Progetto Icon 2021-2022

Sistema diagnostico dei problemi cardiaci

Angelo Cassano 648427 a.cassano64@studenti.uniba.it

Luca Depalo mat. 646958 l.depalo8@studenti.uniba.it

Il progetto è disponibile su questa repository GitHub: lucadepalo/icon

Per informazioni sul primo avvio consultare il **readme.md**

In cosa consiste il progetto?

Il progetto consiste in due parti: la prima parte consiste nell'utilizzare un dataset contenente i vari indicatori chiave associati ai problemi al cuore di un campione di persone. Il dataset specifica anche coloro che hanno effettivamente riscontrato questo tipo di problemi. La seconda parte si compone di un sistema basato su regole, implementato sotto forma di questionario per l'utente in modo da aiutarlo a capire se soffre di una patologia cardiaca o di disturbi di natura ansiosa. È accompagnata da un sistema di interrogazione di un'ontologia che fornisce ulteriori delucidazioni all'utente riguardo i sintomi.

Parte 1: Apprendimento supervisionato

Il progetto prevede l'uso di una parte di questi dati per addestrare un classificatore utilizzando diversi modelli e utilizzando una parte dei dati come test per calcolare e confrontare le performance dei modelli utilizzati. Di seguito verranno illustrate caratteristiche e scopi del dataset.

Il dataset di riferimento è disponibile qui: link

Cosa contiene questo dataset?

Questo dataset contiene i risultati del sondaggio annuale (2020) del Center for Disease Control and prevention (CDC) su un campione di quattrocentomila americani adulti relativo al loro stato di salute

Di cosa tratta questo dataset?

Secondo il CDC, le patologie cardiache sono una delle principali cause di morte per le persone di varie etnie presenti negli Stati Uniti (afroamericani, nativi americani, nativi d'Alaska e caucasici). Quasi la metà degli americani (47%) presenta almeno uno dei tre fattori chiave di rischio di patologie cardiache: ipertensione, ipercolesterolemia e tabagismo. Altri fattori chiave includono diabete, obesità, scarsa attività fisica e consumo elevato di bevande alcoliche. Determinare e prevenire quei fattori che hanno il maggior impatto nelle patologie cardiache è molto importante in ambito sanitario. Per questo motivo, negli ultimi anni parte della ricerca in ambito informatico si è proposta di sviluppare soluzioni innovative che permettano l'applicazione di metodi di machine learning al fine di determinare "pattern" dai dati che possano predire le condizioni di un paziente.

Da dove proviene il dataset e quali operazioni sono state applicate?

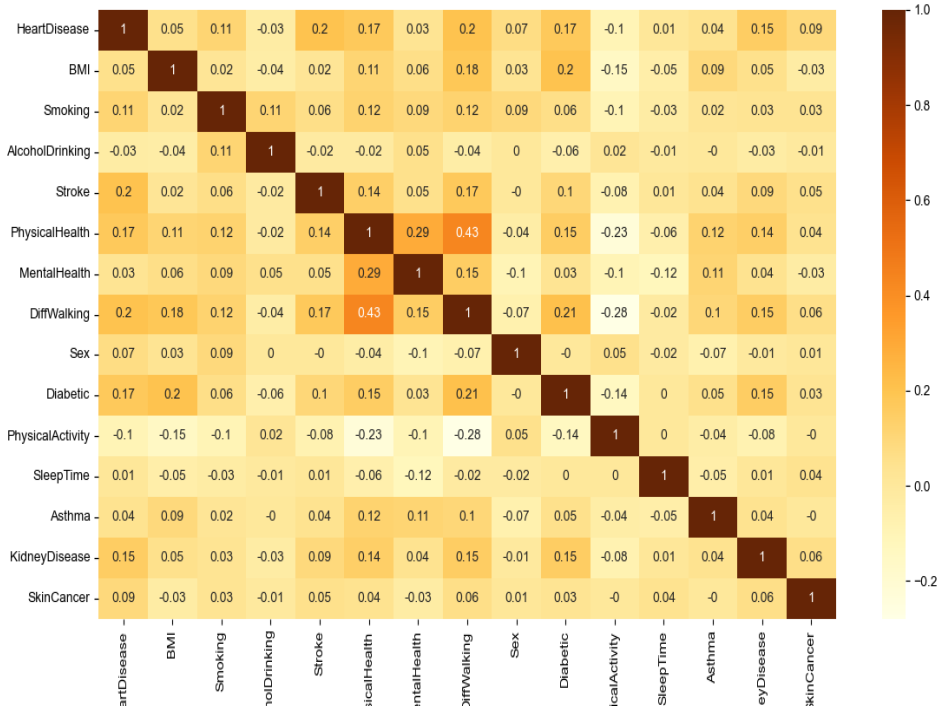
Il dataset originale proviene dal CDC ed è frutto del lavoro del Behavioral Risk Factor Surveillance System (BRFSS), che conduce sondaggi telefonici annuali per raccogliere dati sullo stato di salute dei cittadini degli U.S.A. Come affermato dal CDC: "Stabilito nel 1984 in 15 stati, il BRFSS adesso raccoglie dati in tutti i 50 stati.

Il BRFSS compila più di 400,000 interviste ogni anno, rendendolo il più grande e il più costante sistema di questionari al mondo.". Il dataset più recente include dati del 2020. Consiste di 401.958 righe e 279 colonne. La maggioranza delle colonne sono domande poste agli intervistati circa la loro situazione di salute, ad esempio "Hai seri problemi a camminare o a salire le scale?" o "Hai fumato almeno 100 sigarette nella tua vita? ". In questo dataset, si notano diversi fattori che direttamente o indirettamente influenzano le patologie cardiache, e per questo si è deciso di selezionare le variabili più rilevanti al fine di rendere il dataset più usabile per un progetto di machine learning.

Cosa si può fare con questo dataset?

Come descritto in precedenza, il dataset originale di quasi trecento variabili è stato ridotto alle venti variabili più significative. In aggiunta all'analisi esploratoria dei dati, questo dataset può essere utilizzato su diversi metodi di machine learning, in particolar modo classificatori. È necessario trattare la variabile "HeartDisease" come booleana ("Yes" - la persona che ha risposto in tal modo soffre di patologie cardiache; "No" - la persona che ha risposto così non soffre di patologie cardiache).

Di seguito mostriamo la matrice di correlazione delle feature presenti nel dataset:




Quali modelli di machine learning sono stati utilizzati su questo dataset?

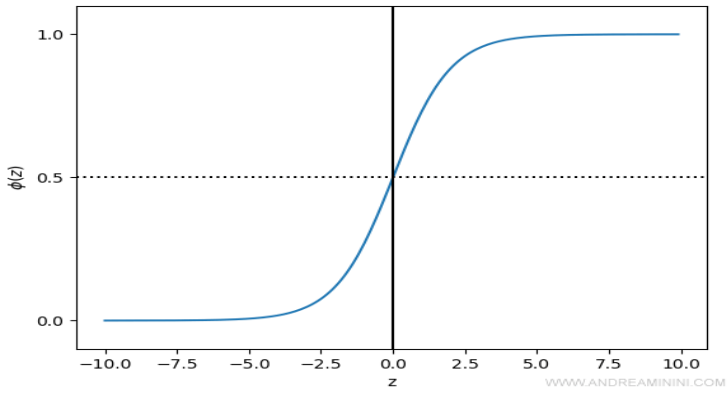
Per classificare i casi fra positivi e negativi in merito alla presenza di patologie cardiache sono state utilizzate le seguenti metodiche:

- Regressione logistica
- Classificatore bayesiano naive gaussiano
- Classificatore K nearest neighbors
- Classificatore random forest
- Classificatore AdaBoost

Di seguito una spiegazione di ciascun algoritmo:

REGRESSIONE LOGISTICA:

La regressione logistica è un algoritmo di apprendimento supervisionato che costruisce un modello probabilistico lineare di classificazione dei dati. È usata nel machine learning per l'addestramento di un algoritmo nella classificazione supervisionata dei dati. La classificazione può essere binaria (due classi) o multiclasse (più classi). Nella fase di addestramento l'algoritmo di regressione logistica prende in input n esempi da un insieme di training (X). Ogni singolo esempio è composto da m attributi e dalla classe corretta di appartenenza. Durante l'addestramento l'algoritmo elabora una distribuzione di pesi (W) che permetta di classificare correttamente gli esempi con le classi corrette. Poi calcola la combinazione lineare z del vettore dei pesi W e degli attributi xm z = → x · → w + b z=x→·w→+b z = x 0 · w 0 + . . . + x m · w m + b z=x0·w0+...+xm·wm+b La combinazione lineare z viene passata alla funzione logistica (sigmoid) che calcola la probabilità di appartenenza del campione alle classi del modello. ϕ (z) = 1 1 + e − z ϕ(z)=11+e−z La funzione sigmoid calcola la probabilità di appartenenza del campione alle classi. È la tipica curva a S della distribuzione delle probabilità.




GAUSSIAN NAIVE BAYES

Naive Bayes è un algoritmo per risolvere problemi di classificazione e apprendimento automatico (machine learning) che utilizza il teorema di Bayes. Il teorema di Bayes permette di calcolare per ogni istanza la probabilità di appartenenza a una classe. Il teorema di Bayes calcola la probabilità di un evento A condizionata a un altro evento B.

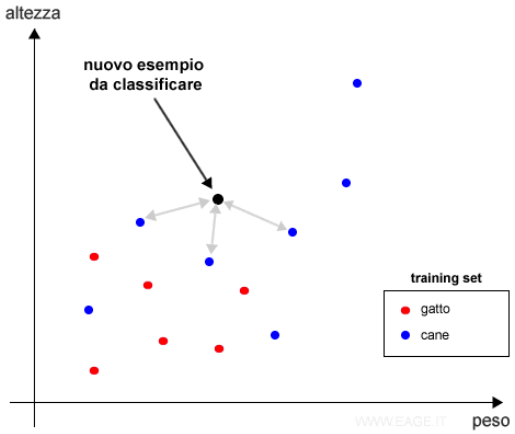
P(A|B) = P(B|A) * P(A) / P(B)

Si usa per calcolare la probabilità condizionata di un evento, in base alle informazioni disponibili su altri eventi correlati. Per praticità si assume l'indipendenza degli eventi (da cui Naive) e la distribuzione gaussiana di essi.

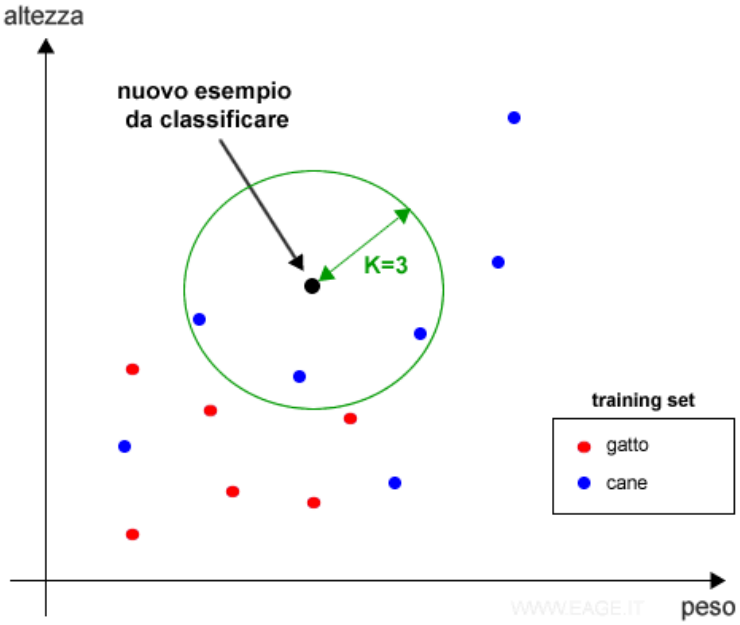
K-NEAREST NEIGHBORS

K-NN (K-Nearest Neighbors) è un algoritmo di riconoscimento dei pattern basato sulla vicinanza dei dati. Se un oggetto A ha caratteristiche simili a un oggetto B, probabilmente A appartiene alla stessa classe (categoria) di B. È usato nei problemi di classificazione. Si tratta di un algoritmo di apprendimento supervised ML perché necessita di un gruppo di esempi di training già classificati (training set).


Quando l'algoritmo analizza i nuovi dati (non classificati) li classifica in base alla distanza rispetto agli esempi del training set.




Per il calcolo della distanza si possono adottare diversi criteri: la distanza euclidea o la distanza Manhattan per i dati numerici, distanza di Hamming per le stringhe. Se le caratteristiche sono soltanto due, l'algoritmo prende in considerazione i K esempi di training più vicini all'esempio da classificare.



La classe più frequente nei K esempi più vicini determina la classificazione del nuovo esempio.

Nell'algoritmo K-NN la configurazione del parametro K è fondamentale:

- Se K è troppo piccolo, l'esempio è assegnato in base ai pochi esempi più vicini. La classificazione è influenzata dal rumore nei dati.

- Se K è troppo grande, l'esempio viene classificato nella categoria più numerosa nella popolazione. Inoltre, la complessità computazionale dell'algoritmo aumenta perché il calcolo della distanza è un'operazione molto onerosa in termini di risorse e tempo.

RANDOM FOREST

Il modello random forest è composto da un certo numero di alberi di decisione (da cui la denominazione "foresta"), al fine di aumentare le performance di classificazione del modello ad albero di decisione introducendo della casualità. La sua caratteristica principale è di addestrare un numero predefinito di alberi di decisione sul dataset in modo che ciascun albero faccia la sua predizione per ciascun esempio; a questo punto si combinano le diverse predizioni effettuate per ciascun esempio a formare la predizione finale per tale esempio. Per ogni albero viene usato un sottoinsieme casuale delle feature e ad ogni nodo viene scelta la migliore feature in un insieme ristretto; perciò, ciascun albero opera su sottoinsiemi del training set contenenti ciascuno esempi estratti casualmente con rimpiazzo (bagging) dal dataset originale. Il random forest è tanto più efficace quanto più sono diversi gli alberi generati in quanto essi forniranno diverse predizioni e al momento del confronto si potrà stabilire la bontà della predizione finale effettuando la media fra le singole.

ADABOOST

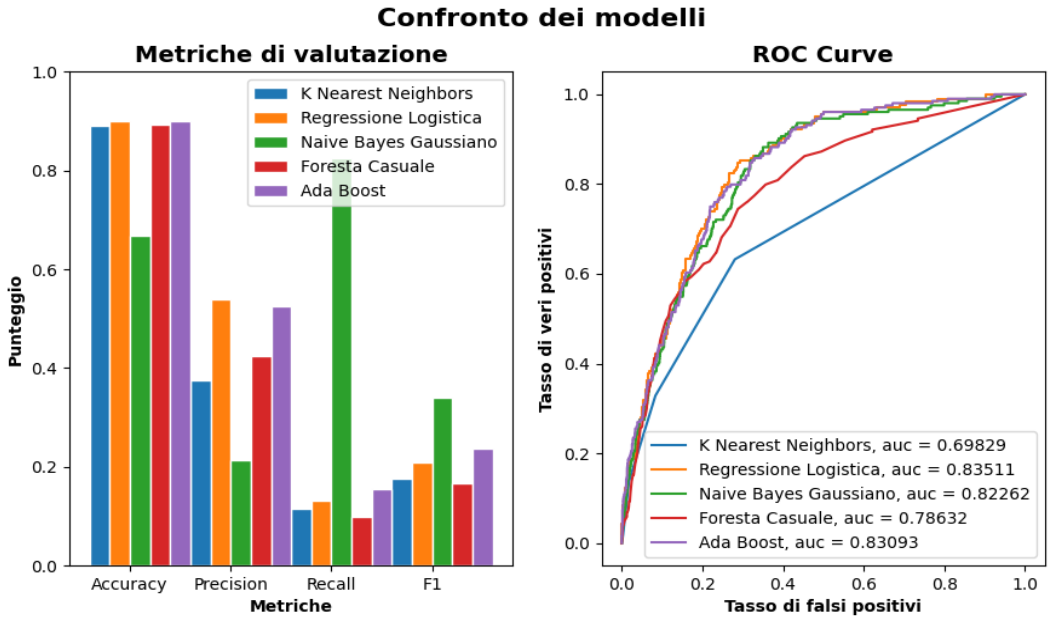
Il modello AdaBoost (adaptive boosting) consiste nell'utilizzare la tecnica di boosting, ossia impiegare un classificatore base sul dataset iterando il procedimento un prefissato numero di volte e avendo cura di ripetere il passaggio successivo sugli esempi su cui sono stati riscontrati degli errori di classificazione nel passaggio precedente, ordinando gli esempi in base all'errore; la predizione finale avverrà aggregando (con la media, la somma o con un'altra operazione ritenuta adatta allo scopo e al tipo di predizione) le predizioni dei singoli modelli prodotti nelle iterazioni susseguitesi. In particolare, nell'AdaBoost si usa la funzione d'errore $E(x)=e^{-y\hat{Y}(x)}$ e un classificatore di base ad albero di decisione.

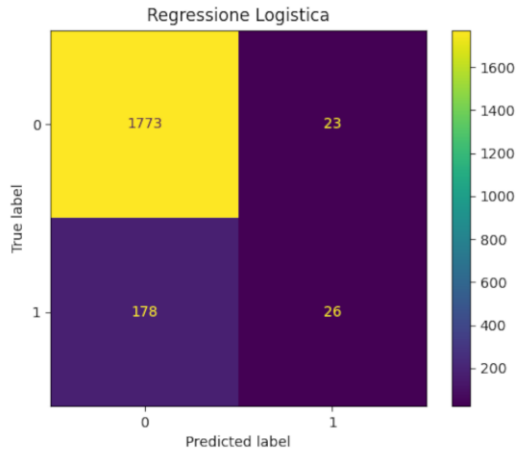
K-FOLD CROSS VALIDATION

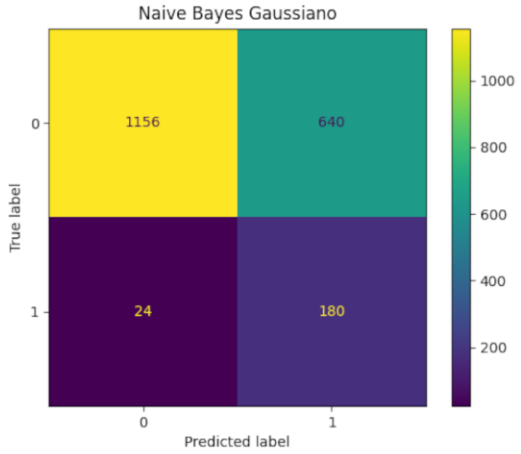
Ognuno di questi algoritmi sono stati utilizzati sfruttando la tecnica della k-fold cross validation la quale consiste nel suddividere gli esempi in k sottoinsiemi ed utilizzare k-1 sottoinsiemi per allenare i modelli di classificazione e il restante sottoinsieme da utilizzare come test. Questo procedimento verrà effettuato finché ciascuna delle k partizioni non verrà utilizzata per il testing, questa tecnica è usata per contrastare l'overfitting del modello sui dati. In questo modo abbiamo ottenuto k(=5) risultati e relative metriche di score di cui abbiamo effettuato la media prima di mettere a confronto i modelli.

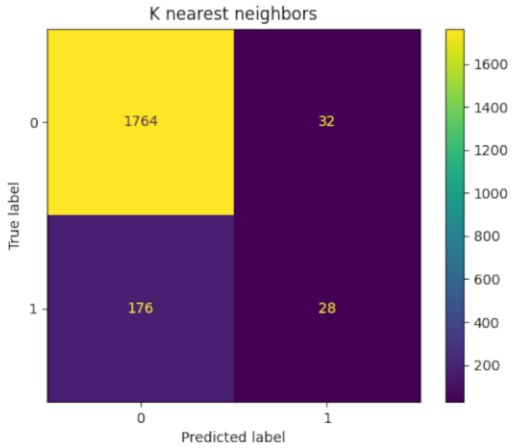
CONFRONTO DELLE TECNICHE UTILIZZATE

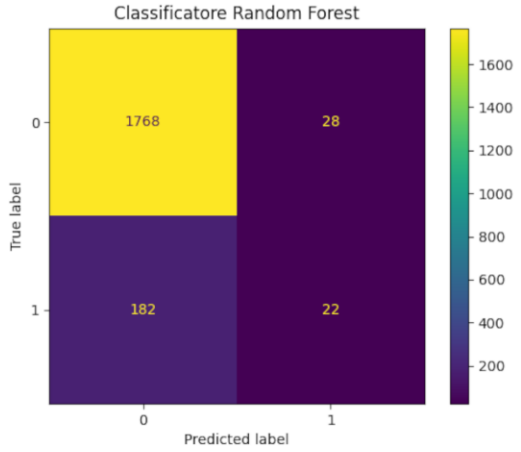
Di seguito un confronto delle performance e le rispettive matrici di confusione:

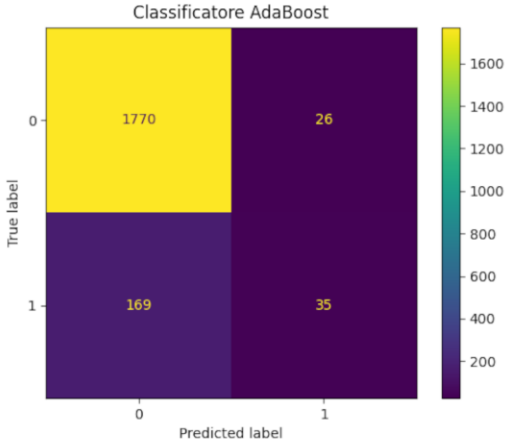

Confronto dei modelli
Metriche di valutazione
Punteggio
Metriche
Accuracy
Precision
Recall
F1
K Nearest Neighbors
Regressione Logistica
Naive Bayes Gaussiano
Foresta Casuale
Ada Boost
ROC Curve
Tasso di veri positivi
Tasso di falsi positivi
K Nearest Neighbors, auc = 0.69829
Regressione Logistica, auc = 0.83511
Naive Bayes Gaussiano, auc = 0.82262
Foresta Casuale, auc = 0.78632
Ada Boost, auc = 0.83093




Per un confronto più accurato si riportano le metriche complete usate per la valutazione delle metodiche implementate:

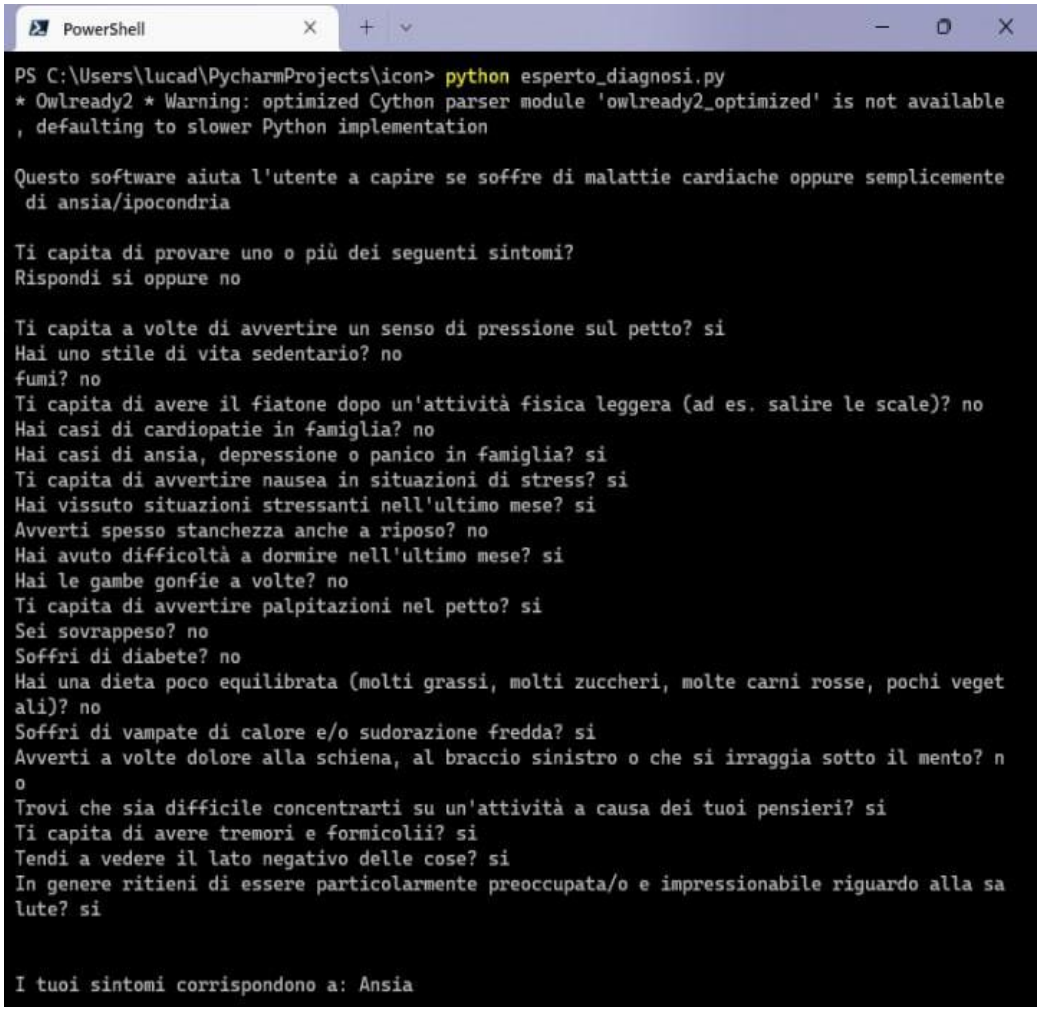
```
Regressione Logistica
Accuracy: 0.8999499999999999
Precision: 0.5394258753779362
Recall: 0.131767925561029
F1 Score: 0.2091353823662899
Area Under Curve: 0.8351101576488056
```

```
Naive Bayes gaussiano
Accuracy: 0.6678499999999999
Precision: 0.21461221219138293
Recall: 0.8262446025664417
F1 Score: 0.3392655598833234
Area Under Curve: 0.8226150705270974


classificatore K nearest neighbors
Accuracy: 0.88995
Precision: 0.37638011746943933
Recall: 0.11549352307972997
F1 Score: 0.1753972673479862
Area Under Curve: 0.6982892266037818


classificatore random forest
Accuracy: 0.8935000000000001
Precision: 0.42396257507069135
Recall: 0.10610715806118105
F1 Score: 0.16761733133955198
Area Under Curve: 0.7927434058255819


classificatore Ada Boost
Accuracy: 0.8994
Precision: 0.5243032346741208
Recall: 0.15545824971112326
F1 Score: 0.23649215752952907
Area Under Curve: 0.8309273876588497
```

Parte 2: Sistema Esperto basato su regole e Ontologia

Può capitare molto di frequente che delle persone con sintomi riconducibili a problemi cardiaci in realtà sono persone soggette ad ansia che, per la natura stessa della patologia, specialmente in presenza di ipocondria (paura di ammalarsi), tendono a scambiare i sintomi di ansia e panico con sintomi di natura cardiaca, data anche la loro somiglianza.

Il nostro sistema utilizza un questionario che viene somministrato all'utente al fine di trovare il disturbo che più corrisponde alla sintomatologia presentata. In ogni caso all'utente verranno forniti consigli su come comportarsi e quali esami effettuare per giungere ad una diagnosi più approfondita. Per realizzarlo abbiamo utilizzato la liberia python di nome Experta, realizzata come porting delle funzionalità di CLIPS in python. Abbiamo optato per questa scelta in modo da rendere le due metà del progetto coerenti e integrate fra loro sfruttando le potenzialità di python in ambito scientifico.


PowerShell
PS C:\Users\lucad\PycharmProjects\icon> python esperto_diagnosi.py
* Owlready2 * Warning: optimized Cython parser module 'owlready2_optimized' is not available
, defaulting to slower Python implementation
Questo software aiuta l'utente a capire se soffre di malattie cardiache oppure semplicemente
di ansia/ipocondria
Ti capita di provare uno o più dei seguenti sintomi?
Rispondi si oppure no
Ti capita a volte di avvertire un senso di pressione sul petto? si
Hai uno stile di vita sedentario? no
fumi? no
Ti capita di avere il fiatone dopo un'attività fisica leggera (ad es. salire le scale)? no
Hai casi di cardiopatie in famiglia? no
Hai casi di ansia, depressione o panico in famiglia? si
Ti capita di avvertire nausea in situazioni di stress? si
Hai vissuto situazioni stressanti nell'ultimo mese? si
Avverti spesso stanchezza anche a riposo? no
Hai avuto difficoltà a dormire nell'ultimo mese? si
Hai le gambe gonfie a volte? no
Ti capita di avvertire palpitazioni nel petto? si
Sei sovrappeso? no
Soffri di diabete? no
Hai una dieta poco equilibrata (molti grassi, molti zuccheri, molte carni rosse, pochi veget
ali)? no
Soffri di vampate di calore e/o sudorazione fredda? si
Avverti a volte dolore alla schiena, al braccio sinistro o che si irraggia sotto il mento? n
o
Trovi che sia difficile concentrarti su un'attività a causa dei tuoi pensieri? si
Ti capita di avere tremori e formicolii? si
Tendi a vedere il lato negativo delle cose? si
In genere ritieni di essere particolarmente preoccupata/o e impressionabile riguardo alla sa
lute? si
I tuoi sintomi corrispondono a: Ansia

Ecco una breve descrizione della patologia :

Che cos'Ã¨ l'Ansia?
Spesso descritta come una sensazione di tensione psicofisica, di preoccupazione e di inquiet
udine che talvolta sconfina nella paura, l'ansia non sempre Ã¨ sinonimo di malattia. A quest
o proposito, quindi, Ã¨ importante stabilire i confini tra ansia normale e ansia patologica.


Ansia Fisiologica o Ansia Patologica?
L'ansia normale - fisiologica o d'allarme - Ã¨ uno stato di tensione psicologica e fisica ch
e implica un'attivazione generalizzata di tutte le risorse dell'individuo, consentendo cosÃ¬
 l'attuazione di iniziative e comportamenti utili all'adattamento. Essa Ã¨ diretta contro un
o stimolo realmente esistente, spesso ben conosciuto, rappresentato da condizioni difficili
ed inusuali.

L'ansia Ã¨ invece patologica quando disturba in maniera piÃ¹ o meno notevole il funzionament
o psichico, determinando una limitazione delle capacitÃ  di adattamento dell'individuo. Ã¨ c
aratterizzata da uno stato d'incertezza rispetto al futuro, con la prevalenza di sentimenti
spiacevoli.

Ecco cosa fare quando si sospetta di avere questa patologia:

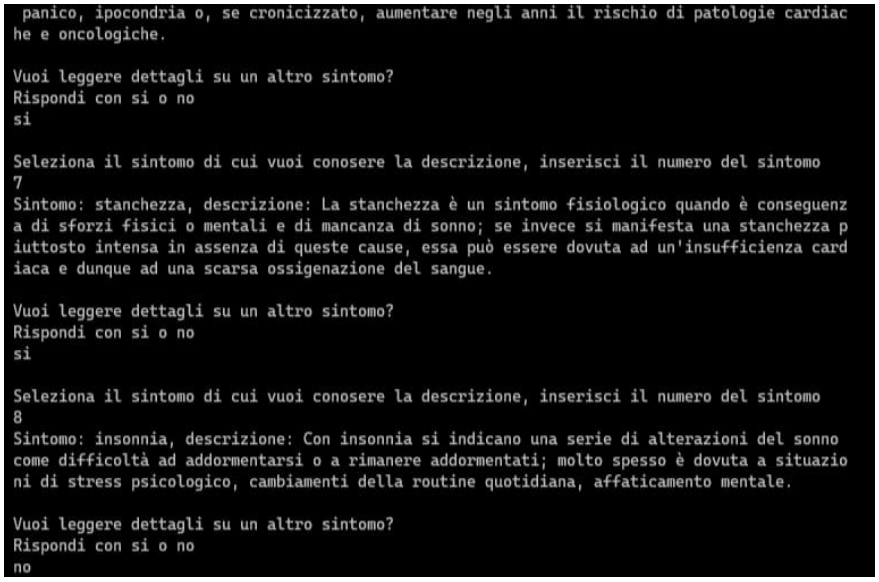
Quando gli stati ansiosi costituiscono l'anticamera di un problema piÃ¹ grave o cronico, Ã¨
necessario prevenirli o intraprendere un percorso medico specifico.

Il primo intervento Ã¨ di rivolgersi al medico di base, che, se lo riterrÃ  opportuno (trami
te un'anamnesi), somministrerÃ  una breve e leggera terapia ansiolitica o indirizzerÃ  la pe
rsona da uno psichiatra (per una diagnosi specifica e una terapia farmacologica piÃ¹ mirata)
.

Chi non accetta di buon grado la somministrazione di farmaci puÃ² rivolgersi direttamente ad
 uno psicologo-psicoterapeuta. Dopo aver identificato la causa scatenante, egli stabilirÃ  q
uale psico-terapia utilizzare.

Vuoi ripetere il test?
 Rispondi con si o no

```
no
Vuoi visualizzare ulteriori informazioni sui sintomi presenti nelle domande appena lette?
 Rispondi con si o no
si
Sintomo [1]: Nome: sedentarieta
Sintomo [2]: Nome: pressione_petto
Sintomo [3]: Nome: fumatore
Sintomo [4]: Nome: dispnea
Sintomo [5]: Nome: nausea
Sintomo [6]: Nome: stress
Sintomo [7]: Nome: stanchezza
Sintomo [8]: Nome: insonnia

Seleziona il sintomo di cui vuoi conosere la descrizione, inserisci il numero del sintomo
1
Sintomo: sedentarieta, descrizione: Uno stile di vita sedentario promuove l'insorgenza di ma
lattie cardiovascolari e di tutta una serie di disturbi dovuti all'impigrimento dei tessuti
corporei.

Vuoi leggere dettagli su un altro sintomo?
Rispondi con si o no
si

Seleziona il sintomo di cui vuoi conosere la descrizione, inserisci il numero del sintomo
2
Sintomo: pressione_petto, descrizione: La sensazione di pressione sul petto è spesso descrit
ta come un senso di peso o talvolta di indolenzimento e riporta subito alla mente un infarto
; in realtà si tratta di un sintomo ben più innocuo dovuto ad uno stato di ansia, stress o n
ervosismo.

Vuoi leggere dettagli su un altro sintomo?
Rispondi con si o no
si

Seleziona il sintomo di cui vuoi conosere la descrizione, inserisci il numero del sintomo
3
Sintomo: fumatore, descrizione: Il fumo di sigaretta causa diversi tipi di cancro, specialme
nte ai polmoni, ma è anche una delle principali cause di morte per malattie cardiache e vasc
olari poiché danneggia pesantemente il cuore e i vasi sanguigni impedendone il funzionamento
 fisiologico.

Vuoi leggere dettagli su un altro sintomo?
Rispondi con si o no
si

Seleziona il sintomo di cui vuoi conosere la descrizione, inserisci il numero del sintomo
4
Sintomo: dispnea, descrizione: La dispnea è un'alterazione del respiro manifestata come fame
 d'aria, respiro corto, affanno e può essere indicativa di una patologia polmonare o cardiac
a che impedisce il corretto scambio di ossigeno nel sangue. Se improvvisa e avvertita insiem
e a una sensazione di panico e respiri molto frequenti e profondi invece si tratta di attacc
o di panico con iperventilazione.

Vuoi leggere dettagli su un altro sintomo?
Rispondi con si o no
si

Seleziona il sintomo di cui vuoi conosere la descrizione, inserisci il numero del sintomo
5
Sintomo: nausea, descrizione: La nausea può manifestarsi in situazioni di stress intenso o p
aura e può accompagnarsi a tosse secca e stizzosa e in taluni casi anche a vomito.

Vuoi leggere dettagli su un altro sintomo?
Rispondi con si o no
si

Seleziona il sintomo di cui vuoi conosere la descrizione, inserisci il numero del sintomo
6
Sintomo: stress, descrizione: Lo stress psicofisico è una delle più importanti cause di mala
ttie, in passato spesso sottovalutata. Un evento traumatico può innescare processi di ansia,
```




Come eseguire il sistema esperto?
Per eseguire il sistema esperto, aprire un terminale nella directory principale del progetto e digitare il comando: **`python esperto_diagnosi.py .`**
A questo punto l'utente si troverà di fronte il questionario preposto alla diagnosi differenziale.

In cosa consiste?
Esso si compone di 20 domande incentrate sui sintomi sia del disturbo d'ansia sia di patologie cardiache. Ogni domanda è collegata alla presenza o meno di un sintomo e ogni sintomo è dichiarato come fatto. Il valore di verità del fatto è direttamente immesso dall'utente che avrà a disposizione come risposte "si" oppure "no". Le domande vengono poste all'utente nell'ordine stabilito in base al parametro `salience` che indica la priorità di ciascuna domanda.

Come calcola la risposta?
Il sistema confronta i valori di verità dei sintomi immessi dall'utente con quelli presenti nei file che identificano le due patologie. In caso di corrispondenza completa esso comunicherà l'esito preciso all'utente; in caso contrario il sistema sceglierà la patologia in base alla maggior presenza di sintomi positivi e dunque rilevanti per ciascuna patologia. Il sistema in ogni caso mostra all'utente una breve descrizione della patologia e dei possibili trattamenti, invitando comunque l'utente a rivolgersi ad un medico per ottenere una diagnosi certificata da un professionista.

In seguito il sistema chiede all'utente se desidera ripetere il processo di diagnosi. In caso di risposta negativa, all'utente viene chiesto se desidera ricevere ulteriori informazioni in merito alle domande che gli sono state poste. Queste informazioni sono contenute in un'ontologia che abbiamo creato sfruttando il W3C Web Ontology Language (OWL), un linguaggio per il web semantico che permette di rappresentare ontologie, cioè rappresentazioni della conoscenza di un particolare ambito rese per mezzo di una logica descrittiva del primo ordine.