# Progetto per Laboratorio di Big Data, Data Mining e Data Analytics

**PRESENTAZIONE DEL DATASET:**

Il dataset scelto è stato scaricato da: https://www.kaggle.com/datasets/hummaamqaasim/jobs-in-data

Al suo completo rappresenta il salario registrato tra gli anni 2020-2023 in base al lavoro svolto, classificandolo in categorie differenti, e specificandone la valuta.

Inoltre per ogni lavoratore specifica la residenza, il livello e il tipo di esperienza lavorativa, la modalità di lavoro, dov’è situata e la grandezza della compagnia per cui lavora.

I dati utilizzati sono stati raccolti da utenti kaggle e provengono da diverse aziende ed enti di ricerca che condividono annualmente i loro dati.

**OBBIETTIVI DEL PROGETTO:**

1)Evidenziare le differenze salariali medie tra l’anno 2020 e l’anno 2023, raggruppandole prima nelle categorie di lavoro a disposizione, poi nei singoli lavori.

2)Progettare poi uno script che tramite i salari e le categorie lavorative di appartenenza registrate nel 2023 riesca, grazie alle tecniche viste durante questo corso, ad indovinare il livello di esperienza del lavoratore, in base alle probabilità ottenute.

Rappresentare poi la Confusion Matrix e indicarne l’accuratezza.

I salari e le categorie per i test dovranno essere ottenuti dividendo la lista salariale del 2023 del dataset in 2 parti: salari e categorie per il Train del nostro modello, salari e categorie per i Test.

Tutti i salari dovranno avere come valuta l’euro.

**CONCLUSIONI:**

1)Confrontando le **categorie** e il loro salario medio tra il 2020 e il 2023 notiamo che:

**Data Analysis** passa da un salario medio di circa 34300€ a 54500€, abbiamo perciò un **aumento** di circa 20200€.

**Data Science and Research** passa da un salario medio di circa 55300€ a 50200€, abbiamo perciò una **diminuzione** di circa 5100€.

**Data Engineering** passa da un salario medio di circa 62600€ a 62100€, abbiamo perciò una **diminuzione** di circa 500€.

**Machine Learning and AI** passa da un salario medio di circa 32700€ a 80600€, abbiamo perciò un **aumento notevole** di circa 47900€.

Dal 2020 al 2023 Data Science and Research e Data Engineering hanno subito una diminuzione del salario, mentre Machine Learning and AI e Data Analysis sono aumentate.

Inoltre nel 2023 sono state registrate tre categorie lavorative in più:

**Leadership and Management** con circa 52400€ di stipendio medio, **BI and Visualization** con circa 60000€ e **Data Quality and Operations** con circa 22000€.

Confrontando invece i **lavori** e il loro stipendio medio tra il 2020 e il 2023 notiamo che:

**Data Analyst** passa da un salario medio di circa 34300€ a 43400€, abbiamo perciò un **aumento** di circa 9100€.

**Data Scientist** passa da un salario medio di circa 48500€ a 47100€, abbiamo perciò una **diminuzione** di circa 1400€.

**Data Engineer** passa da un salario medio di circa 56400€ a 61400€, abbiamo perciò un **aumento** di circa 5000€.

**Machine Learning Infrastructure Engineer** passa da un salario medio di circa 44000€ a 85000€, abbiamo perciò un **aumento notevole** di circa 41000€.

**ML Engineer** ha un salario medio di circa 14000€ , nel 2023 non esistono dati associati a questo lavoro con valuta in EURO.

**Big Data Engineer** passa da un salario medio di circa 100000€ a 45000€, abbiamo perciò una **notevole** **diminuzione** di circa 55000€.

**Principal Data Scientist** passa da un salario medio di circa 130000€ a 80000€, abbiamo perciò una **notevole diminuzione** di circa 50000€.

**Machine Learning Engineer** passa da un salario medio di circa 40000€ a 85300€ abbiamo perciò un **aumento** di circa 45300€.

Nel 2023 sono stati registrati 26 lavori in più rispetto all’anno 2020, per visualizzare il loro stipendio si consiglia di utilizzare il grafico a barre, oppure di inserire il codice “print(“%s ha come salario: %d” % (lavori2[j], y[“salary”].mean()))” alla fine del for di riga 71(fare una lista di 26 lavori con il loro salario medio renderebbe la lettura molto complicata).

2)Abbiamo fatto eseguire alla regressione logistica una previsione su 39 valori casuali contenuti nel nostro dataset (ovvero il 25% dei dati totali contenuti al suo interno), di questi 39 dati le previsioni corrette sono state 26, abbiamo avuto perciò una accuratezza del circa 67%.

In particolare su 0 “Executive” ne ha previsti correttamente 0, su 2 “Entry\_level” ne ha previsto 1 correttamente, mentre l’altro è stato classificato come “Executive”.

Ha avuto inoltre una notevole difficoltà nel prevedere la categoria “Mid\_level”, in quanto su 9 Mid\_level ne ha previsti 2 “Entry\_level” e 7 “Senior”, mentre su 28 “Senior” ne ha previsti correttamente 25, classificandone poi 2 “Executive” e uno “Entry\_level”.

Abbiamo fatto fare una previsione molto difficile alla nostra regressione logistica e perciò ci potevamo aspettare una bassa accuratezza, in quanto doveva provare a classificare i lavoratori in ben 4 livelli di esperienza differenti, tramite un salario e 7 categorie lavorative altrettanto differenti.

Possiamo dedurre che **non conviene** utilizzare la regressione logistica quando si devono ottenere previsioni su dati molto differenti tra loro.