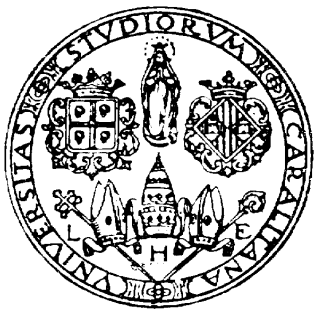

UNIVERSITA' DEGLI STUDI DI CAGLIARI

FACOLTA' DI SCIENZE ECONOMICHE, GIURIDICHE E POLITICHE

Data Science, Business Analytics e Innovazione

PROGETTO LABORATORIO DI BIG DATA

CLASSIFICAZIONE NEI CASI DI DIABETE

A cura di: Michela Concas

2022/2023

Obiettivo

Il progetto ha come obiettivo quello di prevedere, tramite una serie di aspetti legati alla salute degli individui, se questi verranno valutati come soggetti portatori di Diabete o soggetti non portatori di Diabete da parte del personale medico. Si è cercato inoltre di indagare quali siano i fattori che possano influire nella diagnosi o meno della malattia.

Descrizione dei dati

Il dataset utilizzato è stato estrapolato dal sito Kaggle.com

(https://www.kaggle.com/datasets/alexteboul/diabetes-health-indicators-dataset?select=diabetes_binary_5050split_health_indicators_BRFSS2015.csv),

Lo stesso contiene informazioni ottenute tramite un sondaggio telefonico che dal 1984 viene riproposto ogni anno dalla Behavioral Risk Factor Surveillance System (BRFSS). Il sondaggio raccoglie gli esisti delle domande a cui sono stati sottoposti 400.000 americani, domande inerenti ai comportamenti a rischio legati alla salute, domande circa le loro condizioni e sull'utilizzo di servizi di prevenzione.

Il file contenente i dati è stato caricato come Dataframe composto da 70692 osservazioni e 22 variabili qui elencate:

1. Diabetes_binary: La variabile presenta due classi, 0 indica l'assenza del diabete mentre 1 indica prediabete o la presenza di diabete;

2. HighBP: La variabile binaria indica la diagnosi di pressione alta da parte di un medico, infermiere o altro professionista sanitario. 1 indica la diagnosi di pressione alta ricevuta dal soggetto, 0 la situazione in cui non è stata diagnosticata la pressione alta;

3. HighChol: La variabile indica la diagnosi da parte di un medico, infermiere o altro operatore sanitario di un livello di colesterolo nel sangue alto. 1 indica la diagnosi di colesterolo alto, 0 l'assenza della stessa;

4. CholCheck: La variabile indica con la classe 1 se il soggetto ha eseguito un controllo del colesterolo negli ultimi 5 anni, 0 se il soggetto non ha eseguito un controllo del colesterolo negli ultimi 5 anni;

5. BMI: La variabile indica l'indice di massa corporea;

6. Smoker: La variabile risponde alla domanda: 'Hai fumato almeno 100 sigarette in tutta la tua vita?' 1 indica una risposta affermativa alla domanda, 0 una risposta negativa alla domanda;

7. Stroke: Hai mai avuto un ictus? 0 = no 1 = sì;

8. HeartDiseaseorAttack: La variabile indaga sul fatto che un soggetto abbia mai avuto una malattia coronarica (CHD) o un infarto del miocardio (MI), 0 = no mentre 1 = sì;

9. PhysActivity: La variabile fa riferimento allo svolgimento di attività fisica nei 30 giorni precediti oltre il normale lavoro eseguito dal soggetto adulto. 0 indica il mancato svolgimento dell'attività fisica, 1 lo svolgimento della stessa;

10. Fruits: 1= Il soggetto consuma frutta più di una volta al giorno, 0 altrimenti;

11. Veggies: 1=Il soggetto consuma verdure 1 o più volte al giorno, 0 altrimenti;

12. HvyAlcoholConsump: I soggetti che sono forti bevitori, ovvero gli uomini adulti che bevono più di 14 drink a settimana e le donne adulte che bevono più di 7 drink a settimana, rientrano nella classe 1, gli altri nella classe 0:

13. AnyHealthcare: Hai qualche tipo di copertura sanitaria, piani prepagati come HMO o piani governativi come Medicare o Indian Health Service? 1=Si, 0=No;

14. NoDocbcCost: C'è stato un momento negli ultimi 12 mesi in cui hai avuto bisogno di vedere un medico ma non hai potuto a causa dei costi?1=Si, 0=No;

15. GenHlth: Indica il livello di salute che un soggetto ritiene di avere in un intervallo da 1 a 5

16. MentHlth: I soggetti sono stati sottoposti alla domanda: "Pensando alla tua salute mentale, che include stress, depressione e problemi con le emozioni, per quanti giorni negli ultimi 30 giorni la tua salute mentale non è stata buona?"

17. PhysHlth: Questa variabile invece fa riferimento alle risposte in merito alla domanda: "Pensando alla tua salute fisica, che include malattie fisiche e infortuni, per quanti giorni negli ultimi 30 giorni la tua salute fisica non è stata buona?"

18. DiffWalk: 1= Il soggetto ha difficoltà a salire le scale, 0=il soggetto non ha difficoltà a salire scale

19. Sex: La variabile indica il sesso del paziente, 0 uomini, 1 donne

20. Age: Indica le età dei soggetti suddivise in 14 livelli

21. Education: Indica il livello di istruzione del soggetto

22. Income: Il tuo reddito familiare annuo proviene da tutte le fonti: (Se l'intervistato rifiuta a qualsiasi livello di reddito, codifica "Rifiutato"

Analisi dei dati

Dopo aver importato le varie librerie utilizzate per la realizzazione del progetto, aver verificato l'ampiezza del DataFrame, come passo successivo, attraverso la funzione

describe(), si è presentato un riepilogo statistico dei dati presenti nel DataFrame al fine di restituire le statistiche descrittive per ogni colonna. Si è provveduto tramite un'apposita funzione a verificare la presenza di valori nulli e successivamente è stata indagata la presenza di valori duplicati attraverso una funzione. Dall'indagine sono emerse 1168 righe duplicate e si è deciso di non eliminare le seguenti osservazioni in quanto potrebbero essere dovute a individui che semplicemente hanno fornito le stesse risposte nel test e in ogni caso la loro presenza non impatterebbe sul risultato.

Essendo presenti nel DataFrame numerose variabili codificate come double, ma che in realtà presentano valori pari a 0 oppure a 1, si è provveduto alla loro ricodifica in stringhe. Si è codificata come stringa anche quella che sarà la nostra **variabile target (Diabetes_binary)** e successivamente si è provveduto ad osservare quante osservazioni cadessero nelle diverse modalità della variabile di risposta. Il DataFrame si presenta così bilanciato:

```
+---------------+-----+
|Diabetes_binary|count|
+---------------+-----+
|      Diabetico|35346|
|  Non Diabetico|35346|
+---------------+-----+
```

Analisi Grafica

Attraverso le librerie di pandas, matplotlib e seaborn si è proceduto ad effettuare l'analisi grafica. L'intento di quest'ultima era farsi un'idea su quelle che potessero essere le relazioni tra la nostra variabile target, ovvero la variabile Diabetes_binary e gli altri feature.

Per quanto concerne le variabili categoriche sono stati utilizzati dei grafici a barre, mentre per le restanti variabili numeriche sono stati utilizzati dei boxplot.

Analizzando il rapporto tra la variabile target e *HighBP*, ovvero la diagnosi di pressione alta, è emerso che tra i soggetti diabetici o nella fase di prediabete, prevalgono quelli a cui è stata diagnosticata la pressione alta da parte del personale sanitario. Tra i soggetti non diabetici invece la pressione alta è meno frequente; infatti, prevalgono i soggetti a cui non è mai stata diagnosticata.

La variabile *HighChol* indaga la diagnosi di colesterolo alto, notiamo grazie all'analisi grafica che tra i soggetti non diabetici la maggior parte non presenta un livello di colesterolo alto, mentre, tra i soggetti con diabete e prediabete, il colesterolo è nella maggior parte dei soggetti alto.

Sempre con riferimento al colesterolo (*CholCheck*) è emerso che entrambi i soggetti non hanno eseguito negli ultimi 5 anni un controllo dello stesso. Essendo l'andamento dei dati pressoché identico per entrambe le categorie e non si ritiene che la variabile stessa possa influire sulla diagnosi.

Il grafico eseguito sulla variabile *Smoker* vede nei soggetti non diabetici prevalere quelli che nella loro vita hanno fumato sicuramente meno di 100 sigarette, mentre tra i soggetti diabetici

e prediabetici risultano essere in leggera maggioranza quelli che nella loro vita hanno fumato un ammontare di sigarette superiore a 100.

Dall'analisi grafica è emerso che la variabile *Stroke* non sembrerebbe essere correlata con la diagnosi di diabete, in quanto, in entrambe le categorie, prevalgono i soggetti che non hanno mai avuto un ictus.

Stessa affermazione può essere effettuata per la variabile *HeartDiseaseorAttack*, per cui nessuna delle categorie presenta malattie coronariche, la variabile *Fruits*, presenta dati molto simili e in entrambi i casi prevalgono i soggetti che consumano almeno una volta al giorno della frutta, stesse considerazioni possono essere fatte per la variabile *Veggies*, *HvyAlcoholConsump*, *AnyHealthcare*, *NoDocbcCost*, *Sex*, *DiffWalk*, *Education* ed *Income*.

Per quanto concerne *PhysActivity* notiamo che i risultati sono molto simili sia per diabetici che non diabetici, in entrambe le categorie prevalgono i soggetti che svolgono attività fisica ma, nel momento in cui un soggetto è diabetico, sono numerosi anche coloro che non svolgono attività fisica.

Analizzando *GenHlth*, variabile che fa riferimento allo stato di salute che ogni soggetto ritiene di avere, notiamo come tra i non diabetici siano comunque poche le persone che ritengono di avere un'ottima salute, in prevalenza gli stessi ritengono di avere uno stato addirittura insufficiente. Per chi è affetto da diabete o si trova in una situazione di prediabete, lo stato di salute si presenta sufficiente nella maggior parte dei casi e in pochissimi pessimo.

Per quanto riguarda le variabili quantitative, queste sono state esplorate tramite l'utilizzo dei boxplot. La variabile *BMI* ha mostrato come i soggetti con diabete o in fase di prediabete, presentano un 'indice di massa corporea leggermente più elevato, ma la variabile non sembrerebbe influenzare quella che è la nostra variabile di risposta. Stessa considerazione può essere assunta per la variabile *MenHlth*.

Possiamo notare come la variabile *PhysHlth* incida sulla diagnosi in quanto i soggetti diabetici presentano valori più alti rispetto ai soggetti non diabetici.

Data Preprocessing

Una volta conclusa l'analisi grafica i dati sono stati processati attraverso diverse funzioni. Innanzitutto, la variabile target, ovvero **Diabetes_binary**, è stata riconvertita in variabile intera; successivamente, si sono separati i campi dello schema in due liste distinte: una con i campi con tipo StringType chiamata Categoriche e una per i campi con tipo di dati DoubleType chiamata Numeriche.

Attraverso la classe StringIndexer si sono convertite le variabili categoriche in numeriche, assegnando a ciascuna categoria un valore intero univoco. Il codice prende in ingresso la colonna di input da codificare e restituisce una colonna che contiene i valori codificati concatenando "_index" al nome originale della colonna.

È stata poi creata una lista chiamata cols che inizialmente conteneva al suo interno unicamente le colonne numeriche presenti nel dataset, successivamente, il codice, attraverso un ciclo for, ha concatenato le colonne seguite da "_index" per le variabili categoriche. Alla fine, la lista contiene al suo interno i nomi delle colonne numeriche originali e i nomi delle colonne codificate con indice corrispondenti alle variabili categoriche.

Ho creato poi un oggetto VectorAssembler per la lista di colonne presenti nella lista cols, restituendo un output denominato 'features', combinando così le caratteristiche dei valori delle colonne in un unico vettore. L'operazione è stata assegnata alla lista "assemblers" che contiene le combinazioni di tutte le colone in una singola colonna chiamata "features".

La lista creata successivamente, di nome Scalers, contiene un oggetto StandardScaler che scala i dati nella colonna "features" e produce i dati scalati nella colonna "scaled_features"

Attraverso la classe Pipeline ho concatenato le liste "indexers", "assemblers" e "scalers". Attraverso il metodo fit ho addestrato la pipeline applicando tutte le trasformazioni definite nelle istanze di StringIndexer, VectorAssembler e StandardScaler. Ho applicato la pipeline addestrata al dataset assegnando tutto ad un nuovo dataset "scaled". Il dataset "scaled" è stato poi trasformato in un nuovo dataset che contiene solo le colonne "features" (che prima era chiamata "scaled_features") e "target" (che prima era chiamata "Diabetes_binary"). Questa operazione è stata utile per preparare il dataset all'addestramento dei modelli di machine learning in cui "features" rappresenta le caratteristiche di input e "target" rappresenta l'etichetta di output che andremo a predire.

Algoritmi di classificazione

Una volta conclusa l'attività di pre-processing ho diviso il dataset, tramite il metodo randomSplit, in training set e test set, il primo mi servirà per addestrare il mio modello e conterrà il 70% delle osservazioni, il secondo invece mi servirà per validare il modello e conterrà al suo interno il restante 30% delle osservazioni.

• REGRESSIONE LOGISTICA

È stato addestrato un modello di regressione logistica per classificazione binaria con l'obiettivo di prevedere la variabile di output sulla base di una serie di variabili di input. Una volta addestrato il modello sui dati di training, si sono valutate le prestazioni del modello stesso sul test set calcolandone l'accuratezza. Confrontando quindi le previsioni del modello con i valori reali presenti nel Dataframe delle previsioni è emerso che **l'accuracy del modello è pari al 73.07%**

• RANDOM FOREST

Il secondo modello che è stato addestrato sul training set è una Random Forest, algoritmo che combina un insieme di alberi decisionali per prendere decisioni di classificazione. Il modello è stato poi validato, come in precedenza, sui dati di test ed è emersa un **accuracy pari al 74.25%**

• NAIVE BAYES CLASSIFIER

Il terzo ed ultimo modello è stato il naive bayes, tale modello si basa sull'assunzione di indipendenza condizionale tra le caratteristiche di input dato il valore dell'etichetta di output. Una volta addestrato il modello sui dati di training è stato successivamente validato sui dati di test, producendo un **accuracy del 70.81%**

Conclusioni

In conclusione, poiché il dataset risulta essere bilanciato, si è ritenuta l'accuracy una misura adeguata al caso corrente. I risultati migliori si sono ottenuti con la Random Forest, a seguire con la Regressione Logistica e infine con il Naive Bayes Classifier.