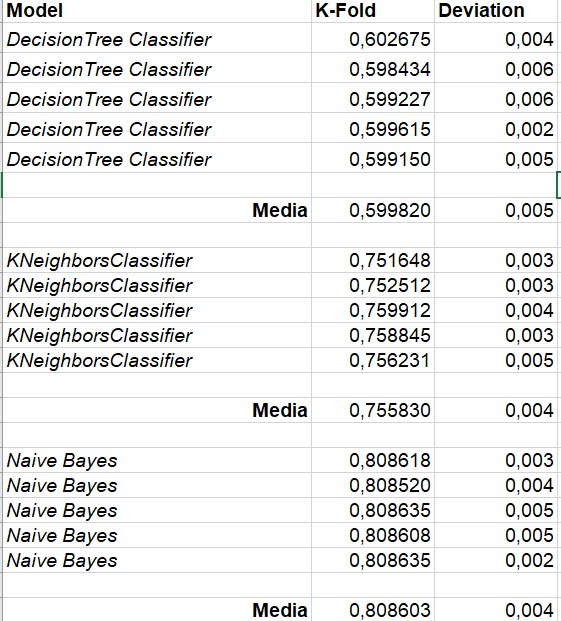
Abbiamo utilizzato il dataset dal seguente link: [Heart Disease Health Indicators Dataset](https://www.kaggle.com/datasets/alexteboul/heart-disease-health-indicators-dataset) per addestrare il nostro sistema.

I modelli e le tecniche utilizzate:

* Prima di passare direttamente all’apprendimento supervisionato abbiamo osservato il dataset scelto, e mediante una matrice di correlazione, abbiamo notato alcune features che risultavano poco coerenti (poiché riferiti ad un contesto statunitense) nei confronti delle features obiettivo e pertanto abbiamo effettuato *features selection.*



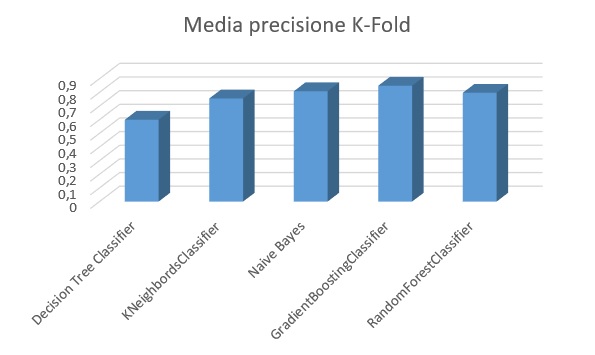
* Calcolando accuratezza del sistema prima e dopo l’eliminazione delle features abbiamo notato una leggera diminuzione della precisione (circa 0.01), ma che abbiamo pensato fosse accettabile, al fine di ricevere risposte reali e non date a caso dall’utente.
* abbiamo utilizzato il Gradient Boosting come classificatore, per la predizione della feature obiettivo.  
  Questa scelta non è stata casuale ma ci siamo basati sui seguenti dati che mostrano l’accuratezza riscontrata in 5 esecuzioni del sistema, ognuna delle quali utilizzando una   
  K-fold cross validation, mostra la media riscontrata tra le precisioni calcolate nelle varie convalide, *deviation* indica la deviazione standard:



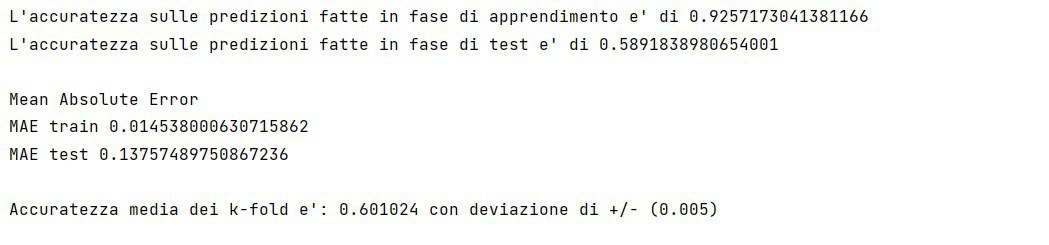
Questo classificatore è stato scelto a causa delle alte prestazioni rispetto agli altri e

non abbiamo preso affatto in considerazione i regressori a causa del contenuto del

dataset, che contiene soli valori discreti.  
 Il file di cui è stato mostrato l’estratto è disponibile nel repository fornito, denominato “test\_models.ods” del quale di seguito un breve confronto.



* Come strumento per misurare l’accuratezza del sistema, ci siamo avvalsi della  
  K-Fold cross-validation. In particolare usavamo 10 fold e come metrica per l’accuratezza la “Compute Area Under the Receiver Operating Characteristic Curve”(roc\_auc), abbiamo scelto questa a causa della sua compatibilità con features binarie.  
  Abbiamo implementato anche una funzione che permette di visionare la precisione del sistema, nell’esecuzione corrente. Anche qui come metrica per le precisione abbiamo la roc\_auc, che utilizza di default una macro-media.
* cosi come per l’accuratezza, abbiamo pensato potesse tornare utile tenere conto dell'errore che il sistema presenta, quindi sempre mediante un'apposita funzione, è possibile osservare il “Mean Absolute Error”, cioè l’errore assoluto riscontrato tra le features predette e quelle reali sia in fase di training che di test.  
  Es: del Decision Tree Regressor;



Architettura del sistema

Il sistema presenta la seguente struttura:

* **Progetto**  
  - ***dataset***, contenente i dataset utilizzati. In particolare, *heart\_disease\_health\_indicators\_BRFSS2015* è quello originale, cioè così come reperito da Internet, che abbiamo pensato fosse bene conservare, e il secondo H*eartDiseaseHealthIndicators* che rappresenta il dataset effettivamente utilizzato in fase di apprendimento;

- ***src***, contenente tutto il codice sorgente.

- *model\_performance.py*, contiene le funzioni utili a osservare le

prestazioni del sistema (accuratezza, K-Fold e l’errore assoluto). Per

quanto riguarda queste funzioni, abbiamo deciso di implementarle,

non renderle utilizzabili all’utente ma conservarle per

applicazioni/modifiche future, in cui possano tornare utili.

- *system.py*, costituisce il cuore del progetto, si occupa di richiamare

le librerie necessarie, aprire il dataset, creare un modello, addestrare il

sistema e in seguito al richiamo dei metodi necessari all’acquisizione

dei sintomi dell'utente, effettuare la predizione vera e propria.

- *request.py*, contiene la parte di interazione con l’utente in cui

vengono poste una serie di domande;

- *user.py*, una classe con metodi e attributi necessari ad asserire i

sintomi comunicati dall’utente;

Processo e sviluppo dell’organizzazione del lavoro

Il nostro progetto è stato svolto sull’IDE PyCharm, abbiamo collaborato attraverso la

piattaforma Microsoft Teams nel caso di problemi relativi a scelte implementative o codice e

usato github come piattaforma di code hosting.

Conclusioni

Abbiamo riscontrato delle difficoltà nell’interfacciarci con questo nuovo linguaggio mai utilizzato. Nonostante questo, siamo soddisfatti del lavoro svolto che ci ha permesso di applicare le conoscenze ottenute durante questo corso e di utilizzare la nostra curiosità per superare gli ostacoli che si sono presentati in fase di progettazione.