



UGotTheJob

Artificial Intelligence DOCUMENT

"Dipartimento di Informatica anno 2022/2023"

"Professore: Fabio Palomba"

Autori	Matricola
Giulio Incoronato	0512111363
Antonio Mazzearella	0512112830

Contents

1	Introduzione	3
1.1	Link Utili	3
2	Specifiche P.E.A.S.	3
2.1	Proprietà dell'Ambiente	3
3	Machine Learning	4
4	CRISP-DM	4
4.1	Business Understanding	5
4.2	Data Understanding	5
4.3	Data Preparation	6
4.3.1	Data cleaning	6
4.3.2	Feature scaling	6
4.3.3	Feature selection	6
4.3.4	Data balancing	10



1 Introduzione

Quante volte hai avuto l'ansia di essere preso o pure no in uno specifico lavoro? Quante volte ti sei domandato se fossi giusto tu per quel lavoro? Con la fine del proprio percorso di studio ci si pongono tante domande e dubbi se si viene presi in un determinato lavoro oppure no.

Tutto questo sorge perchè dopo diversi anni di studio si vuole avere la sicurezza di essere presi in un determinato lavoro, magari il lavoro dei propri sogni. Sarebbe utile avere un tool in grado di prevedere, attraverso dei dati, se sarai preso.

Il nostro team mira a combattere tutte queste ansie creando un tool chiamato "**UGotTheJob**" che integrerà algoritmi di machine learning supervisionato che andranno ad analizzare un dataset in modo da prevedere i risultati in base agli input forniti dall'utente.

1.1 Link Utili

1. Questo è il link alla repository ufficiale di **UGotTheJob**: [Link](#)
2. Questo è il link dove abbiamo preso i dataset usati per l'addestramento: [Link](#)
3. Qui è dove è stata presa l'immagine: [Link](#)

2 Specifiche P.E.A.S.

Performance	Capacità dell'agente di prevedere se l'utente sarà preso o meno per un lavoro.
Enviroment	L'ambiente in cui l'agente opera rappresentato da un form di cui l'utente scriverà i dati necessari.
Actuators	Interfaccia utente dell'applicazione dove uscirà il valore predetto.
Sensors	Form nell'interfaccia utente.

2.1 Proprietà dell'Ambiente

L'ambiente possiede le seguenti proprietà:

- **Completamente osservabile**: l'agente ha accesso completo a tutte le informazioni fornite dall'utente.
- **Deterministico**: lo stato dell'ambiente dipende dall'azione intrapresa dall'agente.
- **Sequenziale**: le decisioni dell'agente dipendono dagli input dell'utente.
- **Statico**: nel momento in cui l'agente sta elaborando la sua previsione l'utente non può modificare il form dato in partenza.
- **Discreto**: le previsioni dell'agente dipendono soprattutto dagli input inseriti dall'utente, oltretutto c'è un numero limitato e preciso di informazioni che l'utente può inserire.
- **Singolo-agente**: esiste solo un agente che opera nell'ambiente.



3 Machine Learning

Il machine learning (apprendimento automatico) è una tecnologia dell'intelligenza artificiale che consente alle macchine di imparare dai dati, senza essere esplicitamente programmate. In altre parole, il machine learning si basa sulla costruzione di algoritmi che possono imparare da un insieme di dati e migliorare la loro capacità di risolvere compiti specifici con l'esperienza.

Ci sono tre tipi principali di apprendimento automatico:

- **Apprendimento supervisionato:** in questo tipo di apprendimento, il modello è addestrato su un insieme di dati che includono sia le caratteristiche di input che le relative etichette di output. Il modello usa queste etichette per adattarsi ai dati di input e fare previsioni su dati simili.
- **Apprendimento non supervisionato:** in questo tipo di apprendimento, il modello è addestrato su un insieme di dati senza etichette di output. Il modello cerca di scoprire pattern o strutture nei dati di input.
- **Apprendimento per rinforzo:** in questo tipo di apprendimento, il modello impara attraverso l'interazione con un ambiente dinamico. Il modello prende decisioni in base allo stato attuale dell'ambiente e riceve feedback sulle sue azioni.

Il machine learning viene utilizzato in molte applicazioni, tra cui la classificazione di immagini, la traduzione automatica, la diagnosi medica, la rilevazione di frodi e molto altro ancora. Per il nostro tool abbiamo utilizzato un algoritmo di machine learning ad apprendimento supervisionato.

4 CRISP-DM

Per progettare una soluzione basata su machine learning bisogna avere un approccio **data and software engineering**. Per la creazione di tale software abbiamo utilizzato il modello *CRISP-DM* (*CRISP-DM* è l'acronimo di Cross-Industry Standard Process for Data Mining.), che rappresenta il ciclo di vita di progetti basati su intelligenza artificiale e data science.

Possiamo paragonare il modello *CRISP-DM* ad un modello a cascata con feedback utilizzato per lo sviluppo di sistemi software tradizionali. Presenta anche un modello **non sequenziale** in cui le diverse fasi possono essere eseguite un numero illimitato di volte. Esistono diverse fasi rappresentate nell'immagine di seguito (Immagine 1):

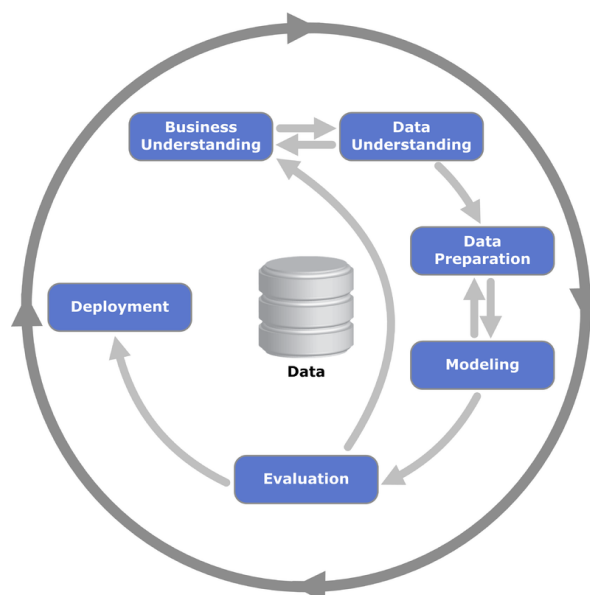


Immagine 1



4.1 Business Understanding

In questa fase si raccolgono e definiscono gli obiettivi di Business che si vogliono raggiungere, oltre a determinare la disponibilità delle risorse, stimare i rischi, indicare tecnologie e gli strumenti utilizzati per raggiungere gli obiettivi prefissati.

- **Obiettivi di Business:** L'obiettivo principale di **UGotTheJob** è la realizzazione di un tool con cui l'utente interagisce inserendo dei dati chiesti in partenza sul suo percorso di studi, il tutto verrà analizzato e processato per poi dare in output la previsione inerente al percorso di studio in questione. Lo scopo di questo tool sarà anche quello di rimanere in un giudizio che non sarà in base al sesso della persona o altri tratti somatici.
- **Disponibilità delle risorse:** La risorsa che utilizzeremo per il nostro software sarà un dataset che conterrà le informazioni sui collocamenti in base ai vari percorsi di studio e esperienze pregresse. Per reperire questo dataset utilizzeremo una piattaforma importante che è Kaggle.
- **Stima dei rischi:** I rischi che incontreremo saranno di tipo perlopiù Etico/Morale in quanto il dataset non fornisce una bilanciata percentuale di dati tra persone di sesso differente.
- **Tecnologie e Strumenti:** Per analizzare, acquisire e modellare il dataset utilizzeremo il linguaggio *Python* che presenta alcune librerie come **Pandas**.

4.2 Data Understanding

Come già discusso nella *Stima dei rischi (par.4.1 Business Understanding)* il problema da noi riscontrato è stata la poca imparzialità che l'agente potesse avere con il dataset da noi utilizzato. Il dataset avendo un discreto bilanciamento dei dati relativi al gender (*vedi Figure 1*), avrebbe portato al nostro agente una poco corretta previsione del piazzamento di una persona, rischiando quindi di cadere in una discriminazione di tipo Etico/Morale di gender.

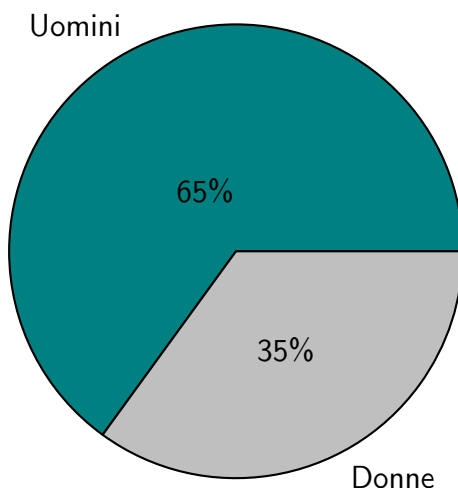


Figure 1: Gender Dataset

Il dataset inoltre presenta dati presi da un ambiente Americano, il che porta un cattivo adattamento generale dovuto al differente sistema scolastico e di votazione (*vedi Table 1*).



Voto Americano	Voto Italiano
Grado A	70-110
Grado B	60-70
Grado C	50-60
...	...

Table 1: Esempio di voti

4.3 Data Preparation

In questa sezione, tratteremo le adottate per preparare i dati acquisiti in modo che il nostro machine learner non dia problemi.

Il data preparation si articola nei seguenti quattro passaggi:

1. Data cleaning;
2. Feature scaling;
3. Feature selection;
4. Data balancing;

4.3.1 Data cleaning

Il *Data Cleaning*, definito come "*Pulizia dei dati*", si occupa di rimediare a problemi quando ci sono righe di dati mancanti ma più in generale ha come obiettivo quello di fornire un dataset dotato di una qualità adeguata. Nel nostro dataset non sono presenti dati mancanti e di conseguenza non abbiamo effettuato la fase di *Data Imputation*, che sarebbe la fase dove si stimano i valori dei dati mancanti. Oltretutto non è stato necessario effettuare una normalizzazione sui dati.

4.3.2 Feature scaling

4.3.3 Feature selection

La *Feature selection* rientra nell'ambito del feature engineering, che sarebbe il processo nel quale il progettista utilizza la propria conoscenza del dominio per determinare le caratteristiche (feature) dai dati grezzi estraibile tramite tecniche di data mining.

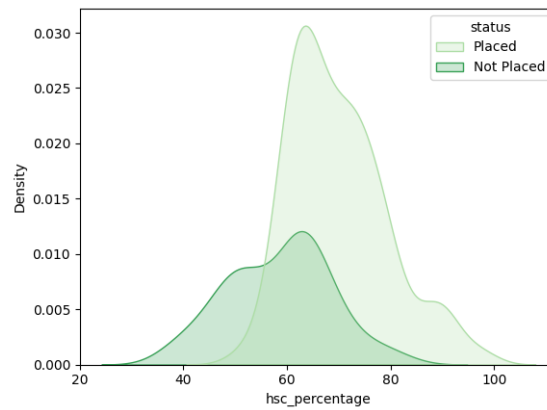
Nel nostro caso abbiamo pensato di rimuovere colonne che non erano adeguate per il nostro obiettivo ovvero, quello di creare un software che preveda un piazzamento nel mondo del lavoro quanto più etico possibile e adeguato per studenti che studiano in Italia.

	gender	ssc_percentage	ssc_board	hsc_percentage	hsc_board	hsc_subject	degree_percentage	undergrad_degree	work_experience	emp_test_percentage	specialisation	mba_percent	status
0	M	67.00	Others	91.00	Others	Commerce	58.00	Sci&Tech	No	55.0	Mkt&HR	58.80	Placed
1	M	79.33	Central	78.33	Others	Science	77.48	Sci&Tech	Yes	86.5	Mkt&Fin	66.28	Placed
2	M	65.00	Central	68.00	Central	Arts	64.00	Comm&Mgmt	No	75.0	Mkt&Fin	57.80	Placed
3	M	56.00	Central	52.00	Central	Science	52.00	Sci&Tech	No	66.0	Mkt&HR	59.43	Not Placed
4	M	85.80	Central	73.60	Central	Commerce	73.30	Comm&Mgmt	No	96.8	Mkt&Fin	55.50	Placed

Table 2: Esempio dataset.csv

Come si può vedere dalla *Table 2* che rappresenta il dataset che abbiamo scelto per il nostro progetto. Analizziamo l'influenza di ogni singola feature di questo dataset:

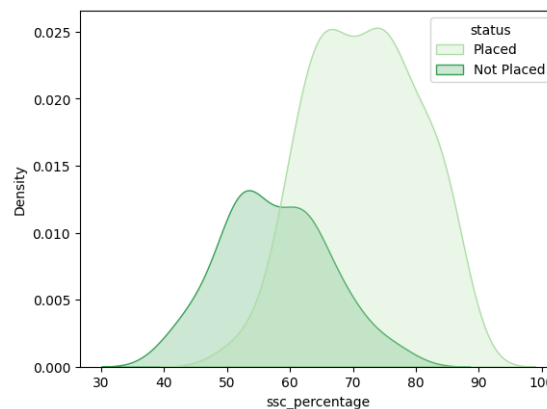




HSC Percentage

Da questo grafico possiamo vedere che la variabile *HSC Percentage* influisce nel dataset, perchè all'aumentare del valore (voto) può incidere sulla previsione del modello.

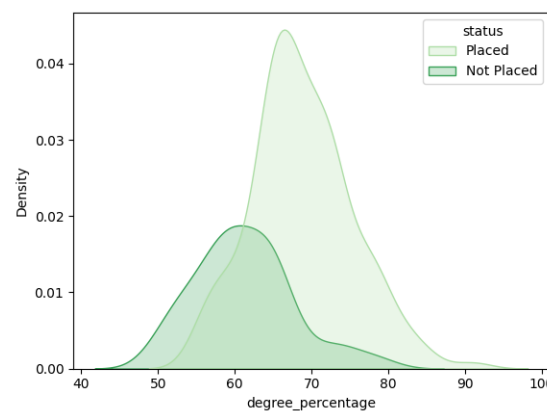
Da questa considerazione abbiamo deciso di non eliminare questa variabile.



SSC Percentage

Da questo grafico si può notare che la variabile *SSC Percentage* influisce notevolmente nel dataset perchè come abbiamo visto con la variabile *HSC Percentage*, all'aumentare del valore aumenta anche la possibilità di essere piazzati.

Da queste considerazioni abbiamo deciso di non eliminare questa variabile.

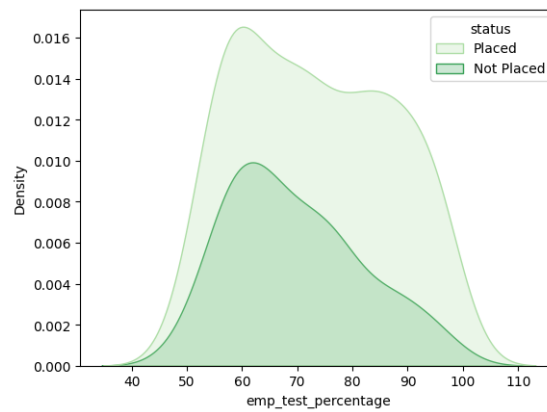


Degree Percentage

Da questo grafico si può notare che la variabile del *Degree Percentage* influisce nel dataset, perchè all'aumentare del valore aumenta anche la possibilità di essere piazzati.

Da questa considerazione abbiamo deciso di non rimuovere questa variabile.

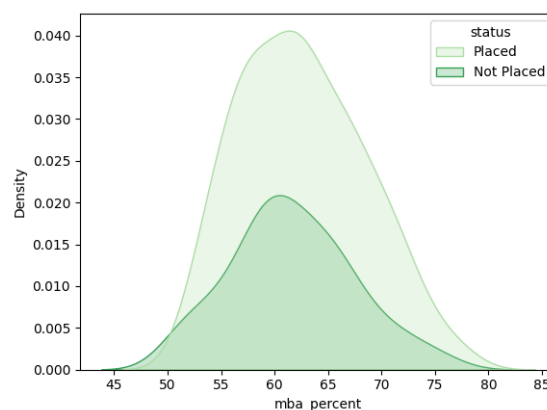




EmpTest Percentage

Da come si nota da questo grafico la variabile *EmpTest Percentage* influisce di poco nel dataset, perchè all'aumentare del valore aumenta leggermente la possibilità di essere presi.

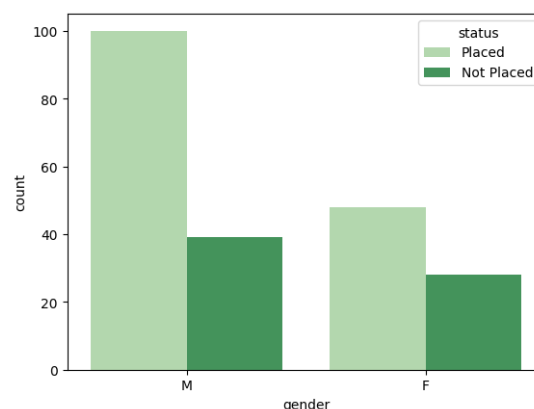
Da queste considerazioni abbiamo deciso di rimuoverlo.



MBA Percentage

Come si nota da questo grafico la variabile *MBA Percentage* influisce minimamente nel dataset, perchè non esiste una grande differenza di probabilità di essere presi o non in base al valore.

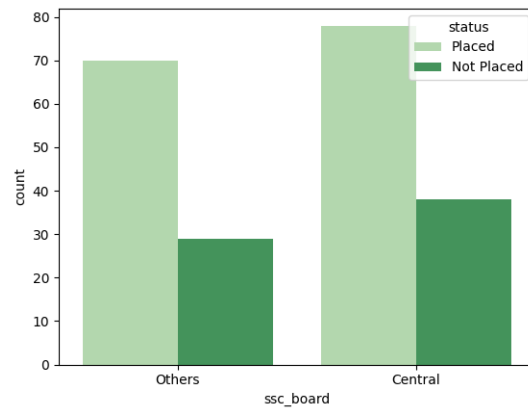
Da questa considerazione abbiamo deciso di rimuovere la variabile.



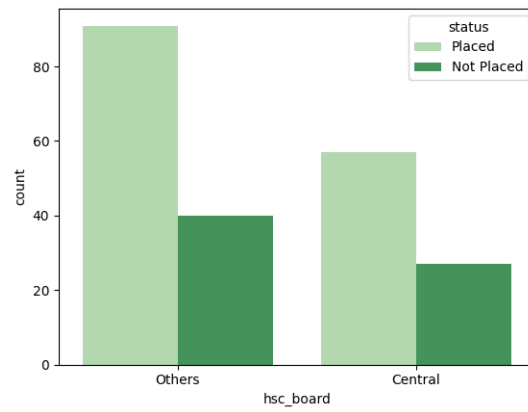
Gender

La prima colonna che abbiamo eliminato è quella riguardante il **Gender**, in quanto abbiamo pensato che non convenisse porre dei limiti riguardanti il sesso di una persona visto che il dataset in considerazione presentava più istanze con soggetto maschile rispetto a quelle femminili e che poteva portare ad un machine learner che descriminasse i vari gender.

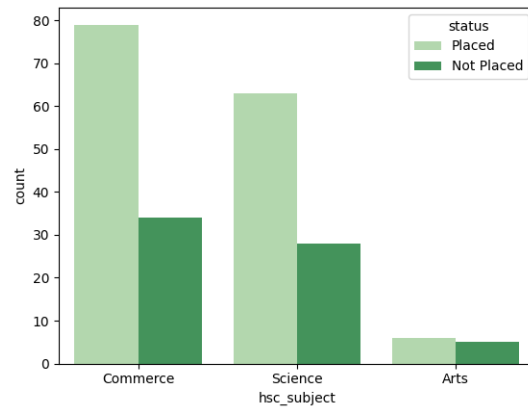




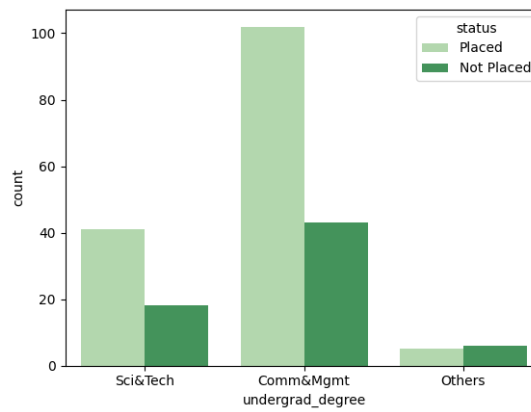
SSC Board



HSC Board

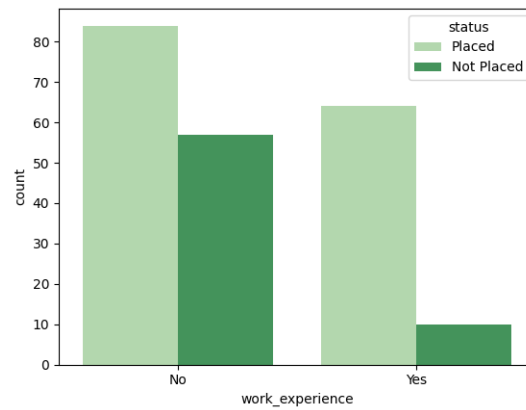


HSC Subject

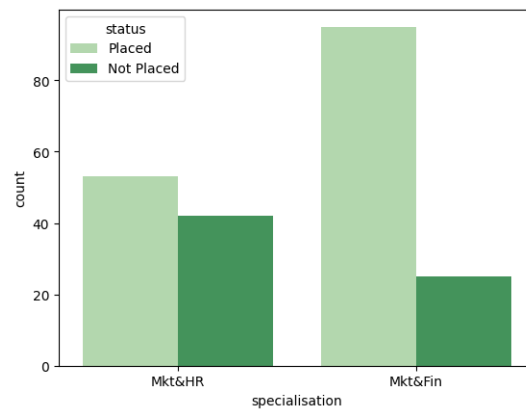


Undergrad





WORK EXP



Specialization

Le successive colonne presentavano dati che non erano adeguati in modo universale rispetto tutti i tipi di studio nelle varie lingue tra cui, in questo caso, l'Italiano. Il dataset presentava solo voti e riferimenti ai vari studi che riguardavano solo ed esclusivamente l'ambito scolastico Americano. Oltretutto la tabella presenta la percentuale di riuscita di un test di ingresso aziendale, che nel nostro caso abbiamo ritenuto giusto lasciare poichè solitamente nella maggior parte dei casi è presente.

4.3.4 Data balancing

