



UNIVERSITÀ DEGLI STUDI DI SALERNO

Dipartimento di Informatica

Corso di Laurea Triennale in Informatica

TESI DI LAUREA

# **DIA: Un Nuovo Approccio per Spiegare l'Output di Modelli di Predizione del Diabete**

RELATORE

Prof. Fabio Palomba

Dott. Antonio Della Porta

Università degli Studi di Salerno

CANDIDATO

**Francesco Festa**

Matricola: 0512113547

Anno Accademico 2023-2024

*Questa tesi è stata realizzata nel*

sesa<sup>lab</sup>  
SOFTWARE ENGINEERING  
SALERNO

*Alle Stelle,  
che da quando ne ho memoria  
mi indicano il cammino*

## Abstract

Negli ultimi anni l'attenzione verso l'Intelligenza Artificiale (IA) è cresciuta più che mai grazie al rapido e inimmaginabile progresso che questa sta vivendo. Tuttavia questo progresso viene ostacolato da molti problemi, spesso non di natura tecnica ma sociale. Primo fra questi è la fiducia che gli utenti ripongono nelle previsioni dei modelli di ultima generazione, i quali non permettono di comprendere pienamente il motivo per cui viene restituito uno specifico output, e sono per questo detti "black box". Questo problema è particolarmente rilevante in ambiti come quello medico, dove una diagnosi accurata può avere conseguenze significative. Per affrontare questa sfida è nata l'eXplainable Artificial Intelligence (XAI), che ha come obiettivo quello di fornire spiegazioni al comportamento dei modelli. Tuttavia la XAI presenta un grande limite: le spiegazioni fornite sono difficili da interpretare per utenti non esperti, in quanto spesso rappresentate da numeri privi di un significato esplicito. L'obiettivo di questa tesi è quello di rendere più accessibili e comprensibili le spiegazioni degli output di modelli di classificazione del diabete. Per farlo è stato utilizzato un Large Language Model per trasformare queste spiegazioni in testi in linguaggio naturale, attraverso un processo detto "Summarization", rendendole così leggibili anche ad utenti non esperti. È stato per questo sviluppato DIA, un tool che permette di inserire i propri dati per ottenere una diagnosi sulla presenza del diabete, accompagnata da una spiegazione testuale del perché il modello ha fornito quella risposta, superando quindi i limiti di usabilità della XAI. Questa ricerca dimostra che fornire spiegazioni comprensibili all'utente aumenta la sua fiducia nei confronti dei modelli, specialmente in ambiti critici e delicati come quello medico. Per favorire l'integrazione dell'Intelligenza Artificiale nella quotidianità è quindi necessario lavorare non solo sull'explainability dei modelli, ma anche sulla comprensibilità di queste spiegazioni.

---

# Indice

---

<b>Elenco delle Figure</b>	<b>iii</b>
<b>Elenco delle Tabelle</b>	<b>iv</b>
<b>1 Introduzione</b>	<b>1</b>
1.1 Contesto applicativo . . . . .	1
1.2 Motivazioni e obiettivi . . . . .	3
1.3 Risultati ottenuti . . . . .	3
1.4 Struttura della tesi . . . . .	4
<b>2 Background e Stato dell'Arte</b>	<b>5</b>
2.1 IA in Ambito Medico . . . . .	5
2.1.1 IA per la previsione del Diabete . . . . .	6
2.2 eXplainable Artificial Intelligence . . . . .	7
2.2.1 SHAP . . . . .	10
2.2.2 LIME . . . . .	11
2.2.3 CIU . . . . .	12
2.2.4 Anchor . . . . .	12
2.3 XAI in ambito medico . . . . .	13
2.3.1 XAI per la previsione del Diabete . . . . .	14
2.4 Il problema della XAI . . . . .	15

<b>3</b>	<b>Metodologia di Ricerca</b>	<b>16</b>
3.1	Domande di Ricerca . . . . .	16
3.2	Dataset e Feature Engineering . . . . .	17
3.3	Costruzione e valutazione dei modelli . . . . .	21
3.4	Implementazione e valutazione delle tecniche di Explainability . . .	25
3.5	Scelta del modello per Anchor . . . . .	29
3.6	Definizione della Summarization strategy . . . . .	30
<b>4</b>	<b>DIA: un tool per spiegare l'output di modelli black box</b>	<b>40</b>
<b>5</b>	<b>Conclusioni</b>	<b>45</b>
5.1	Impatto . . . . .	45
5.2	Future Work . . . . .	45
	<b>Bibliografia</b>	<b>47</b>

---

## Elenco delle figure

---

2.1	Grafico che mostra la correlazione tra performance e interpretabilità.	9
2.2	Grafico che mostra i valori di Shapley per una previsione. . . . .	11
3.1	Istogrammi prima di gestire i valori mancanti. . . . .	18
3.2	Istogrammi dopo aver gestito i valori mancanti. . . . .	19
3.3	Grafico a torta per il bilanciamento del dataset. . . . .	20
3.4	Pairplot delle feature del dataset. . . . .	21
3.5	Heatmap delle feature del dataset. . . . .	22
4.1	Schermata iniziale di DIA, in cui l'utente può inserire i propri dati. .	41
4.2	Schermata con la previsione del modello e la spiegazione in forma testuale. . . . .	42
4.3	Schermata con una previsione errata e spiegazioni insensate. . . . .	43

---

## Elenco delle tabelle

---

3.1	Feature del dataset PIDDD . . . . .	17
3.2	Performance dei modelli . . . . .	24
3.3	Performance dei modelli dopo lo scaling del dataset. . . . .	25
3.4	Coverage media dei modelli . . . . .	30



# CAPITOLO 1

---

## Introduzione

---

### 1.1 Contesto applicativo

Negli ultimi anni l'Intelligenza Artificiale (IA) è entrata sempre di più a far parte della nostra quotidianità. In particolare il progresso ha preso una forte impennata a partire dal 30 Novembre 2022, quando ChatGPT è stato rilasciato, diventando il primo Large Language Model (LLM) utilizzabile dal pubblico. Da allora basta una veloce ricerca su un qualunque sito di informazione per venir sommersi da notizie riguardanti l'IA, raramente però di carattere scientifico o per annunciare un nuovo modello dalle ottime prestazioni. Difatti i media preferiscono di gran lunga soffermarsi su notizie di carattere sociale, creando spesso allarmismo o incertezza, come la paura di venir sostituiti.

Per questo possiamo affermare che i più grandi ostacoli nella diffusione e integrazione massiva dell'IA non risiedono nei laboratori di ricerca o nelle grandi aziende di prodotto, ma negli utenti finali. Il più grande problema è, infatti, la fiducia che questi ultimi sono disposti a riporre nei sistemi intelligenti. Gli esseri umani nel corso della storia sono sempre stati diffidenti nei confronti dell'innovazione, poiché spesso non è semplice comprendere il funzionamento delle nuove invenzioni. L'IA non è di certo esente da questo ragionamento, soprattutto quando consideriamo che i modelli

più performanti nascondono al loro interno talmente tanta complessità da essere del tutto incomprensibili, tanto da venir definiti "black box".

In opposizione a questa problematica, è nato il campo dell'eXplainable Artificial Intelligence (XAI), il quale obiettivo è quello di rendere i modelli di IA interpretabili per gli umani. In particolare la sfida risiede nei modelli più complessi, per i quali è quasi impossibile spiegare come e perché é stata raggiunta una certa conclusione.

É facile intuire come un modello spiegabile possa trasmettere maggiore sicurezza e fiducia negli utenti, soprattutto in ambiti critici, in cui una previsione errata ha un costo elevato, come quello medico [1]. Un'errata diagnosi da parte di un modello infatti potrebbe costare la vita di un paziente, rendendo l'integrazione dell'IA in questo campo ancora più difficile, nonostante questa possa salvare numerose vite. É stato infatti dimostrato che l'IA in ambito medico può raggiungere prestazioni sensibilmente migliori degli esperti del settore, oltre che velocizzare le procedure di diagnosi e diminuirne i costi [2].

Uno dei campi di ricerca più trattati nell'IA in ambito medico è la diagnosi del Diabete Mellito. Esistono difatti numerosi articoli a riguardo, dai quali traspare l'enorme quantità di dati e informazioni a riguardo [3]. Il Diabete Mellito è una malattia cronica caratterizzata da livelli elevati di glucosio (zucchero) nel sangue, causati da un'anomalia nella produzione o nell'azione dell'insulina, l'ormone che regola il glucosio. Anche qui però, nonostante l'estensiva ricerca, esistono poche soluzioni di spiegabilità, che rendono questi modelli non adatti all'utilizzo nel pratico [4].

Rendere i modelli spiegabili é quindi necessario per abilitare l'integrazione di questi nelle professioni sanitarie, costruendo strumenti a supporto delle decisioni dei medici e aumentando la qualità della sanità su scala globale. Nonostante questo però, la XAI possiede intrinsecamente un enorme problema: l'usabilità. Difatti le principali tecniche di XAI non forniscono una spiegazione pronta all'uso, ma restituiscono dei valori numerici che non possiedono un significato intrinseco, ma vanno contestualizzati nei singoli casi, rendendo paradossalmente la spiegazione di una previsione, non comprensibile ad un utente non esperto o non addestrato nell'utilizzo di tali tecniche [5]. Essendo i modelli di IA comunque dei software, anch'essi rispettano le leggi dell'Ingegneria del Software: un prodotto inutilizzabile non avrà utenti, ed è per questo fallimentare, indipendentemente dalle sue prestazioni.

## 1.2 Motivazioni e obiettivi

Questo lavoro di tesi nasce dall'esigenza di avere spiegazioni dell'output di modelli comprensibili per tutta l'utenza, così da facilitare l'accettazione sociale dell'utilizzo dell'IA nel campo medico, in particolare quello del Diabete Mellito.

Viene quindi proposta una soluzione innovativa a questo problema, sfruttando le avanzate capacità dei LLM per produrre un tool in grado di spiegare attraverso un testo le motivazioni che si celano dietro la previsione di un modello black box del diabete. Ciò avviene grazie ad un processo detto "Summarization", che permette di trasformare gli output in forma testuale. Inoltre implementando un sistema di spiegabilità del genere, è più facile individuare previsioni errate, aumentando ulteriormente l'affidabilità del modello. Infatti un medico esperto può reputare la spiegazione del modello priva di senso rispetto alla sua esperienza, e questo è spesso indice di un'errata previsione [5].

## 1.3 Risultati ottenuti

È stato implementato il tool DIA, il quale appare come una webapp che permette agli utenti (pazienti o medici) di inserire i risultati di analisi standardizzate a livello internazionale per la diagnosi del diabete e ottenere la previsione di un modello corredata da una spiegazione in linguaggio naturale.

Durante lo sviluppo sono state inoltre prodotte una serie di osservazioni sull'efficacia e sull'usabilità di varie tecniche di explainability in relazione a diversi modelli black box per il problema della previsione del diabete.

## **1.4 Struttura della tesi**

Il resto della tesi è diviso in quattro capitoli:

1. Il capitolo due tratta dello stato dell'arte, scendendo nel dettaglio dell'IA in ambito medico, parlando nello specifico del diabete, ed esponendo le principali tecniche di XAI oltre che i loro problemi.
2. Il capitolo tre esprime la metodologia di ricerca nel dettaglio, descrivendo tutti gli step e il razionale dietro le scelte effettuate.
3. Il capitolo quattro mostra il funzionamento del tool oltre che l'analisi dei risultati ottenuti.
4. Il capitolo cinque indica l'impatto del lavoro oltre che future migliorie in relazione a nuove ricerche e tecnologie.

---

### Background e Stato dell'Arte

---

#### 2.1 IA in Ambito Medico

Al giorno d'oggi nessun aspetto della società è al sicuro dall'impatto dell'IA. Tra questi non è di certo escluso il campo della Medicina, il quale, negli ultimi anni, ha subito una vera e propria rivoluzione grazie alle nuove tecnologie. Difatti, nonostante i sistemi di supporto alle decisioni non siano di certo una novità, è solo ora che possono davvero mostrare tutte le loro capacità grazie ad importanti rivoluzioni informatiche, come l'avvento di cartelle cliniche elettroniche, sistemi per la gestione efficace dei Big Data e, soprattutto, nuove tecniche di apprendimento automatico, come il deep learning [2].

In particolare quest'ultimo ha aperto le porte alla classificazione di immagini, implementata grazie alle reti neurali convoluzionali, le quali sono in grado di analizzare quello che è il tipo di dato più comune in questo dominio [2]. Basta infatti pensare alle radiografie o alle risonanze magnetiche, utilizzate come strumenti diagnostici in quasi tutti gli ambiti medici, come Dermatologia, Oculistica e Patologia Clinica.

Yu et al. hanno dimostrato che questi nuovi approcci sono migliori degli esperti del settore dal punto di vista di performance, costi e disponibilità di orari, ma sono notevolmente peggiori dal punto di vista della comprensibilità [2]. Difatti questo è il

più grande freno all'impiego su larga scala di queste tecniche: nessuno accetterebbe una diagnosi senza una spiegazione valida sul come è stata prodotta [1]. Risolvere questo problema significherebbe facilitare l'implementazione di sistemi in grado di salvare un numero di vite maggiore, permettendo inoltre l'accesso alla sanità anche in Stati più in difficoltà, grazie ai costi di impiego sensibilmente minori rispetto a quelli di un medico esperto.

Vi sono però anche importanti altri utilizzi dell'IA che non sono prettamente diagnostici, ma variano dall'impiego di robot intelligenti per effettuare interventi di precisione, al monitoraggio della salute attraverso dispositivi indossabili, ma anche questi hanno bisogno di essere spiegabili e comprensibili all'essere umano.

Tra le altre sfide tecniche da affrontare nell'ambito dell'IA medica si ha, ad esempio, la gestione di grosse quantità di dati provenienti da diverse fonti, per far sì che il modello sia libero da bias, oppure la gestione della privacy dei pazienti. Infatti le cartelle cliniche, per loro natura, contengono numerose informazioni sensibili, quindi vi sono rischi elevati di incorrere in problemi di natura legale [2].

Non vi sono però solo le sfide di natura tecnica, ma anche quelle di natura sociale. Difatti l'impiego dell'IA nel workflow clinico potrebbe non esser visto come una rivoluzione, ma come un pericolo, sia da parte dei pazienti che dei medici. Tra questi ultimi infatti c'è chi sostiene che una scorretta diagnosi da parte del modello possa facilmente creare bias nella diagnosi reale del medico, rendendo l'imputabilità degli errori un discorso estremamente complicato. Infine non bisogna dimenticare che i medici, nella maggior parte dei casi, non sono informatici, e che quindi è necessario spendere fondi in formazione per insegnare il corretto funzionamento dei sistemi di ultima generazione ad un'utenza che, in parte, non è nemmeno disposta ad accettarli [2].

### **2.1.1 IA per la previsione del Diabete**

Uno dei principali campi della medicina in cui l'IA sta trovando numerose applicazioni è quello della diagnosi del Diabete Mellito, una malattia cronica caratterizzata da un'elevata concentrazione di glucosio nel sangue (che prende il nome di iperglicemia), causata da un'insufficiente produzione di insulina, l'ormone che regola il livello

di zucchero nel sangue. Questa malattia si presenta sotto forma di sete eccessiva, fame intensa, perdita di peso inspiegabile, affaticamento o vista offuscata. Se non trattato però, il diabete può portare anche a complicazioni più gravi, come problemi cardiovascolari, danni ai reni o perdita della vista.

Questa malattia non può essere curata, ma può essere gestita attraverso cambiamenti nello stile di vita, che includono terapie farmacologiche, monitoraggio regolare della glicemia e una dieta che permetta di mantenere i livelli di glicemia sotto controllo. È per questo importante diagnosticare quanto prima la malattia, così da iniziare immediatamente i trattamenti necessari ed evitare complicazioni.

Ad oggi vi è molto interesse nella ricerca e nelle applicazioni dell'IA per la previsione del Diabete, rendendo il quantitativo di dati e le soluzioni proposte numerosi e vari [3]. Un esempio degno di nota è un nuovo sistema utilizzato per diagnosticare la Retinopatia Diabetica, una complicanza del diabete che può causare la perdita della vista, basato su immagini della retina, il quale è ampiamente diffuso e utilizzato in India [2].

Nonostante l'enorme interesse nel campo del deep learning per la diagnostica da immagini (il quale rappresenta essere il principale argomento di ricerca sull'ambito) non sono assenti gli approcci su dati tabulari, sui quali è stato possibile raggiungere ottime prestazioni attraverso l'impiego di tecniche molto complesse come ad esempio reti neurali [3]. Questi modelli black box non sono però spiegabili, e questo li rende inadeguati ad un ambiente di produzione.

Questo significa che, nonostante la ricerca sia abbondante, essa risulta essere lontana dalla realtà, rendendo molte delle scoperte effettuate poco utili al miglioramento delle procedure diagnostiche. Diventa quindi importante comprendere come è possibile migliorare gli approcci proposti e renderli più vicini alle esigenze reali dei pazienti. La risposta a questa domanda si trova nel campo dell'eXplainable Artificial Intelligence.

## **2.2 eXplainable Artificial Intelligence**

L'eXplainable Artificial Intelligence (XAI) è un insieme di tecniche e algoritmi che permette di rendere i modelli di intelligenza artificiale più trasparenti e comprensibili

per gli esseri umani. Questo permette di costruire fiducia, facilita l'adozione in settori critici e permette un uso responsabile dell'IA.

La XAI nasce dalla necessità di comprendere come funzionano i modelli di IA più complessi, definiti black box, poiché non è normalmente possibile individuare il modo in cui questi sono giunti ad una conclusione. Molte delle tecniche più basilari e semplici sono di natura interpretabili, basti pensare agli alberi decisionali, attraverso i quali è possibile produrre un insieme di regole percorrendo il cammino dal nodo radice alla foglia raggiunta [5]. Per questo motivo non tutti i modelli necessitano di complessi algoritmi di XAI, ma di una semplice attenzione aggiuntiva in fase di implementazione, per permettere di visualizzare in modo corretto il processo decisionale del modello.

Apparentemente basterebbe limitarsi ad utilizzare le tecniche spiegabili di natura per rispettare i requisiti di spiegabilità, ma purtroppo questa non è sempre la scelta migliore. Esiste infatti un'importante correlazione tra la performance e l'interpretabilità di un modello: più il modello è performante, meno questo sarà interpretabile (Figura 2.1). Questo significa che se si desiderano performance migliori, bisogna necessariamente utilizzare modelli black box, ed è su questi che è possibile applicare le tecniche di XAI più interessanti [5].

È possibile quindi suddividere la XAI in due insiemi di tecniche ben distinte:

- **Modelli interpretabili di natura:** sono modelli progettati per essere comprensibili, la cui struttura è trasparente e intuitiva. Tra questi vi sono i già citati alberi di decisione, ma anche la regressione lineare o i modelli di classificazione statistica.
- **Tecniche post-hoc:** queste tecniche vengono applicate dopo l'addestramento di un modello complesso, per permettere di comprenderne il funzionamento interno. Tra queste è possibile distinguere le tecniche per semplificazione, che costruiscono nuovi modelli più semplici e spiegabili che approssimano il funzionamento del modello originale (ad esempio LIME), oppure le tecniche di feature relevance, che misurano quanto ogni feature ha impatto nella produzione di un dato output (ad esempio SHAP).





**Figura 2.1:** Grafico che mostra la correlazione tra performance e interpretabilità.

Poiché si desiderano sempre le performance migliori per i modelli, è chiaro che le tecniche post-hoc siano quelle più utili e sulle quali vi è maggiore ricerca [5]. I modelli principali che richiedono tecniche di spiegabilità aggiuntive sono gli ensemble di alberi, come Random Forest e XGBoost oppure le tecniche di deep learning, come è possibile vedere in Figura 2.1.

La XAI è estremamente importante, in quanto gli umani sono diffidenti di natura dalle tecniche che non comprendono o che, peggio ancora, non si possono comprendere in primo luogo. Per questo motivo tra gli obiettivi della XAI vi è la possibilità di informare su come funziona un modello non spiegabile gli utenti non esperti, i quali però hanno tuttora ancora molte difficoltà dovute all'usabilità di queste tecniche [5].

Un altro importante obiettivo della XAI è la riduzione dell'insicurezza da parte degli utenti del sistema, in quanto se una spiegazione non risulta convincente, questo è probabilmente indice di una previsione errata, e che quindi è possibile rifiutare [5].

Infine tra gli altri obiettivi della XAI vale la pena citare la possibilità di trasferire la conoscenza di un modello ad un altro o la possibilità di controllare che un modello

sia equo ed esente da bias [1].

É opportuno ora procedere con la descrizione delle principali tecniche di spiegabilità post-hoc per dati tabulari. Queste sono tutte definite model-agnostic, poiché funzionano a priori dal modello di partenza, e che quindi non fanno riferimento ad uno specifico algoritmo di apprendimento.

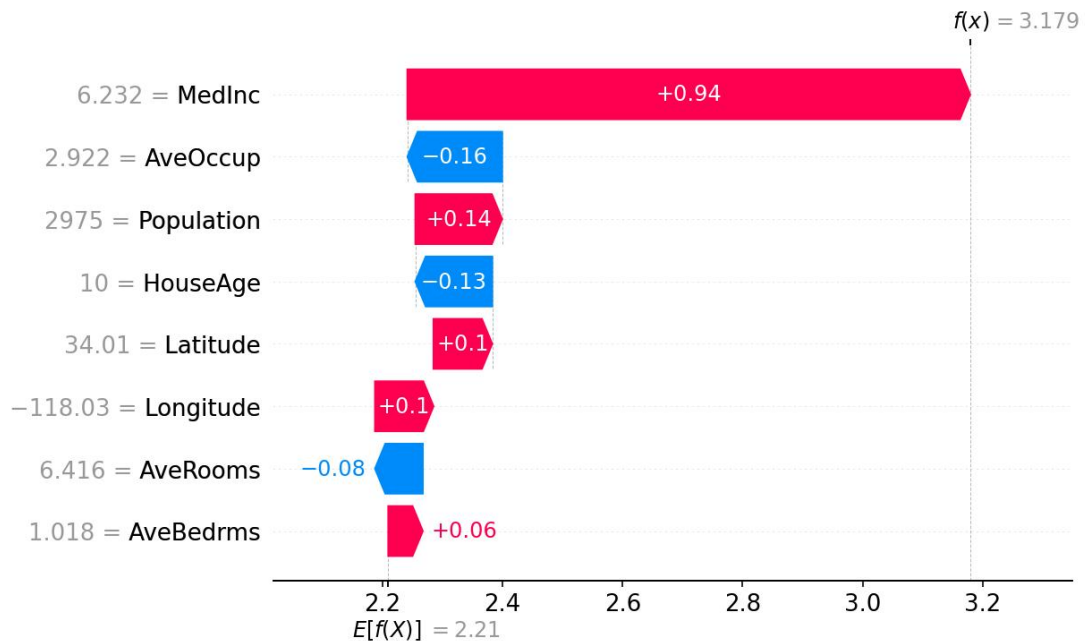
### 2.2.1 SHAP

SHAP (SHapley Additive exPlanations) è una tecnica di spiegabilità post-hoc proposta da Lundberg e Lee basata sulla teoria dei giochi [6]. Alla base di questa tecnica vi sono i valori di Shapley, una misura che distribuisce in modo equo il contributo di ciascun input alla previsione del modello, trattando le feature come "giocatori" che collaborano per produrre una previsione.

SHAP inizia confrontando la previsione del modello per un'istanza specifica con una previsione definita "di base", la quale è pari alla media delle previsioni su tutto il dataset. Dopodiché viene calcolato il contributo di ogni feature considerando tutte le possibili combinazioni in cui queste possono essere aggiunte o rimosse dal modello. Questo permette di determinare quanto la presenza di una feature e il suo valore cambiano la previsione rispetto alla media. Dopo aver calcolato tutti i contributi, viene effettuata la media, generando così un valore di Shapley per ogni feature.

Una delle proprietà più interessanti di SHAP è che la sommatoria dei valori di Shapley per tutte le feature corrisponde esattamente alla differenza tra la previsione del modello per l'istanza e la previsione di base. Questo significa che é facile comprendere quanto ogni singola feature ha contribuito alla previsione, rendendo il modello trasparente.

SHAP viene solitamente utilizzata per produrre grafici sui valori di Shapley, i quali risultano essere molto comprensibili anche per utenti non esperti, in quanto intuitivi di natura (Figura 2.2). Inoltre vale la pena notare che SHAP eccelle nel produrre spiegazioni globali dei modelli, che permettono di comprenderne il funzionamento nella loro interezza.



**Figura 2.2:** Grafico che mostra i valori di Shapley per una previsione.

## 2.2.2 LIME

LIME (Local Interpretable Model-agnostic Explanations) è una tecnica di spiegabilità post-hoc proposta da Ribeiro et al. che si basa sulla produzione di un modello lineare surrogato locale e interpretabile [7].

LIME crea un modello interpretabile e approssimato attorno alla singola previsione del modello originale per una determinata istanza. L'obiettivo è quello di trovare un modello quanto più fedele a quello originale possibile, ma solo nell'area attorno all'istanza di interesse. Per costruire tale modello, LIME genera delle istanze sintetiche leggermente perturbate rispetto all'istanza originale, per poi effettuare una previsione su ognuna di queste. Utilizzando poi le previsioni del modello complesso per le istanze campionate, LIME costruisce il modello surrogato in modo tale che questo sia una fedele approssimazione del modello originale ma solo attorno all'istanza di interesse. Questo viene quindi usato per comprendere quali feature influenzano maggiormente la previsione e in che modo.

Questa tecnica è ad oggi molto utilizzata nonostante sia una tecnica di approssi-

mazione, e che quindi rischia di non approssimare correttamente il modello originale. Difatti nei casi in cui si hanno modelli troppo complessi, potrebbe non essere possibile approssimarli correttamente in modo lineare.

### 2.2.3 CIU

CIU (Contextual Importance and Utility) è una tecnica di spiegabilità post-hoc proposta da Frimling, che fornisce una valutazione dell'importanza e dell'utilità delle feature in una singola previsione [8]. Come SHAP, CIU non costruisce un modello surrogato, ma valuta il comportamento del modello esistente.

CIU calcola due valori, la Contextual Importance (CI) e la Contextual Utility (CU) attraverso i quali è possibile comprendere il modo in cui è stato raggiunto un determinato output:

- **Contextual Importance (CI):** La CI misura quanto una particolare feature influenzi la previsione del modello. Viene calcolata confrontando la previsione attuale con la previsione che si otterrebbe modificando i valori della feature in esame all'interno di un determinato intervallo definito dal contesto.
- **Contextual Utility (CU):** La CU misura quanto il valore di una feature contribuisca positivamente o negativamente al raggiungimento della previsione corrente. In altre parole, misura quanto il valore di una feature sia "buono" nel contesto della previsione. Viene calcolata normalizzando la previsione corretta rispetto ai valori di massimo e minimo che si ottengono modificando solo la feature di interesse.

Attraverso questi due concetti è possibile fornire una spiegazione intuitiva dell'output: la CI indica quali feature sono importanti per la previsione, mentre la CU spiega se i valori di queste sono favorevoli o meno nel contesto del risultato previsto.

### 2.2.4 Anchor

Anchor è una tecnica di spiegabilità post-hoc proposta da Ribeiro et al., che permette di generare regole interpretative chiamate ancore, che spiegano il comporta-

mento del modello [9]. L'idea è che un'ancora rappresenta un insieme di condizioni sufficienti per garantire una certa predizione su un sottoinsieme di dati simili.

Anchor costruisce le regole combinando valori specifici delle feature rilevanti per una determinata previsione (ad esempio un'ancora per la diagnosi positiva del diabete potrebbe essere "Glucosio > 150 AND BMI > 30"). Una qualunque istanza che soddisfa tali condizioni avrà la stessa classificazione con altissima probabilità, indipendentemente dal valore delle altre feature. Per generare le regole Anchor perturba i dati circostanti per poi prevedere gli output di questi: se la previsione rimane costante per i nuovi esempi, allora l'ancora è considerata una buona spiegazione locale. Difatti vengono considerate ancora solo le regole che hanno una precisione molto alta, maggiore del 95

Anchor utilizza un approccio di costruzione incrementale delle regole: parte con una condizione semplice (una singola feature) e, se la precisione non è sufficiente, aggiunge nuove feature fino a raggiungere la precisione desiderata.

Questa tecnica permette di produrre spiegazioni estremamente intuitive e comprensibili, anche per utenti non esperti. Un altro aspetto interessante è anche la capacità delle regole di coprire una regione del modello piuttosto che un singolo punto, rendendo la spiegazione più robusta. Per questo motivo è anche possibile calcolare una metrica detta "Coverage", ovvero la percentuale di istanze coperte da una determinata ancora.

## 2.3 XAI in ambito medico

Ora che è chiaro cosa è la XAI, bisogna comprendere cosa è stato fatto per integrarla al meglio nell'ambito medico. Tra i vari ricercatori, Band et al. hanno sottolineato l'enorme importanza di quest'area di ricerca, oltre che evidenziare l'enorme quantitativo di studi in merito [1].

In particolare, poiché molta classificazione è effettuata attraverso immagini, molte delle tecniche più innovative sono applicate per la spiegabilità di reti neurali convoluzionali. Tra queste è bene citare NeuroXAI, una tecnica per spiegare l'output di sistemi per l'individuazione di tumori al cervello, oppure GradCAM, una tecnica per

interpretare l'output attraverso la visualizzazione delle aree che hanno influenzato maggiormente la decisione del modello [1].

Per le spiegazioni di output di modelli addestrati su dati tabulari invece, le principali tecniche utilizzate sono quelle già citate, SHAP, LIME, CIU e Anchor, nonostante su queste ultime due, la ricerca sia sensibilmente minore rispetto alle altre, essendo più recenti [1].

Band et al. si sono anche soffermati sull'usabilità delle tecniche su dati tabulari, che risulta non essere soddisfacente per l'utente medio, vista la loro complessità. Viene infatti ribadito che ad oggi non esistono molte soluzioni automatizzate e affidabili che portino a spiegazioni convincenti in ambito medico [1].

### **2.3.1 XAI per la previsione del Diabete**

Nonostante il Diabete sia un'area molto attiva nella ricerca in ambito IA, non si può dire lo stesso dell'ambito XAI, dove le ricerche risultano essere scarse [1].

Una delle soluzioni proposte più interessanti, è quella di Chen et al., i quali hanno ideato un'architettura IA composta da una rete neurale, utilizzata per prevedere la probabilità che un paziente abbia il diabete, la quale viene approssimata da un albero decisionale, in grado di produrre una serie di semplici regole di decisione. Nel tentativo di rendere l'approccio più usabile e comprensibile per i pazienti è stata infine applicata una tecnica innovativa per aggregare e ordinare le regole per renderle più leggibili [4].

Nonostante ciò, la tecnica utilizza un modello surrogato, e per questo rischia di generalizzare male la rete neurale originaria, producendo regole potenzialmente errate. Inoltre viene anche dichiarato che l'approccio, seppur innovativo, non è ancora molto usabile per i pazienti [4].

Da qui emerge quindi un buco nella letteratura scientifica, in quanto non esistono ad oggi tecniche efficaci di spiegabilità per il diabete che siano anche comprensibili per gli utenti medi.

## 2.4 Il problema della XAI

É ormai chiaro che la spiegabilità è uno dei requisiti fondamentali in tutti i sistemi IA, specialmente quelli in ambito medico, poiché senza di essa gli utenti non potrebbero riporre fiducia nel modello [1].

A prima vista sembrerebbe un problema "risolto", infatti apparentemente basterebbe aggiungere le adeguate tecniche di spiegabilità ai modelli per soddisfare il requisito. Il problema però si cela nell'usabilità di queste tecniche: un utente non esperto posto di fronte ai risultati dei principali algoritmi di XAI si ritroverebbe una serie di numeri apparentemente senza alcun significato. Inoltre non aiuta di certo il fatto che ogni tecnica interpreta i valori in modo diverso, causando negli utenti ulteriore confusione, nel caso in cui questi si siano interfacciati a diverse tecniche di XAI in passato [5].

Una spiegazione incomprensibile è utile quanto una spiegazione assente. A tal proposito vale la pena citare la definizione di modello spiegabile data da Arrieta et al. [5]: "Dato un pubblico, un modello spiegabile è tale se produce dettagli o motivazioni che rendono il suo funzionamento chiaro e facile da comprendere". Se il pubblico sono però persone non esperte di informatica, e che quindi non conoscono il funzionamento delle tecniche, possiamo davvero definire i nostri modelli spiegabili?

Seguendo la definizione sopracitata, le tecniche di spiegabilità che la rispettano sono pochissime e poco utilizzate, almeno per quanto riguarda i dati tabulari. É per questo necessario lavorare e produrre ulteriore ricerca sul campo dell'usabilità dei metodi di XAI, in quanto questa è il vero collo di bottiglia per la messa in uso su larga scala dell'IA nella medicina [1].

---

### Metodologia di Ricerca

---

#### 3.1 Domande di Ricerca

Avendo evidenziato gli importanti gap nello Stato dell'Arte, questa tesi si propone di fornire una nuova soluzione che rispetti i parametri di affidabilità e comprensibilità, i quali hanno valore centrale in questo contesto. Per fare ciò sono state prodotte le seguenti domande di ricerca:

**Q RQ<sub>1</sub>.** *È possibile utilizzare i LLM per rendere le spiegazioni degli algoritmi di XAI più comprensibili?*

**Q RQ<sub>2</sub>.** *È possibile utilizzare spiegazioni in linguaggio naturale per aumentare la fiducia degli utenti nei confronti delle previsioni di un modello?*

**Q RQ<sub>3</sub>.** *È possibile utilizzare spiegazioni in linguaggio naturale per aumentare l'affidabilità delle previsioni di un modello?*

Per rispondere a queste domande sono stati effettuati i passi descritti nei prossimi paragrafi, fino ad arrivare all'implementazione del tool DIA, una webapp che fornisce spiegazioni testuali dell'output di modelli black box per la previsione del diabete.



**Tabella 3.1:** Feature del dataset PIDD

Feature	Descrizione	Tipo di dato
Pregnancies	Numero di gravidanze	Intero
Glucose	Concentrazione di glucosio nel plasma	Intero
BloodPressure	Pressione sanguigna	Intero
SkinThickness	Spessore della plica cutanea del tricipite	Intero
Insulin	Insulina sierica a 2 ore di diguno	Intero
BMI	Indice di massa corporea	Decimale
DiabetesPedigree Function	Funzione che quantifica la probabilità di sviluppare il diabete basandosi sulla storia clinica familiare	Decimale
Age	Età del paziente	Intero
Outcome	La diagnosi sulla presenza del diabete. Sarà la fea- ture da prevedere	Booleano

## 3.2 Dataset e Feature Engineering

Prima di poter implementare le tecniche di explainability e utilizzare un LLM per produrre un testo, abbiamo bisogno di modelli da poter spiegare, e per ottenere questi dobbiamo partire dai dati.

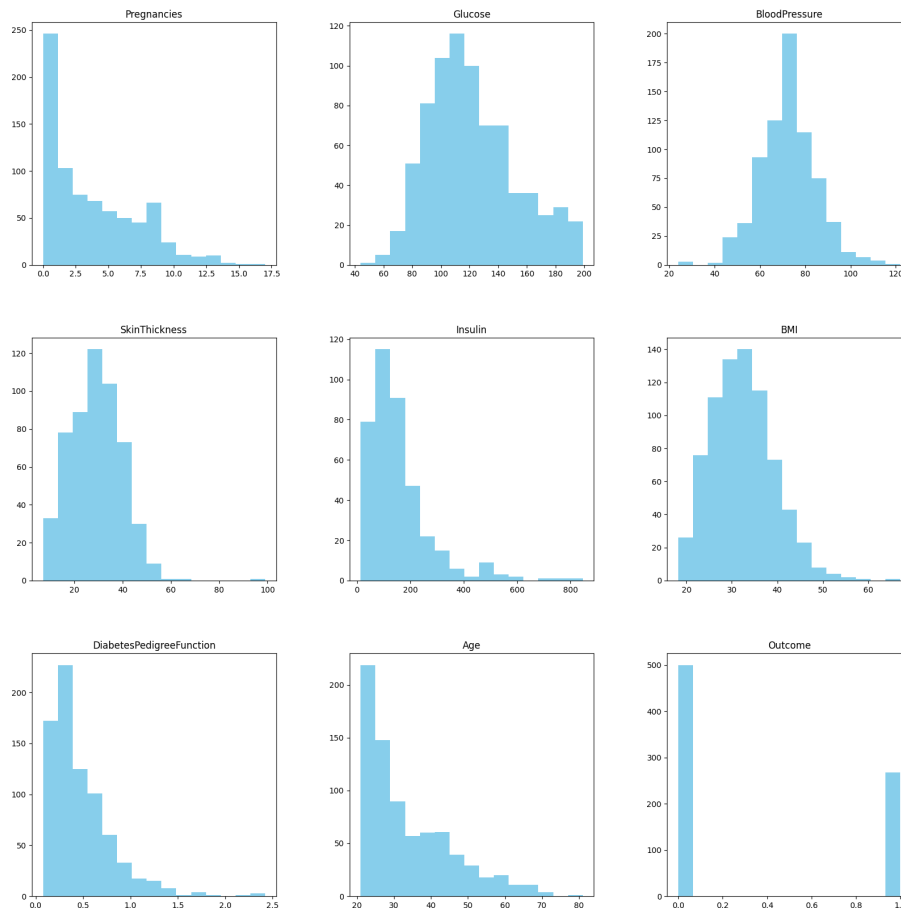
Il dataset scelto è il Pima Indians Diabetes Database (PIDD), uno dei più utilizzati quando si parla di previsione del diabete da dati tabulari [3]. Il dataset è reperibile su Kaggle al seguente url: <https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database> e contiene dati esclusivamente di donne di almeno 21 anni di discendenza degli Indiani Pima.

Esso è formato da 768 righe e 9 colonne, le quali sono descritte nella Tabella 3.1.

Indagando sugli intervalli delle colonne, notiamo che alcune righe hanno dei valori pari a 0 per feature in cui questo è logicamente impossibile, un problema riportato anche da Chaki et al. [3]. Questi sono valori mancanti non apparenti, e per questo vanno gestiti, per evitare che il classificatore apprenda erroneamente. Le colonne che soffrono di tale problema sono le seguenti:

- **Glucose** con 5 valori mancanti.
- **BloodPressure** con 35 valori mancanti.
- **SkinThickness** con 227 valori mancanti.
- **Insulin** con 374 valori mancanti.
- **BMI** con 11 valori mancanti.

Per decidere come gestire l'imputazione di questi valori sono stati prodotti gli istogrammi di ogni feature (Figura 3.1), così da studiarne le distribuzioni.

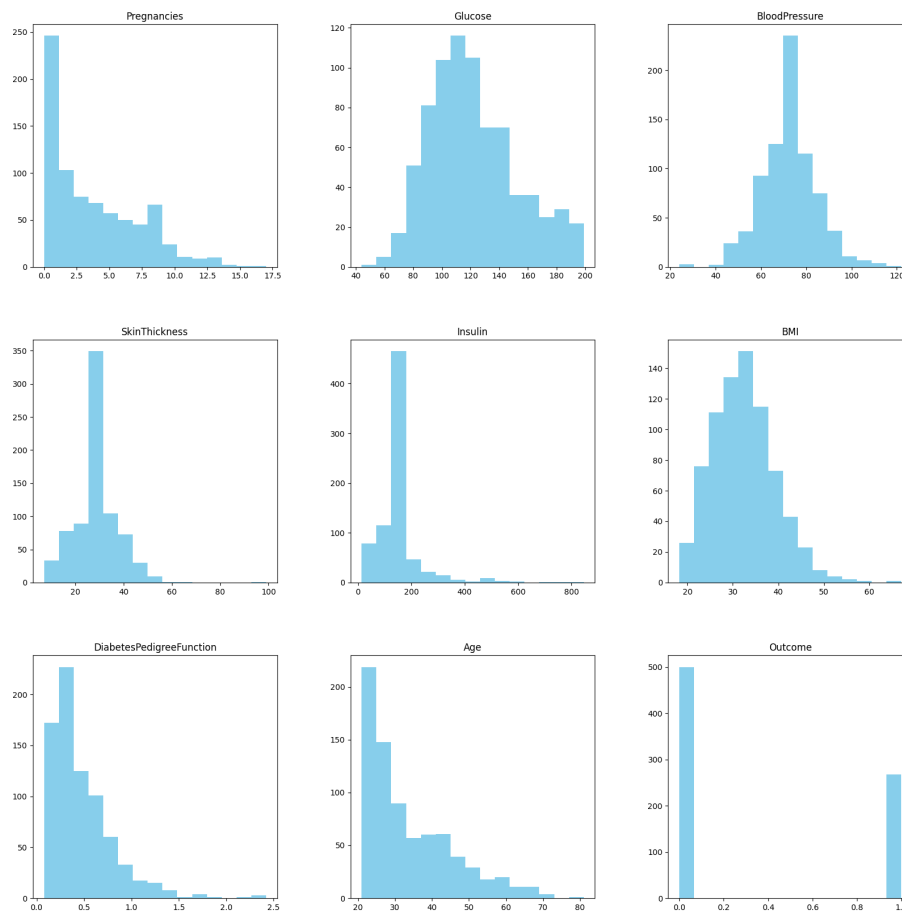


**Figura 3.1:** Istogrammi prima di gestire i valori mancanti.

Da questi sono emerse le seguenti osservazioni:

- Glucose e BloodPressure hanno distribuzione gaussiana, quindi è opportuno sostituire i valori mancanti con la media.
- SkinThickness, Insulin e BMI seguono invece una distribuzione asimmetrica, quindi è opportuno sostituire i valori mancanti con la mediana.

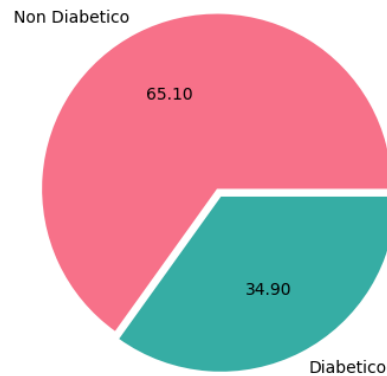
Dopo aver sostituito i valori mancanti sono stati prodotti nuovamente gli istogrammi (Figura 3.2) per accettarsi che le distribuzioni non siano cambiate.



**Figura 3.2:** Istogrammi dopo aver gestito i valori mancanti.

Dopo aver imputato i valori mancanti è stato controllato il bilanciamento del dataset e da qui è emerso che i dati non sono bilanciati. Difatti vi sono 500 istanze di

persone sane (pari al 65,10% del dataset) e 268 istanze di persone diabetiche (pari al 34,90% del dataset) (Figura 3.3).



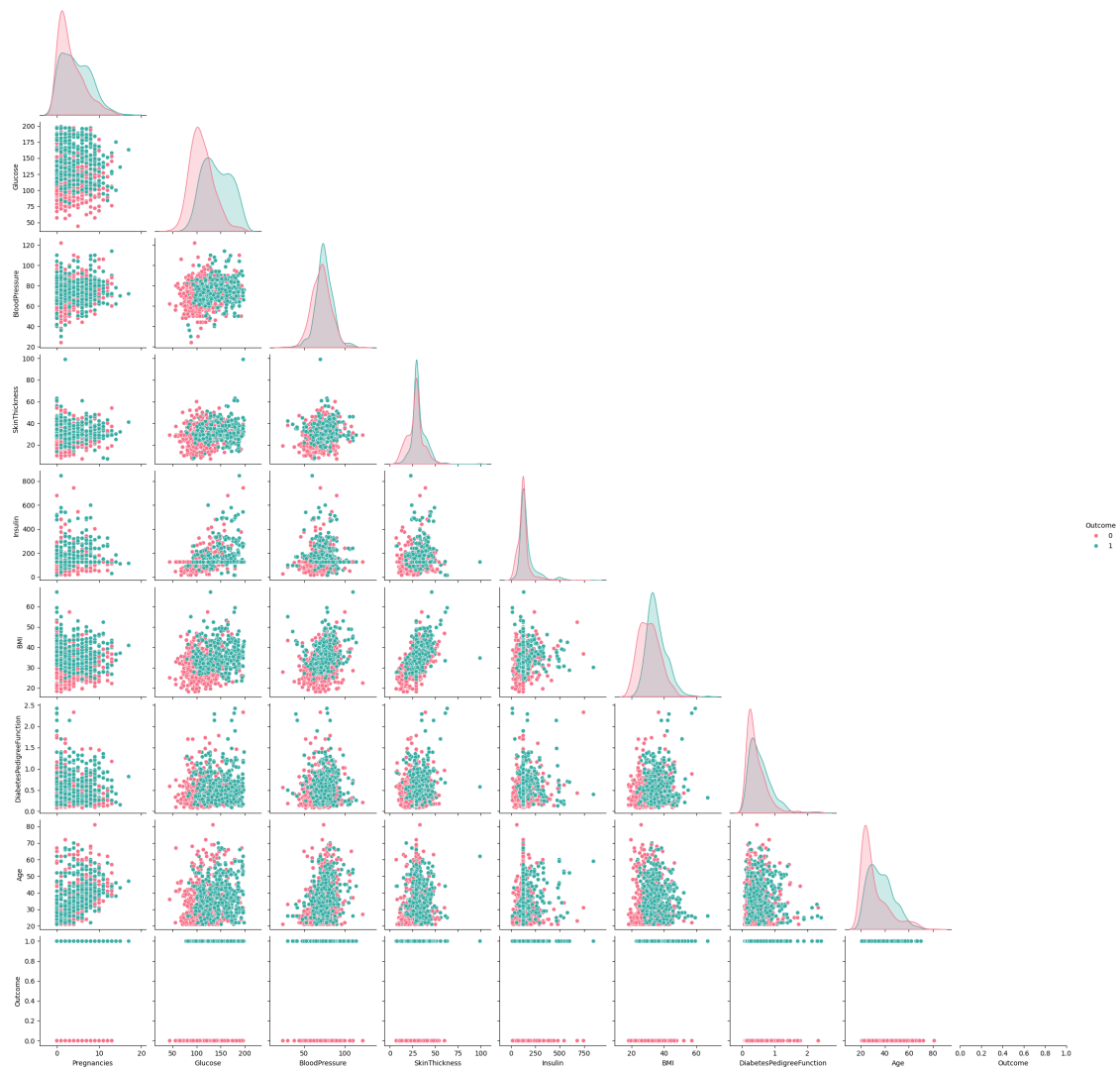
**Figura 3.3:** Grafico a torta per il bilanciamento del dataset.

Per risolvere il problema è stata utilizzata la tecnica SMOTE (Synthetic Minority Over-sampling Technique). Questa tecnica permette di generare nuovi campioni sintetici della classe minoritaria, ovvero le persone diabetiche, creando istanze intermedie tra i campioni reali. In questo modo è possibile ridurre il bias verso la classe maggioritaria, aumentando così le prestazioni dei modelli.

Dopo aver applicato SMOTE il dataset ha raggiunto le 1000 righe, 500 per ogni classe, diventando perfettamente bilanciato.

Il dataset è ora pronto per il training dei modelli. Sono stati quindi prodotti dei pair plot (Figura 3.4) e una heatmap (Figura 3.5), così da studiare la correlazione delle varie feature. Questo è stato fatto principalmente per diagnosticare eventuali problemi nelle spiegazioni prodotte dagli algoritmi di Explainability, ad esempio individuando quando questi restituiscono correlazioni fasulle.

Dal pairplot è possibile notare che le feature che hanno maggiore correlazione con Outcome sono Glucose, BMI ed Age, in linea con le procedure diagnostiche e la sintomatologia del Diabete Mellito. Difatti il diabete si presenta attraverso una quantità maggiore di glucosio nel sangue, ma è anche molto più comune nelle persone in sovrappeso o obese e in età adulta o anziana.



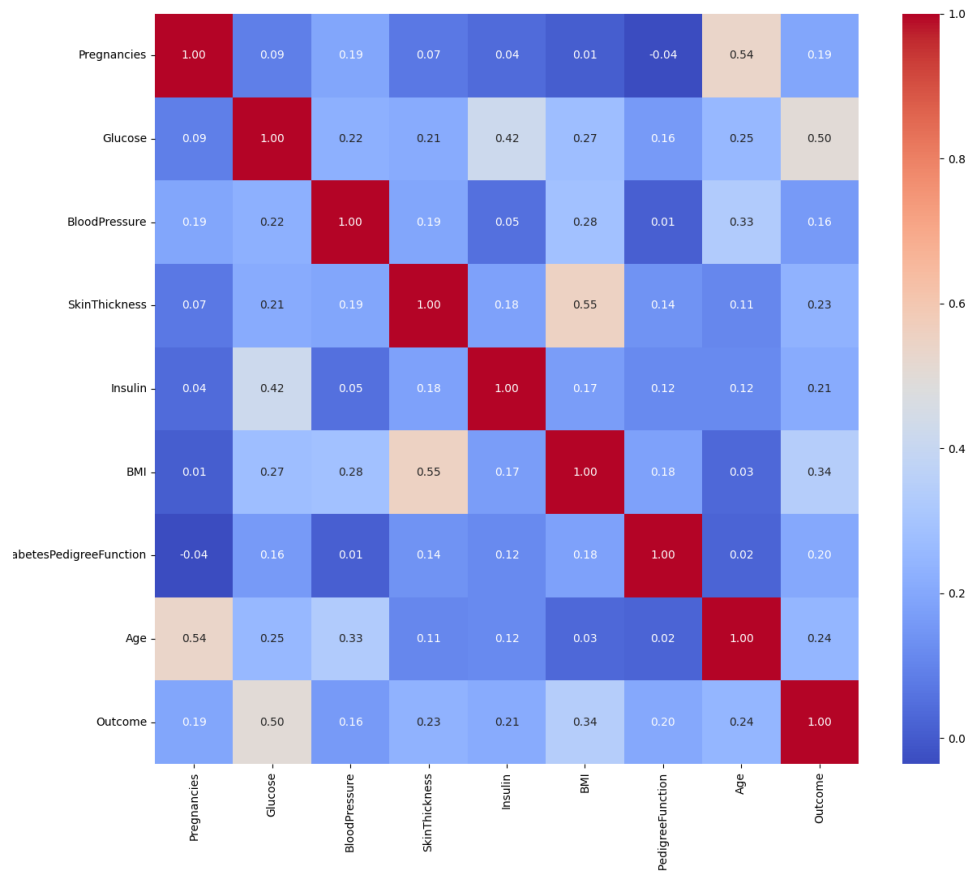
**Figura 3.4:** Pairplot delle feature del dataset.

Dalla heatmap invece è possibile confermare la correlazione di Glucose e BMI con Outcome, oltre che notare che le altre feature sono tendenzialmente indipendenti le une dalle altre.

### 3.3 Costruzione e valutazione dei modelli

Il dataset è stato diviso in 80% training set e 20% test set seguendo il principio di Pareto e mantenendo il bilanciamento delle classi, così da non perdere troppi dati di addestramento.

Sono stati selezionati i seguenti modelli da addestrare:



**Figura 3.5:** Heatmap delle feature del dataset.

- **Random Forest:** un algoritmo di ensemble che utilizza un insieme di alberi decisionali. Funziona creando una "foresta" di alberi, dove ognuno di questi è un classificatore addestrato su un diverso sottoinsieme casuale di dati, utilizzando inoltre un sottoinsieme casuale di feature in ogni divisione dei nodi. Nelle task di classificazione ciascun albero effettua la sua previsione, e la classe finale è determinata dalla maggioranza dei voti tra gli alberi. L'estensivo impiego di casualità nella fase di training permette di ridurre l'overfitting, ma ha come effetto collaterale quello di rendere il modello non spiegabile, a differenza dei singoli alberi decisionali.
- **Support Vector Machine (SVM):** un algoritmo che cerca un iperpiano ottimale che separa i dati delle classi in zone distinte. Questo viene fatto massimizzando la distanza tra i punti più vicini delle classi e l'iperpiano stesso. Questi punti

vicini sono detti support vectors e sono cruciali per determinare la posizione e l'orientamento dell'iperpiano di separazione. Infatti l'algoritmo si concentra solo su questi punti, ignorando quelli più lontani, così da ottimizzare la velocità di calcolo. Nel caso in cui i dati non siano linearmente separabili, SVM utilizza una funzione di kernel per mappare i dati in uno spazio di dimensioni superiori, dove invece è possibile trovare un iperpiano lineare. Questo algoritmo funziona molto bene in problemi in cui le classi sono ben divise ma, poiché ragiona utilizzando piani multidimensionali incomprensibili per l'essere umano, risulta essere non spiegabile.

- **XGBoost:** un algoritmo di ensemble che si basa sulla tecnica del boosting, una metodologia che costruisce un modello forte combinando una serie di modelli deboli, tipicamente alberi decisionali. Il principio è quello di costruire alberi in modo sequenziale, dove ogni nuovo albero cerca di correggere gli errori di quelli precedenti. Questo processo di apprendimento incrementale consente al modello di affinare continuamente le sue previsioni, migliorandone le prestazioni complessive. L'enorme quantità di alberi decisionali e di interazioni tra questi rende però il modello non spiegabile, in quanto è impossibile seguire il percorso di decisione complessivo per comprendere come il risultato finale viene generato.
- **Neural Network:** un particolare modello di apprendimento automatico ispirato al funzionamento del cervello umano. Funziona attraverso un numero variabile di strati di neuroni, ognuno dei quali possiede una funzione di attivazione che trasforma l'input e permette al modello di catturare relazioni complesse e non lineari tra le feature. Durante l'addestramento il modello viene ottimizzato attraverso un processo chiamato backpropagation. Applicandolo, l'errore tra le previsioni del modello e le etichette reali viene propagato all'indietro attraverso la rete per aggiornare i pesi delle connessioni tra i neuroni, così da permettere al modello di fare previsioni più accurate. La loro complessa struttura, che può arrivare a contenere anche milioni di parametri, rende difficile comprendere come i dati vengano trasformati, rendendo anche questo modello non spiegabile. In particolare la rete neurale utilizzata è composta da quattro strati,

rispettivamente di 32, 16, 8 e 1 neurone, tutti con funzione di attivazione Relu tranne l'ultimo, il quale ha una funzione Sigmoidale.

Questi modelli sono stati scelti poiché hanno tutti delle ottime performance ma non sono spiegabili, costringendo l'utilizzo di tecniche di spiegabilità post-hoc.

I modelli sono stati valutati sulla base di:

- **Accuracy:** misura la percentuale di previsioni corrette sul numero di previsioni totali effettuate.
- **Precision:** misura la percentuale di previsioni positive corrette (persone col diabete) sul totale delle previsioni positive effettuate.
- **Recall:** misura la percentuale di veri positivi individuati rispetto al totale dei positivi esistenti.
- **F1 Score:** è la media armonica tra Precision e Recall.

A seguito del testing i modelli hanno ottenuto le metriche indicate nella Tabella 3.2. Le performance non sono certamente pessime, ma è sicuramente possibile effettuare un tentativo per migliorarle.

	Accuracy	Precision	Recall	F1 Score
Random Forest	0.7945	0.7485	0.8300	0.7871
SVM	0.7841	0.7700	0.7485	0.7590
XGBoost	0.7808	0.7695	0.8000	0.7845
Neural Network	0.8010	0.7978	0.7410	0.7684

**Tabella 3.2:** Performance dei modelli

Un primo tentativo è stato fatto attraverso il fine tuning dei modelli utilizzando una Grid Search, una tecnica di ottimizzazione che permette di individuare la migliore combinazione di iperparametri di ogni modello. Questa consiste nell'esplorare in modo sistematico una griglia predefinita di valori per ogni iperparametro, per poi valutare ogni combinazione possibile. Infine viene selezionata quella che offre le migliori performance. A seguito di questa operazione però i modelli non sono



migliorati, lasciando intuire che la miglior combinazione è quella di default per ciascuno di essi.

È stato quindi deciso di provare a lavorare sui dati. Difatti le varie feature possiedono diversi ordini di grandezza, e questo può potenzialmente portare i modelli a generare bias errati, per esempio dando maggiore importanza a una feature piuttosto che ad un'altra sulla base del loro ordine. Per questo motivo il dataset è stato scalato utilizzando uno Standard Scaler, per poi riaddestrare e testare i modelli.

Questa volta abbiamo avuto un miglioramento sostanziale nelle performance, indicato nella Tabella 3.3.

	Accuracy	Precision	Recall	F1 Score
<b>Random Forest</b>	0.8250	0.7982	0.8700	0.8325
<b>SVM</b>	0.8200	0.7909	0.8700	0.8286
<b>XGBoost</b>	0.8100	0.7981	0.8300	0.8137
<b>Neural Network</b>	0.8350	0.8131	0.8700	0.8406

**Tabella 3.3:** Performance dei modelli dopo lo scaling del dataset.

È evidente che il modello con le prestazioni migliori è la Neural Network. Tuttavia, gli altri modelli mostrano performance molto simili, quindi non è opportuno escluderli a priori. Pertanto la scelta del classificatore non può esser determinata esclusivamente dalle sue performance, ma anche e soprattutto da quale tra questi produrrà le spiegazioni migliori.

## 3.4 Implementazione e valutazione delle tecniche di Explainability

Una volta prodotti i modelli è stato necessario implementare degli algoritmi per spiegare gli output di questi. Tutte le tecniche selezionate hanno implementazioni open-source, ma spesso queste non possono essere utilizzate direttamente, richiedendo modifiche nel codice sorgente. Nei casi in cui siano state effettuate operazioni di modifica, queste verranno riportate esplicitamente.

Sono state valutate le seguenti tecniche sulla base dei loro vantaggi e svantaggi:

- **SHAP:** permette di individuare i singoli contributi di ogni feature alla previsione finale, senza però creare un modello surrogato, garantendo una spiegazione che si applica al modello effettivo, piuttosto che ad un'approssimazione. I valori però non sono immediatamente interpretabili per l'utente, difatti solitamente SHAP viene utilizzato per la produzione di grafici facilmente leggibili, ma l'obiettivo di questa ricerca è quello di produrre una spiegazione testuale. Inoltre SHAP eccelle in particolare nel produrre la spiegazione globale del modello, ma poiché il tool che si vuole produrre deve funzionare sulle singole diagnosi, siamo interessati principalmente alle spiegazioni locali delle previsioni. L'implementazione di questa tecnica è risultata semplice, in quanto è presente già una libreria perfettamente funzionante che la implementa attraverso degli oggetti `Explainer`. L'unica attenzione necessaria è stata quella di utilizzare il corretto `Explainer` per ogni modello (ad esempio nel caso di Random Forest e XGBoost è stato necessario un `TreeExplainer`, mentre per la Neural Network ed SVM è stato utilizzato un `KernelExplainer`).
- **LIME:** come SHAP permette di calcolare i singoli contributi delle caratteristiche per una previsione, creando però un modello locale approssimato ma interpretabile. Questo significa che i modelli surrogati potrebbero non catturare accuratamente il comportamento del modello originario, essendo questi più semplici. Inoltre per la natura dell'algoritmo, LIME può produrre spiegazioni inconsistenti per gli stessi dati, generando confusione nell'utente. Inoltre, come SHAP, i valori numerici restituiti non sono immediatamente interpretabili. L'implementazione è risultata semplice grazie all'utilizzo della libreria ufficiale, che permette in poche righe di eseguire l'algoritmo. L'unica operazione effettuata è stata quella di convertire l'output in un formato tabulare più leggibile, in quanto di default viene restituito un formato confusionario, in cui sono presenti degli indici numerici al posto dei nomi delle feature, e questi non risultano essere sempre nell'ordine originario delle colonne del dataset.
- **CIU:** permette di calcolare due valori per ogni feature, la Contextual Importance (CI), che valuta quanto è importante la feature per quella previsione, e

la Contextual Utility (CU), che valuta quanto incide il valore numerico della feature. Questo approccio è molto più dettagliato rispetto alle prime due tecniche, poiché permette di valutare sia l'importanza della feature in quanto tale, sia l'importanza del valore di questa. Il problema nasce dal fatto che in questo caso stiamo lavorando col doppio dei valori, compresi tutti fra 0 ed 1, portando facilmente a confusione o interpretazioni sbagliate. Inoltre la definizione propria di CI e CU può risultare confusionaria di natura, diminuendo ancor di più il grado di usabilità. L'implementazione questa volta è risultata più onerosa, in quanto non è presente una libreria nativa, costringendo ad importare il pacchetto manualmente e a studiare il funzionamento del codice, che risulta essere perlopiù non commentato. Una volta compreso è stato semplice utilizzare il pacchetto per calcolare i valori delle previsioni.

- **Anchor:** permette di produrre delle regole di ancoraggio, che mostrano quali condizioni devono essere soddisfatte affinché una previsione rimanga stabile. Queste sono estremamente interpretabili di natura anche per un utente non esperto, in quanto non sono valori numerici privi di significato, ma indicano esplicitamente una lista di regole che, se soddisfatte, garantiscono che il modello darà la stessa previsione per input simili. Il problema nasce però in modelli complessi, dove le regole possono diventare troppe e troppo specifiche, diminuendone l'interpretabilità. Un ulteriore vantaggio di questa tecnica è la possibilità di produrre una percentuale di Coverage, ovvero la percentuale di istanze del dataset che rispettano un determinato set di regole. Questo ci permette anche di ottenere un grado di affidabilità per la spiegazione, cosa che gli altri algoritmi non possono fare. L'implementazione di questa tecnica è stata sicuramente la più onerosa, in quanto, così come CIU non ha una libreria, ma è stato necessario importare il codice all'interno del progetto, che risulta essere privo di commenti. Dopo aver compreso il codice è emersa una problematica interessante: poiché il dataset è stato precedentemente scalato, le regole facevano riferimento ai valori scalati piuttosto che agli input reali, rendendole inutili alla comprensione dell'utente. Per risolvere è stato quindi prodotto un algoritmo che permette la conversione delle regole alla scala originale.

Questi quattro algoritmi sono stati selezionati poiché sono tutti model-agnostic oltre che essere tra i più centrali nel campo della XAI per dati tabulari. Nonostante questo però, CIU e Anchor, essendo più recenti, sono ancora poco utilizzati in campo medico, nonostante sembrano essere i più utili e interessanti [1].

Per valutare le performance delle varie tecniche è stato preparato uno script che permette di selezionare una riga del dataset da dare in input a tutti e quattro i modelli, per poi ottenere le seguenti informazioni:

- La previsione di ogni modello.
- L'effettivo valore di Outcome (usato per individuare previsioni errate).
- I valori di SHAP, LIME e CIU per ogni previsione di ogni modello.
- Le regole Anchor e la Coverage per ogni previsione di ogni modello.

Questo script è stato quindi utilizzato per effettuare una valutazione delle tecniche attraverso un testing manuale, dal quale sono emersi molti risultati interessanti:

1. Quando i modelli predicono correttamente, i risultati di tutte le tecniche coincidono, nella maggior parte dei casi, con gli studi sul dominio e sui dati precedentemente mostrati in Figura 3.5 e Figura 3.4. In un paziente diabetico infatti le feature che solitamente hanno maggiore impatto sulla previsione sono l'elevato quantitativo di Glucosio o un elevato BMI.
2. Quando i modelli predicono erroneamente, i risultati delle spiegazioni sono apparentemente insensati, aiutando enormemente nell'individuazione di errori dei modelli. La tecnica che più risente di questo fenomeno è Anchor, difatti sugli output errati vengono spesso prodotte moltissime regole con intervalli anomali. Questo può essere visto come un vantaggio in quanto, anche grazie alla presenza della Coverage (che in questi casi risulta essere tendente allo 0), è possibile facilmente individuare errori, aumentando l'affidabilità del modello.
3. Talvolta è presente una forte inconsistenza tra le spiegazioni di tutte le tecniche e lo studio del dominio, principalmente per quanto concerne la Neural Network, facendoci intendere che, nonostante le elevate prestazioni, probabilmente il

modello non ha appreso i pattern corretti, ma ha trovato correlazioni errate tra i dati. Un esempio lampante sono alcune spiegazioni che su degli input indicavano il numero di gravidanze come la feature più importante, nonostante non vi siano studi che provino una correlazione diretta tra questo e la presenza del Diabete Mellito.

4. Il modello più interpretabile è risultato essere Anchor, in quanto le regole sono comprensibili a colpo d'occhio, mentre le altre tre tecniche spesso hanno richiesto più tempo per essere correttamente interpretate. Inoltre Anchor risulta essere anche quello che restituisce le spiegazioni più in linea con lo studio del dominio. Infine un ultimo vantaggio di Anchor è che le spiegazioni sono poste in un linguaggio più vicino a quello naturale piuttosto che in forma numerica, lasciando intendere che anche un LLM possa comprenderle meglio

Da questo studio è chiaro che Anchor risulta essere la tecnica migliore per il problema del diabete, producendo spiegazioni già naturalmente chiare oltre che aiutare nella diagnostica di errori.

Nonostante ciò in fase di testing è stato notato che i modelli non producono sempre le stesse regole, è stato quindi definito un modo per valutare e scegliere quale classificatore è in grado di produrre le spiegazioni migliori.

## 3.5 Scelta del modello per Anchor

Per selezionare il modello migliore per Anchor è stato prodotto un algoritmo multithread di valutazione che esegue i seguenti passi:

1. Divide il dataset in cinque parti uguali da 200 righe ciascuna, assegnando ogni parte ad un thread diverso.
2. Per ogni riga viene eseguita una previsione con ognuno dei quattro modelli.
3. Per ogni previsione viene eseguito Anchor e ne viene salvata la Coverage.
4. Quando tutti i thread hanno terminato l'esecuzione, vengono uniti tutti i risultati e viene calcolata la media delle Coverage per ogni modello.

5. Queste medie vengono salvate in un file csv che viene esportato.

Il risultato dell'esecuzione dell'algoritmo è riportato nella Tabella 3.4. È possibile notare subito che XGBoost e Random Forest non hanno una buona Coverage, quindi, nonostante le loro performance, sono stati esclusi. Neural Network e SVM hanno invece Coverage simili, con SVM leggermente superiore. Considerati inoltre i risultati precedenti, in cui Neural Network tendeva a mostrare pattern errati, è stato deciso di procedere con SVM nonostante l'Accuracy leggermente inferiore (Tabella 3.3), in virtù di produrre spiegazioni più precise.

	Random Forest	SVM	XGBoost	Neural Network
Coverage	0.068	0.106	0.058	0.097

**Tabella 3.4:** Coverage media dei modelli

È stato quindi deciso di utilizzare SVM con Anchor all'interno del tool.

## 3.6 Definizione della Summarization strategy

Per rendere le regole di Anchor un testo leggibile da un essere umano è necessario utilizzare un LLM, un particolare modello di intelligenza artificiale addestrato su enormi quantità di dati testuali in grado di comprendere e generare testi.

I LLM sono una delle innovazioni più importanti degli ultimi anni nell'ambito AI, e devono il loro enorme successo al lavoro di Vaswani et al., i quali hanno inventato l'architettura a trasformatore, un modello di deep learning basato sul meccanismo di self-attention, che permette di pesare l'importanza di ogni parola nella sequenza rispetto alle altre, migliorando la capacità di catturare relazioni nel linguaggio naturale [10].

Il modello scelto è ChatGPT 3.5 Turbo, il quale possiede un ottimo rapporto performance-costo per le task di produzione testuale.

Per utilizzare il modello, l'API necessita di un prompt di sistema, che descrive il comportamento da seguire alla ricezione di un messaggio. È stato quindi necessario definire un prompt per permettere al LLM di prendere in input la previsione e il risultato di Anchor e produrre in output un testo comprensibile agli utenti.

Per individuare il prompt migliore è stato utilizzato un approccio iterativo, cercando di migliorare di volta in volta la qualità delle spiegazioni. Le varie evoluzioni sono qui di seguito riportate, e per ognuna di esse viene indicato:

- Il prompt attuale. É bene notare che questo è scritto in lingua inglese, per rendere il tool internazionale e non circoscritto al territorio Italiano.
- Un output di esempio, il quale è stato sempre eseguito sulle regole Anchor "Glucose > 147.034 AND BMI > 32.772" per un paziente diabetico per il quale il modello effettua una previsione corretta.
- Una serie di osservazioni sull'output, dalle quale scaturisce l'iterazione successiva.

### **Prima Iterazione**

In questa iterazione il prompt è stato prodotto senza particolari regole, cercando di descrivere nel miglior modo possibile il comportamento desiderato.

**Prompt:** Your role is to give a diabetes diagnosis based on an integer (1 = Has Diabetes, 0 = Doesn't have diabetes), a set of rules, which are numerical intervals for certain variables (do your best effort to explain each variable meaning) divided by AND, and a coverage percentage, which the higher it is, the more reliable the rules are (A coverage of 0.20 is extremely good for example). You should always remind the user that the diagnosis is not 100% accurate and that they should consult a doctor for a professional opinion.

**Output:** Based on the rules provided, if the individual's glucose level is greater than 147.034 AND their BMI is greater than 32.772, the prediction is that they have diabetes (represented by 1). However, it is important to note that this prediction is not 100% accurate. The coverage of 0.16 indicates that these rules only apply to 16% of the cases studied, which means there could be other factors or conditions not considered in this prediction. It is highly recommended that anyone concerned about diabetes or any other health condition consult a healthcare professional (such as a doctor or a specialist) for a proper diagnosis and appropriate medical advice.

💡 **Osservazioni:** Il modello non restituisce un testo coeso, ma si limita a riportare l'output di Anchor senza aggiungere spiegazioni utili sui significati delle singole feature e senza una formattazione propria. Inoltre la percentuale di coverage indicata così com'è potrebbe causare confusione nell'utente, ed è quindi opportuno nascondere inserendo un'informazione più ad alto livello sull'affidabilità della spiegazione.

### Seconda Iterazione

In questa iterazione il prompt è stato rafforzato inserendo più regole e limitazioni, nella speranza che questo produca spiegazioni di qualità migliore.

**Prompt:** Your role is to give a diabetes diagnosis based on an integer (1 = Has Diabetes, 0 = Doesn't have diabetes), a set of rules, which are numerical intervals for certain variables divided by the string AND that explain why a patient has or hasn't diabetes, and a coverage percentage, which the higher it is, the more reliable the rules are (this variable can be between 0 and 1, with values around 0.20 being extremely good). You should always try to explain what each variable means given its name. You must never say the coverage percentage and the diagnosis integer numbers directly, only the deductions that come from them. You should give the answer in a human-readable fashion by citing and explaining the rules, basing your explanation also on medical studies. You should always remind the user that the diagnosis is not 100% reliable and that they should always consult a doctor for a professional opinion.

**Output:** Based on the rules provided, a patient is diagnosed with diabetes if their glucose level is greater than 147.034 and their BMI is greater than 32.772. Glucose levels are a key indicator for diabetes as elevated blood sugar levels can be a sign of diabetes. A fasting blood sugar level of 126 milligrams per deciliter (mg/dL) or higher indicates diabetes. A level between 100 and 125 mg/dL is considered prediabetes, which indicates a high risk of developing diabetes. BMI, or Body Mass Index, is a measurement that takes into account a person's weight and height. A higher BMI is often associated with a greater risk of developing type 2 diabetes. This is because excess body fat, particularly around the abdomen, can lead to insulin resistance, a key



factor in the development of type 2 diabetes. It's important to note that the coverage percentage provided is 0.17, which indicates that the rules have a lower reliability in accurately predicting diabetes. Therefore, it is crucial for individuals who suspect they may have diabetes to consult a healthcare professional for a proper diagnosis and appropriate management.

💡 **Osservazioni:** Questa volta il modello riesce a spiegare correttamente il significato delle varie features, ma utilizza termini presenti nel prompt e non adeguati alla spiegazione. Difatti parla di "regole", un termine tecnico di Anchor, quando dovrebbe invece parlare di osservazioni mediche. Inoltre la Coverage viene ancora mostrata, ignorando le istruzioni indicate.

### Terza Iterazione

In questa iterazione il prompt è stato meglio strutturato per favorirne la comprensibilità per il modello. Inoltre sono stati sostituiti i termini che facevano riferimento ad Anchor con termini di natura medica, così da favorire una spiegazione più fluida.

**Prompt:** You are a Diabetes Diagnoser, and your job is to give a human-readable explanation for a diabetes diagnosis. You will receive: an integer that represents the diagnosis (1 = Has Diabetes, 0 = Doesn't have diabetes); a set of observations, which are numerical intervals for certain variables divided by the string AND, that should give insight on why the patient has or hasn't diabetes; a coverage percentage, which tells how reliable the observations are.

You must: always explain what each variable means given its name; never say the coverage percentage directly and the diagnosis integer number directly, only the deductions that come from them; say how much the explanation is reliable based on the coverage percentage without saying the coverage number directly (remembering that numbers close to 0.15 should indicate a reliable explanation); give the answer in a simple human-readable fashion by citing and explaining each rule, also basing your explanation on medical knowledge; remind the user that the diagnosis is not 100% reliable and that they should always consult a doctor for a professional opinion.

**Output:** Based on the observations provided, it is likely that the patient has diabetes.

Glucose levels higher than 147.034 indicate that there is an excess of sugar in the blood, which is a common indicator of diabetes. Additionally, a Body Mass Index (BMI) greater than 32.772 is associated with an increased risk of developing diabetes, as obesity is a significant risk factor for the disease.

The reliability of this explanation is somewhat reduced given the coverage percentage of 0.16, which indicates some uncertainty in the observations. Therefore, it is important to remember that this diagnosis is not 100% reliable, and it is crucial for the patient to consult a healthcare professional for a comprehensive evaluation and confirmation of the diagnosis.

💡 **Osservazioni:** Il modello produce ora testi meglio strutturati, ma la formattazione della risposta può essere ulteriormente migliorata. Inoltre anche questa volta non riesce a comprendere come gestire correttamente le istruzioni sulla Coverage, indicando non solo la percentuale esplicitamente, ma ignorando anche le istruzioni indicate (viene detto che una Coverage di 0.16 non è buona mentre nel prompt viene detto che superiore al 0.15 è ottima).

### Quarta Iterazione

In questa iterazione è stato deciso di implementare il framework RISEN, un approccio di prompt engineering che divide il prompt in cinque parti ben definite, così da garantire spiegazioni migliori e più consistenti [11]. Il prompt viene diviso in:

- **Role:** il ruolo del modello, indica come deve comportarsi.
- **Instruction:** le istruzioni che il modello deve seguire.
- **Steps:** descrive nel dettaglio tutte le fasi delle istruzioni prima indicate.
- **End Goal:** il risultato che ci aspettiamo da parte del modello.
- **Narrowing:** eventuali vincoli che il modello deve seguire.

**Prompt:** Role: Diabetes Diagnoser

Input: You will receive a string formatted like this: 'Diabetes: x, Observations: y, Reliability: z' where x is a binary value (0 or 1) that represents the diagnosis (0 = Doesn't have diabetes, 1 = Has diabetes), y is a set of observations, which are

numerical intervals for certain features divided by the string AND, that should give insight on why the patient has or hasn't diabetes, and  $z$  is a reliability percentage, which tells how reliable the observations are.

**Steps:** Start by saying if the patient has or hasn't diabetes, then explain why citing and explaining each observation. Make sure to include a short explanation of what each feature represents and add medical facts to sustain the observation. Then you should say how much the explanation is reliable based on the reliability percentage. Remember that a reliability around 0.15 is a pretty reliable explanation. Finally, remind the user that the diagnosis can't be 100% reliable and that they should always consult a doctor for a professional opinion.

**Expectation:** A human-readable explanation for a diabetes diagnosis.

**Narrowing:** Be clear and complete in your explanation, but avoid using complex medical terms. You will never tell the user the reliability percentage explicitly.

**Output:**Based on the provided observations, it seems that the patient has diabetes. Let's break down the observations to understand why:

Glucose > 147.034: This observation indicates that the patient's blood glucose level is above 147.034. Elevated glucose levels are a common indicator of diabetes. In a healthy individual, the body regulates blood glucose levels effectively. However, in diabetes, the body either does not produce enough insulin (Type 1) or is unable to use insulin effectively (Type 2), leading to high glucose levels in the blood.

BMI > 32.772: BMI (Body Mass Index) is a measure of body fat based on height and weight. A BMI over 32.772 indicates obesity, which is a significant risk factor for type 2 diabetes. Excess body fat, especially around the abdomen, can lead to insulin resistance, where cells don't respond effectively to insulin, increasing the risk of diabetes.

Considering these observations, it appears likely that the patient has diabetes. The reliability of this explanation is moderately high at 0.16, indicating that the observations provide a somewhat reliable basis for the diagnosis.

It's important to note that while these observations point towards diabetes, a definitive diagnosis can only be made by a healthcare professional. Therefore, it's crucial for the patient to consult a doctor for further evaluation and management.

💡 **Osservazioni:** La spiegazione è molto più leggibile e chiara, segue tutte le istruzioni tranne che per la gestione del Coverage (ora definito "Reliability" per renderlo più chiaro all'utente). È chiaro che il modello non riesce a ragionare con i numeri, quindi è opportuno cambiare approccio.

### Quinta Iterazione

In questa iterazione è stato deciso di modificare la definizione di Reliability nel codice. Ora la percentuale di Coverage viene sostituita da una stringa che può essere "High" quando la coverage è maggiore di 0.15, "Medium" quando è compresa tra 0.7 e 0.15 e "Low" quando è inferiore a 0.7. Questo dovrebbe permettere al modello di seguire correttamente tutte le istruzioni indicate, evitando comportamenti non voluti. Inoltre il testo ora presenta le metriche di valutazione del modello, così da validare ulteriormente l'affidabilità delle previsioni.

**Prompt:** Role: Diabetes Diagnoser

Input: You will receive a string formatted like this: 'Diabetes: x, Observations: y, Reliability: z' where x is a binary value (0 or 1) that represents the diagnosis (0 = Doesn't have diabetes, 1 = Has diabetes), y is a set of observations, which are numerical intervals for certain features divided by the string 'AND', that should give insight on why the patient has or hasn't diabetes, and z is how reliable the explanation is.

Steps: Start by saying if the patient has or hasn't diabetes. Then explain why, citing and explaining each observation in a numbered list. Make sure to include a short explanation of what each feature represents and add medical facts to sustain the observation. Then you should say how much the explanation is reliable. Finally, remind the user that the diagnosis can't be 100% reliable and that they should always consult a doctor for a professional opinion and that the accuracy of the model used is 82%, the precision is 79% and the recall is 87%, explaining what this means.

Expectation: A human-readable explanation for a diabetes diagnosis.

Narrowing: Be clear and complete in your explanation, but avoid using complex medical terms.

**Output:** Based on the provided information, the patient has diabetes.

Here is the breakdown of the observations:

Glucose > 147.034: Glucose levels refer to the amount of sugar in the blood. A high glucose level, especially above 147.034, indicates hyperglycemia, a common indicator of diabetes. When the body doesn't produce enough insulin or becomes resistant to insulin, glucose can't enter cells and builds up in the bloodstream, leading to high blood sugar levels.

BMI > 32.772: BMI (Body Mass Index) is a measure of body fat based on height and weight. A BMI over 32.772 indicates obesity, which is a significant risk factor for type 2 diabetes. Excess body fat, especially around the abdomen, can lead to insulin resistance, where cells don't respond effectively to insulin, increasing the risk of diabetes.

The reliability of this explanation is high, meaning that the combination of high glucose levels and a high BMI strongly suggests the presence of diabetes in the patient. It's essential to remember that while this explanation is based on statistical analysis, no model can provide a 100% accurate diagnosis. For proper medical advice and treatment, it is crucial to consult a healthcare professional. The model used has an accuracy of 82%, which means it is correct 82% of the time. The precision is 79%, indicating that when it predicts diabetes, it is 79% accurate. The recall is 87%, showing that it correctly identifies 87% of actual diabetes cases.

**💡 Osservazioni:** Dopo aver cambiato l'approccio per la Reliability, la spiegazione è ora perfettamente leggibile, comprensibile e informativa, seguendo in maniera consistente il pattern indicato nel prompt e aggiungendo informazioni mediche inerenti e coerenti col problema. Il prompt si può quindi considerare adeguato.

Una volta terminata la pipeline di Summarization, questa è stata inserita all'interno di una semplice webapp, DIA, il principale risultato di questa ricerca, la quale verrà descritta nel dettaglio al successivo capitolo.

Il tool implementato rispetta alla perfezione gli obiettivi prefissati, essendo sia molto semplice da usare, sia in grado di fornire spiegazioni in linguaggio naturale che risultano essere perfettamente comprensibili anche per utenti non esperti. Inoltre grazie a queste spiegazioni, anche l'individuazione di problemi nelle previsioni risulta essere più semplice e immediata.

É quindi ora possibile rispondere alle domande di ricerca:

**Q RQ<sub>1</sub>.** *É possibile utilizzare i LLM per rendere le spiegazioni degli algoritmi di XAI più comprensibili?*

**👉 Answer to RQ<sub>1</sub>.** Sì, i LLM sono in grado di trasformare con successo le spiegazioni degli algoritmi di XAI in testi in linguaggio naturale, che risultano essere molto più comprensibili e chiari per gli esseri umani rispetto a una sfilza di dati numerici senza apparente significato.

Questa scoperta apre le strade all'impiego dei LLM in ambito XAI, con l'obiettivo di aiutare gli utenti a comprendere le spiegazioni nascondendo la complessità intrinseca degli algoritmi di spiegabilità, mostrando invece un semplice testo in linguaggio naturale, il quale può perfino essere arricchito di ulteriori informazioni se necessario, come dettagli di natura medica o avvertimenti al paziente.

**Q RQ<sub>2</sub>.** *É possibile utilizzare spiegazioni in linguaggio naturale per aumentare la fiducia degli utenti nei confronti delle previsioni di un modello?*

**👉 Answer to RQ<sub>2</sub>.** Sì, attraverso spiegazioni in linguaggio naturale è possibile aumentare la fiducia dell'utenza nei confronti del modello, permettendo facilmente a chiunque di comprendere il processo decisionale dietro un determinato output.

Questa scoperta implica che attraverso l'impiego di processi di Summarization è possibile abbattere la barriera della comprensibilità dei complessi algoritmi black box e delle tecniche di XAI, portando a una maggiore fiducia da parte degli utenti, poiché questi sono finalmente in grado di comprendere correttamente come un modello arriva ad una risposta.

**Q RQ<sub>3</sub>.** *É possibile utilizzare spiegazioni in linguaggio naturale per aumentare l'affidabilità delle previsioni di un modello?*

**👉 Answer to RQ<sub>3</sub>.** Sì, attraverso spiegazioni in linguaggio naturale è possibile individuare più facilmente output sbagliati, che solitamente restituiscono spiegazioni insensate o estremamente complesse, permettendo quindi di riconoscere previsioni potenzialmente errate.

Questa scoperta conferma che l'impiego del linguaggio naturale per produrre spiegazioni non solo migliora la comprensibilità e la fiducia del modello, ma permette anche di individuare più facilmente errori, migliorando quindi l'affidabilità di questo e producendo di conseguenza prestazioni migliori in ambienti di produzione.

---

### **DIA: un tool per spiegare l'output di modelli black box**

---

Tutto il codice prodotto, insieme ai grafici e al tool, è pubblicamente disponibile al seguente url: <https://github.com/MonTheDog/XAI-for-Diabetes>.

Il principale risultato della ricerca effettuata è, come anticipato, il tool DIA, una semplice webapp che ha come design goal principale l'usabilità e la semplicità. Per l'implementazione è stato utilizzato il framework streamlit, che permette di sviluppare velocemente semplici interfacce eleganti in Python, e l'API di OpenAI, che permette di effettuare richieste a uno dei numerosi LLM dell'azienda direttamente dall'applicativo.

L'architettura è monolitica, in quanto, essendo Python un linguaggio di scripting più che di web development, tutto il codice, sia di front end che di back end è contenuto in un unico script, il quale contiene anche la connessione al LLM e il prompt finale utilizzato.

Lo sviluppo è stato effettuato con l'intenzione di tenere l'interfaccia quanto più minimale e pulita possibile, così da non confondere l'utente durante l'utilizzo. Il deployment non è stato effettuato, ma attraverso il framework utilizzato è possibile caricare l'applicativo direttamente dalla repository al cloud, rendendone potenzialmente possibile l'accesso a chiunque sia provvisto di una rete internet.

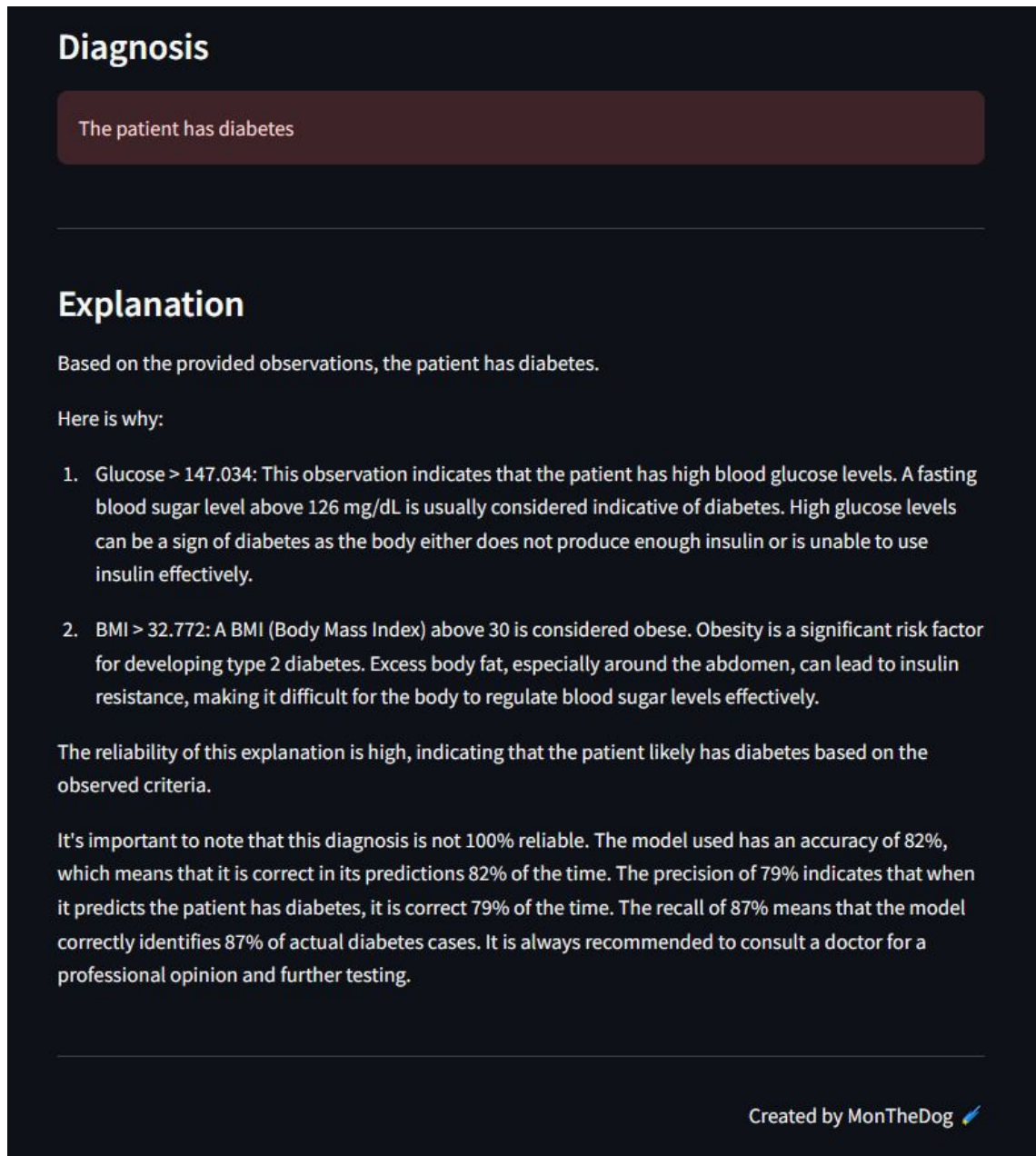


All'avvio l'utente può inserire le proprie informazioni e i risultati di analisi per la diagnosi del diabete per ricevere l'output del modello. Nella Figura 4.1 è possibile visualizzare l'interfaccia iniziale, composta da una serie di input box numeriche, ognuna delle quali indica una delle feature predittive. È inoltre possibile ottenere informazioni su cosa rappresenta ogni feature ponendo la freccetta sui punti interrogativi a lato di ogni nome.

Feature	Value
Pregnancies	6
Glucose	148
BloodPressure	72
SkinThickness	35
Insulin	125
BMI	33,6
DiabetesPedigreeFunc	0,627
Age	50

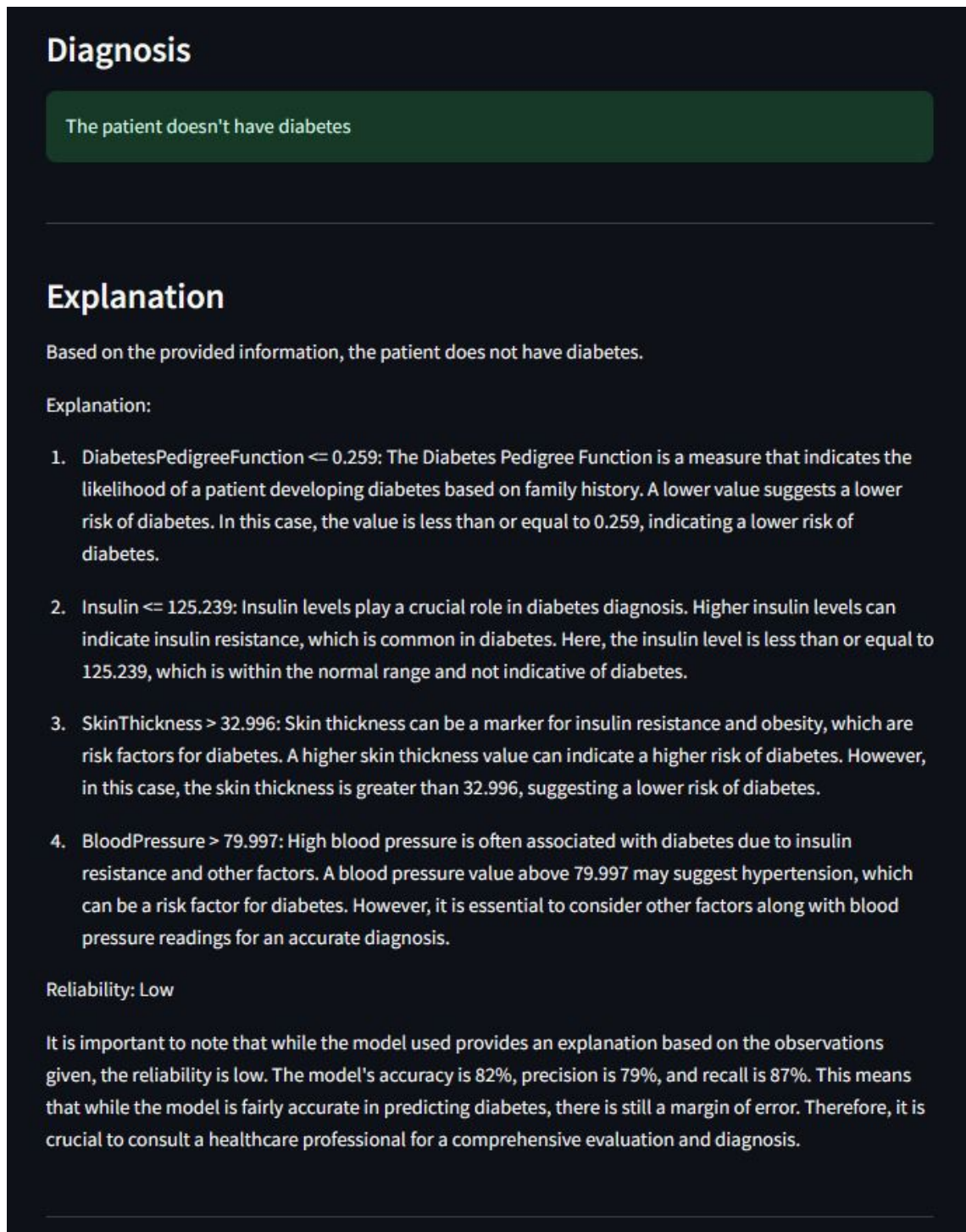
**Figura 4.1:** Schermata iniziale di DIA, in cui l'utente può inserire i propri dati.

Dopo aver inserito le informazioni, l'utente deve semplicemente premere il tasto "Submit" per ricevere il risultato del modello e la spiegazione. Nella Figura 4.2 è possibile vedere un output tipo di questo flusso di interazione: in cima si ha la diagnosi del modello, subito seguita dalla spiegazione prodotta utilizzando il prompt indicato alla fine del capitolo precedente. Questo viene formattato a posteriori da streamlit, per facilitarne la leggibilità, dividendolo in paragrafi e inserendo un elenco numerato per ogni regola di Anchor. Se l'utente vuole interagire nuovamente col modello può farlo cambiando i valori nelle input box precedenti, senza necessità di ricaricare la pagina.



**Figura 4.2:** Schermata con la previsione del modello e la spiegazione in forma testuale.

Quando queste regole risultano essere numerose e con intervalli anomali, questo è probabilmente indice di una previsione sbagliata. Un esempio si ha nella Figura 4.3 in cui il modello ha restituito una spiegazione insensata, lasciando intendere che l'output non è affidabile. A supporto di questo ragionamento abbiamo il sistema della Reliability, che in questo caso è bassa, rafforzando ulteriormente l'ipotesi dell'errore del modello. Infatti, dopo un controllo nel dataset, questi dati corrispondono ad un paziente diabetico, mentre il modello lo ha diagnosticato come sano.



**Figura 4.3:** Schermata con una previsione errata e spiegazioni insensate.

La webapp non contiene ulteriori funzionalità, risultando essere quindi estremamente semplice e diretta, così da permetterne l'utilizzo a chiunque, anche utenti non esperti. Le spiegazioni risultano essere molto chiare e informative, nascondendo completamente l'enorme complessità delle tecniche sottostanti. Ovviamente essendo il modello non perfetto, all'utente verrà sempre suggerito di parlare con un medico a priori dalla previsione.

#### 5.1 Impatto

Tutte le domande di ricerca sono state risposte in maniera positiva, dimostrando che l'utilizzo dei LLM rappresenta una nuova frontiera nell'ambito della XAI, producendo sistemi maggiormente interpretabili e di conseguenza affidabili.

Il tool implementato permette di chiudere finalmente il gap che esisteva tra le tecniche di explainability e la comprensione di queste ultime da parte degli utenti. Inoltre la scelta di calarsi nel contesto medico permette di validare l'approccio anche in ambiti critici, in cui la spiegabilità è un requisito fondamentale.

Le tecniche e le procedure utilizzate possono infine essere facilmente generalizzate per tutti i problemi di classificazione con dati tabulari, rendendo i risultati di questa ricerca universalmente utili per la diffusione e l'integrazione di sistemi intelligenti su larga scala, migliorandone l'affidabilità e la fiducia degli utenti in questi.

#### 5.2 Future Work

L'IA è in continuo sviluppo, e tutt'ora vengono prodotte continuamente innovazioni che rivoluzionano velocemente lo Stato dell'Arte. Tra queste abbiamo ad

esempio nuovi LLM sempre più potenti ed economici, come ChatGPT 4o-mini, il quale è stato rilasciato dopo la conclusione del lavoro [12]. Attraverso l'impiego di nuovi LLM i testi prodotti potrebbero essere di qualità maggiore, migliorando ulteriormente il grado di comprensibilità delle spiegazioni.

Inoltre c'è da considerare che la disciplina del prompt engineering è ancora nelle sue prime fasi, e per questo motivo vengono continuamente prodotte nuove scoperte affascinanti. Tra queste vale la pena citare l'EmotionPrompt, una nuova tecnica che utilizza stimoli emotivi all'interno dei prompt per produrre output migliori sia dal punto di vista delle prestazioni (ottenendo un miglioramento fino al 10%), sia dal punto di vista dell'empatia nei confronti dell'utente, risultando in interazioni più "umane" [13]. Nell'ambito della diagnosi medica, questo rappresenta sicuramente un importante vantaggio, poiché trasmettere calma o sicurezza ad un paziente afflitto da una malattia, non può che migliorare la qualità delle interazioni col modello.

---

## Bibliografia

---

- [1] S. S. Band, A. Yarahmadi, C.-C. Hsu, M. Biyari, M. Sookhak, R. Ameri, I. Dehzangi, A. T. Chronopoulos, and H.-W. Liang, "Application of explainable artificial intelligence in medical health: A systematic review of interpretability methods," *Informatics in Medicine Unlocked*, vol. 40, p. 101286, 2023. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2352914823001302> (Citato alle pagine 2, 6, 10, 13, 14, 15 e 28)
- [2] K.-H. Yu, A. L. Beam, and I. S. Kohane, "Artificial intelligence in healthcare," *Nature Biomedical Engineering*, vol. 2, pp. 719 – 731, 2018. [Online]. Available: <https://api.semanticscholar.org/CorpusID:80789013> (Citato alle pagine 2, 5, 6 e 7)
- [3] J. Chaki, S. Thillai Ganesh, S. Cidham, and S. Ananda Theertan, "Machine learning and artificial intelligence based diabetes mellitus detection and self-management: A systematic review," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 6, Part B, pp. 3204–3225, 2022. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1319157820304134> (Citato alle pagine 2, 7 e 17)
- [4] T.-C. T. Chen, H.-C. Wu, and M.-C. Chiu, "A deep neural network with modified random forest incremental interpretation approach for diagnosing diabetes in smart healthcare," *Applied Soft Computing*, vol. 152, p. 111183, 2024.

- [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1568494623012012> (Citato alle pagine 2 e 14)
- [5] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information Fusion*, vol. 58, pp. 82–115, 2020. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1566253519308103> (Citato alle pagine 2, 3, 8, 9 e 15)
- [6] S. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," 2017. [Online]. Available: <https://arxiv.org/abs/1705.07874> (Citato a pagina 10)
- [7] M. T. Ribeiro, S. Singh, and C. Guestrin, "'why should i trust you?': Explaining the predictions of any classifier," 2016. [Online]. Available: <https://arxiv.org/abs/1602.04938> (Citato a pagina 11)
- [8] K. Främling, "Explainable ai without interpretable model," 2020. [Online]. Available: <https://arxiv.org/abs/2009.13996> (Citato a pagina 12)
- [9] M. T. Ribeiro, S. Singh, and C. Guestrin, "Anchors: High-precision model-agnostic explanations," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, Apr. 2018. [Online]. Available: <https://ojs.aaai.org/index.php/AAI/article/view/11491> (Citato a pagina 13)
- [10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2023. [Online]. Available: <https://arxiv.org/abs/1706.03762> (Citato a pagina 30)
- [11] S. Manghani, "Prompt engineering, explained," <https://medium.com/electronic-life/prompt-engineering-explained-3b83ba347722>, accessed: 2024-09-08. (Citato a pagina 34)
- [12] OpenAI, "Gpt-4o mini: advancing cost-efficient intelligence," <https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/>, accessed: 2024-09-08. (Citato a pagina 46)



- [13] C. Li, J. Wang, Y. Zhang, K. Zhu, W. Hou, J. Lian, F. Luo, Q. Yang, and X. Xie, "Large language models understand and can be enhanced by emotional stimuli," 2023. [Online]. Available: <https://arxiv.org/abs/2307.11760> (Citato a pagina 46)

---

## Ringraziamenti

---

Arrivato alla conclusione di questo percorso, mi rendo conto che in questi anni ho avuto il piacere e la fortuna di essere accompagnato da tantissime persone fantastiche, e ora che bisogna tirare le somme, è giusto che io ringrazi chi ha colorato le mie giornate. Così come tutti voi mi avete donato qualcosa, io voglio chiudere con delle promesse, nella speranza di ripagare tutto ciò che avete fatto per me.

Ringrazio innanzitutto il SeSa Lab, in particolare il Prof. Fabio Palomba, per avermi indicato la strada che voglio percorrere nella vita grazie ai suoi insegnamenti, e il dott. Antonio Della Porta, per avermi seguito e aiutato in questo lavoro, permettendomi di produrre qualcosa di cui posso andare molto fiero. Vi prometto che farò tesoro di tutto ciò che ho imparato grazie a voi.

Ringrazio mia madre, grazie a te sono diventato la persona che vedo allo specchio, una persona con la testa sulle spalle, che guarda sempre al futuro e che cerca sempre di combinare qualcosa di buono. Ti prometto che me la caverò davanti a tutto, come hai sempre fatto tu.

Ringrazio mio padre, grazie a te ho imparato il valore del lavoro e dell'impegno, capendo cosa significa essere un uomo di famiglia. Ti prometto che proteggerò la mia famiglia come tu hai sempre protetto la nostra.

Ringrazio mia sorella, grazie a te ho sempre avuto un punto di riferimento, un esempio da seguire e una persona da ammirare. Ti prometto che continuerò a dare il massimo, così che tu possa continuare ad esser fiera di me.

---

Ringrazio mio nonno, grazie a te ho capito che nella vita bisogna esser sempre sé stessi, senza aver paura del giudizio altrui. Ti prometto che conserverò per sempre il tuo ricordo e che alla fine, anche io, all'alba vincerò.

Ringrazio tutti i miei parenti, grazie per aver sempre creduto in me e per aver gioito dei miei traguardi durante il mio percorso. Vi prometto che porterò sempre il nostro cognome a testa alta.

Ringrazio la mia Lulù, grazie a te ho compreso il valore dell'amore e ho capito per la prima volta nella vita cosa significhi essere amati incondizionatamente. Ti prometto che continuerò ad amarti fino alla fine, e che riusciremo a realizzare i nostri sogni. Insieme.

Ringrazio Nicolò, grazie a te ho trovato un fratello in grado di rialzarmi anche quando il peso da portare era troppo. Ti prometto che raggiungeremo la cima dandoci forza l'un l'altro, come abbiamo sempre fatto.

Ringrazio tutti gli amici che hanno condiviso con me innumerevoli ore di lezione: Sara, Giacomo, Alessandro, Marica, Sasha e Tod. Grazie per tutti i bei momenti passati insieme, poiché è grazie a questi che ricorderò col sorriso questi anni. Vi prometto che ci sarò sempre per voi e che mi impegnerò per far sì che le nostre strade rimangano incrociate.

Ringrazio tutti gli amici della mia amata Avellino, grazie a voi ho capito che l'amicizia vera non sfuma col tempo e che il calore delle mie radici rimarrà sempre qui ad aspettarmi, ovunque mi porti la vita. Vi prometto di non dimenticarvi mai e che farò di tutto pur di tornare a respirare la nostra aria insieme ogni tanto.

Ringrazio la Speranza, grazie a te ho capito di non dover mai mollare e che, continuando a seguire la mia strada, tutto verrà da sé. Ti prometto che non smetterò mai di guardare le Stelle, e che un giorno riuscirò finalmente a raggiungerle.

Ringrazio infine chiunque sia arrivato fin qui, grazie per aver letto il mio lavoro e spero che vi sia stato utile o che, perlomeno, sia stata una piacevole lettura. Vi prometto che non sarà l'ultima volta che sentirete parlare di me.

*Questa tesi ha contribuito a piantare un albero in Madagascar tramite il progetto Treedom.*

<https://www.treedom.net/it/user/sesalab/event/sesa-random-forest>