UNIVERSITÀ DEGLI STUDI DI SALERNO

Dipartimento di Informatica

Corso di Laurea Triennale in Informatica

TESI DI LAUREA

Explainable Artificial Intelligence: Un'Analisi Empirica sull'Explainability dei Framework SHAP e LIME

RELATORE

Prof. **Fabio Palomba**

Università degli studi di Salerno

CANDIDATO

Rebecca Di Matteo

Matricola: 0512106379

Anno Accademico 2021-2022

“l’intelligenza artificiale è come il cervello: non si può tagliare la testa e vedere come funziona“

-Andy Rubin

Sommario

L'intelligenza artificiale permette di costruire modelli che apprendono in modo indipendente e trovano soluzioni a problemi complessi ciò la rende una delle aree di ricerca più importanti nonostante la complessità. Con la crescita della complessità assume un ruolo maggiore anche la trasparenza per rendere il processo decisionale e i risultati ottenuti dall'intelligenza artificiale il più comprensibile possibile, da ciò nasce l'Explainable Artificial Intelligence. A tal proposito, esistono delle soluzioni automatiche che permettono l'interpretazione delle predizioni dei modelli di machine learning, generalmente implementate per mezzo di librerie software.

Lo scopo di questa tesi è di valutare due librerie Python LIME e SHAP in termini di spiegabilità dei modelli dati in input, andando a costruire un modello di classificazione per la predizione del consumo di alcol degli studenti che sarà fornito in input alle due librerie.

L'output ottenuto dalle due librerie sarà inserito all'interno di un questionario che sarà somministrato a studenti andando a confrontato i diversi risultati al fine di valuatare il grado di spiegabilità delle due librerie in termini di verità, chiarezza e leggibilità.

I risultati ottenuti hanno prodotto un risultato migliore per SHAP nei confronti di LIME.

Indice

Elenco delle figure

Elenco delle tabelle

CAPITOLO 1

Introduzione

1.1 Contesto applicativo

Il campo dell'intelligenza artificiale è oggigiorno sempre più complicato e meno comprensibile. L'intelligenza artificiale permette ai programmi di apprendere in modo indipendente e trovare soluzioni a problemi complessi ciò la rende una delle aree di ricerca più importanti nonostante la complessità. Con la crescita della complessità assume un ruolo maggiore anche la trasparenza per rendere il processo decisionale e i risultati ottenuti dall'intelligenza artificiale il più comprensibile possibile da ciò nasce l'Explainable Artificial Intelligence che significa letteralmente INTELLIGENZA ARTIFICIALE SPIEGABILE. Gli utenti vogliono e devono capire come funziona l'IA (Intelligenza Artificiale) di un programma e come devono essere valutati i risultati ottenuti, altrimenti quest'ultimi non potranno mai guadagnare la fiducia dell'utente. La trasparenza creata dall'Explainable Artificial Intelligence è quindi di enorme importanza per l'accettazione dell'intelligenza artificiale [1] .
Il termine è un neologismo che viene utilizzato nella ricerca e nelle argomentazioni sull'apprendimento automatico dal 2004. Ad oggi non esiste una definizione generalmente riconosciuta dell'Explainable AI.
Il programma XAI della DARPA (Defense Advanced Research Projects Agency) definisce i requisiti degli obiettivi dell'intelligenza artificiale spiegabile:

- deve generare modelli spiegabili senza dover fare a meno delle elevate prestazioni di apprendimento.

- deve mettere gli utenti futuri nelle condizioni di poter comprendere la generazione emergente di partner artificialmente intelligenti, di fidarsi in misura adeguata e di relazionarsi e lavorare con loro in modo efficiente.

1.2 Obiettivi e Metodologia

In generale il lavoro svolto segue una metodologia il cui scopo è quello di analizzare in maniera oggettiva e approfondita le tecniche di Explainable Artificial Intelligence nello specifico saranno analizzate due librerie LIME e SHAP mediante lo sviluppo di :

- domande di ricerca
- un modello di classificazione per la predizione del consumo di alcol degli studenti.
- questionario per il consumo di alcol degli studenti e l'Explainability di LIME e SHAP.
- confronto tra l'analisi dell'output ottenuto da LIME e SHAP e i risultati dei due questionari.

1.3 Struttura della tesi

All'interno di questa sezione verrà esposta la struttura della tesi la quale è organizzata in sei capitoli ognuno dei quali a sua volta è costituito dalle rispettive sotto sezioni.

- Il capitolo 2 descrive in maniera approfondita Explainable Artificial Intelligence, alcuni algoritmi librerie e Tool.
- Il capitolo 3 descrive le diverse fasi di costruzione del modello di predizione per il consumo di alcol degli studenti nello specifico Data Preparation, Data Modeling.
- Il capitolo 4 descrive l'obiettivo e il contesto dello studio, la metodologia con le successive fasi di Data Extraction, Data Collection e Data Analysis.
- Il capitolo 5 si focalizza sull'osservazione e l'analisi dei risultati ottenuti dallo studio e sulla descrizione di quest'ultimi.
- Il capitolo 6 si focalizza sulle conclusioni e sviluppi futuri del lavoro svolto.

CAPITOLO 2

Background

All'interno di questo capitolo viene descritta l'importanza dell'Explainable Artificial Intelligence nello specifico cosa rappresenta, una panoramica sugli algoritmi di Explainable Artificial Intelligence e le librerie e tool utilizzati.

> *"Ciò che è fondamentale è rendere qualsiasi cosa sull'Intelligenza artificiale spiegabile, equa, sicura e con linguaggio, il che significa che chiunque potrebbe vedere molto semplicemente come si è sviluppata qualsiasi applicazione dell'Intelligenza artificiale e perché"[2].*
> **-Ginni Rometty**

2.1 Explainable Artificial Intelligence

Explainable Artificial Intelligence è un insieme di metodi e processi che consentono agli utenti di comprendere e considerare attendibili i risultati e l'output creati dagli algoritmi di machine learning. L'intelligenza artificiale spiegabile viene utilizzata per descrivere un modello di intelligenza artificiale, il relativo impatto previsto ed i potenziali errori.
Permette di caratterizzare la precisione, la correttezza, la trasparenza e i risultati del modello nel processo decisionale, inoltre è fondamentale per un'organizzazione nello sviluppo della fiducia e della sicurezza nei confronti dell'utente quando vengono messi in produzione i modelli permettendo di adottare uno sviluppo responsabile [3] . Con il miglioramento dell'intelligenza artificiale, gli esseri umani vengono sfidati a comprendere e ripercorrere il

modo in cui l'algoritmo è arrivato ad un risultato.
L'intero processo di calcolo si trasforma in quello che viene comunemente definito una BLACK-BOX ovvero una SCATOLA NERA che è impossibile interpretare. Questi modelli di scatola nera vengono creati direttamente dai dati e gli ingegneri o i data scientist che creano l'algoritmo non possono comprendere o dimostrare cosa avviene precisamente all'interno di tali scatole nere o come l'algoritmo d'intelligenza artificiale è arrivato ad un risultato specifico.
Un noto problema nello sviluppo dei modelli di intelligenza artificiale è il PREGIUDIZIO. Le prestazioni dei modelli possono migliorare o peggiorare perché i dati prodotti sono differenti dai dati iniziali utilizzati per l'addestramento del modello, dunque si devono monitorare continuamente i modelli per rendere promovibile Explainable Artificial Intelligence, quest'ultima porta ad avere una fiducia maggiore negli utenti finali, la controllabilità del modello e l'uso produttivo dell'intelligenza artificiale. La fiducia dell'uomo nell'utilizzare un modello d'intelligenza artificiale si basa sulla comprensione di esso [4] .

2.1.1 Esempio Explainable Artificial Intelligence

Un esempio di spiegabilità della previsione di un modello è rappresentato da un modello di previsione dei prezzi delle abitazioni di Boston quest'ultimo è stato addestrato utilizzando l'algoritmo XGBOOST. La spiegabilità si ottiene utilizzando la libreria SHAP.
La spiegabilità mostrata di seguito rappresenta le caratteristiche che contribuiscono ciascuna a spingere l'output del modello dal valore di base ovvero l'output medio del modello sul set di dati di training che abbiamo passato all'output del modello.
Le caratteristiche che contribuiscono maggiormente alla predizione sono mostrate in rosso, quelle che contribuiscono in modo minore sono in blu [5] .

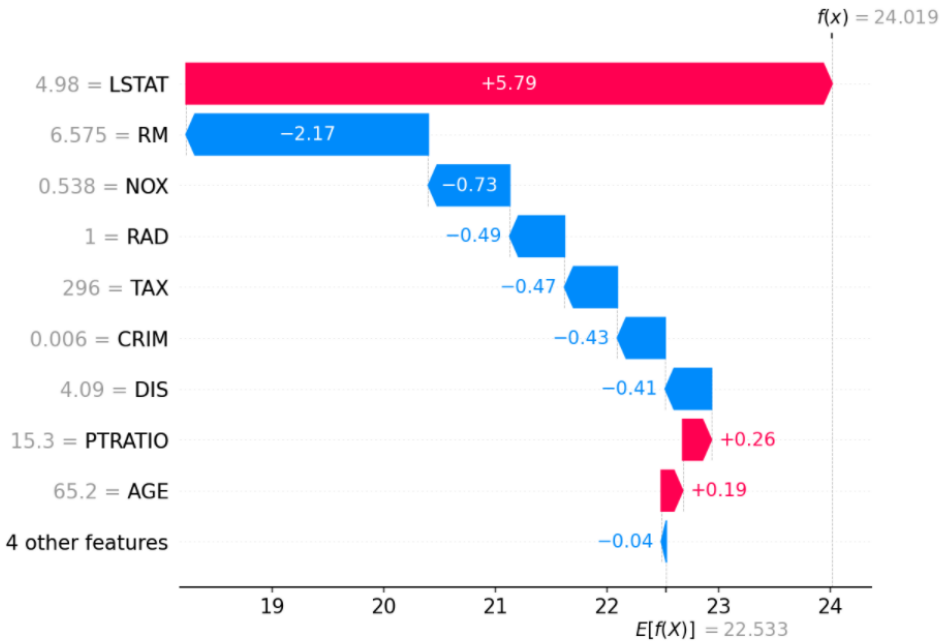


Figura 2.1: Previsione SHAP prezzi case Boston

2.1.2 Alcuni casi d'uso Explainable Artificial Intelligence

- **Assistenza sanitaria:** quando si diagnosticano i pazienti con malattia, l'IA spiegabile può spiegare la loro diagnosi. Può aiutare i medici a spiegare la loro diagnosi ai pazienti e spiegare come un piano di trattamento aiuterà. Ciò contribuirà a creare una maggiore fiducia tra i pazienti e i loro medici, mitigando al contempo eventuali problemi etici. Uno degli esempi in cui le previsioni dell'IA possono spiegare le loro decisioni potrebbe comportare la diagnosi di pazienti con polmonite. Un altro esempio in cui l'IA spiegabile può essere estremamente utile è nell'assistenza sanitaria con dati di imaging medico per la diagnosi del cancro.

- **Veicoli autonomi:** l'IA spiegabile sta diventando sempre più importante nel settore automobilistico a causa di eventi altamente pubblicizzati che coinvolgono incidenti causati da veicoli autonomi. Ciò ha posto l'accento sulle tecniche di spiegabilità per gli algoritmi di intelligenza artificiale, specialmente quando si tratta di casi d'uso che coinvolgono decisioni critiche per la sicurezza [6] . L'intelligenza artificiale spiegabile può essere utilizzata per i veicoli autonomi in cui la spiegabilità fornisce una maggiore consapevolezza situazionale in incidenti o situazioni impreviste, che potrebbe portare a un funzionamento più responsabile della tecnologia.

- **Rilevamento delle frodi:** l'IA spiegabile è importante per il rilevamento delle frodi nei servizi finanziari. Questo può essere utilizzato per spiegare perché una transazione è stata contrassegnata come sospetta o legittima, il che aiuta a mitigare potenziali sfide etiche associate a pregiudizi ingiusti e problemi di discriminazione quando si tratta di identificare transazioni fraudolente.

2.2 Differenze Explainable Artificial Intelligence e Artificial Intelligence

La differenza fondamentale tra XAI e IA è che Explainable Artificial Intelligence implementa tecniche e metodi specifici per garantire che ogni decisione presa durante il processo di machine learning possa essere tracciata e spiegata. L'Artificial Intelligence, invece, spesso arriva a un risultato utilizzando un algoritmo di marchine learning, ma gli architetti dei sistemi AI non comprendono completamente in che modo l'algoritmo ha raggiunto tale risultato.
Questo rende difficile il controllo della precisione e determina una perdita di controllo, responsabilità e controllabilità [7] . Le tecniche di spiegabilità utilizzate nell'IA spiegabile sono fortemente influenzate dal modo in cui gli esseri umani fanno inferenze e formano conclusioni, il che consente loro di essere replicati all'interno di un sistema di intelligenza artificiale spiegabile.
L'intelligenza artificiale spiegabile può aiutare a spiegare il ragionamento alla base della decisione di un modello di apprendimento automatico addestrato utilizzando un qualsiasi algoritmi come ad esempio la regressione lineare, Random Forest, gli alberi decisionali. Questa è una delle tecniche di spiegabilità più comuni utilizzate nella pratica e ci sono molti strumenti che forniscono questa funzionalità.

2.3 Overview sugli algoritmi di Explainable Artificial Intelligence

Esistono diverse interpretazioni del concetto di Explainable Artificial Intelligence ben rappresentate dai principi su cui stanno avanzando le ricerche su questo settore che sono:

- Understandability
- Comprehensibility

- Interpretability
- Explainability
- Transparency

L'interesse dei ricercatori si sofferma su Interpretability e Explainability interrogandosi su cosa rende davvero spiegabile un sistema di intelligenza artificiale. Tale percorso può essere affrontato agendo su due livelli e partendo dall'interpretabilità intesa come la possibilità di mettere in relazione causale i dati in ingresso con quelli in uscita spiegando perché un modello ha compiuto una certa scelta o ha fornito una determinata previsione. Nel secondo step si passa dal *what* al *why* mirando all'Explainability ovvero alla possibilità di fornire una spiegazione accessibile a tutti di come un sistema di AI sia arrivato ad una determinata scelta o previsione. Giunti a questo livello di HUMAN STYLE INTERPRETATIONS si può distinguere tra comprensibilità GLOBALE, quando si comprende come le variabili impattano su risultati forniti dal modello in relazione all'intero insieme di dati di addestramento, oppure LOCALE se la spiegazione è valida solo per uno specifico output.
La configurazione delle tecniche XAI è composto da tre metodologie principali.

- Precisione della previsione: la precisione è una componente fondamentale del successo dell'utilizzo dell'AI nelle operazioni quotidiane. Eseguendo simulazioni e confrontando l'output dell'XAI con i risultati nel dataset di formazione, è possibile determinare la precisione della previsione. La tecnica più utilizzata per questa operazione è denominata LIME (Local Interpretable Model-Agnostic Explanations), che spiega la previsione dei classificatori tramite algoritmi di machine learning.
- Tracciabilità: la tracciabilità è un'altra tecnica fondamentale per realizzare l'XAI. Questa si ottiene, ad esempio, limitando le modalità in cui è possibile prendere le decisioni e impostando un ambito più ristretto di regole e funzioni di machine learning.
- Comprensione della decisione: questo è il fattore umano. Molte persone non hanno fiducia nell'AI, eppure per poterla utilizzare in modo efficiente, hanno bisogno di imparare a fidarsi. Questo obiettivo viene raggiunto addestrando il team che lavora con l'AI, in modo che possa comprendere come e perché l'AI prende decisioni.

La spiegabilità può essere ante-hoc ovvero modelli white-box direttamente interpretabili o post-hoc tecniche per spiegare un modello precedentemente addestrato o la sua previsione. I modelli ante-hoc includono reti neurali spiegabili (xNN), macchine di amplificazione spiegabili (EBM), modelli interi lineari supersparsi (SLM), modello di attenzione a tempo invertito

(RETAIN) e deep learning bayesiano (BDL).
I metodi di spiegabilità post-hoc includono spiegazioni locali interpretabili indipendenti dal modello (LIME) e visualizzazioni locali e globali di previsioni del modello come grafici a effetto locale accumulato (ALE), grafici a dipendenza parziale (PDP) unidimensionali e bidimensionali, grafici di aspettativa condizionale individuale (ICE) e modelli surrogati dell'albero decisionale.

2.4 Librerie e Tool Explainable Artificial Intelligence

In particolare ci focalizzeremo maggiormente su due librerie Local interpretable model-agnostic explanations (LIME) e SHapley Additive exPlanations (SHAP) sono strumenti di spiegabilità che consentono di spiegare le decisioni prese da un modello di machine learning utilizzando interpretazioni locali.

- **LIME** è una tecnica post-hoc per spiegare le previsioni di qualsiasi classificatore di machine learning perturbando le caratteristiche di un input ed esaminando le previsioni. Parliamo di metodo AGNOSTICO in quanto può essere applicato su qualsiasi modello a differenza dei metodi specifici che sono stati progettati per essere applicati su modelli specifici, agisce localmente ovvero ingrandendo l'area locale della singola previsione possiamo spiegare qualcosa che abbia senso in quella regione locale dunque non ci preoccupiamo del resto del modello ma di quelle singole aree [8] .
  L'idea alla base è che il modello dimentica i dati di training e immagina di avere solo il modello black box dove inserisce i dati e ottiene in output le previsioni. L'obiettivo è capire perché il modello di machine learning fa una determinata previsione, verifica cosa succede alle previsioni quando si forniscono variazioni dei dati nel modello di machine learning, genera un nuovo set di dati costituito da campioni perturbati e le corrispondenti previsioni del modello black box. Su questo nuovo set di dati LIME addestra quindi un modello interpretabile, che viene ponderato dalla vicinanza delle istanze campionate all'istanza di interesse. Il modello appreso dovrebbe essere una buona approssimazione delle previsioni del modello di machine learning a livello locale, ma non deve essere una buona approssimazione globale. Questo tipo di precisione è anche chiamato fedeltà locale. L'intuizione chiave alla base racchiude la facilità di poter approssimare un modello black box con un modello semplice localmente.

- **SHAP** è un metodo per spiegare le previsioni individuali, applica un metodo locale ma può essere ampliato al metodo globale aggregando le soluzioni locali, si basa sui valori di Shapley [9] . I valori di Shapley sono un approccio ampiamente utilizzato dalla teoria dei giochi cooperativi che hanno proprietà desiderabili. I valori delle feature di un'istanza dati fungono da attori in una coalizione. Il valore di Shapley è il contributo marginale medio di un valore di caratteristica in tutte le possibili coalizioni.
  Consideriamo un gioco con dei giocatori e una vincita, l'idea alla base è di considerare il gioco come il nostro modello, la vincita come la previsione e trattare ogni caratteristica come un giocatore [3] .
  Calcoliamo i valori di ogni giocatore per scoprire il loro contributo nel modello BLACK-BOX e andiamo a considerare l'eliminazione di un giocatore attuando ciò comprendiamo come quest'ultimo ha contribuito al gioco, dobbiamo inoltre considerare le iterazioni di quest'ultimo con gli altri giocatori dunque terremo conto anche di sottoinsiemi calcolando il contributo di ogni giocatore nel sottoinsieme e successivamente sviluppando la media su tutti i contribuiti ciò ci poterà ad avere il contributo marginale o anche detto valore marginale del giocatore all'interno della squadra.
  Proiettando quando detto precedentemente nell'ambito dell'intelligenza artificiale ogni caratteristica viene considerata singolarmente per comprendere il suo contributo all'interno della predizione black-box, come quest'ultime interagiscono tra loro e il loro contribuiscono alla predizione, consideriamo dei sottoinsiemi e calcoliamo il contributo di ogni caratteristica per il sottoinsieme inerente, sviluppando la media su tutti i contributi otteniamo il contributo marginale o anche appunto denominato valore marginale.

CAPITOLO 3

Creazione Modello

3.1 Data Understanding

Una prima fase è stata sviluppata ricercando e analizzando i dataset su `https://www.kaggle.com/` tra cui:

- Predizione consumo di alcol negli studenti di matematica e lingua portoghese
- Predizione del risultato di uno studente ad un test
- Predizione dell'utilizzo di WhatsApp per gli studenti durante il Covid-19

Da un'accurata analisi si è deciso di utilizzare i dataset inerenti al consumo di alcol negli studenti di matematica e lingua portoghese in quanto oltre a risultare molto interessante essi risultano essere formattati in maniere maggiormente adeguata e corretta rispetto agli altri. I dataset sono costituiti da 33 feature:

Tabella 3.1: Tabella feature dataset

Feature	Descrizione	Tipo di dato
School	scuola	stringa
Sex	genere dello studente	stringa
Age	età dello studente	intero
Address	indirizzo	stringa
Famsize	numero di componenti	intero
Pstatus	stato convivenza del genitore	stringa
Medu	livello istruzione madre	stringa
Fedu	livello istruzione padre	stringa
Mjob	lavoro della madre	stringa
Fjob	lavoro del padre	stringa
Reason	motivo per scegliere questa scuola	stringa
Guardian	tutore dello studente	stringa
Traveltime	tempo di viaggio da casa a scuola	intero
Studytime	tempo di studio settimanale	intero
Failures	numero di errori di classe passati	intero
Schoolsup	supporto educativo extra	stringa
Famsup	supporto educazione familiare	stringa
Paid	lezioni extra pagate all'interno della materia del corso (matematica o portoghese)	intero
Activities	attività extra-curriculari	stringa
Nursery	scuola materna frequentata	stringa
Higher	vuole prendere l'istruzione superiore	stinga
Romantic	relazione romantica	stringa
Famrel	qualità delle relazioni familiari	stringa
Freetime	tempo libero dopo la scuola	stringa
Goout	uscire con gli amici	stringa
Dalc	consumo di alcol al giorno lavorativo	intero
Walc	consumo di alcol nel fine settimana	intero
Health	stato di salute attuale	stringa
Absences	numero di assenze scolastiche	intero
G1	voti primo periodo scolastico	stringa
G2	voti secondo periodo scolastico	stringa
G3	voto finale periodo scolastico	stringa

3.2 Data Preparation

Analizzando i dataset è stato deciso di intraprendere un processo di feature engineering seguendo le diverse fasi di:

- Data Cleaning
- Feature Scaling
- Feature Selection
- Data Balancing

3.2.1 Data Cleaning

Nella fase di Data Cleaning i dati sono stati valutati andando a considerare se ci fossero valori nulli o non validi.
Da un analisi preliminare dei dati a disposizione è risultato che non sono presenti dati nulli o invalidi tra quelli presenti, per cui non sono state necessarie tecniche di Data cleaning.

3.2.2 Correlazione tra variabili

Successivamente si è visualizzata la correlazione tra le diverse variabili ottenendo i seguenti risultati:
In particolare notiamo:

- una maggiore correlazione tra le variabili G1, G2, G3 dunque si è deciso di eliminare le variabili G1 e G2 lasciando solo la variabile G3 essa infatti è rappresentativa del voto finale dello studente racchiude entrambe le variabili precedentemente elencate.
- che le variabili DALC e WALC rappresentative del consumo di alcol dello studente nel weekend e durante la settimana possono essere racchiuse in un'unica variabili AWK rappresentativa del consumo di alcol inerente all'intera settimana ciò avviene calcolando la somma e la media delle due variabili precedentemene elencate.
- che le variabili SCHOOL, REASON, INTERNET non risultavano avere una grande importanza nella predizione dunque si è deciso di eliminarle.

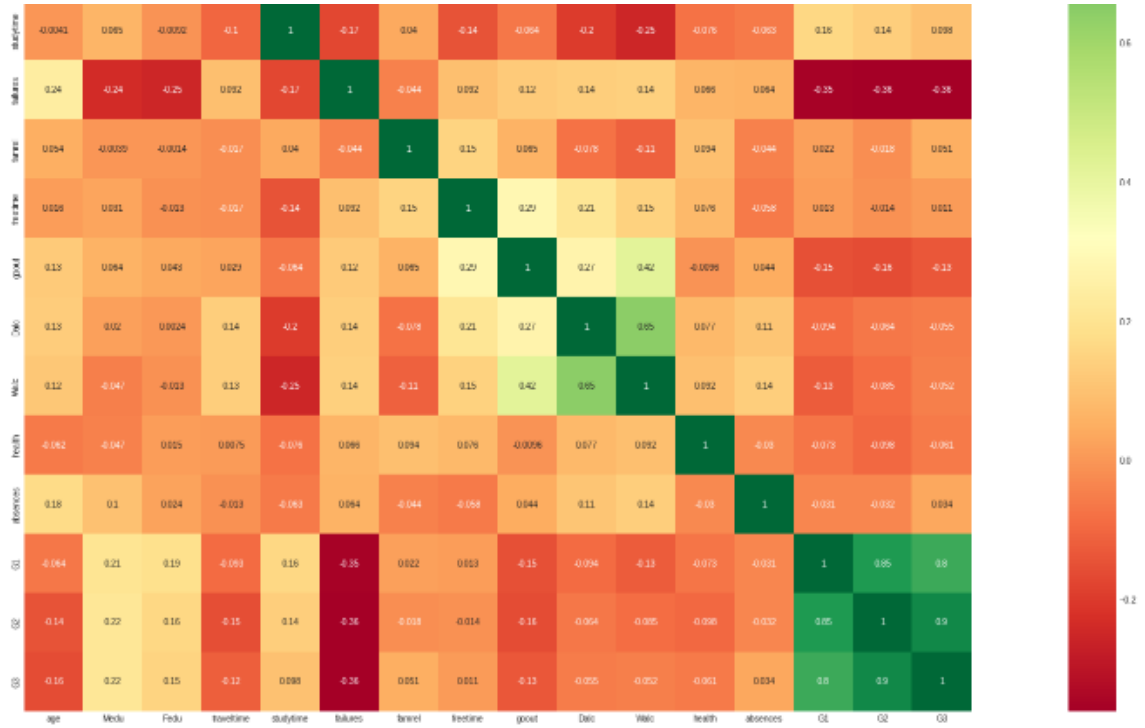


Figura 3.1: Correlazioni tra le diverse feature

3.2.3 Feature Scaling

Nella seconda fase si è sviluppata la Feature Scaling ovvero l'insieme di tecniche che consentono di normalizzare o scalare l'insieme di valori di una caratteristica, ciò è avvenuto utilizzando un metodo della libreria "*sklearn*" la tecninca presa in esame al fine di svolgere al meglio questa fase è stata la tecninca del MINMAX NORMALIZATION il quale normalizza i valori dei dati in valori compresi fra a e b.

Feature Selection

Nella terza fase si è sviluppata la Feature Selection con l'obiettivo di definire delle caratteristiche, anche chiamate feature, metriche, o variabili indipendenti che possano caratterizzare gli aspetti principali del nostro problema in esame e, quindi, avere una buona potenza predittiva. Per l'individuazione delle feature da mantenere è stata utilizzato l'algoritmo SELECTKBEST, che estrae le migliori K caratteristiche del nostro dataset analizzando la loro varianza, andando dunque a selezionare le migliori 10 variabili indipendenti dei dataset.

3.2.4 Data balancing

Nella quarta fase si è sviluppato il Data balancing, ovvero un insieme di tecniche per convertire un dataset sbilanciato in un dataset bilanciato. Notando che alcune classi risultavano più numerose rispetto ad altre dunque si è andato ad effettuare successivamente alla fase di Training e Test una tecnica di Oversampling per bilanciare il nostro datset ottenendo tutti i dati bilanciati in modo equilibrato.

3.3 Data Modeling

Nella prima fase di Data Modeling diversi algoritmi di Machine Learning sono stati testati, andando infine a trattare quest'ultimo come un problema di classificazione. Nella seconda fase sono state stabilite le metriche e le tecniche di validazione delle prestazioni del modello da costruire. Occorre suddividere l'insieme dei dati finora analizzato in due insiemi: il *training set*, composto dalle istanze di dati che saranno utilizzate per l'addestramento, e il *test set*, composto dalle istanze di dati per cui l'agente dovrà predire il valore della variabile dipendente, per effettuare questa suddivisione si è utlizzato la tecnica del REPEATED K-FOLD VALIDATION.
Le metriche che sono state impiegate per valutare la bontà delle previsioni effettuate sono state:

- Precision = $\frac{TP}{(TP+FP)}$
- Recall = $\frac{TP}{(TP+FN)}$
- Accuracy = $\frac{TP+TN}{(TP+TN+FP+FN)}$
- MCC = $\frac{TP*TN+FP*FN}{\sqrt{(TP+FP)*(TP+FN)*(TN+FP)*(TN+FN)}}$

3.3.1 Valutazione algoritmi

Gli algoritmi che sono stati valutati sono:

- **Gaussian Naive Bayes** Questo tipo di algoritmo assume che le caratteristiche non siano correlate l'una all'altra. E' stato quindi necessario verificare che questa condizione fosse verificata anche nel nostro dataset.

- **Decision Tree** Questo tipo di algoritmo non si limita a predire dati con vincoli di linearità, ma permette anche di predire valori che si presentano sotto forma di curve.

- **Random Forest** Questo è un algoritmo di tipo "*ensemble*", cioè esegue n volte l'algoritmo Decision Tree Classifier, ogni albero decisionale è creato in modo autonomo ed effettua le sue personali predizioni.
  In seguito, le predizioni finali sono poi ottenute tramite una media di quelle effettuate dai singoli alberi.

Tabella 3.2: Valutazione Algoritmi

Tipo	Modello	Precision	Recall	Accurancy	Mcc
DS MAT	Gaussian Naive Bayes	0.1497	0.149	0.757	0.052
DS LIG	Gaussian Naive Bayes	0.162	0.1628	0.760	0.049
DS MAT	Decision Tree	0.282	0.282	0.794	0.050
DS LIG	Decision Tree	0.273	0.273	0.792	0.039
DS MAT	Random Forest	0.322	0.322	0.805	0.805
DS LIG	Random Forest	0.312	0.312	0.803	0.070

Da un attento confronto e analisi delle valutazioni degli algoritmi riportante successivamente possiamo notare che per lo sviluppo del modello di predizione inerente al dataset degli studenti di matematica risulta essere migliore nonchè più vantaggioso utilizzare l'algoritmo Random Forest Classifier mentre per il dataset degli studenti di Lingua portoghese è più vantaggioso utilizzare l'algoritmo Decision Tree Classifier.

CAPITOLO 4

Studio Empirico

4.1 Obiettivo dello studio

L'obiettivo dello studio consiste nell'analisi della spiegabilità di un modello di Machine Learning per mezzo di due librerie LIME e SHAP, prendendo in esame alcuni tipi di dataset al fine di comprendere quanto sia giusta, chiara e migliore la predizione e la spiegabilità tra quest'ultimi.

4.2 Domande di ricerca

Le domande di ricerca sono state sviluppate al fine di comprendere in primis il grado explainability di LIME e SHAP sviluppando un confronto tra quest'ultimi per mettere in evidenza quale tra i due risulta avere un explainability più chiara e semplice, in secondo luogo per comprendere il grado di verità di predizione,nello specifico quanto queste predizioni risultano essere vere rispetto alla realtà.

- RQ1 Quanto è il grado di explainability di LIME in confronto a SHAP?
- RQ2 Quanto è veritiero il grado di predizione di LIME in confronto a SHAP?

4.3 Contesto dello studio

4.3.1 Framework di Explainable AI analizzati

Una prima fase nel contesto dello studio è stata sviluppata andando ad analizzare i diversi tipi di fremework di Explainable AI tra questi sono stati analizzati:

- Explainable neural networks
- Deep learning bayesiano
- LIME
- SHAP

Explainable neural networks (xNN) si basano su modelli di indice additivo, che possono approssimare funzioni complesse. Gli elementi di questi modelli sono chiamati indici di proiezione e funzioni di cresta. Gli xNN sono reti neurali progettate per apprendere modelli di indice additivo, con sottoreti che apprendono le funzioni di cresta. Il primo livello nascosto utilizza funzioni di attivazione lineare, mentre le sottoreti sono in genere costituite da più livelli completamente connessi e utilizzano funzioni di attivazione non lineari, possono essere utilizzati da soli come modelli predittivi spiegabili costruiti direttamente dai dati, possono anche essere utilizzati come modelli surrogati per spiegare altri modelli non parametrici come metodi basati su alberi e reti neurali feedforward.
Bayesian deep learning (BDL) offre stime di incertezza basate su principi da architetture di deep learning. Fondamentalmente, BDL aiuta a porre rimedio al problema che la maggior parte dei modelli di deep learning non può modellare la loro incertezza modellando un insieme di reti con pesi tratti da una distribuzione di probabilità appresa, in genere raddoppia solo il numero di parametri.
Successivamente ad un'attenta analisi di quest'ultimi si è stabilito di utilizzare i framework LIME e SHAP, essi risultano essere interessanti in quanto tra di loro rappresentano due tipi di spiegabilità differenti in diversi fattori, ciò ci permette di fornire una spiegabilità completa sotto diverse caratteristiche.
Nello specifico LIME è una tecnica post-hoc ovvero fornisce spiegazioni locali dunque fornisce informazioni su ogni singola osservazione di un dataset interpretabile di un qualsiasi classificatore di machine learning analizzando le caratteristiche di un input e le previsioni. SHAP è un metodo per spiegare le previsioni individuali, basato sui valori shapley teoricamente ottimali del gioco. Il valore di Shapley è il contributo marginale medio di un valore

di una caratteristica in tutte le possibili coalizioni con le altre caratteristiche ad esempio se prendiamo in considerazione un gioco con dei giocatori ed una vincita il valore di Shapley rappresenta il contributo di quel giocatore nel gioco al fine di arrivare alla vittoria.
Sia LIME che SHAP risultano esprimere una spiegazione a livello locale ma con SHAP possiamo ottenere anche una spiegazione a livello globale fornendo informazioni sull'insieme delle osservazioni di un dataset interpretabile, andando ad aggregare tutte le spiegazioni locali. L'output di LIME può essere presentato graficamente con il metodo "*show in notebook*" quest'ultimo mostra nella parte destra le diverse features e il valore assegnato ad esse, nella parte sinistra la probablità della predizione della variabile dipendente. SHAP rappresenta in output la feature più promettente al fine della predizione, per rappresentare ciò vi è la possibilità di utilizzare diversi tipi di grafici [3] .

4.3.2 Sviluppo Questionario

Una seconda fase nel contesto dello studio consiste nello sviluppare due tipi di questionari. Il primo viene sviluppato al fine di comprendere l'explainbility di LIME e SHAP ed è indirizzato ad un target di persone con una minima conoscenza del problema dunque il segeunte questionario sarà compilato dagli studenti del Dipartimento Informatica dell'Università degli Studi di Salerno. Il secondo questionario viene realizzato per comprendere quanto giusta sia la predizione dei due framework dunque quanto quest'ultimi abbiano dato una giusta predizione del consumo di alcol degli studenti di matematica e lingua esso verrà compilato da un target di persone inerenti al dataset come gli studenti del dipartimento di matematica e lingua dell'Università degli Studi di Salerno. I questionari saranno sviluppati utilizzando GOOGLE MODULI `https://docs.google.com/forms/u/0/?tgif=d`.

Tabella 4.1: Questionario Consumo Alcol Studenti

Id	Domanda	Tipo di risposta
1	Qual è il tuo corso di studi?	lingue, matematica, altro
2	Qual è il tuo genere?	uomo, donna, non binario, altro
3	Quanti anni hai?	Numerico 18-20, 21-23, 23-26
4	In che tipo di zona abiti?	Città, Periferia
5	Quante volte alla settimana esci con i tuoi amici?	Numerico
6	Da quante persone è composta la tua famiglia?	Numerico
7	Qual è la media dei tuoi esami universitari?	Numerico 18-20, 21-25, 25-30
8	Quante volte in una settimana ti assenti alle lezioni?	Numerico
9	Come valuteresti da 0 a 5 il tuo consumo di alcol in una settimana?	Numerico 0-5
10	Quanto tempo (ore/minuti) impieghi nello studio in un giorno?	Numerico
11	Quante tempo (ore/minuti) impieghi per arrivare all'università?	Numerico
12	Hai una relazione affettiva?	Binario si/no
13	Qual è lo stato di convivenza dei tuoi genitori?	conviventi, separati

Tabella 4.2: Questionario Explainability

Id	Domanda	Tipo di risposta
1	Qual è il tuo corso di laurea in informatica ?	Triennale, Magistrale
2	Da quanti anni sei iscritto ad informatica ?	Numerico
3	Considerando l'immagine di seguito come valuteresti la leggibilità su una scala da 1 a 5 ?	Numerico 1-5
4	Considerando l'immagine di seguito riusciresti a definire quale feature è più rilevante ?	Testuale
5	Considerando l'immagine di seguito riusciresti ad identificare la probabilità di predizione del consumo di alcol degli studenti ?	Testuale
6	Considerando l'immagine di seguito riusciresti ad identificare a cosa sono associati i colori nel grafico ?	Testuale
7	Considerando l'immagine di seguito come valuteresti la chiarezza su una scala da 1 a 5 ?	Numerico 1-5
8	Quanto valuteresti rilevante su una scala da 1 a 5 l'utilizzo del grafico rappresentato nell'immagine sottostante per spiegare una predizione ?	Numerico 1-5
9	Quanto valuteresti rilevante su una scala da 1 a 5 l'utilizzo del grafico rappresentato nell'immagine sottostante per spiegare quale "features" più promettente per la predizione ?	Numerico 1-5
10	Pensi sia stato utile questo questionario per rilevare il grado di spiegabilità delle librerie prese in considerazione ? Scrivi le tue osservazioni e considerazioni	Testuale

4.4 Metodologia

4.4.1 Test Pilota

Precedentemente all'invio dei relativi questionari ai studenti del Dipartimento di Informatica, Lingua e Culture Straniere e Matematica dell'Università degli Studi di Salerno i questionari sono stati inviati ad un campione di studenti del Dipartimento di Informatica dell'Università degli Studi di Salerno i quali sono stati contattati mediante un gruppo Telegram. Su 40 studenti 10 hanno partecipato al test con lo scopo di comprendere se i questionari sviluppati fossero adeguati e chiari in termini di strutturazione, chiarezza e leggibilità delle domande e delle immaggini presenti.
Le domande che sono state poste agli studenti sono le seguenti:

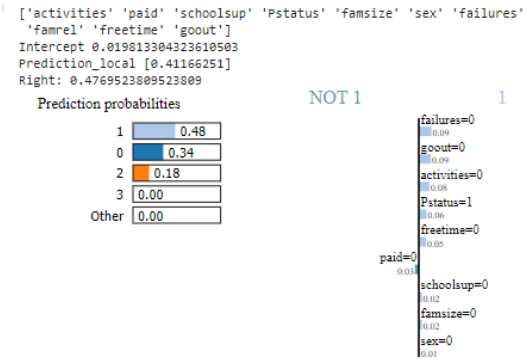
- Come risulta essere la struttura del questionario in termini di chiarezza e leggibilità delle domande?
- Come risultano essere le immagini in termini di leggibilità e chiarezza dei dati?
- Risulta essere chiaro lo scopo del questionario ?

Il test ha dato dei risultati soddisfacenti pertanto non sono state apportate modifiche.

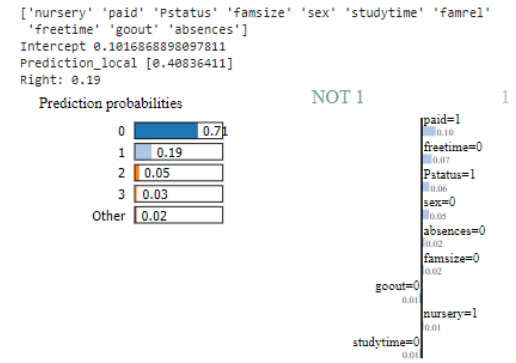
4.4.2 Data Extraction/Collection

Successivamente alle fasi di Data Preparation e Data Modeling sono stati eseguiti i modelli di classificazione per la predizione del consumo di alcol negli studenti sui due framework LIME e SHAP ottenendo i seguenti output visualizzabili al seguente link `https://github.com/rebeccadimatteo/ExplainableAI`. Dall'esecuzione di LIME si sono ottenuti 20 grafici sia per il dataset degli studenti di matematica che per gli studenti di lingua portoghese rappresentativi della predizione del consumo di alcol degli studenti.
Visualizzando e analizzando i grafici successivi possiamo in una prima fase spiegare come poter leggere quest'ultimi in una seconda fase estrapolare e collezionare diversi dati che saranno analizzati nel paragrafo successivo.
Nei grafici rappresentativi di LIME notiamo nella parte sinistra del grafico le diverse feature con i valori assegnati ad esse mentre nella parte destra abbiamo la probabilità della predizione espressa in un range da 0 a 3 nel quale 0 rappresenta che lo studente non consuma alcol 3 che lo studente consuma molto alcol.

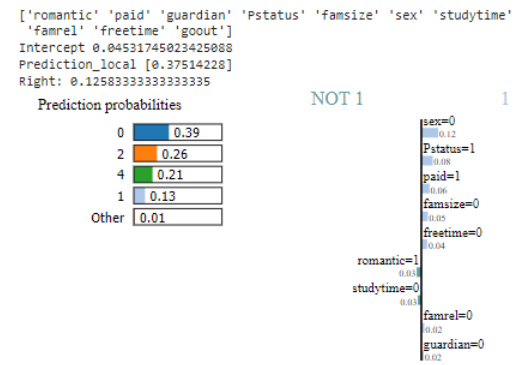
4.4.3 LIME Grafici Risultati Predizione



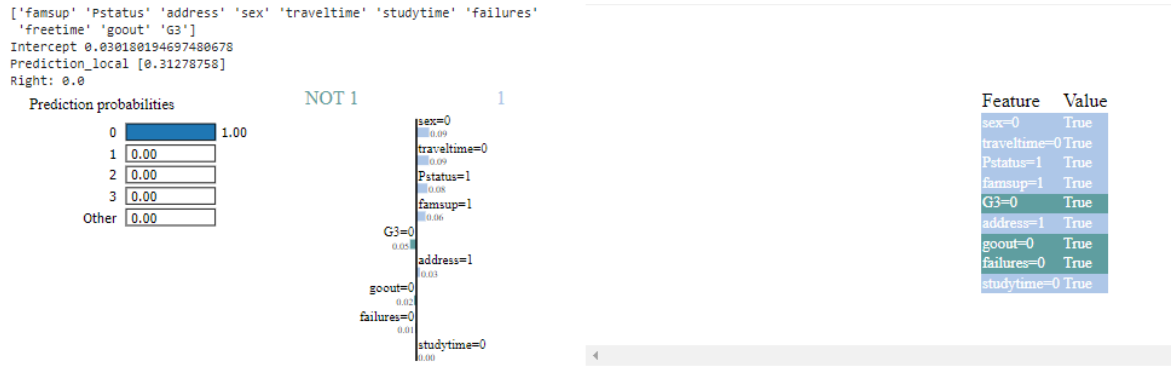
Feature	Value
failures=0	True
goout=0	True
activities=0	True
Pstatus=1	True
freetime=0	True
paid=0	True
schoolsup=0	True
famsize=0	True
sex=0	True



Feature	Value
paid=1	True
freetime=0	True
Pstatus=1	True
sex=0	True
absences=0	True
famsize=0	True
goout=0	True
nursery=1	True
studytime=0	True



Feature	Value
sex=0	True
Pstatus=1	True
paid=1	True
famsize=0	True
freetime=0	True
romantic=1	True
studytime=0	True
famrel=0	True
guardian=0	True



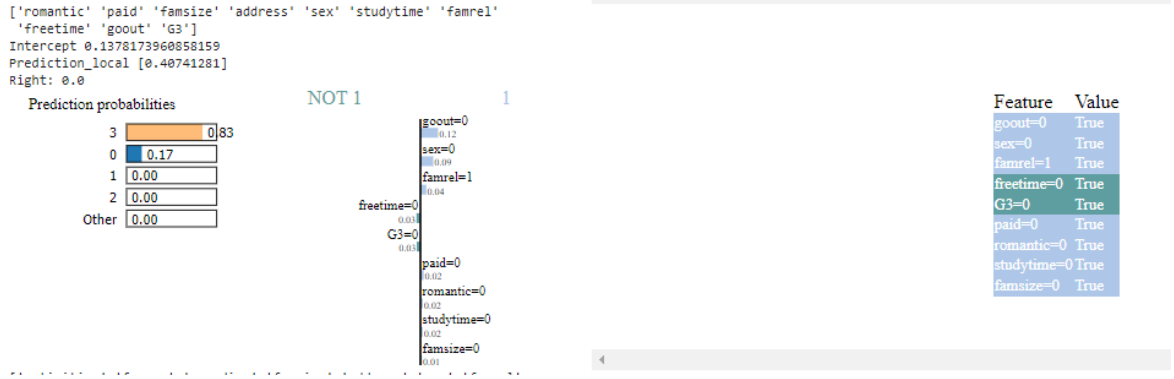
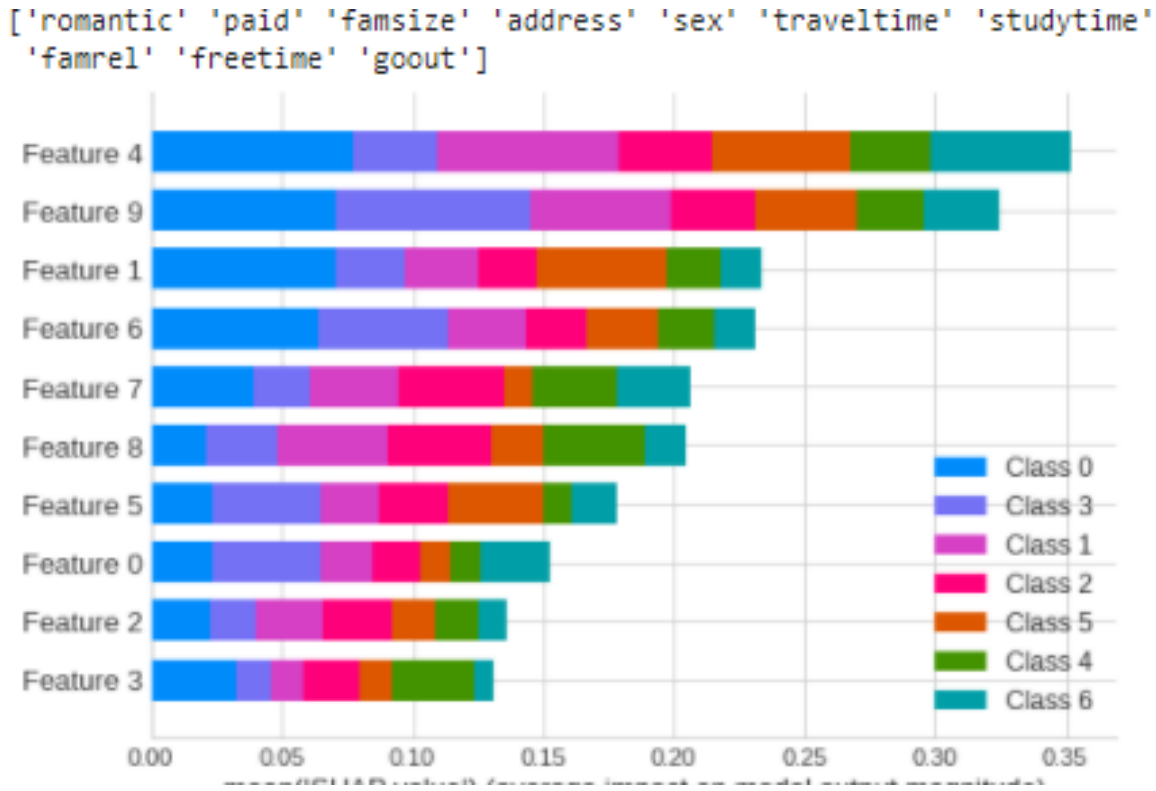


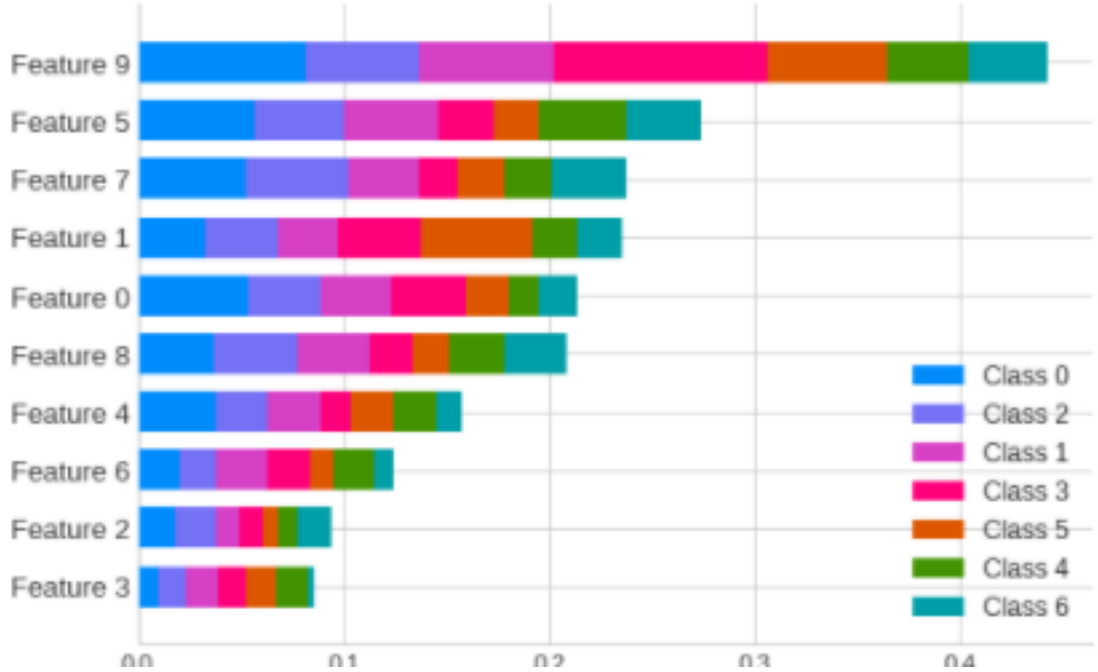
Figura 4.1: Grafici dei risultati LIME

4.4.4 SHAP Grafici Risultati Predizione

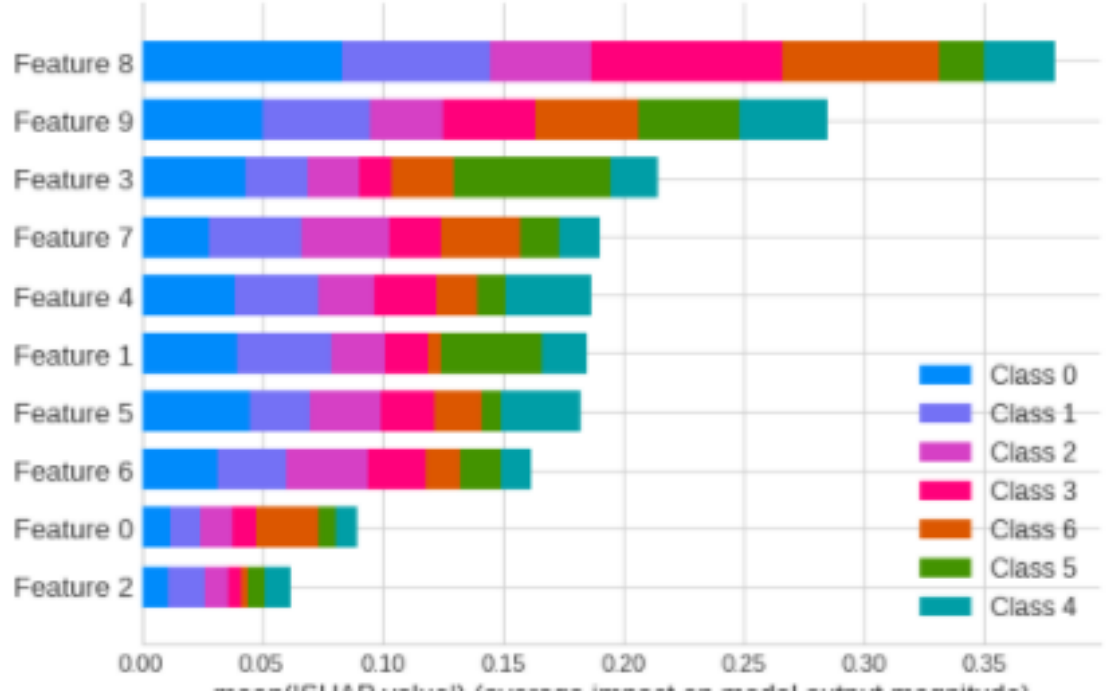
Dall'esecuzione di SHAP si sono ottenuti 21 grafici sia per il dataset degli studenti di matematica che per gli studenti di lingua portoghese rappresentanti le features più influenti per la predizione del consumo di alcol negli studenti.
Nei grafici rappresentativi di SHAP notiamo le features che contribuiscono maggiormente a spingere l'output del modello dal valore di base verso la predizione. Nello specifico nella rappresentazione del grafico possiamo notare a destra una legenda con le diverse classi di output rappresentate da un colore, prendendo in considerazione il valore assoluto medio dei valori SHAP per ogni feature otteniamo un grafico a barre standard nel quale possiamo visualizzare in base al legame percentuale/colore quanto quella determinata features è influente per la relativa classe di output.



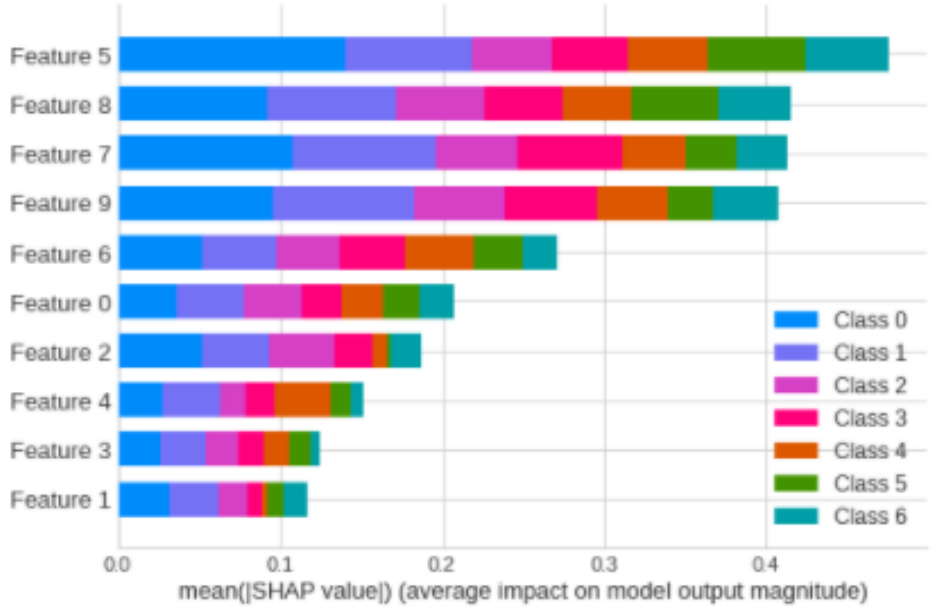
['activities' 'paid' 'schoolsup' 'Pstatus' 'famsize' 'sex' 'failures'
 'famrel' 'freetime' 'goout']



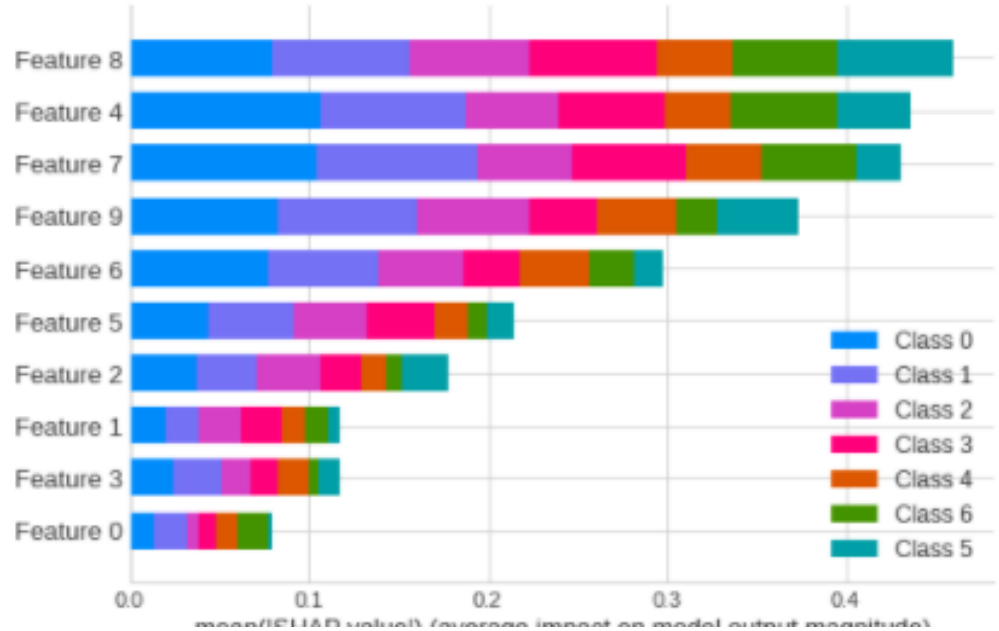
['nursery' 'paid' 'Pstatus' 'famsize' 'sex' 'studytime' 'famrel'
 'freetime' 'goout' 'absences']



['romantic' 'nursery' 'famsup' 'famsize' 'address' 'sex' 'age' 'goout'
 'absences' 'G3']



['paid' 'guardian' 'famsize' 'address' 'sex' 'age' 'studytime' 'goout'
 'absences' 'G3']



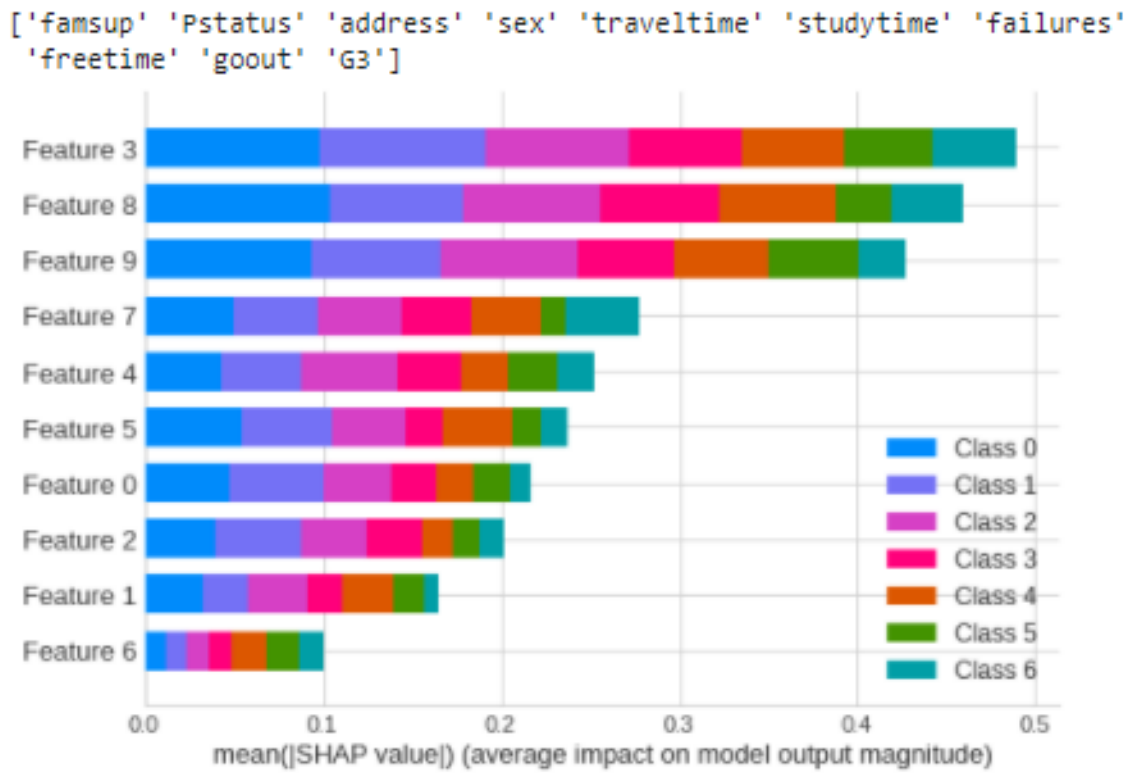


Figura 4.2: Grafici dei risultati SHAP

4.4.5 Data Analysis

Considerando i diversi dati collezionati dai grafici dati in output dai due framework possiamo analizzare quest'ultimi in due modi a seconda del framework:

- LIME Risultati Predizione

  Analizzando i risultati dati in output dai grafici in relazione alle diverse feature è stata calcolata la mediana delle percentuali di consumo di alcol degli studenti.

- SHAP Risultati Predizione

  Analizzando i risultati delle migliori 10 features date in output in relazione alle diverse classi stabilendo il range di percentuale per cui una features è piu influente di un'altra all'interno della predizione.

Tabella 4.3: Percentuale LIME consumo alcol studenti

Tipo	Consumo	Percentuale
Dataset Matematica	0	25%
Dataset Matematica	1	30%
Dataset Matematica	2	36%
Dataset Matematica	3	9%
Dataset Lingue	0	60%
Dataset Lingue	1	20%
Dataset Lingue	2	15%
Dataset Lingue	3	5%

Tabella 4.4: Percentuale SHAP consumo alcol studenti

Feature	DS MAT	DS LIG
Goout	⋆⋆⋆⋆	⋆⋆
Address	⋆⋆	⋆
G3	⋆	⋆⋆⋆⋆
Sex	⋆	⋆⋆⋆⋆
Absences	⋆	⋆⋆⋆
Famsize	⋆	⋆
Pstatus	⋆	⋆
Studytime	⋆	⋆
Paid	⋆	⋆
Romantic	⋆	⋆

Note: Le stelle all'interno della tabella rappresentano le percentuali secondo le quali una feature è più influente rispetto ad un altra per la predizione. Ogni stella corrisponde ad un influenza del 10% all'interno della predizione.

CAPITOLO 5

Osservazioni e Risultati

All'interno di questo capitolo vengono descritti i risultati e le osservazioni dell'analisi dei dati che sono stati estrapolati dai due questionari precedentemente descritti nel capitolo 4. Nello specifico il primo questionario relativo all'Explainability delle librerie Python LIME e SHAP per comprendere quale tra le due abbia un grado di predizione maggiore in base a due caratteristiche fondamentali la leggibilità e la chiarezza. Il secondo questionario relativo al consumo di alcol negli studenti dei corsi di studi di Lingue e Culture Straniere e Matematica al fine di comprendere quale sia il grado di predizione più veritiero tra le due librerie LIME e SHAP.
In particolare mediante l'analisi dei dati potremmo rispondere alle domande di ricerca che ci siamo posti precedentemente:

- RQ1 Quanto è il grado di explainability di LIME in confronto a SHAP?
- RQ2 Quanto è veritiero il grado di predizione di LIME in confronto a SHAP?

5.1 Raccolta risultati questionario consumo alcol studenti

Il questionario relativo al consumo di alcol degli studenti è stato sottoposto ad un campione di studenti dell'Università degli Studi di Salerno, nello specifico è stato inviato su due gruppi Telegram con all'interno 60 studenti frequentati il corso di studi di Matematica dei quali 42 hanno partecipato al questionario e 40 frequentanti il corso di studi Lingue e Culture Straniere dei quali 10 hanno partecipato al questionario.

I risultati sono visualizzabili al seguente link `https://github.com/rebeccadimatteo/ExplainableAI`.

5.2 Analisi dei risultati consumo alcol degli studenti

I risultati sono stati estrapolati da un attento studio delle risposte fornite dagli studenti.

- **Prima Fase:**
  La prima fase è stata sviluppata andando a suddividere l'insieme delle risposte in due sottoinsiemi in base al corso di studi degli studenti nello specifico Matematica e Lingue e Culture Straniere.

- **Seconda fase:**
  La seconda fase è stata sviluppata calcolando la mediana del consumo di alcol dei due sottoinsiemi andando ad utilizzare le risposte fornite dalla domanda identifica con id 9 presente nel capitolo precedente 4.

- **Terza Fase:**
  Nella terza fase è stato sviluppato un confronto tra la mediana fornita dalla seconda fase e la predizione fornita da LIME, tale confronto ci ha permesso di valutare il grado di verità di LIME in quanto la predizione fornita da LIME risulta essere conforme al risultato fornito dal questionario visualizzabile di seguito, ciò implica che il gardo di verità di LIME in questo caso è stato del 100%.

Tabella 5.1: Risultati confronto Lime-Questionario

Tipo DS	Consumo	LIME	Questionario	Tipo DS	LIME	Questionario
MAT	0	25%	17%	LIG	60%	58%
MAT	1	30%	35%	LIG	20%	22%
MAT	2	36%	37%	LIG	15%	10%
MAT	3	9%	11%	LIG	5%	10%

- **Quarta fase:**

  La quarta fase è stata sviluppata andando ad analizzare le risposte delle domande identificate con id da 1 a 13 escluso id 9 presenti nel capitolo precedente 4 .
  Possiamo da quest'ultime osservare che tra le feature più rilevanti per la predizione fornite in output da SHAP nello specifico GOOUT, ADDRESS, SEX, ABSENCES,G3 solo due risultano essere molto rilevanti al fine di una predizione per il consumo di alcol, nello specifico GOOUT, ABSENCES infatti si è notato che gli studenti consumanti maggior alcol avevano un maggior numero di uscite con gli amici identificato dalla feature GOOUT e maggiori assenze all'università identificato dalla feature ABSENCES ugualmente gli studenti con minor consumo di alcol avevano un minor numero di uscite con gli amici e di assenze, dunque abbiamo potuto osservare che il grado di verità di SHAP in questo caso risulta essere all'incirca del 20%.

Risultati per RQ1

Quanto è veritiero il grado di predizione di LIME in confronto a SHAP?
Basandoci sull'analisi e sulle osservazioni precedentemente descritte possiamo dedurre che il grado di verità delle predizioni di LIME in confronto a SHAP risulta migliore nel caso di LIME.

5.3 Raccolta risultati questionario Explainability

Il questionario relativo all' Explainability delle sue librerie LIME e SHAP è stato sottoposto ad un campione di studenti dell'Università degli Studi di Salerno, nello specifico è stato inviato su un gruppo Telegram con all'interno 100 studenti frequentati il corso di studi di Informatica dei quali 39 hanno partecipato al questionario.
I risultati sono visualizzabili al seguente link `https://github.com/rebeccadimatteo/ExplainableAI`.

5.4 Analisi dei risultati Explainability

I risultati sono stati estrapolati da un attento studio delle risposte fornite dagli studenti.

- **Prima Fase:**
  La prima fase è stata sviluppata andando ad estrapoalre le percentuali dalle risposte fornite dagli studenti relative alla leggibilità, chiarezza e rilevanza nell'utilizzo.

- **Seconda fase:**
  La seconda fase è stata sviluppata calcolando la media tra le percentuali estrapolate precedentemente nella prima fase.

- **Terza Fase:**
  Nella terza fase è stato sviluppato un confronto tra la media delle percentuali di chiarezza, leggibilità e rilevanza nell'utilizzo di LIME e SHAP al fine di comprendere quale tra le due fosse più spiegabile.

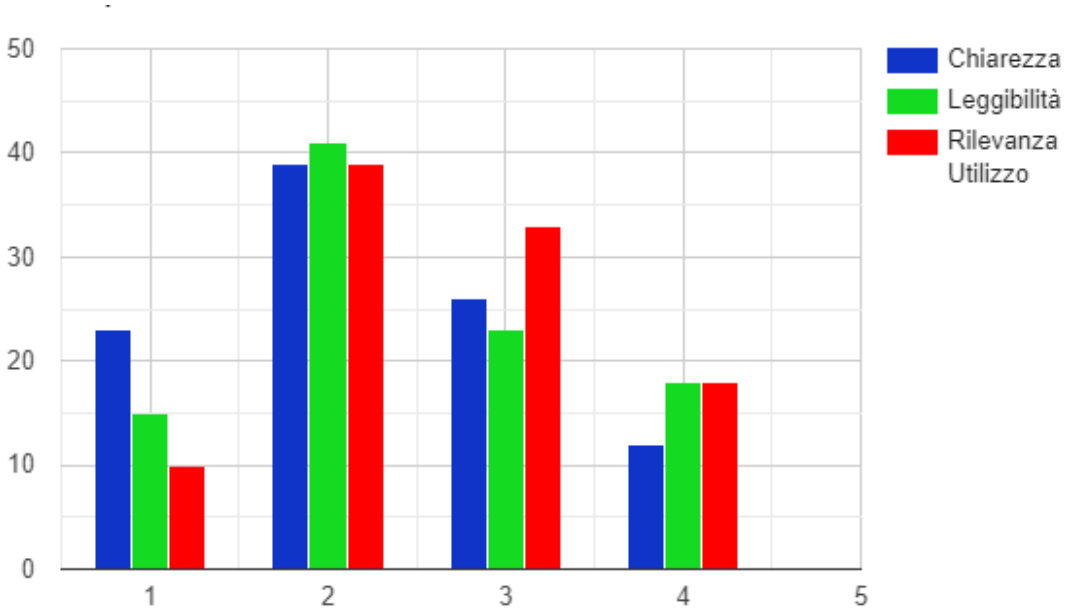


Figura 5.1: Risultati questionario Explainability LIME

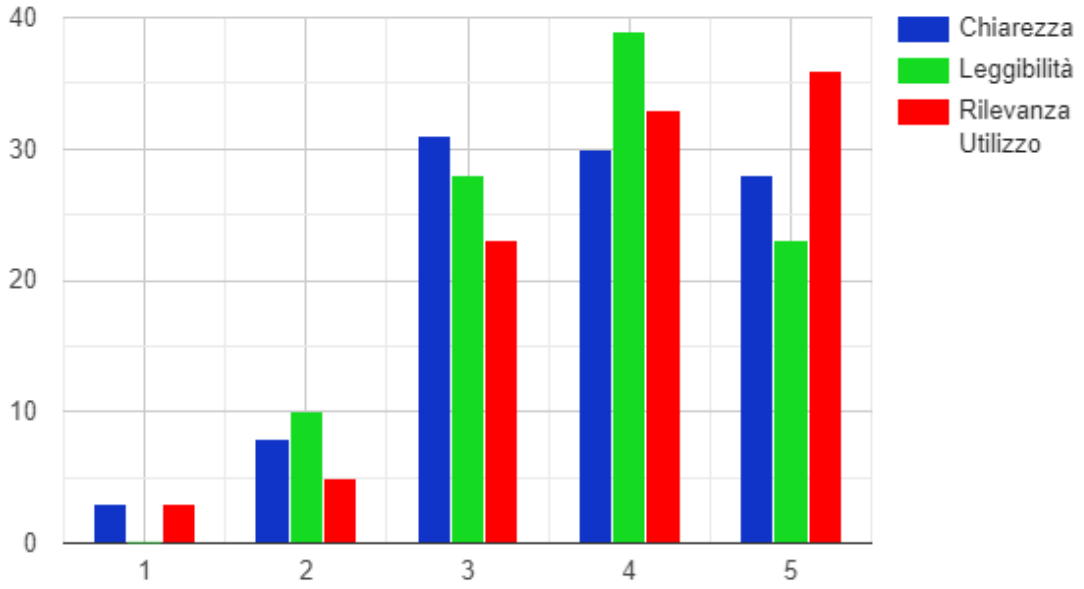


Figura 5.2: Risultati questionario Explainability SHAP

Risultati per RQ2

Quanto è il grado di explainability di LIME in confronto a SHAP?

Basandoci sull'analisi e sulle osservazioni precedentemente descritte possiamo dedurre:

- La chiarezza con cui uno studente riesce a comprendere l'output delle due liberie risulta essere migliore in LIME
- La leggibilità dei dati forniti in output risulta essere migliore in SHAP
- Gli studenti presi in esame preferirebbero utilizzare SHAP rispetto a LIME

Dunque da quest'ultime osservazioni possiamo affermare che il grado di spiegabilità di SHAP è maggiore rispetto a LIME.

CAPITOLO 6

Conclusioni e Sviluppi futuri

L'obiettivo della tesi era quello di valuatre l'Explainability delle due librerie Python LIME e SHAP. La prima fase è stata sviluppata dall'analisi delle diverse librerie Explainable focalizzandosi sulle caratteristiche positive e negative per poi focalizzarsi su due librerie più rilevanti LIME e SHAP.
La seconda fase è stata sviluppata con la costruzione di un modello per la predizione del consumo di alcol degli studenti utilizzato successivamente come input sulle due librerie scelte nella prima fase.
L'output fornito dalle due librerie è stato analizzato e sulla base di quest'ultimo è stato costruito un questionario per valutare il grado di Explaibality e verità delle due librerie rispetto ai dati reali dati in output dal questionario.
Alla fine del percorso, possiamo affermare di essere riusciti nel conseguimento del nostro obiettivo si è analizzato l'Explainability delle due librerie sviluppando un confronto il quale ha prodotto un risultato migliore per SHAP nei confronti di LIME.
Si suggerisce ad utilizzatori futuri di considerare l'idea di utilizzare LIME invece che SHAP nel caso in cui si voglia ottenere una predizione senza focalizzarsi sulle "feature" che hanno contribuito allo sviluppo di quest'ultima.
Sviluppi futuri della tesi potrebbero essere realizzati andando a confrontare i risultati ottenuti dall'utilizzo di LIME e SHAP con i risultati ottenuti lanciando il modello costruito su altre tool dell'Explainability.

Bibliografia

[1] G. Vitti, "L'intelligenza artificiale nelle politiche sanitarie," 2021. (Citato a pagina 1)

[2] P. Donghi, "Limiti e frontiere della scienza," *Limiti e frontiere della scienza*, pp. 0–0, 1999. (Citato a pagina 3)

[3] "https://www.infoworld.com/article/3634602/explainable-ai-explained.html," (Citato alle pagine 3, 9 e 18)

[4] J. D. S. A. B. Alejandro Barredo Arrieta, Natalia Diaz Rodriguez, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," (Citato a pagina 4)

[5] "https://shap.readthedocs.io/en/latest/," (Citato a pagina 4)

[6] "https://tech4future.info/explainable-ai-cose-principi-esempi/," (Citato a pagina 5)

[7] "https://www.ibm.com/it-it/watson/explainable-ai," (Citato a pagina 6)

[8] "https://christophm.github.io/interpretable-ml-book/lime.html," (Citato a pagina 8)

[9] "https://christophm.github.io/interpretable-ml-book/shap.html," (Citato a pagina 9)

Ringraziamenti