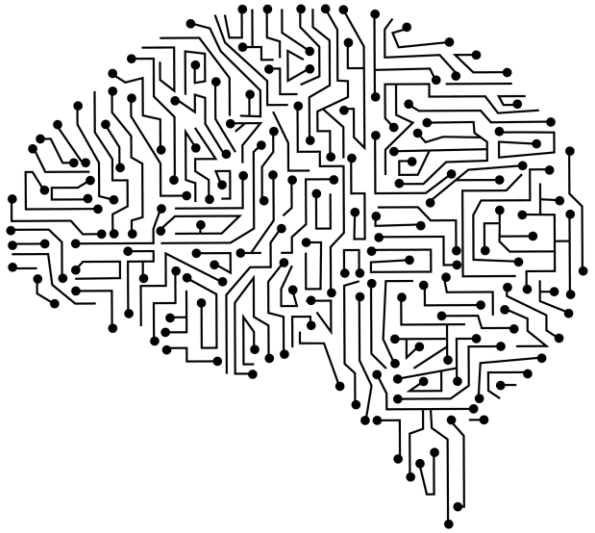
TC Recommender

Progetto di Fondamenti di Intelligenza Artificiale.

A.A. 2023/2024

Una nota iniziale.

Questo è un progetto combinato!



Il progetto di FIA è stato sviluppato come sottosistema di quello di IS.

Il team!

Antonino Lorenzo*

Jacopo Passariello

Claudio Gaudino

**Antonino ha scelto di non fornire una foto*

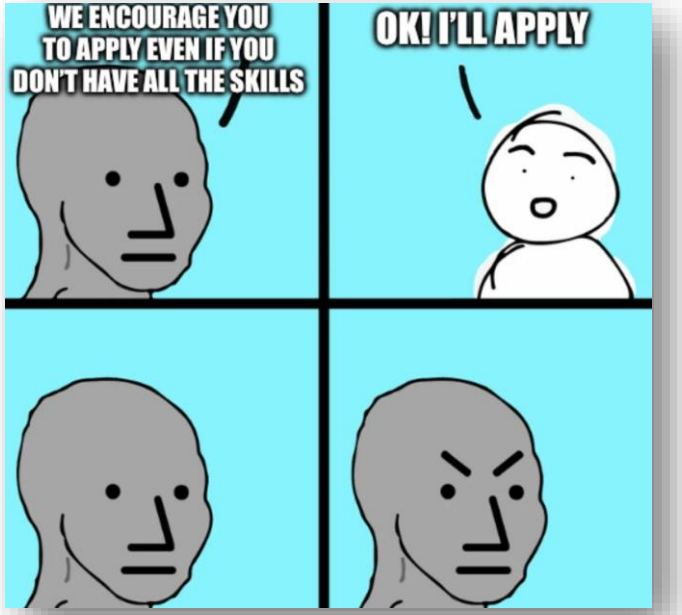
Il team!

E un membro nascosto...

Scopriremo più avanti il misterioso aiutante!

Il problema.

La vita di uno sviluppatore è già molto difficile.
Perché complicarla con estrosi metodi di recruitment?
TUTTI ODIANO IL RECRUITING!



La nostra soluzione!

Il nostro sistema propone una risposta semplice al problema del recruiting:

indeed — Permette a persone e ***aziende*** di ***cercare lavoro***.

GitHub — Permette a ***sviluppatori*** di collaborare.

La nostra soluzione!

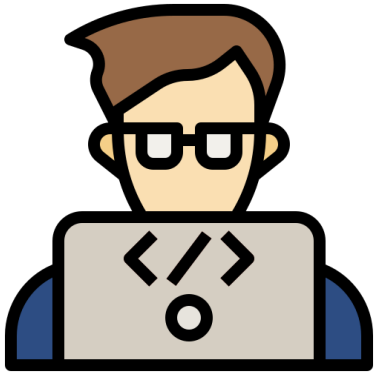
Il nostro sistema propone una risposta semplice al problema del recruiting:

tUring CAREERS

Una piattaforma per ***sviluppatori*** *e* ***software house*** per il recruiting!

Il Ruolo dell'Agente

L'agente sarà integrato come sotto-sistema in Turing Careers, ma qual è il suo ruolo?



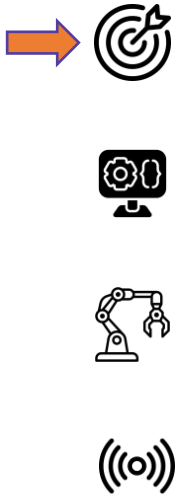
Per gli Sviluppatori:

Raccomandazione di Offerte di Lavoro.

Per i Datori di Lavoro:

Raccomandazione di Profili di Sviluppatori.

La specifica PEAS



Performance:

La capacità di restituire a datori di lavoro i **migliori sviluppatori disponibili** per una data posizione.

Per sviluppatori **la migliore posizione** di lavoro data la loro conoscenza.

La specifica PEAS

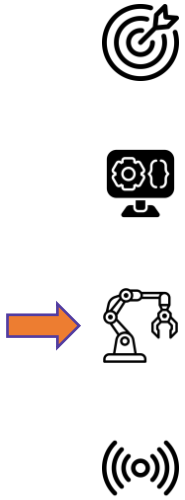
Environment:

Lo spazio delle **offerte di lavoro** e lo **spazio dei profili di sviluppatori** nel sistema.

Caratteristiche dell'Ambiente:

- Completamente Osservabile.
- Deterministico
- Dinamico.
- Discreto.
- Agente Singolo.

La specifica PEAS



Actuators:

L’agente agisce **riportando** le **offerte di lavoro** o i **profili di sviluppatori** all’utente.

La specifica PEAS

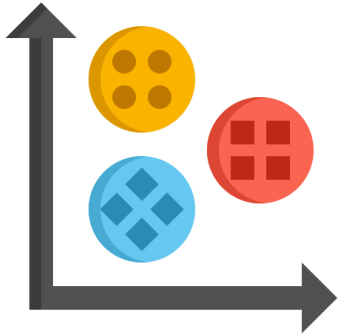
Sensors:

L’agente percepisce input tramite la **barra di ricerca** e le **informazioni dell’utente** o **offerta di lavoro**.

Analisi del Problema

Si è deciso di approcciare il problema tramite **Machine Learning:**

in particolare è un'istanza di **Apprendimento non Supervisionato,**

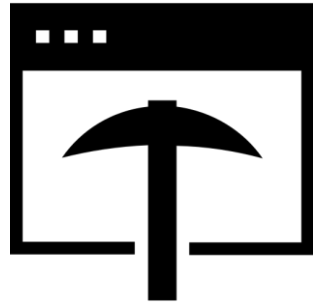


che verrà risolto utilizzando **clustering.**

Acquisizione dei Dati

Come acquisire i dati per il modello? Abbiamo identificato Due Sorgenti.

Abbiamo usato Web Scraping per raccogliere Offerte di Lavoro su **Indeed (592 Istanze)** e una lista di competenze da **StackOverflow (98 Istanze)**.



Abbiamo usato GPT 3.5 per generare un dataset di Sviluppatori (**508 Istanze)**.

*Amici della Postale, siamo **GDPR** compliant!*

Ecco il nostro membro nascosto!

Chat GPT

Non preoccupatevi, ci ha aiutato solo per i dataset.

Data Exploration

Di che tipo sono i dati raccolti?

Offerte: tuple contenenti un titolo, una location e una descrizione.

Name text	Description text	Location text
(Azure) Cloud Solutions Engineer	If you like this offer, please send your CV mentioning the job title to: recruitm...	Valencia, Valencia provincia
(Senior) Fullstack Developer Marketplace (m...	Ready to digitalise retail?Let's Go!(Senior) Fullstack Developer Marketplace (...	Barcelona, Barcelona provincia
18 stagiaires Business Developer APEROL S...	Date: Jan 11, 2024Location: Paris, FRAdditional Location: Moitié Sud de la Fr...	Paris (75)
2024 Intern - Software Development Enginee...	Our CompanyChanging the world through digital experiences is what Adobe's...	Edinburgh EH11
3D Artist	DETAILSDate01/10/2024Contact addressjobs@blackmouthgames.comLoca...	Madrid, Madrid provincia
AI / GenAI Manager (F/H)	L'équipe Accenture Data & AI , est le point focal de l'ensemble des services d...	Paris (75)
AI Developer - MILANO [DIG]	Trasforma le tue aspirazioni professionali in una storia di successo, entra in ...	Italia
AI Engineer	- Paris, Île-de-France, France ↩ - Technology and Research & Development ...	Paris (75)
ARTIFICIAL INTELLIGENCE ENGINEER	Artificial Intelligence Engineer ↩ Location: Rome, Italy ↩ Department: ...	Roma, Lazio
AWS Security Engineer	DescriptionAWS Security EngineerProgramme Name: LCST ↩ Location: B...	(NULL)
AZURE Cloud Engineer (H/F)	ContratCDILocalisationParis La VilletteRéférenceR-35589CatégorieBureaux	La Villette (75)
Advanced Support Engineer- Database	Advanced Support Engineer- Database-2300044XApplicants are required to r...	Madrid, Madrid provincia
Airbus UpNext - Machine Learning Engineer (...	Job Description:The new Airbus UpNext demonstrator will pave the way to s...	Toulouse (31)
Airbus UpNext - Machine Learning Engineer (...	Job Description:The new Airbus UpNext demonstrator will pave the way to s...	Toulouse (31)
AI Engineers-Machine Learning & Deep Lear...	ProgrammazioneScadenza candidatura 29 settembre, 2024Sede lavorativa P...	Padova, Veneto
AI Engineers-Machine Learning & Deep Lear...	ProgrammazioneScadenza candidatura 29 settembre, 2024Sede lavorativa P...	Padova, Veneto
Angular Front-End Developer (Málaga)	What does our company do?:GoldenRace is a global market leader for virtual ...	Málaga, Málaga provincia
Animateur d'atelier - développeur de jeux vid...	Cette offre est en partenariat avec TUMO ↩ Le centre TUMO pour les techn...	13002 Marseille 2e
Appartenente alle categorie protette Softwar...	Annuncio dedicato alle persone appartenenti alle categorie protette - legge 6...	20864 Agrate Brianza
Application Support Technician – Database ...	If you like this offer, please send your CV mentioning the job title to: recruitm...	Brindisi, Puglia
Art Design Project Manager	Department: Art DesignLocation: Barcelona, SpainDublin, IrelandSao Paolo, B...	Barcelona, Barcelona provincia
Assistant Producer	DETAILSDate12/12/2023Contact addressjobs_hr@mercurysteam.comLocati...	Madrid, Madrid provincia
Associate Software Developer	Overview:SITA FOR AIRCRAFT – Associate Software DeveloperLocation: Bar...	08005 Barcelona, Barcelona provincia
Associate Solution Engineer	Here at Appian, our core values of Respect, Work to Impact, Ambition, and Co...	Sevilla, Sevilla provincia
Associate Translation Specialist (German or ...	We are looking for an Associate Translation Specialist to join our talented tra...	Barcelona, Barcelona provincia
Athonet - Elixir Software Engineer for Mobile ...	Athonet - Elixir Software Engineer for Mobile CoreThis role has been designe...	Cernusco sul Naviglio, Lombardia
Azure Cloud Engineer	DP331-2024DPWAY S.r.l. società con esperienza decennale in soluzioni e ser...	Roma, Lazio
Azure Cloud Engineer (Attività di Operation)	by Dolmen Group ↩ ↩ Data: 2024-01-26Luogo: RemoteIn collaborazione ...	(NULL)
BUSINESS DEVELOPER H/F BTOC - BTOB	Wall Street English est un groupe international. Notre réseau de franchises c...	42000 Saint-Étienne

Data Exploration

Di che tipo sono i dati raccolti?

Competenze: tuple contenenti un id, una categoria e un nome.

ID integer	SKILL text	TYPE text
0	JavaScript	Programming Language
1	HTML	Programming Language
2	Python	Programming Language
4	TypeScript	Programming Language
5	Bash	Programming Language
6	Java	Programming Language
7	C#	Programming Language
8	C++	Programming Language
9	C	Programming Language
10	PHP	Programming Language
11	PowerShell	Programming Language
13	Rust	Programming Language
14	Kotlin	Programming Language
15	Ruby	Programming Language

Data Exploration

Di che tipo sono i dati raccolti?

Profili di Sviluppatori: tuple contenenti Informazioni Anagrafiche, una lista di skill, una location (nome, latitudine e longitudine) e una lista di lingue conosciute.

	name	surname	email	password	bio	location	lon	lat	languages	skills
0	Enzo	Brugnaro	enzo.brugnaro@outlook.com	EnzoBrugnaro123_	This is an automatically generated profile for...	Millepini, Italy	11.784508	45.319027	{inglese, spagnolo, italiano}	{angular, oracle, git, react, mysql, webpack, ...
1	Pepito	Ureña	pepito.ureña@gmail.com	PepitoUreña123_	This is an automatically generated profile for...	Murcia, Spain	-1.130654	37.99224	{francese, spagnolo}	{composer, git, django, python, flask, mysql, ...
2	Eugenio	Bernal	eugenio.bernal@yahoo.com	EugenioBernal123_	This is an automatically generated profile for...	Guipúzcoa, Spain	-2.223667	43.07563	{francese, spagnolo}	{typescript, angular, yarn, git, node .js, rea...
3	Marcus	Reed	marcus.reed@gmail.com	MarcusReed123_	This is an automatically generated profile for...	Edinburgh, United Kingdom	-3.188267	55.953252	{inglese}	{python, pandas, pytorch, mongodb, postgresql,...
4	Luc	Paul	luc.paul@gmail.com	LucPaul123_	This is an automatically generated profile for...	Toulouse, France	1.444209	43.604652	{inglese, francese, spagnolo}	{angular, git, css, spring, c++, next .js, kaf...
...	...	...	...	...	...	...	...	...	...	...

Generazione dei Profili

Sono stati generati circa 500 profili di sviluppatore usando Chat GPT.
Si è fatta attenzione a generare profili con competenze in relazione logica.

Aspetti Positivi:

- Non abbiamo infranto la legge.

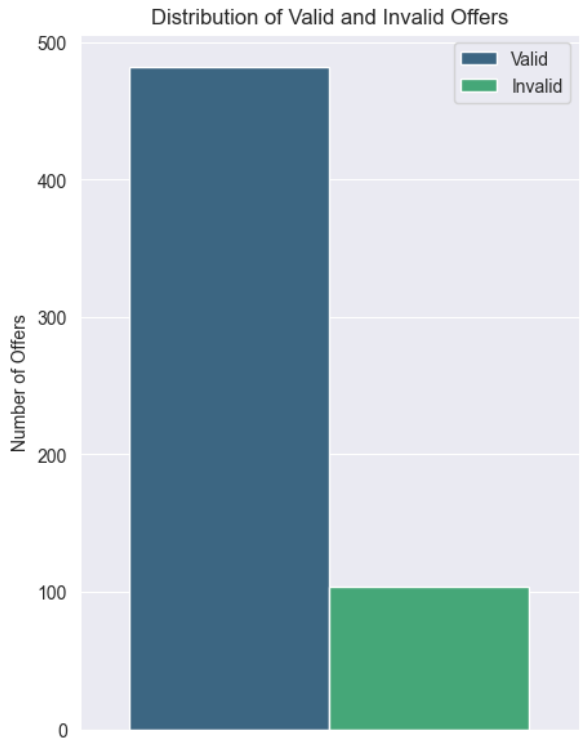
Aspetti Negativi:

- I dati **non** sono totalmente **realistici**.

Data Preparation

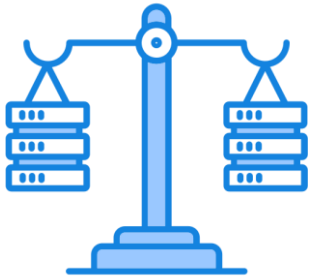
E’ il momento di preparare i dati per essere usati dal modello.

Data Cleaning:
Sono state estratte le skill contenute nelle offerte e rimosse quelle con una lista vuota, lasciandone 482.



Data Preparation

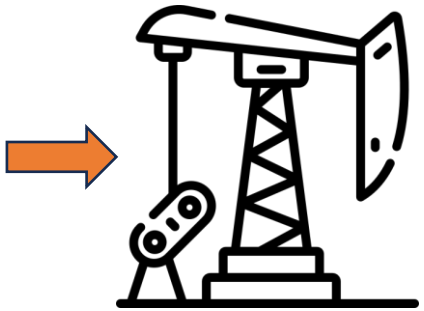
E’ il momento di preparare i dati per essere usati dal modello.



Feature Extraction:

Nelle Offerte:

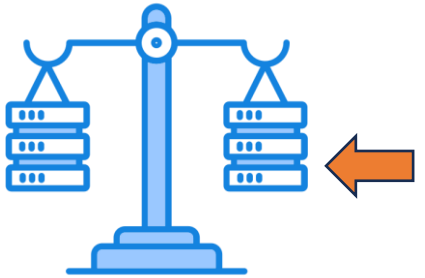
1. E’ stato estratto il **tipo di Location**, che verrà utilizzata nel filtering front-end.
2. Sono state estratte le **Lingue Richieste** dall’offerta, anche queste utilizzate nel filtering front-end.

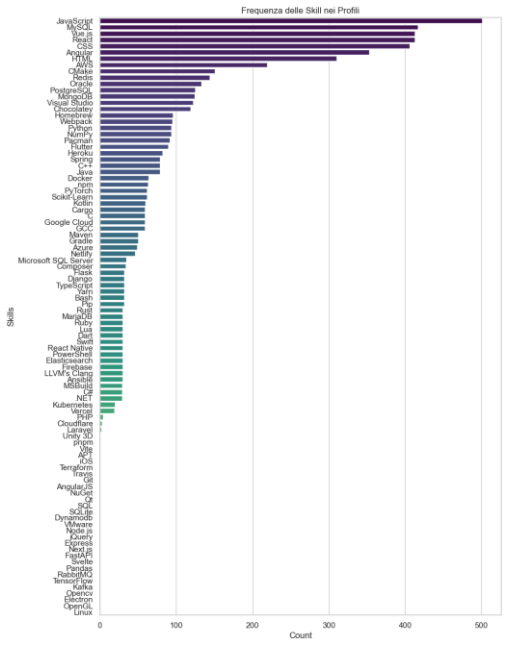


Data Preparation

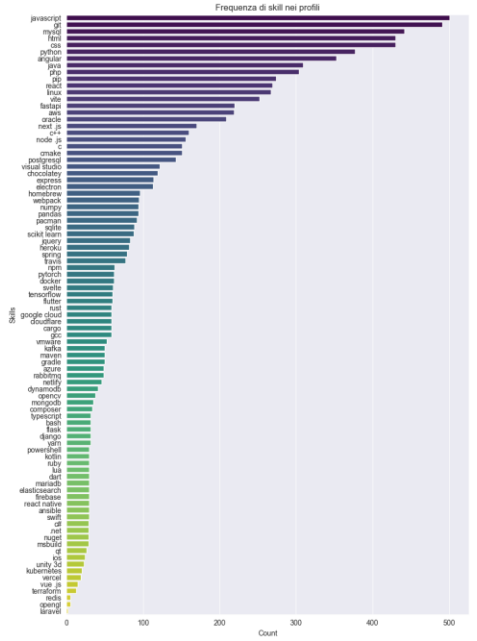
E’ il momento di preparare i dati per essere usati dal modello.

Data Balancing:
Abbiamo identificato distribuzioni di dati irrealistiche nelle competenze e deciso di bilanciare i dati.



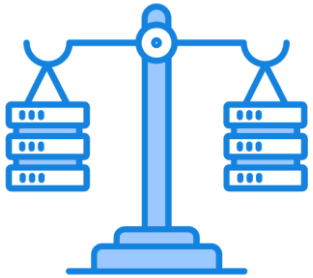


Il bilanciamento ha mantenuto le relazioni logiche tra competenze inizialmente generate.



Data Preparation

E’ il momento di preparare i dati per essere usati dal modello.



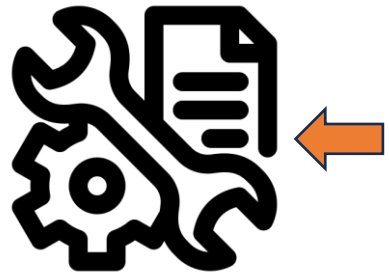
Feature Construction:

Per utilizzare particolari algoritmi, si è scelto di realizzare una matrice delle distanze (sia per offerte che per sviluppatori).

- E’ stata usata come metrica la **Distanza di Jaccard.**
- E’ stata applicata la riduzione della dimensionalità tramite **PCA.**

	0	1	2	3	4	5	6	7	8	9	...	472	473	474	475	476
0	0.0	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	0.666667	1.000000	1.000000	...	1.000000	1.000000	1.000000	1.000000	1.000000
1	1.0	0.000000	0.888889	0.400000	0.846154	0.833333	0.600000	1.000000	1.000000	0.833333	...	1.000000	0.888889	0.777778	0.933333	0.833333
2	1.0	0.888889	0.000000	1.000000	0.636364	0.833333	0.833333	1.000000	0.857143	0.833333	...	0.833333	0.888889	1.000000	0.857143	1.000000
3	1.0	0.400000	1.000000	0.000000	1.000000	1.000000	0.750000	1.000000	1.000000	1.000000	...	1.000000	0.857143	0.714286	0.923077	0.750000
4	1.0	0.846154	0.636364	1.000000	0.000000	0.800000	0.909091	1.000000	0.700000	0.909091	...	0.800000	0.846154	1.000000	0.687500	1.000000
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
477	1.0	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	...	1.000000	1.000000	1.000000	1.000000	1.000000
478	1.0	0.875000	1.000000	0.833333	1.000000	1.000000	0.800000	1.000000	1.000000	1.000000	...	1.000000	0.875000	0.750000	0.846154	0.500000
479	1.0	0.857143	0.857143	1.000000	0.916667	0.750000	0.750000	1.000000	1.000000	0.750000	...	1.000000	1.000000	0.714286	0.923077	0.750000
480	1.0	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	...	1.000000	1.000000	1.000000	1.000000	1.000000
481	1.0	1.000000	1.000000	1.000000	0.900000	1.000000	1.000000	1.000000	1.000000	1.000000	...	1.000000	0.800000	1.000000	0.909091	1.000000

482 rows × 482 columns



Riduzione della Dimensionalità

Perché PCA?

La riduzione della dimensionalità può causare perdita di **Explainability** del modello.
L'abbiamo scelta comunque in quanto restituisce un aumento del **135%** delle performance dell'algoritmo.

Data Modeling

Abbiamo confrontato i due metodi di Clustering (**partizionale e gerarchico**) per vedere quale fosse il più adatto al nostro problema.

Per gli Algoritmi Partizionali: K-Means (K = 3)

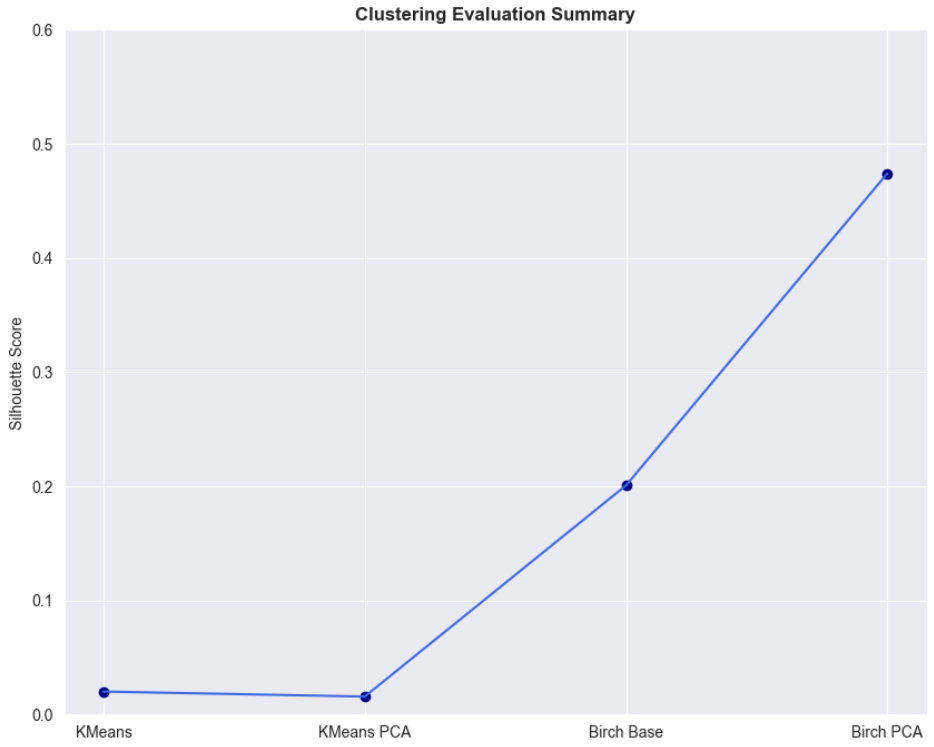
Per gli Algoritmi Gerarchici: BIRCH

Data Modeling

Il confronto tra gli algoritmi:

E’ stata scelta come metrica il **Silhouette Score**:

- Fornisce più informazioni sulla validità dei cluster rispetto ad **Elbow Point**.
- Ottimo per problemi *Intrinsechi* in cui è sconosciuta la **Ground Truth**.

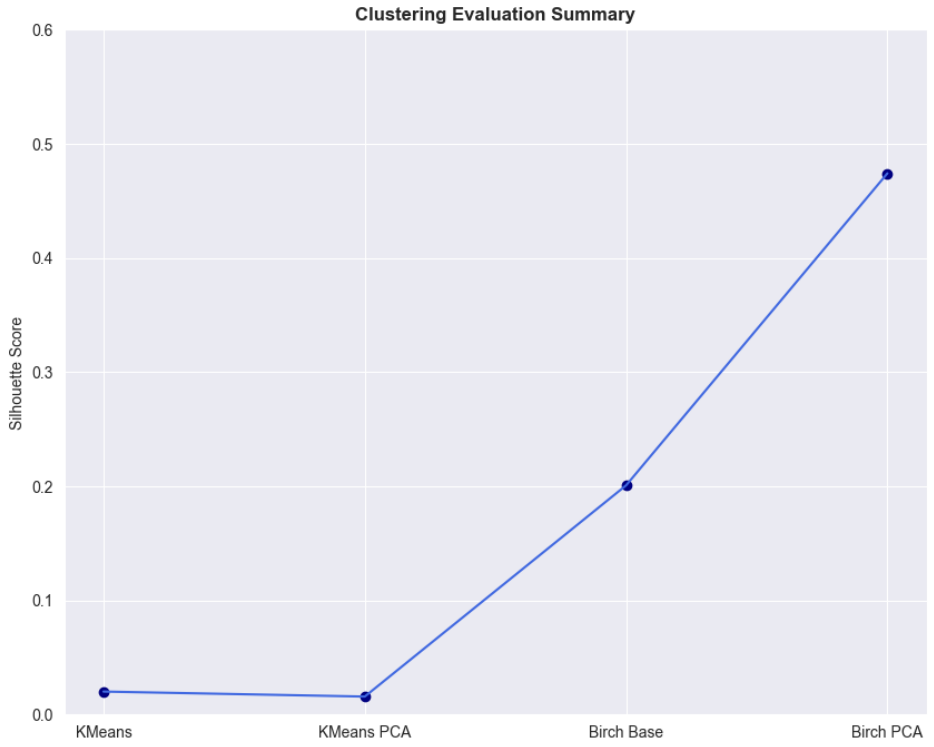


Data Modeling

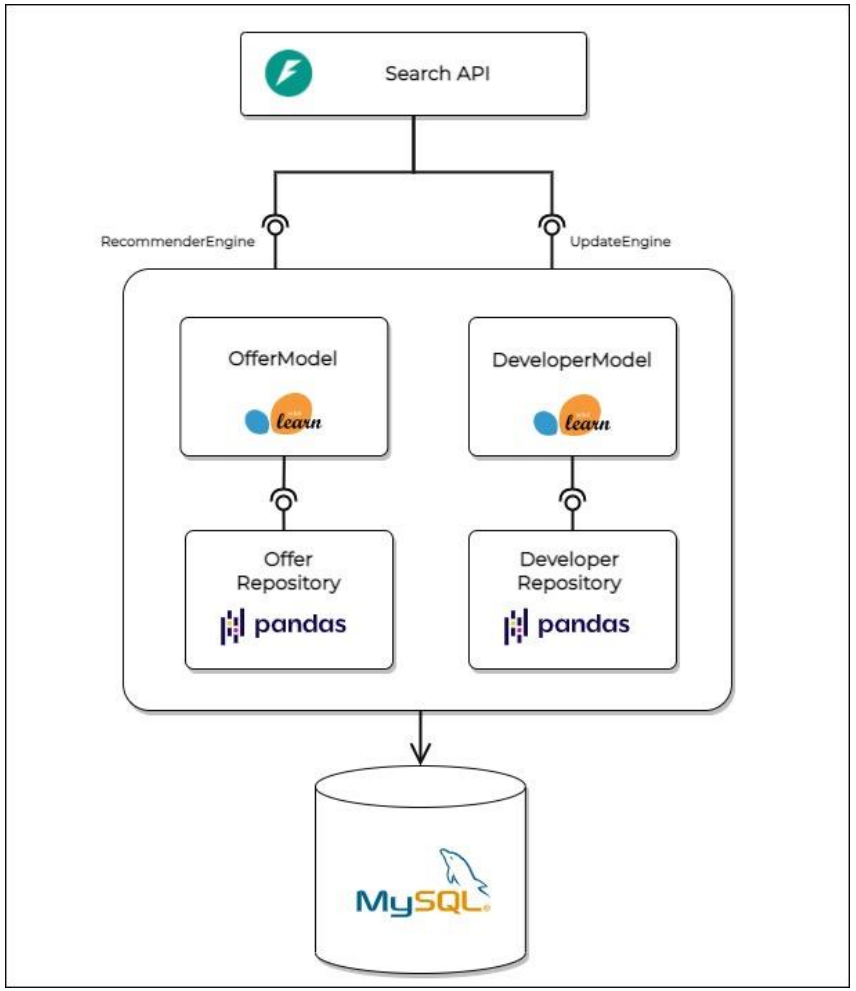
Il confronto tra gli algoritmi:

Il BIRCH vince su K-Means (e di molto).

Si vede inoltre un **massiccio miglioramento** dall'applicazione dell'algoritmo PCA (0,20->0,47).



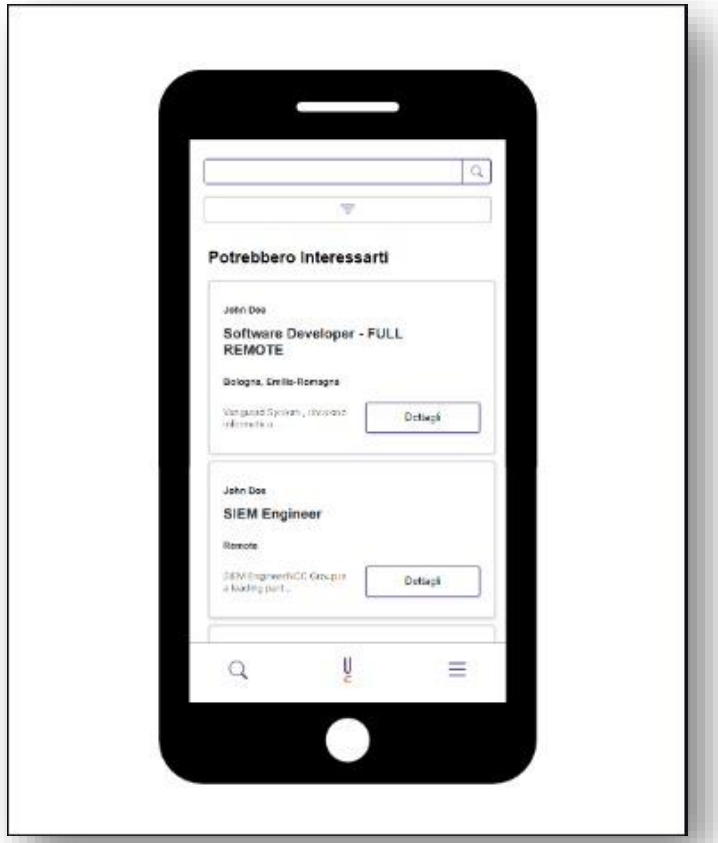
Integrazione con il Sistema

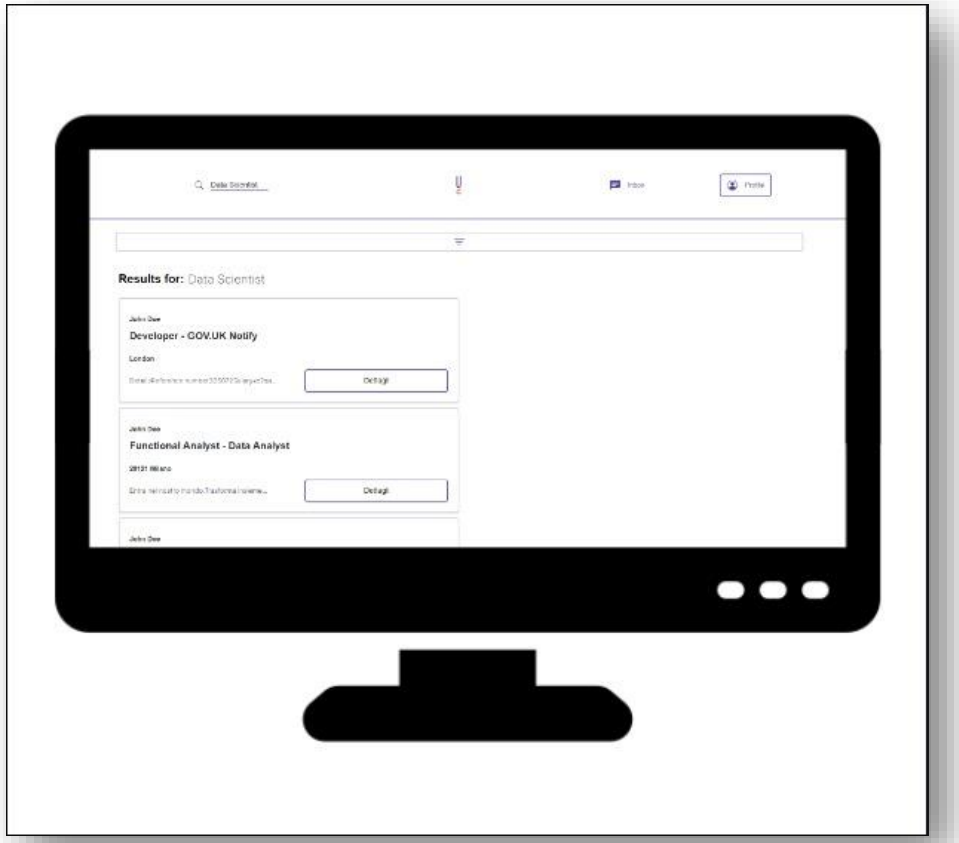


Il deploy del sistema è su un Web Server Uvicorn che espone una **API** alla quale il Sistema Core accede per i servizi di Ricerca e Raccomandazione.

Integrazione con il Sistema

GUI del sistema di Ricerca.





Grazie dell'Attenzione.

Progetto di Fondamenti di Intelligenza Artificiale.

A.A. 2023/2024

Slide Bonus: Risultati del K-Means

Risultati K-Means al variare di K, con e senza PCA:

K Value	Silhouette Score (no PCA)	Silhouette Score (PCA)
3	0.012	0.016
4	0.016	0.005
5	0.020	0.002
6	0.008	0.010

Slide Bonus: Prompt Generativi

Seguono i Prompt utilizzati per la generazione del dataset «Profili di Sviluppatori».

1. **Prompt per la generazione dei linguaggi di programmazione.**

> Genera un dataset di 70 tuple in formato csv di profili di sviluppatori che includano i seguenti attributi: ID, Linguaggi di Programmazione. I linguaggi di programmazione sono in formato di una lista, il cui contenuto varia da 2 a 5 linguaggi di programmazione, correlati tra di loro. I linguaggi di programmazione da utilizzare sono contenuti nella seguente lista:

Lista di Linguaggi di Programmazione

Slide Bonus: Prompt Generativi

Seguono i Prompt utilizzati per la generazione del dataset «Profili di Sviluppatori».

2. Prompt per la generazione dei framework.

> Adesso, partendo dal testo in formato che hai generato, crea una nuova categoria e associa ad essa per ogni tupla da 0 a 3 framework. I framework devono corrispondere alle skill generate. Esempio: uno sviluppatore non può conoscere NumPy se non conosce python. Esempio: se uno sviluppatore conosce Java, allora potrebbe conoscere Spring o JakartaEE.

Ecco la lista:
Lista di Framework

Slide Bonus: Prompt Generativi

Seguono i Prompt utilizzati per la generazione del dataset «Profili di Sviluppatori».

3. Prompt per la generazione dei database.

> Adesso, partendo dal testo in formato che hai generato, crea una nuova categoria "Database" e inserisci da 1 a 2 degli elementi forniti; MySQL e Oracle devono essere i più comuni. Includi gli ultimi tre in almeno 10 tuple.

Lista di Database

Slide Bonus: Prompt Generativi

Seguono i Prompt utilizzati per la generazione del dataset «Profili di Sviluppatori».

4. Prompt per la generazione dei tools.

Adesso, partendo dal testo in formato CSV che hai generato crea una nuova categoria "Tools" e inserisci da 0 a 3 degli elementi forniti:

Lista di Tool

Slide Bonus: Prompt Generativi

Seguono i Prompt utilizzati per la generazione del dataset «Profili di Sviluppatori».

5. Prompt per la generazione delle piattaforme Cloud.

Adesso, partendo dal testo in formato CSV che hai generato crea una nuova categoria "Cloud" e inserisci da 0 a 1 degli elementi forniti:

Lista di Piattaforme Cloud

Slide Bonus: Prompt Generativi

Seguono i Prompt utilizzati per la generazione del dataset «Profili di Sviluppatori».

6. Prompt di Estensione del dataset:

Genera altre 50 tuple ma differenti dalle precedenti, numera gli ID a partire da **ultimo id**.

Slide Bonus: Scraping Guidelines

1. **Agire da buon cittadino digitale**:
   Non sovraccaricare il sito limitando il numero di richieste al minuto.
2. **Disponibilità pubblica dei dati**:
   I dati raccolti devono essere pubblicamente accessibili senza registrarsi al sito.
3. **Informazioni che non violano copyright**:
   La raccolta dati non deve violare il diritto d'autore dell'entità giuridica che possiede il sito.
4. **Dati di natura non personale**:
   È illegale secondo il GDPR effettuare scraping di dati sensibili di individui.
5. I **dati raccolti non devono essere usati per danneggiare il sito dal quale si raccolgono:**
   I dati raccolti vanno usati in modo *trasformativo*, cioè per creare nuovi prodotti e non cannibalizzare quote di mercato dal sito da cui si fa scraping.

Slide Bonus: meme

Questa sessione ci ha resi fumatori.