Advanced Statistical Learning 2 Project

Andrea Mauro [mat.0222400914]

1 Diabete Dataset

Il dataset Diabete è una raccolta di dati medici e demografici dei pazienti, insieme al loro stato di diabete (presente o assente). La tipologia di diabete preso in cosiderazione è il diabete mellito, una malattia cronica caratterizzata da un eccesso di zuccheri (glucosio) nel sangue, nota come iperglicemia. L'iperglicemia può essere causata da un'insufficiente produzione di insulina (ossia l'ormone che regola il livello di glucosio nel sangue) o da una sua inadeguata azione.

Il dataset preso in considerazione è formato da 100.000 osservazioni che registrano valori per 9 variabili, nello specifico :

- *gender*: il sesso del paziente;
- *age*: l'età del paziente;
- *hypertension*: se il paziente è affetto da ipertensione (1 = si, 0 = no);
- *heart_disease*: se il paziente ha una malattia cardiaca (1 = sì, 0 = no);
- *smoking_history*: storia del fumo del paziente (mai, in passato o attualmente, non disponibile);
- *bmi*: indice di massa corporeo;
- *blood_glucose_level*: il livello di glicemia del paziente(a digiuno);
- *diabetes*: se il paziente soffre di diabete (1 = si, 0 = no);
- *HbA1c_level*: livello di HbA1c del paziente;*

* L'emoglobina glicata o glicosilata A1c (HbA1c) è un test che serve per misurare un tipo particolare di emoglobina. I valori sono inerenti alla concentrazione media di glucosio presente nel sangue negli ultimi tre o quattro mesi.

1.1 Data Handling

Prima di soffermarci nell'analisi esplorativa e il successivo sviluppo del problema di classificazione si rendono necessarie alcune operazione di manipolazione del dataset su cui stiamo lavornado, nel dettaglio:

- ridefinizione dei tipi delle feature;
- rimozione dei duplicati;
- rimozione degli NA;
- rimozione degli errori di misurazione della feature BMI;

Dopo aver eseguito le diverse operazioni elencate in precedenza la dimensione del campione si riduce a 90.740 osservazioni.

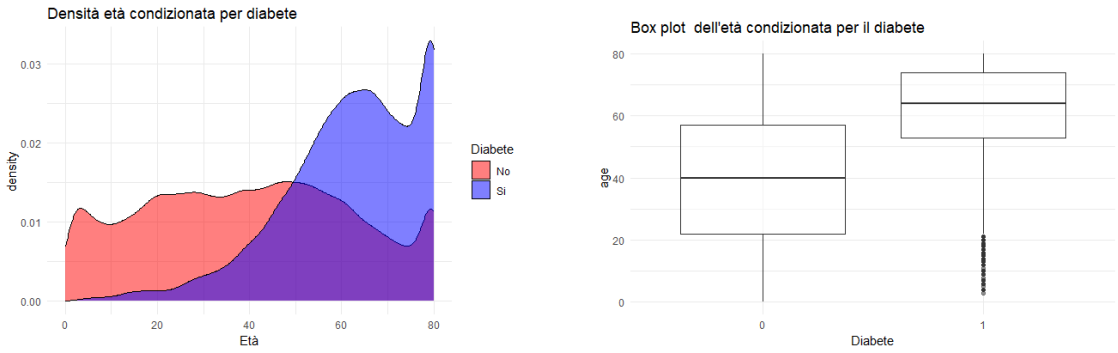
Prime 5 osservazioni del dataset

gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
Female	80	0	1	never	25.19	6.6	140	0
Female	54	0	0	No Info	27.32	6.6	80	0
Male	28	0	0	never	27.32	5.7	158	0
Female	36	0	0	current	23.45	5.0	155	0
Male	76	1	1	current	20.14	4.8	155	0

1.2 Analisi Descrittiva

In queste breve sezione saranno presentate analisi grafiche per approfondire la struttura sottostante le variabili di maggior interesse e come interagiscono tra loro.
Nello specifico l'analisi si soffermerà su quello che è il rapporto della variabile diabete, dove viene registrata la presenza o meno della malattia, con le principali features presenti nel dataset.
Di seguito viene mostrato come varia a differenti livelli di età la presenza della malattia, nello specifico:

- in 1a viene rappresentata la densità della variabile età condizionata per la variabile diabete
- in 1b viene rappresentato il box-plot dell'età condizionato per la feature diabete



(a) (b)

Figure 1

Osservando la funzione di densità(1a) dell'età si nota subito la netta concentrazione di soggetti che soffrono di diabete in età avanzata, evidenza confermata anche da box-plot, infatti guardando la 1b il 50% dei soggetti che soffrono di diabete si trovano in un range di età che va da circa 52 anni a 73. Osserviamo anche dei casi in cui la malattia si presenta in giovane età.

Continuando con l'analisi descrittiva, di seguito vengono rappresentati i box-plot della distribuzione relativa alla feature dell'emoglobina glicata (HbA1c) per i soggetti con e senza diabete.

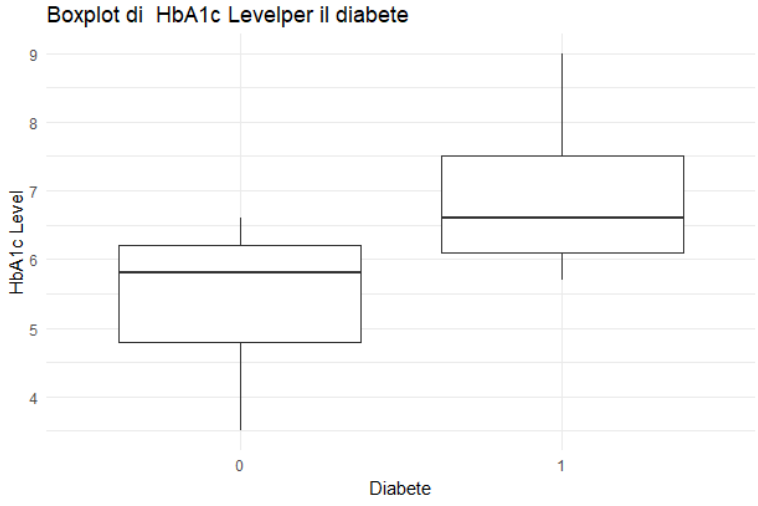


Figure 2

Osservarando i risultati, notiamo come i soggetti affetti dalla patologia presentino valori più elevati nei risultati del test HbA1c, infatti il 50% dei soggetti presenta un esito del test che va da 6.5 a 7.5. I risultati che si osservano non rappresentano una sorpesa per la composizione del test e per ciò che

comporta la patologia, difatti il test recentemente è stato rivalutato anche nella diagnosi della malattia.

Continuando ad approfondire le relazioni tra le feature precedenti, viene considerata nel confronto anche la variabile che rappresenta la presenza di ipertensione nel soggetto.

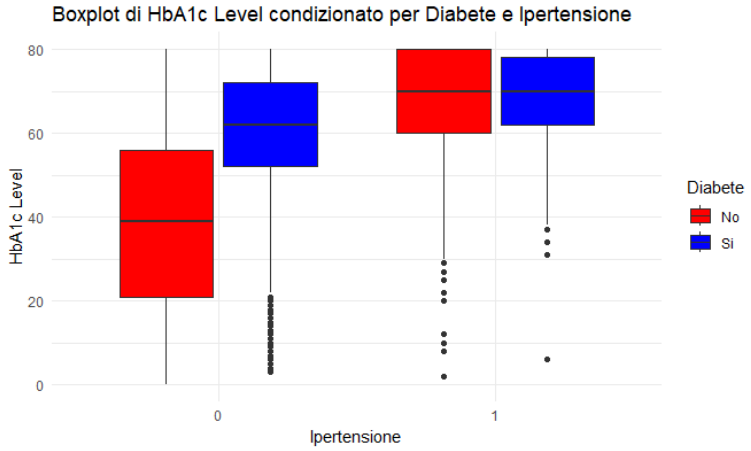
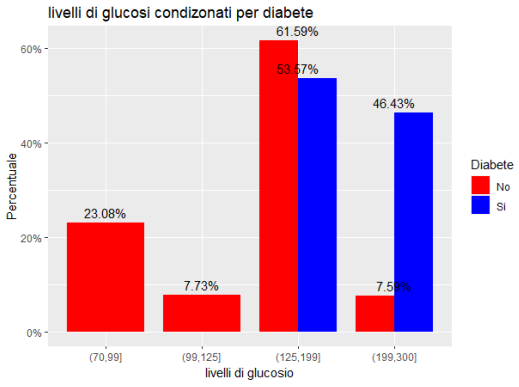


Figure 3

Dal box-plot della distribuzione della variabile HbA1c condzionato per il diabete e l'ipertensione si evincono note molte interessanti, quali:

- la metà dei soggetti che soffrono di ipertensione presentano esiti nel test HbA1c maggiori rispetto a chi non ne soffre e non ha il diabete.
- medianamente non avere il diabete ma soffrire di ipertensiore comporta livelli simili nei test HbA1c rispetto ai soggetti che soffrono di diabete e ipertensione.
- medianamente i livelli HbA1c delle persone affette dal dibaete risultano inferiori dei soggetti che soffrono solo di ipertensione e congiuntamente diabete e ipertensione.

Continuando osserviamo in termini percentuali quali sono i valori registrati di glucosio nel sangue:



(a)

Range.glicemico	stato
(1,70]	Ipoglicemia
(70,99]	Normali
(99,125]	Alterata Glicemia
(125,199]	Intolleranza Glicemica
(199,300]	Diabete

(b)

Figure 4: Livello di glucosio(a digiuno)

Come ci si poteva aspettare i livelli di glucosio nei soggetti diabetici risultano nello stato di intolleranza glicemica(sospetto diabete) e diabete, continuando, nel campione sono presenti diversi soggetti che non soffrono di diabete ma con una potenziale intolleranza glicemica.

Successivamente concentriamo la nostra attenzione sul box-plot della variabile rappresentante l'indice di massa corporeo differenziandolo per il diabete.

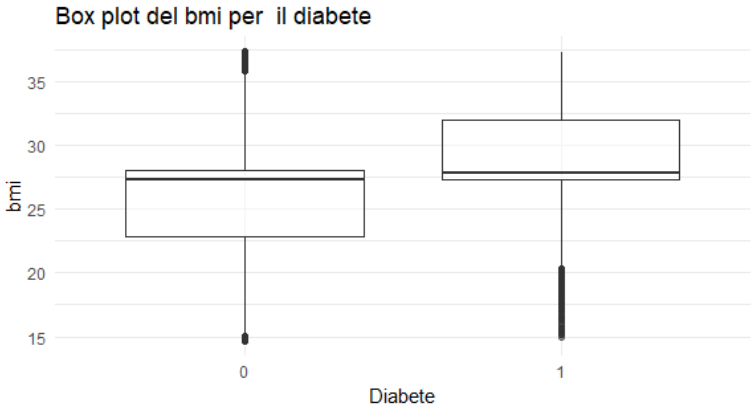


Figure 5

Dalla figura si osserva come i soggetti affetti da diabete presentino degli indici di massa corporei superiori ai soggetti non affetti dalla patologia. Le due distribuzioni appaiono fortemente asimmetriche, nello specifico abbiamo un'asimmetria positiva per i soggetti affetti da diabete ed una negativa per quelli che non ne soffrono.

Per quanto riguarda le altre variabili presenti nel dataset, dopo un'approfondita analisi di esplorazione, non risultano particolari evidenze per cui si è preferito procedere con la stima della rete neurale.

1.3 Classificazione

Le reti neurali artificiali (ANN) sono modelli composti da neuroni artificiali interconnessi tra di loro, organizzati in strati, in cui l'output di un neurone diventa l'input di un altro neurone, generando una risposta non lineare.

Uno dei vantaggi più significativi delle reti neurali rispetto ad altre classi di modelli non lineari è la loro capacità di essere approssimatori universali, in grado di approssimare una vasta gamma di funzioni con un alto grado di precisione.

Le reti neurali sfruttano il potere dell'elaborazione parallela delle informazioni dai dati. Nel processo di costruzione del modello, non è richiesta alcuna assunzione preventiva sulla forma del modello stesso, ma viene determinata principalmente dai dati disponibili. La "Single layer feed forward neural network" è uno dei tipi di reti neurali più utilizzati per la modellazione e la previsione. Questo modello è caratterizzato da tre strati di semplici unità di elaborazione collegate aciclicamente, in cui le informazioni si muovono sempre in una sola direzione, dallo strato di input allo strato di output, passando attraverso lo strato nascosto, senza tornare indietro.

L'utilizzo di una rete neurale feed-forward in questo contesto consente di classificare i pazienti sulla base della loro condizione fisica, in particolare se soffrono di diabete o meno. Dunque in questo problema di classifcazione la variabile dipendente Y assume:

$$\begin{cases} Y = 1, & \text{diabete} \\ Y = 0, & \text{no diabete} \end{cases}$$

Quindi il modello è:

$$Pr(Y = 1) = f(X_1, X_2, \ldots, X_d)$$

Dunque approssimiamo la funzione f con la rete neurale:

$$f(x_1, x_2, ...x_d) = \phi(\sum_{k=1}^{r} c_k \psi(\sum_{j=1}^{d} w_{kj} x_j + w_{k0}) + c_0)$$

La funzione di attivazione considerata per l'hidden layer è quella sigmoide. Chiaramente, data la natura del task affrontato, la scelta della funzione di attivazione del neurone di output è ricaduta sulla funzione sigmoide.

L'architettura della rete è stata selezionata utilizzando una procedura incrociata di 5-cross-validation e congiuntamente tuning del weight decay e la size del layer intermedio.

1.4 Modellazione

Partendo dal campione composto da 90.740 osservazioni ed analizzando la proporzione presente tra i soggetti con la patologia e senza, il campione è risultato fortemente sbilianciato, in quanto risultano:

- il 90% delle osservazioni rappresentano soggetti non diabetici
- il 10% delle osservazioni rappresentano soggetti diabetici

Presa in considerazione la composizone del set di dati, si è scelto di proseguire senza apportare al dataset le modifiche necessarie per bilanciare le classi. La scelta è stata dettata principalmente da due motivi in particolare: non perdere contenuto informativo; osservare come avrebbe reagito la rete neurale single layer feed-forward al problema in questione.

Il set di dati è stato diviso in due parti: il 75% della dimensione del set per la procedura di addestramento e cross-validazione, mentre il restante 25% è stato riservato per la valutazione delle prestazioni della rete neurale più effciente durante le fasi di cross-validazione. Per quanto rigurada il tuning dei parametri, le dimensioni considerate per il layer intermedio vanno da un minimo di un neurone ad un massimo di dieci. Considerando il Weight Decay, il tuning è stato sviluppato su 10 valori in range compreso da 0 a 1.

La cross-validazione scelta è la 5-folds ed è stata effettuata per ciscuna rete neurale combinazione dei due iperparametri specificati in precedenza, di seguito vengono mostrati i risulati di ogni modello rispetto alle metriche: accuratezza, F1-score, kappa, sensibilità, specificità.

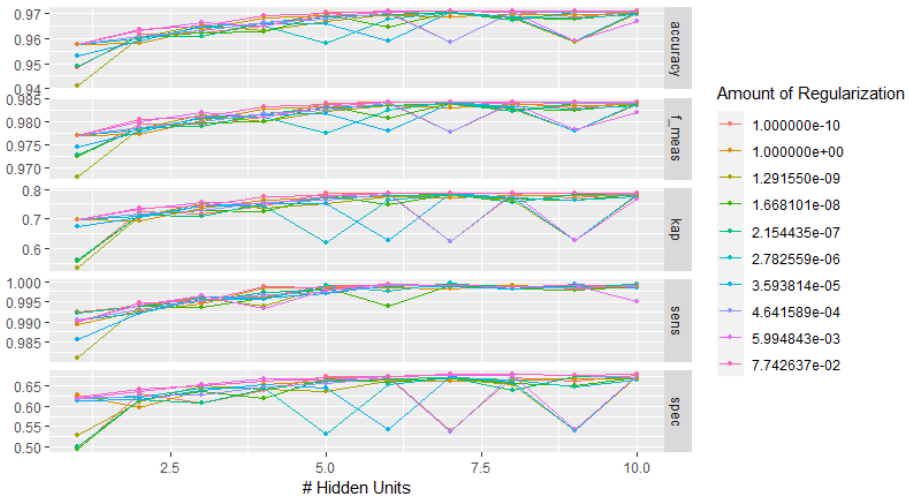


Figure 6

Dal grafico risulta chiaro che per alcuni tipi di penalizzazione quando la dimensione del layer intermedio è maggiore di tre le performance delle metriche considerate risultano instabili.

Successivamente vengono quindi rappresentate le migliori reti neurali in termini di accuratezza sul training.

hidden_units	penalty	.metric	.estimator	mean	n	std_err	.config
7	0.0774264	accuracy	binary	0.9708219	5	0.0003511	Preprocessor1_Model087
6	0.0774264	accuracy	binary	0.9707387	5	0.0001669	Preprocessor1_Model086
10	0.0774264	accuracy	binary	0.9707387	5	0.0002422	Preprocessor1_Model090
8	0.0774264	accuracy	binary	0.9706555	5	0.0002716	Preprocessor1_Model088
9	0.0774264	accuracy	binary	0.9706555	5	0.0003166	Preprocessor1_Model089

Figure 7

Osserviamo come le migliori reti in termini di accuracy siano stabili verso un unico livello di accuratezza, inoltre notiamo il termine di penalità uguale per tutte le reti presenti in figura.
Sebbene le variazioni risultino minime sia in termini di accuracy che di standard error, la rete neurale scelta è quella formata da 7 neuroni nel layer intermedio.

Continuando a considerare la rete neurale che ha presentato performance migliori in termini di accuratezza sul training, di seguito viene rappresentata attraverso un grafo:

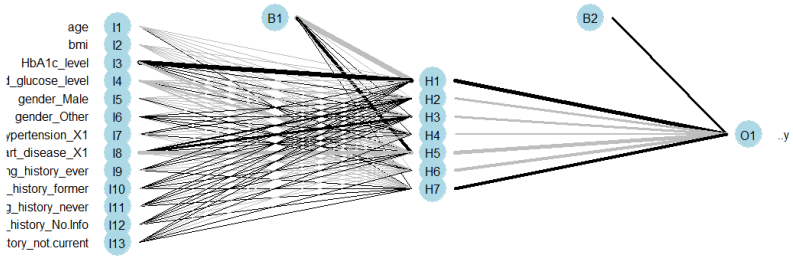


Figure 8

Il grafo rappresenta la rete neurale feedforward con:

- uno strato di ingresso, con 13 neuroni di ingresso;
- uno strato intermedio, con 7 neuroni nascosti;
- uno strato di uscita, con un singolo neurone.

Partendo dalla rete cosi configurata, questa è stata utilizzata per effettuare le previsioni sulla porzione del campione(composta dal 25% delle osservazioni set dei dati iniziale) contente le osservazioni non utilizzate per addestrare la rete.

.metric	.estimator	.estimate	.config
recall	binary	0.9989498	Preprocessor1_Model1
precision	binary	0.9702856	Preprocessor1_Model1
f_meas	binary	0.9844091	Preprocessor1_Model1
accuracy	binary	0.9711694	Preprocessor1_Model1
kap	binary	0.7937725	Preprocessor1_Model1
sens	binary	0.9989498	Preprocessor1_Model1
spec	binary	0.6863296	Preprocessor1_Model1
roc_auc	binary	0.9770592	Preprocessor1_Model1

Figure 9: Performance

Figure 10: Confusion Matrix

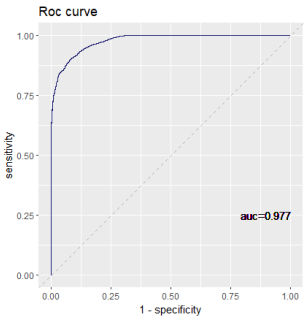


Figure 11: Roc Curve

I risultati del classificatore sviluppato a partire dalla rete neurale per la predizione del diabete sembrano essere molto buoni, con alcune metriche che mostrano risultati eccellenti.

Il modello è in grado di identificare correttamente la stragrande maggioranza dei soggetti che effettivamente hanno il diabete ed ha una bassa percentuale di falsi positivi. Ciò significa che il modello tende a classificare correttamente la maggior parte dei soggetti sani come non affetti da diabete. L'accuratezza del modello è elevata, il che classificatore identifica correttamente il 97.11% dei soggetti, indipendentemente dal fatto che abbiano o meno il diabete. La sensibilità (0.998) rappresenta la percentuale di soggetti affetti da diabete correttamente identificati dal modello, mentre la specificità (0.686) rappresenta la percentuale di soggetti sani correttamente identificati. La sensibilità è molto alta, il che è un buon segno, mentre la specificità potrebbe essere migliorata. Considerato l'elevato sbilanciamento delle classi della variabile dipendente, soffermare la valutazione esclusivamente all'accuratezza potrebbe risultare fuorviante, per tale motivo è stato considerato anche la balanced accuracy, il cui valore resta comunque piuttosto buono in quanto si registra al livello 0,84.
L'area sotto la curva ROC (0.977) suggerisce che il modello ha una buona capacità di discriminazione tra classi positive e negative.

Complessivamente, i risultati del classificatore sono molto positivi, con un'elevata capacità di individuare correttamente i soggetti affetti dal diabete e una buona precisione. Tuttavia, la specificità potrebbe essere migliorata se considerata la condizone di sblianciamento iniziale delle classi.