UGotTheJob

Artificial Intelligence DOCUMENT

"Dipartimento di Informatica anno 2022/2023"

"Professore: Fabio Palomba"

Autori	Matricola
Giulio Incoronato	0512111363
Antonio Mazzarella	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, quantà probabilità hai di essere preso.

Il nostro team mira a combattere tutte queste ansie creando un tool chiamato **"UGotTheJob"** che integrerà un modello di machine learning supervisionato che andrà a prevedere la possibità di essere piazzat0-

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 perché andremo a risolvere un problema di **classificazione**.

4 CRISP-DM

Per progettare una soluzione basata su machine learning bisogna avere un approccio **data and software engeneering**. 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 raffigurate nell'immagine di seguito (Immagine 1):

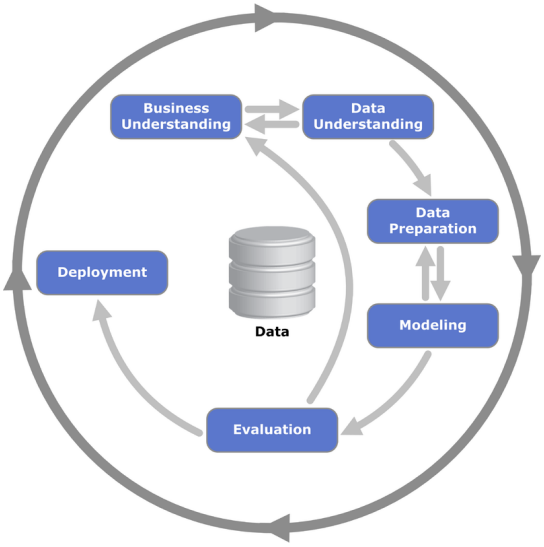


Immagine 1

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 obiettibi prefissati.

- **Obiettivi di Business:** L'obiettivo principale di **UGotTheJob** è la realizzazione di un tool con cui l'utente interagisce inserendo dei dati richiesti in partenza sul suo percorso di studi, il tutto verrà analizzato e processato per poi dare in output la probabilità di essere piazzati o non.
- **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**, **sklearn**, **seaborn** ed etc .

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.

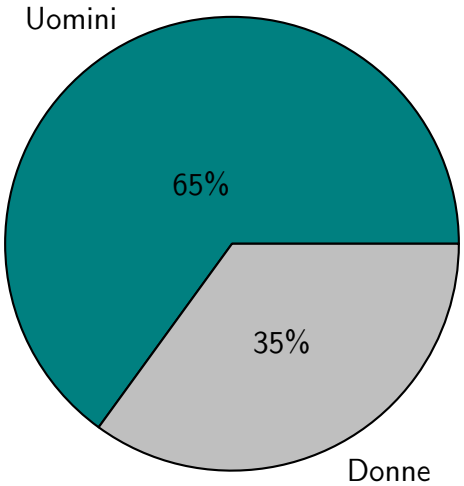


Figure 1: Gender Dataset

Il dataset inoltre presenta dati, come voti o specializzazioni, che non sono inerenti all'ambiente italiano. Possiamo vedere una tabella con le descrizioni delle singole feature presenti:

Questo ci porta a lavorare per un modello che non potrà essere utilizzato in una realtà italiana.

4.3 Data Preparation

In questa sezione, tratteremo le tecniche adottate per preparare i dati acquisiti in modo che il nostro machine learner non darà problemi e sarà quanto più efficiente possibile.

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 sui dati.

4.3.2 Feature scaling

Il *Feature Scaling*, definito come

4.3.3 Feature selection

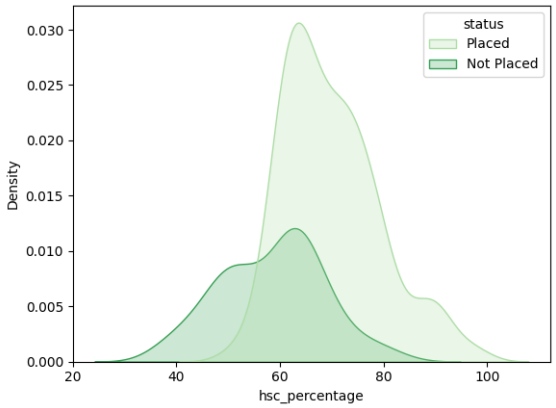
La *Feature selection* rientra nell'ambito del feature engeneering, che sarebbe il processo nel quale il progettista utilizza la propria conoscenza del dominio per determinare le caratteristiche (feature) dai dati grezzi estraibile tramite tecniche di data mining.

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

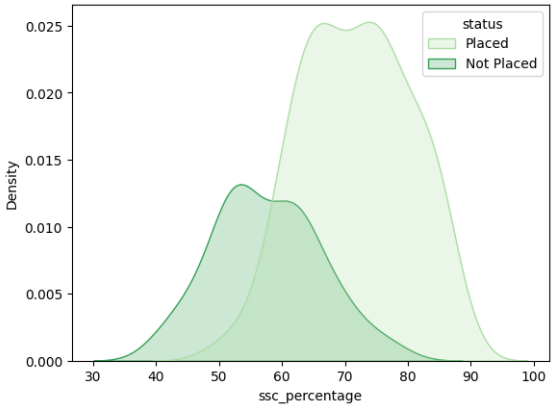
Come si può vedere dalla *Table 2* che rappresenta il dataset che abbiamo scelto per il nostro progetto. Analizziamo l'influenza di ogni singola feature di questo dataset:



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.

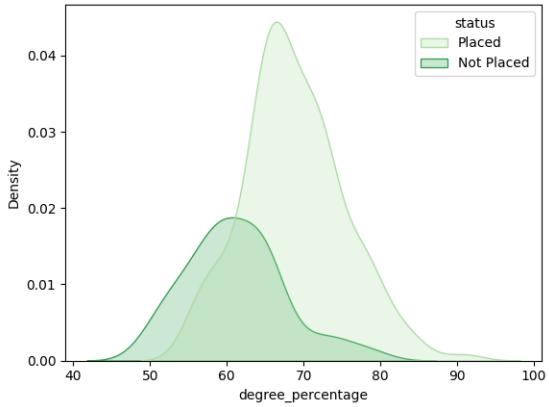
Da questa considerazione abbiamo deciso di non eliminare questa variabile.



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.

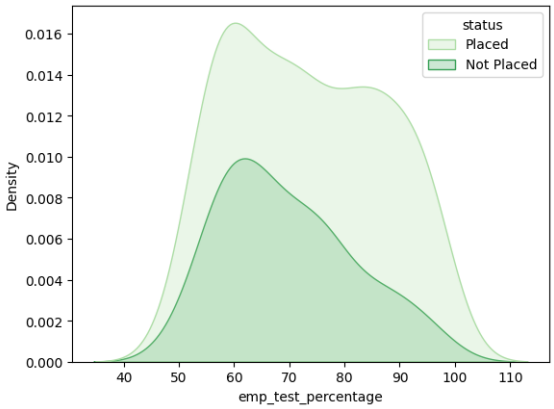
Da queste considerazioni abbiamo deciso di non eliminare questa variabile.



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.

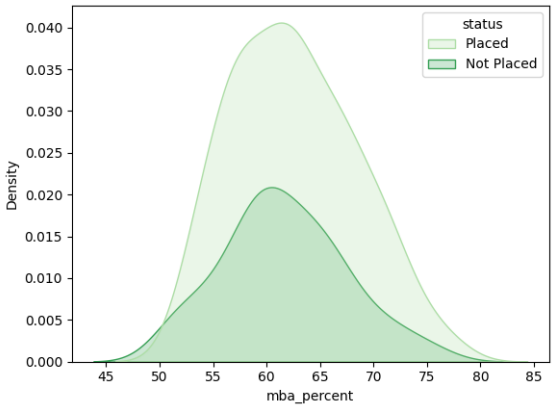
Da questa considerazione abbiamo deciso di non rimuovere questa variabile.



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.

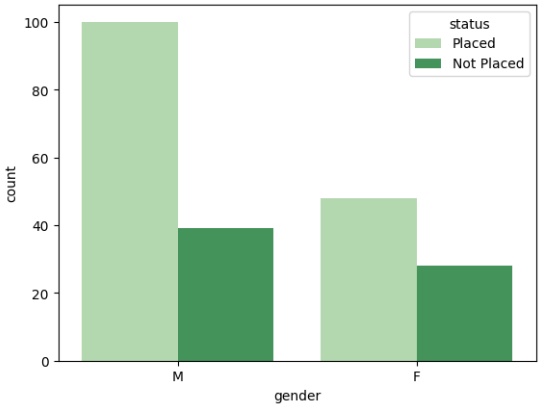
Da queste considerazioni abbiamo deciso di rimuoverla.



MBA Percentage

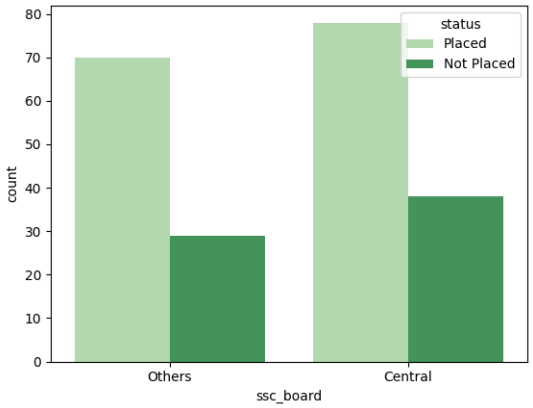
Come si nota da questo grafico la variabile *MBA Percentage* influisce di poco nel dataset, perchè all'aumentare del valore aumenta leggermente essere presi o non.

Da questa considerazione abbiamo deciso di rimuovere la variabile.



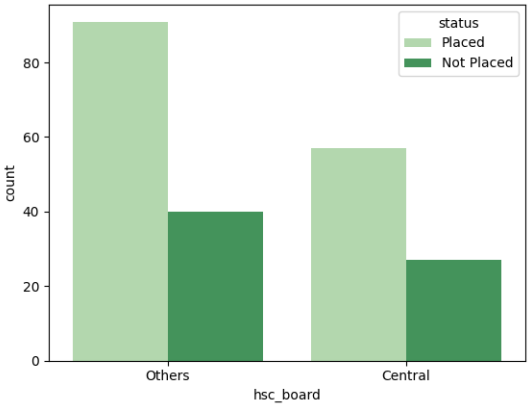
Gender

Come è evidente dal grafico *Gender* influisce notevolmente nel dataset , talmente tanto che il numero di uomini piazzati è maggiore rispetto alle donne in proporzione. Questo influenzerebbe il modello, portandolo anche a dare una previsione discriminatoria per un gender. Alla luce di questo abbiamo deciso di rimuovere la variabile *Gender* così da rendere il modello quanto più imparziare possibile da un punto di vista etico.



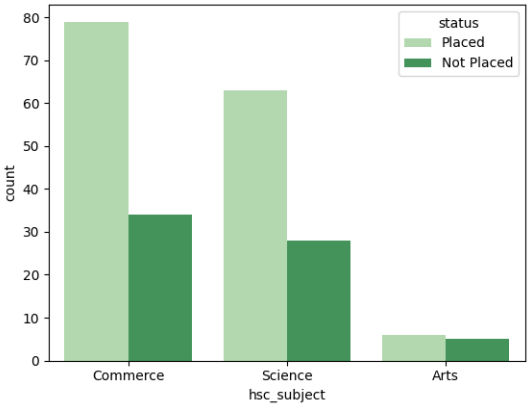
SSC Board

Come si può vedere da grafico, la variabile *SSC Board* influisce discretamente nel dataset, in quanto in base al valore della variabile, non c'è una differenza notevole tra le possibilità di essere preso o non. Quando la variabile ha come valore "Others", abbiamo una probabilità maggiore di essere piazzati rispetto al valore "Central". Nonostante questo, abbiamo deciso di non rimuovere la variabile.



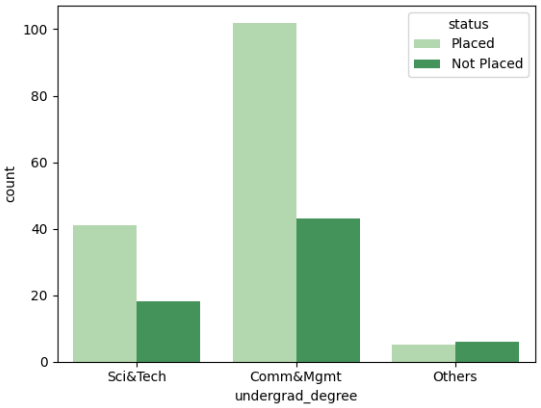
HSC Board

Come si può notare dal grafico, la variabile *HSC Board* influisce notevolmente nel dataset, in quanto in base al valore della variabile, la probabilità di essere piazzato è notevole. In conclusione abbiamo deciso di non rimuovere la variabile.



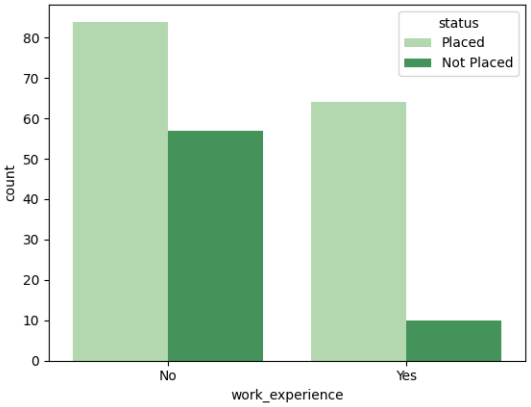
HSC Subject

Come si può vedere dal grafico di sopra, la variabile *HSC Subject* influisce nel dataset, in quanto in base al valore della variabile, si potrebbe avere più possibilità di essere presi o non. In questo caso il valore "Commerce" aumenterebbe le possibilità di essere presi per una persona che ha studiato i questo campo Da questa considerazione abbiamo deciso di non rimuoverla.



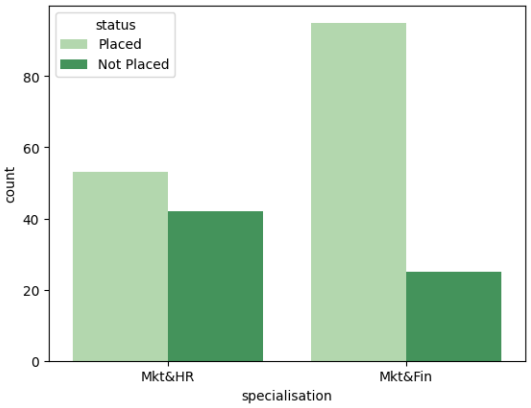
Undergrad

Come si può vedere dal grafico di sopra, la variabile *Undergrade degree* influisce nel dataset, in quanto in base al valore della variabile, si avrà una probabilità maggiore di essere presi o non. Si può notare infatti che le persone che si sono laureate nel campo "Comm&Mgmt" hanno più possibilità di essere presi. Da questa analisi abbiamo deciso di non rimuovere la variabile.



Work Experience

Come si può vedere da questo grafico, la variabile *Work experience* influisce nel dataset. Chi non ha esperienze lavorative, ha più possibilità di essere preso rispetto a chi ne ha. In conclusione abbiamo deciso di non rimuoverla.



Specialisation

In questo grafico la variabile *Specialisation* influisce nel dataset. Chi ha un certo tipo di specializzazione, come "Mkt&Fin", ha più possibilità di essere preso. In conclusione abbiamo deciso di non rimuovere la variabile.

4.3.4 Data balancing