Matteo Lancellotti 574016
Flavio Mangione 575356

Progetto Basi di Dati e Big Data

In un'epoca contraddistinta da progressi tecnologici e dall'importanza sempre maggiore dei dati, è fondamentale comprendere le tendenze e le opportunità lavorative nel campo dei Big Data. Il seguente progetto si focalizza sull'analisi di due dataset principali reperibili su Kaggle, "Cyber Salaries" e "Data Science Salaries", con l'obiettivo di esaminare e confrontare le dinamiche salariali nei settori della sicurezza informatica e della Data Science. Questa analisi si rivela particolarmente rilevante in quanto fornisce informazioni preziose sulle prospettive lavorative a livello sia nazionale che internazionale, contribuendo a scelte consapevoli riguardo al percorso accademico e professionale.

Dettagli dei Dataset
1. Dataset Cyber Salaries

Fonte: Kaggle

VARIABILE	TIPO	DESCRIZIONE
work_year	Integer	Anno lavorativo (2020-2023)
experience_level	String	Livello di esperienza (Junior, Mid-level, Senior, Executive)
employment_type	String	Tipo di contratto di lavoro (Esempio: Tempo pieno, Part-time)
job_title	String	Titolo lavorativo
salary	Integer	Importo del salario annuo nella valuta originale
salary_currency	String	Valuta del salario
salary_in_usd	Integer	Salario annuo in dollari USA.
employee_residence	String	Paese di residenza del lavoratore
remote_ratio	Integer	Percentuale di attività lavorativa svolta in remoto
company_location	String	Ubicazione dell'azienda
company_size	String	Dimensione dell'azienda (Piccola, Media, Grande)

2. Dataset Data Science Salaries

Fonte: Kaggle

VARIABILE	TIPO	DESCRIZIONE
work_year	Integer	Anno lavorativo (2020-2023)
experience_level	String	Livello di esperienza (Junior, Mid-level, Senior, Executive)
employment_type	String	Tipo di contratto di lavoro (Esempio: Tempo pieno, Part-time)
job_title	String	Titolo lavorativo
salary	Integer	Importo del salario annuo nella valuta originale
salary_currency	String	Valuta del salario
salary_in_usd	Integer	Salario annuo in dollari USA.
employee_residence	String	Paese di residenza del lavoratore
remote_ratio	Integer	Percentuale di attività lavorativa svolta in remoto
company_location	String	Ubicazione dell'azienda
company_size	String	Dimensione dell'azienda (Piccola, Media, Grande).

L'analisi comparativa di questi dati è fondamentale per delineare il panorama delle opportunità professionali. Tramite lo studio di questi dati si intende quindi fornire agli studenti uno strumento di orientamento e analisi che li guidi nelle loro scelte educative e professionali, evidenziando le opportunità nel mercato del lavoro italiano e internazionale nel contesto dei settori in rapida evoluzione come Cyber Security e Data Science.

Per l'analisi sono stati impiegati vari strumenti di data analysis, tra cui SQLite e Orange Data Mining. Mediante le query SQL eseguite con SQLite, è stato effettuato un confronto approfondito dei salari medi, espressi in dollari, nei settori della Data Science e della Cyber Security, focalizzandoci sui dipendenti residenti negli stessi paesi. In particolare, sono stati analizzati i salari medi in relazione a diverse variabili: la dimensione delle aziende (company_size), il livello di esperienza dei dipendenti (experience_level), la percentuale di attività lavorativa svolta in remoto (remote_ratio), la posizione delle aziende (company_location) che presentavano una media salariale più elevata e i titoli di lavoro (job_title) con le retribuzioni più alte in ciascun settore. Questi dati sono stati poi visualizzati attraverso metodi di statistica descrittiva, quali istogrammi e diagrammi a barre.

Attraverso l'uso del software Orange Data Mining, è stata effettuata un'analisi approfondita del dataset riguardante i salari nel settore della Data Science.

Questa analisi ha preso in considerazione le correlazioni tra le diverse variabili e ha portato all'implementazione di algoritmi supervisionati, quali gli alberi decisionali, per effettuare analisi predittive. Inoltre, sono state esaminate le regole di associazione derivanti dalle variabili selezionate all'interno del dataset.

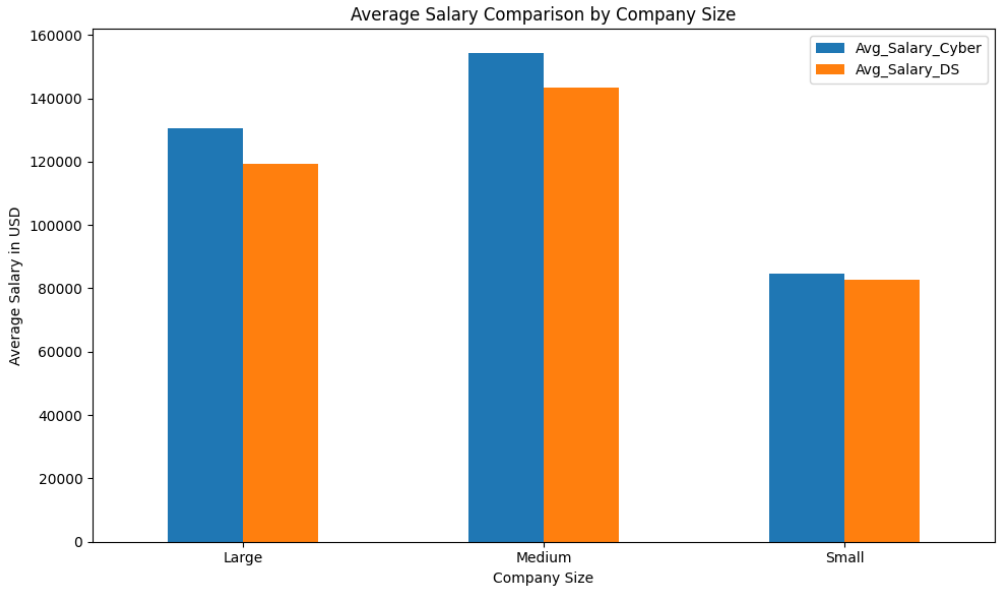


Figura 1 Confronto salari in dollari per dimensione dell'azienda

Nella figura 1, il grafico mostra il confronto dei salari medi tra i settori della Cyber Security e della Data Science suddivisi per dimensione dell'azienda. I salari medi del settore Cyber Security tendono ad essere più alti di quelli del settore Data Science, specialmente per le aziende di grandi e medie dimensioni.

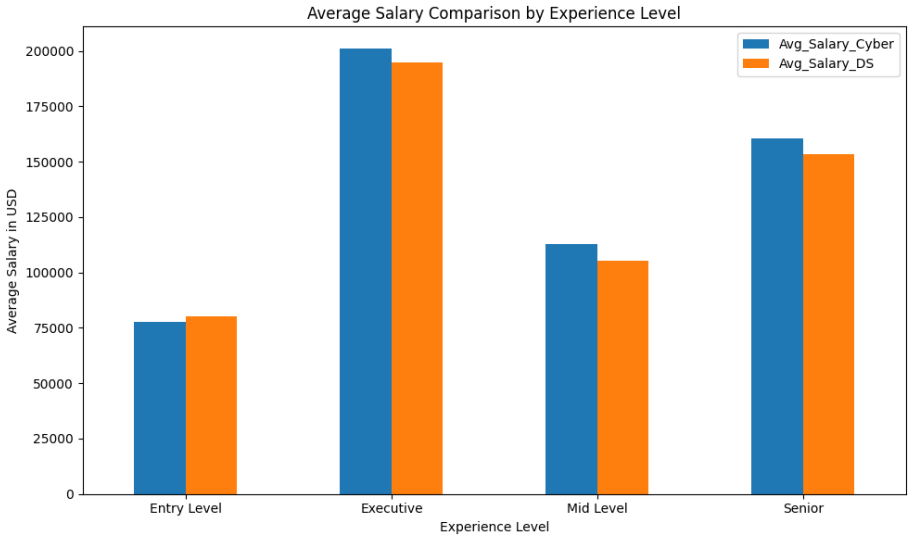


Figura 2 Confronto salari in dollari per i diversi livelli esperienza

Nella figura 2 viene rappresentato il salario medio in dollari classificato per livello di esperienza. Per i livelli Executive, Mid e Senior il settore Cyber Security ha un salario maggiore rispetto al settore Data Science mentre per gli Entry level avviene il contrario.

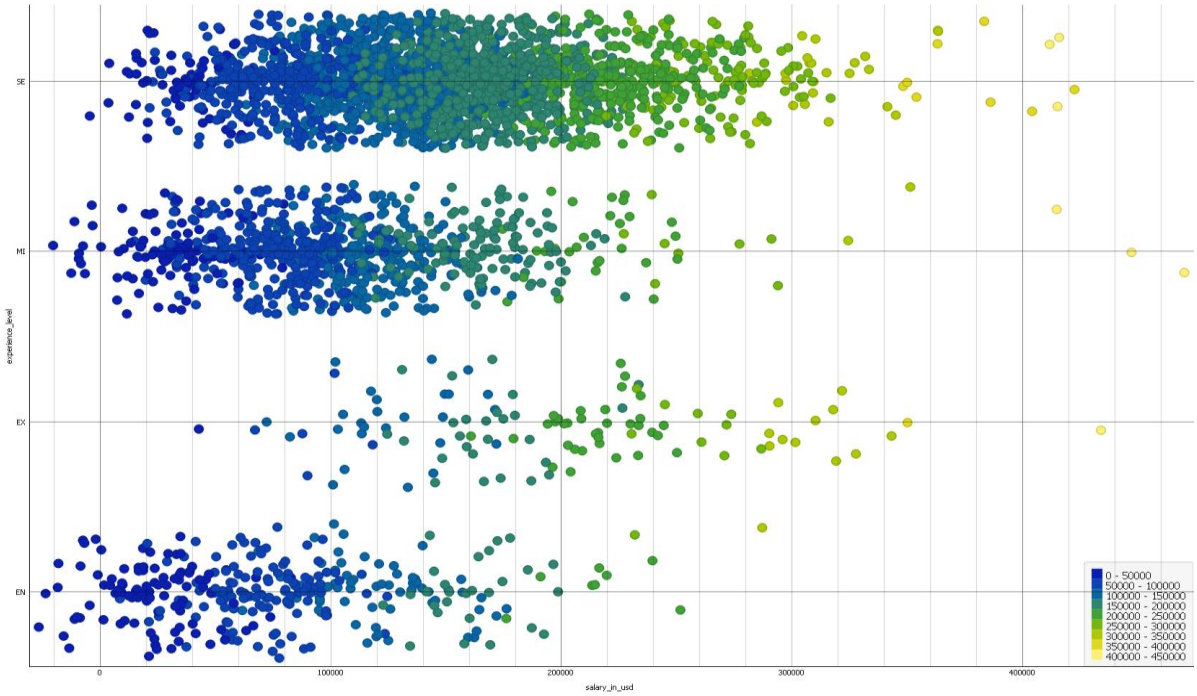


Figura 3 Scatter plot dei salari del settore Data Science per i diversi livelli di esperienza

Nella figura 3, il grafico rappresenta uno Scatter plot avente sull'asse delle ordinate i livelli di esperienza per i dipendenti (experience_level) e sull'asse delle ascisse i valori dei salari in dollari (salary_in_usd). Si evidenzia che i dati sembrano essere più densamente concentrati nella fascia di esperienza Senior e Mid Level, suggerendo che la maggior parte dei dati riguarda individui a metà e fine carriera.

C'è una tendenza generale che mostra come lo stipendio aumenti con l'aumentare dell'esperienza. Tuttavia, c'è una notevole variazione negli stipendi all'interno di ciascuna fascia di esperienza specialmente per la fascia Executive, indicando che l'esperienza non è l'unico fattore che determina lo stipendio. I punti colorati in giallo rappresentano le fasce di stipendi più alti (da 350.000 dollari a 450.000 dollari), sono prevalentemente posizionati nelle fasce di esperienza Senior. Ciò suggerisce che livelli di esperienza più elevati possono portare a stipendi significativamente più alti, sebbene ci siano delle eccezioni.

Si denota che ci sono alcuni valori anomali (outlier), specialmente nella fascia Senior che rappresentano stipendi eccezionalmente alti rispetto alla maggior parte dei dati. I seguenti outlier potrebbero rappresentare ruoli altamente specializzati o posizioni di leadership all'interno di aziende che pagano molto più della media.

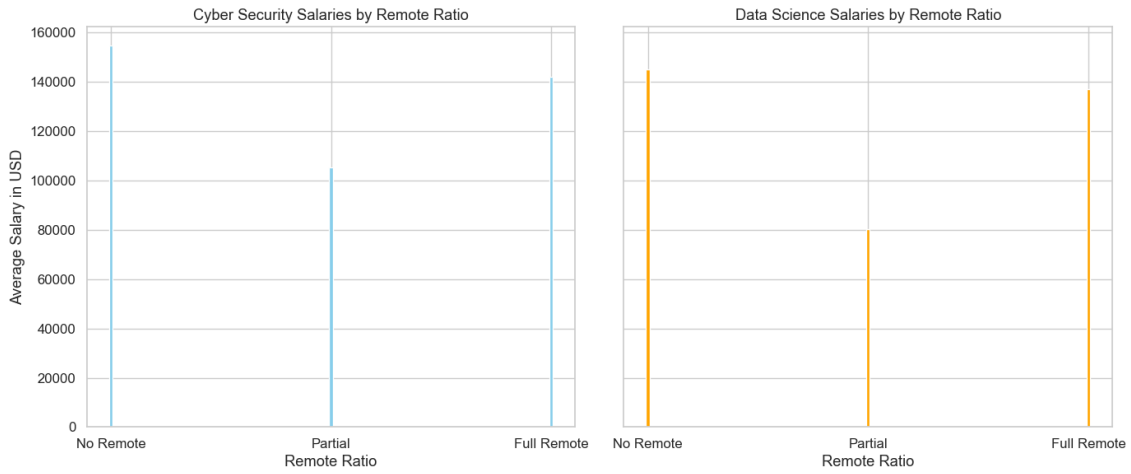


Figura 4 Grafico a Barre confronto salari in dollari in base al rapporto di lavoro remoto

Nella figura 4, il grafico compara i salari medi in dollari in base alla proporzione di lavoro remoto per entrambi i settori. Si evidenzia che sia per il settore della Cyber Security che Data Science, il tipo di lavoro in azienda corrisponde a salari medi più alti rispetto a una tipologia di lavoro mista mentre rispetto a un tipo di lavoro completamente remote, le differenze non sono particolarmente marcate.

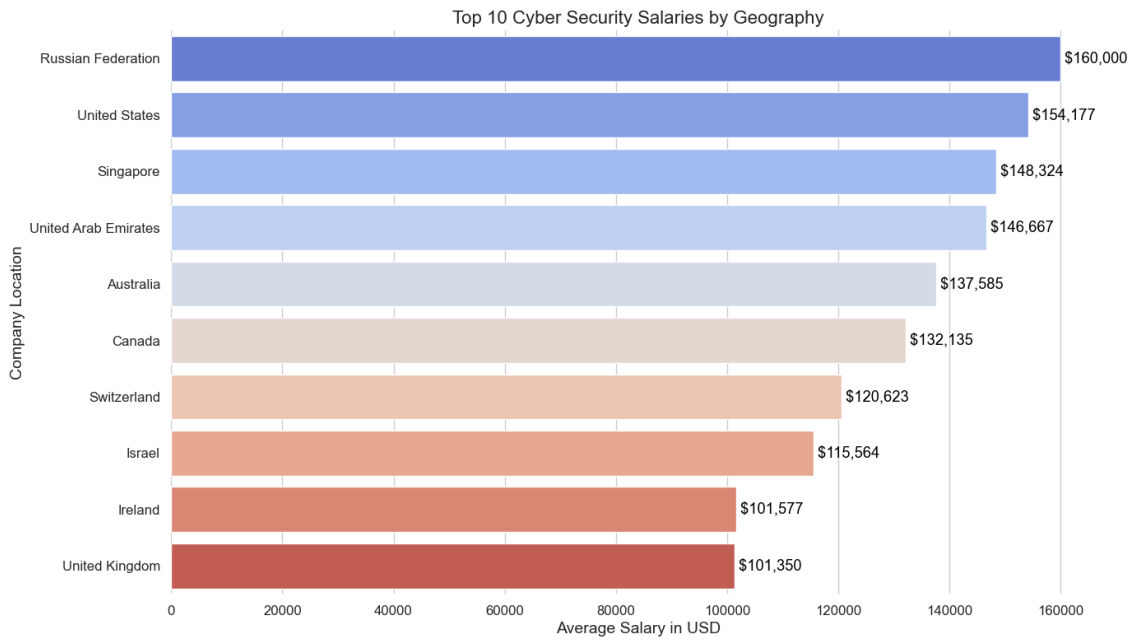


Figura 5 Top 10 dei paesi del settore Cyber Security che guadagnano di più

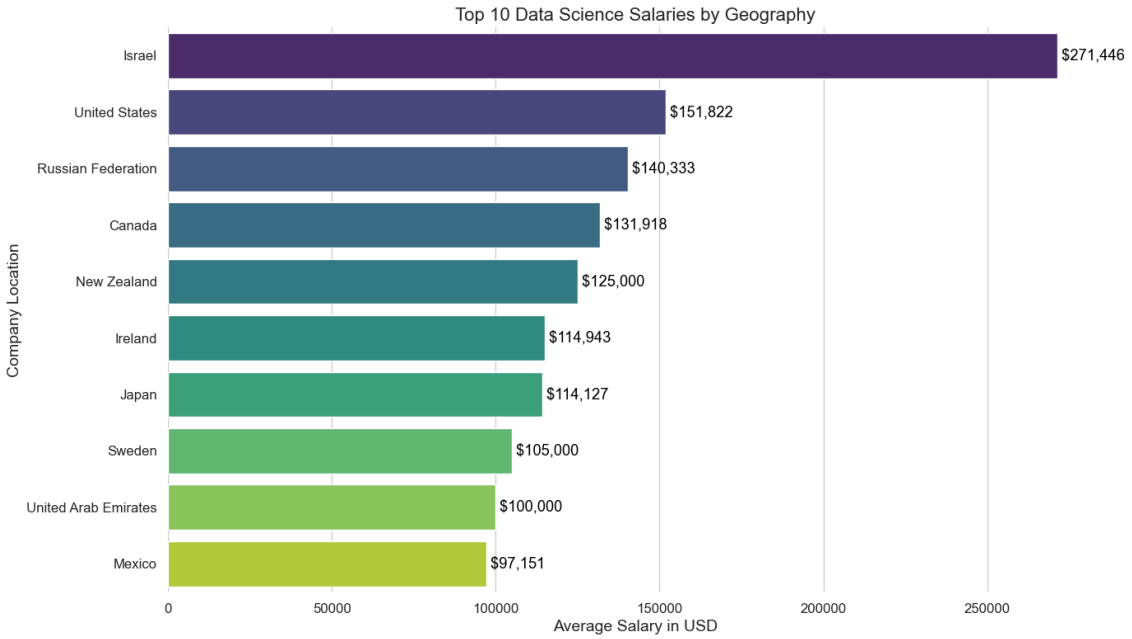


Figura 6 Top 10 dei paesi del settore Data Science che guadagnano di più

I grafici delle figure 5 e 6 offrono uno sguardo dettagliato sul panorama internazionale, svelando le notevoli differenze geografiche nei livelli di remunerazione che possono essere interpretate come conseguenza delle diverse strategie politiche attuate, del livello di investimento in ricerca e sviluppo in ciascun paese.

La Russia emerge come leader salariale nel campo della Cyber Security, un risultato che può essere correlato alla sua consolidata infrastruttura tecnologica che pone la sicurezza informatica al centro delle priorità nazionali.

Gli Stati Uniti mantengono una posizione dominante in entrambi i campi, dimostrando il loro impegno nel promuovere l'innovazione e nel sostenere un ecosistema tecnologico che attrae talenti da tutto il mondo.

Israele a livello di salari si distingue nettamente nel settore Data Science rispetto al settore Cyber Security, dove la combinazione di forti investimenti statali in tecnologia e un ambiente imprenditoriale dinamico ha creato le condizioni ideali per lo sviluppo di una vera e propria industria di analisi dati.

Nel complesso i dati spiegano l'attuale tendenza di mercato che vede una domanda in forte crescita per professionisti altamente qualificati in settori tecnologici avanzati. Una domanda che è stata ulteriormente amplificata dal bisogno di trasformazione digitale scaturito dalla pandemia COVID-19. Diverse organizzazioni in tutto il mondo si sono trovate a dover accelerare il passo verso la digitalizzazione, aumentando la necessità di misure di sicurezza informatica robuste e di analisi di dati sofisticate per migliorare le decisioni aziendali e strategiche.

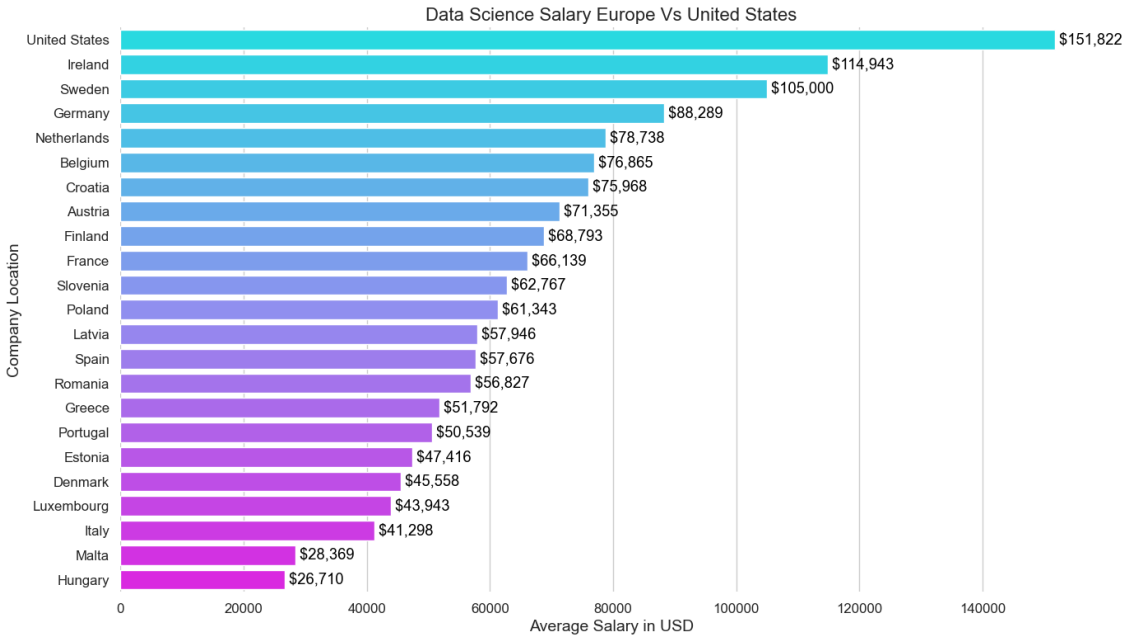


Figura 7 Confronto salari in dollari europei e Stati Uniti del settore Data Science

Il grafico in figura 7 mette a confronto gli stipendi medi lordi nel settore della Data Science in vari paesi europei e negli Stati Uniti. Si può osservare che gli Stati Uniti sono in cima alla classifica con uno stipendio medio lordo annuo di 151.822 dollari, seguiti da Irlanda e Svezia con importi notevolmente inferiori. Gli stipendi medi nei paesi europei variano tra i 50.000 e gli 88.000 dollari lordi annui, con Italia, Malta e Ungheria che si posizionano ai livelli più bassi. Queste differenze possono essere attribuite agli investimenti in tecnologia e innovazione, alla domanda e offerta di competenze in quel campo, oltre alle diverse condizioni economiche, al costo della vita e alle politiche salariali locali che influiscono sulle retribuzioni dei data scientist. Il grafico mette in evidenza la marcata differenza esistente tra i salari nel settore del Data Science negli Stati Uniti e quelli dei paesi appartenenti all'Unione Europea.

Negli ultimi anni, questo settore ha riscontrato una notevole evoluzione, in gran parte dovuta alla rivoluzione dei Big Data. Tale sviluppo ha portato a un'ampia diversificazione dei ruoli e delle competenze richieste, indicando la crescente domanda di professionisti qualificati in grado di gestire e analizzare enormi quantità di dati. Inoltre, la figura 7 mostra l'urgente necessità per i paesi europei di rivalutare le proprie politiche salariali per attrarre e trattenere talenti nel del settore Data Science, un settore chiave per l'innovazione e la crescita economica futura.

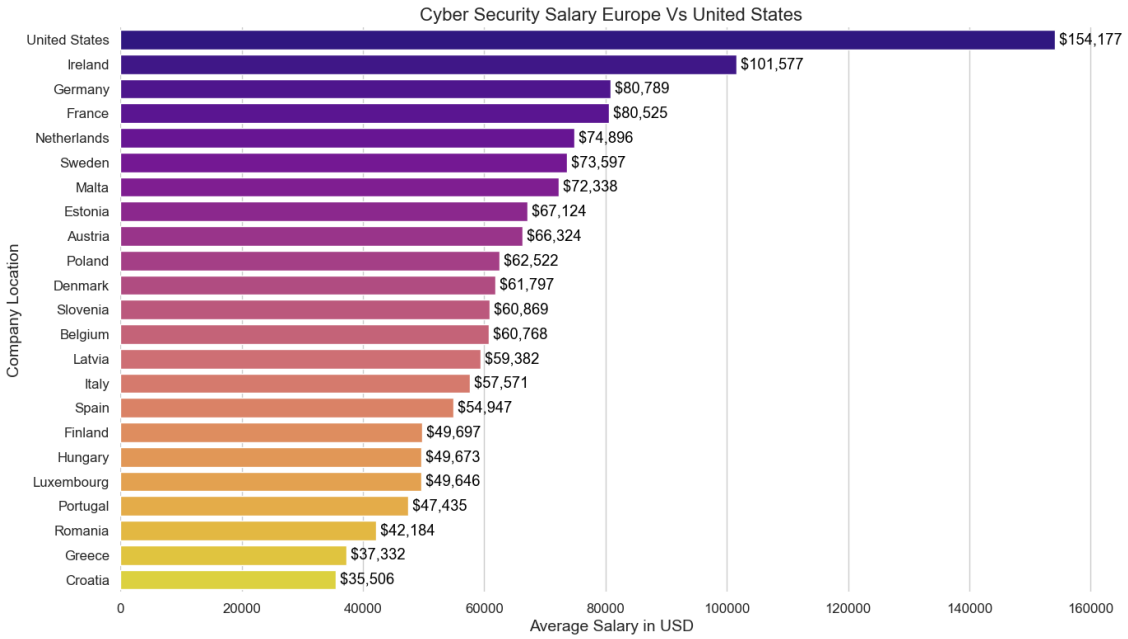


Figura 8 Confronto salari in dollari europei e Stati Uniti del Cyber Security

Il grafico della figura 8 mostra un confronto tra gli stipendi medi lordi nel campo della Cyber Security in diversi paesi europei e negli Stati Uniti. Si può notare che gli Stati Uniti presentano il salario medio lordo più elevato, pari a 154.177 dollari, superando di gran lunga l'Irlanda e la Germania. Nel grafico, l'Italia si colloca nella fascia medio-bassa tra i paesi europei, con uno stipendio medio lordo di 57.571 dollari. Questo suggerisce un'opportunità di crescita nel settore, in linea con l'attuale necessità a livello globale di potenziare la sicurezza informatica nelle infrastrutture.

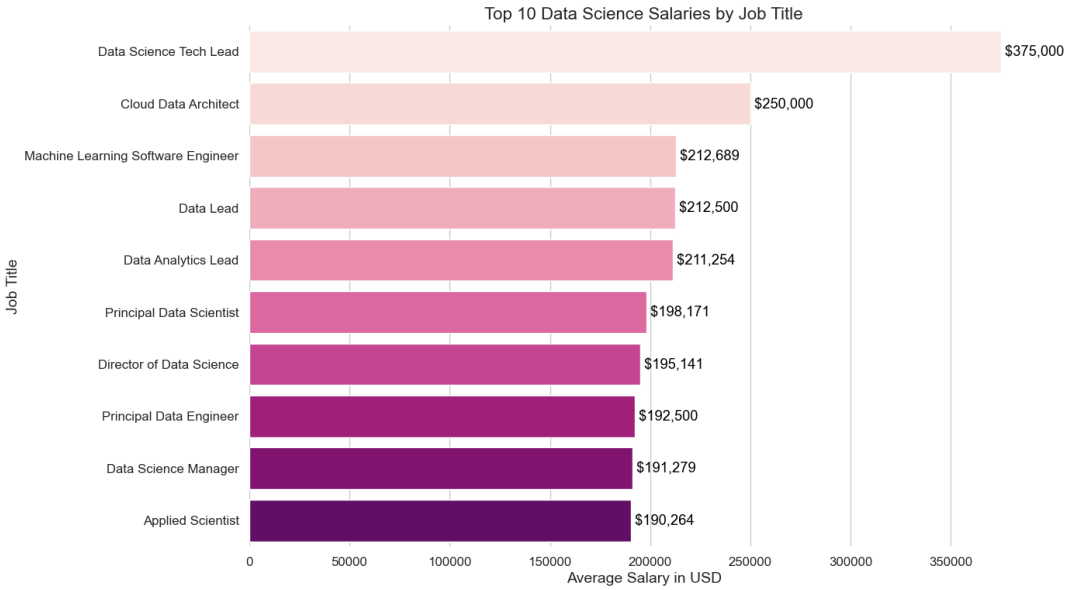


Figura 9 Top 10 dei salari del settore Data Science in base al titolo di lavoro

Nella figura 9, la scala salariale varia considerevolmente, con la posizione del "Data Science Tech Lead" che rappresenta la posizione più remunerativa, con uno stipendio medio di 375.000 dollari. Questo indica un forte valore di mercato per i professionisti in grado di combinare competenze tecniche con capacità di leadership nel settore Data Science. Inoltre la presenza di ruoli tecnici avanzati come "Cloud Data Architect" e "Machine Learning Software Engineer" nella parte superiore della classifica suggerisce che l'industria attribuisce un premio alle competenze specialistiche, in particolare in aree che supportano l'infrastruttura cloud e l'intelligenza artificiale che negli ultimi mesi sta facendo grandi miglioramenti ad esempio come il robot Aloha di Google e il robot Optimus di Elon Musk , i telefoni Android di ultima generazione con l'intelligenza artificiale integrata e i test condotti con il chip della multinazionale Neuralink.

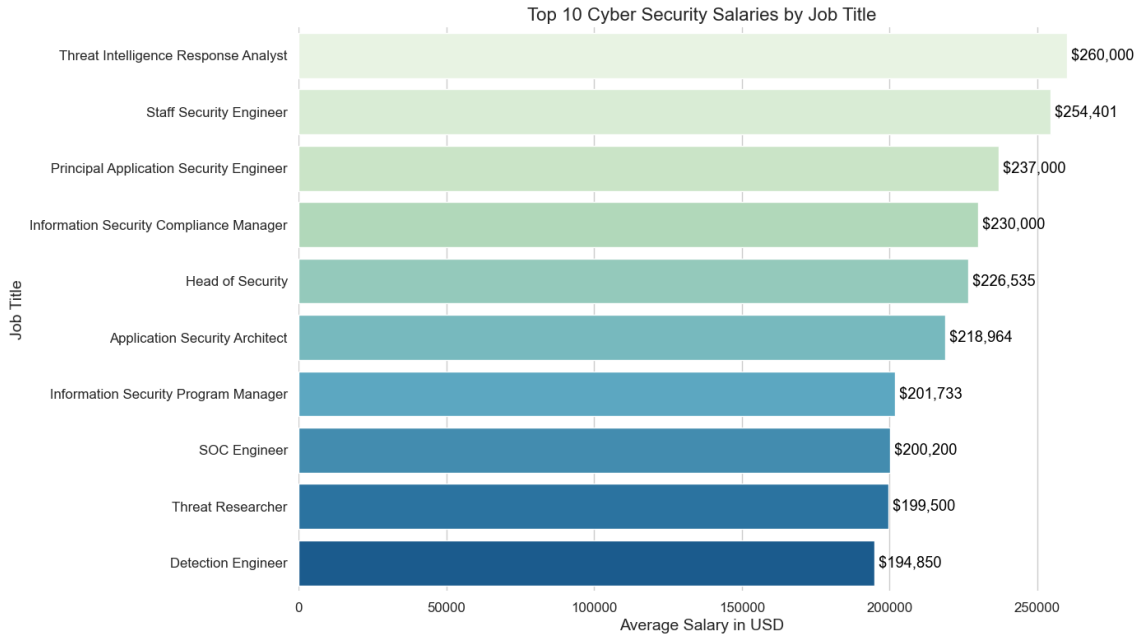
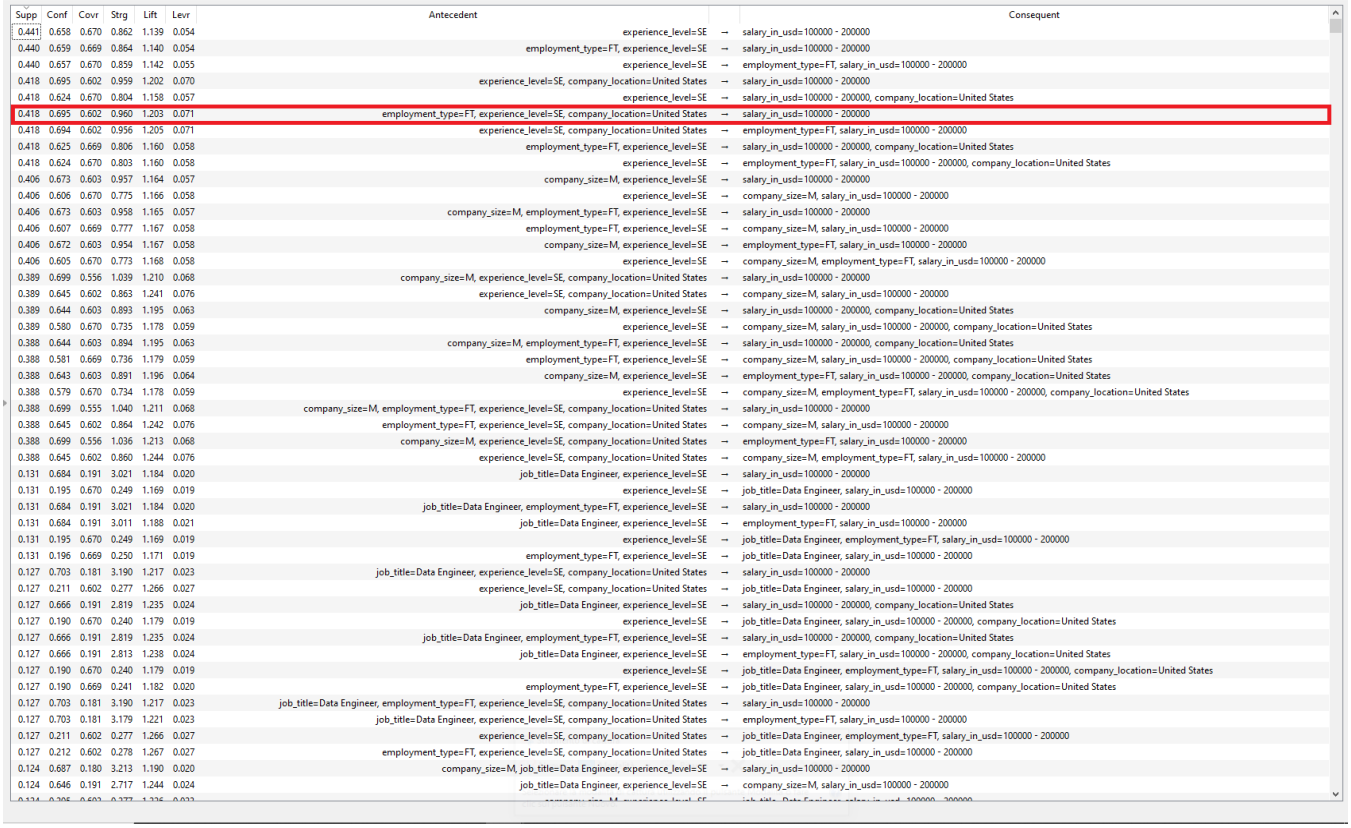


Figura 10 Top 10 dei salari del settore Cyber Security in base al titolo di lavoro

Nella figura 10, si analizza la scala salariale nel settore della Cyber Security. Il ruolo di "Threat Intelligence Response Analyst" emerge come il più remunerativo, con uno stipendio medio di 260.000 dollari, evidenziando l'importanza di una risposta rapida ed efficace alle minacce informatiche.

Ruoli specializzati nel campo della sicurezza, quali "Staff Security Engineer" e "Principal Application Security Engineer", presentano stipendi considerevoli, pari rispettivamente a 254.401 e 237.000 dollari. Ciò indica una crescente domanda di professionisti capaci di ideare e implementare architetture più sicure. Ruoli di responsabilità elevata, come "Head of Security" e "Information Security Compliance Manager", sono anche ben compensati, indicando la priorità delle aziende di garantire la conformità relativa alle normative e la protezione contro le violazioni della sicurezza.

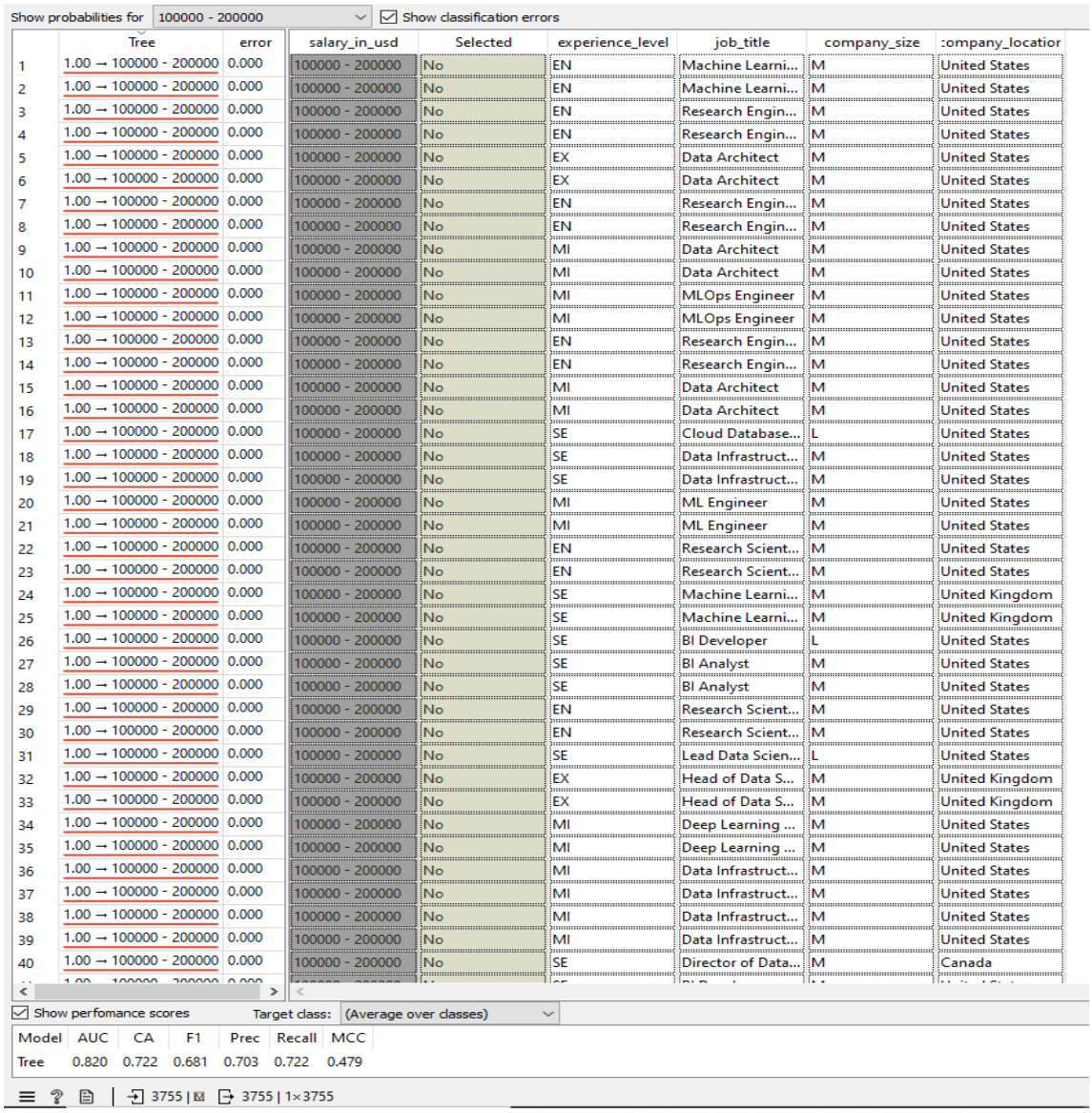


Supp	Conf	Covr	Strg	Lift	Levr	Antecedent		Consequent
0.441	0.658	0.670	0.862	1.139	0.054	experience_level=SE	→	salary_in_usd=100000 - 200000
0.440	0.659	0.669	0.864	1.140	0.054	employment_type=FT, experience_level=SE	→	salary_in_usd=100000 - 200000
0.440	0.657	0.670	0.859	1.142	0.055	experience_level=SE	→	employment_type=FT, salary_in_usd=100000 - 200000
0.418	0.695	0.602	0.959	1.202	0.070	experience_level=SE, company_location=United States	→	salary_in_usd=100000 - 200000
0.418	0.624	0.670	0.804	1.158	0.057	experience_level=SE	→	salary_in_usd=100000 - 200000, company_location=United States
0.418	0.695	0.602	0.960	1.203	0.071	employment_type=FT, experience_level=SE, company_location=United States	→	salary_in_usd=100000 - 200000
0.418	0.694	0.602	0.956	1.205	0.071	experience_level=SE, company_location=United States	→	employment_type=FT, salary_in_usd=100000 - 200000
0.418	0.625	0.669	0.806	1.160	0.058	employment_type=FT, experience_level=SE	→	salary_in_usd=100000 - 200000, company_location=United States
0.418	0.624	0.670	0.803	1.160	0.058	experience_level=SE	→	employment_type=FT, salary_in_usd=100000 - 200000, company_location=United States
0.406	0.673	0.603	0.957	1.164	0.057	company_size=M, experience_level=SE	→	salary_in_usd=100000 - 200000
0.406	0.606	0.670	0.775	1.166	0.058	experience_level=SE	→	company_size=M, salary_in_usd=100000 - 200000
0.406	0.673	0.603	0.958	1.165	0.057	company_size=M, employment_type=FT, experience_level=SE	→	salary_in_usd=100000 - 200000
0.406	0.607	0.669	0.777	1.167	0.058	employment_type=FT, experience_level=SE	→	company_size=M, salary_in_usd=100000 - 200000
0.406	0.672	0.603	0.954	1.167	0.058	company_size=M, experience_level=SE	→	employment_type=FT, salary_in_usd=100000 - 200000
0.406	0.605	0.670	0.773	1.168	0.058	experience_level=SE	→	company_size=M, employment_type=FT, salary_in_usd=100000 - 200000
0.389	0.699	0.556	1.039	1.210	0.068	company_size=M, experience_level=SE, company_location=United States	→	salary_in_usd=100000 - 200000
0.389	0.645	0.602	0.863	1.241	0.076	experience_level=SE, company_location=United States	→	company_size=M, salary_in_usd=100000 - 200000
0.389	0.644	0.603	0.893	1.195	0.063	company_size=M, experience_level=SE	→	salary_in_usd=100000 - 200000, company_location=United States
0.389	0.580	0.670	0.735	1.178	0.059	experience_level=SE	→	company_size=M, salary_in_usd=100000 - 200000, company_location=United States
0.388	0.644	0.603	0.894	1.195	0.063	company_size=M, employment_type=FT, experience_level=SE	→	salary_in_usd=100000 - 200000, company_location=United States
0.388	0.581	0.669	0.736	1.179	0.059	employment_type=FT, experience_level=SE	→	company_size=M, salary_in_usd=100000 - 200000, company_location=United States
0.388	0.643	0.603	0.891	1.196	0.064	company_size=M, experience_level=SE	→	employment_type=FT, salary_in_usd=100000 - 200000, company_location=United States
0.388	0.579	0.670	0.734	1.178	0.059	experience_level=SE	→	company_size=M, employment_type=FT, salary_in_usd=100000 - 200000, company_location=United States
0.388	0.699	0.555	1.040	1.211	0.068	company_size=M, employment_type=FT, experience_level=SE, company_location=United States	→	salary_in_usd=100000 - 200000
0.388	0.645	0.602	0.864	1.242	0.076	employment_type=FT, experience_level=SE, company_location=United States	→	company_size=M, salary_in_usd=100000 - 200000
0.388	0.699	0.556	1.036	1.213	0.068	company_size=M, experience_level=SE, company_location=United States	→	employment_type=FT, salary_in_usd=100000 - 200000
0.388	0.645	0.602	0.860	1.244	0.076	experience_level=SE, company_location=United States	→	company_size=M, employment_type=FT, salary_in_usd=100000 - 200000
0.131	0.684	0.191	3.021	1.184	0.020	job_title=Data Engineer, experience_level=SE	→	salary_in_usd=100000 - 200000
0.131	0.195	0.670	0.249	1.169	0.019	experience_level=SE	→	job_title=Data Engineer, salary_in_usd=100000 - 200000
0.131	0.684	0.191	3.021	1.184	0.020	job_title=Data Engineer, employment_type=FT, experience_level=SE	→	salary_in_usd=100000 - 200000
0.131	0.684	0.191	3.011	1.188	0.021	job_title=Data Engineer, experience_level=SE	→	employment_type=FT, salary_in_usd=100000 - 200000
0.131	0.195	0.670	0.249	1.169	0.019	experience_level=SE	→	job_title=Data Engineer, employment_type=FT, salary_in_usd=100000 - 200000
0.131	0.196	0.669	0.250	1.171	0.019	employment_type=FT, experience_level=SE	→	job_title=Data Engineer, salary_in_usd=100000 - 200000
0.127	0.703	0.181	3.190	1.217	0.023	job_title=Data Engineer, experience_level=SE, company_location=United States	→	salary_in_usd=100000 - 200000
0.127	0.211	0.602	0.277	1.266	0.027	experience_level=SE, company_location=United States	→	job_title=Data Engineer, salary_in_usd=100000 - 200000
0.127	0.666	0.191	2.819	1.235	0.024	job_title=Data Engineer, experience_level=SE	→	salary_in_usd=100000 - 200000, company_location=United States
0.127	0.190	0.670	0.240	1.179	0.019	experience_level=SE	→	job_title=Data Engineer, salary_in_usd=100000 - 200000, company_location=United States
0.127	0.666	0.191	2.819	1.235	0.024	job_title=Data Engineer, employment_type=FT, experience_level=SE	→	salary_in_usd=100000 - 200000, company_location=United States
0.127	0.666	0.191	2.813	1.238	0.024	job_title=Data Engineer, experience_level=SE	→	employment_type=FT, salary_in_usd=100000 - 200000, company_location=United States
0.127	0.190	0.670	0.240	1.179	0.019	experience_level=SE	→	job_title=Data Engineer, employment_type=FT, salary_in_usd=100000 - 200000, company_location=United States
0.127	0.190	0.669	0.241	1.182	0.020	employment_type=FT, experience_level=SE	→	job_title=Data Engineer, salary_in_usd=100000 - 200000, company_location=United States
0.127	0.703	0.181	3.190	1.217	0.023	job_title=Data Engineer, employment_type=FT, experience_level=SE, company_location=United States	→	salary_in_usd=100000 - 200000
0.127	0.703	0.181	3.179	1.221	0.023	job_title=Data Engineer, experience_level=SE, company_location=United States	→	employment_type=FT, salary_in_usd=100000 - 200000
0.127	0.211	0.602	0.277	1.266	0.027	experience_level=SE, company_location=United States	→	job_title=Data Engineer, employment_type=FT, salary_in_usd=100000 - 200000
0.127	0.212	0.602	0.278	1.267	0.027	employment_type=FT, experience_level=SE, company_location=United States	→	job_title=Data Engineer, salary_in_usd=100000 - 200000
0.124	0.687	0.180	3.213	1.190	0.020	company_size=M, job_title=Data Engineer, experience_level=SE	→	salary_in_usd=100000 - 200000
0.124	0.646	0.191	2.717	1.244	0.024	job_title=Data Engineer, experience_level=SE	→	company_size=M, salary_in_usd=100000 - 200000

Figura 11 Regole di associazione Dataset Data Science per un dipendente di livello senior in USA

La figura 11 presenta una tabella di regole di associazione relative agli stipendi nel settore della Data Science. La regola in evidenza suggerisce che, negli Stati Uniti, i Data Scientist con un'ampia esperienza (senior) che lavorano a tempo pieno hanno il 60,2% di probabilità di percepire uno stipendio compreso tra 100.000 e 200.000 dollari, una tendenza riscontrata nel 41,8% del dataset. Il lift di 1,203 e il leverage di 0,071 indicano che questa associazione si verifica più spesso di quanto ci si aspetterebbe in caso di indipendenza tra antecedente e conseguente, sottolineando una correlazione significativa tra queste caratteristiche professionali e la fascia salariale considerata.

Tenendo conto dell'economia statunitense, in particolare dei settori tecnologico e dei Big Data, queste cifre possono essere interpretate come un mercato del lavoro in continua evoluzione, con una crescente domanda di esperti in Data Science con esperienza medio-avanzata e con retribuzioni molto vantaggiose. In questo contesto, le competenze tecniche e analitiche di alto livello vengono premiate, incoraggiando lo sviluppo e l'aggiornamento continuo delle abilità nel campo del Data Science.



	Tree	error	salary_in_usd	Selected	experience_level	job_title	company_size	:ompany_locatior
1	1.00 → 100000 - 200000	0.000	100000 - 200000	No	EN	Machine Learni...	M	United States
2	1.00 → 100000 - 200000	0.000	100000 - 200000	No	EN	Machine Learni...	M	United States
3	1.00 → 100000 - 200000	0.000	100000 - 200000	No	EN	Research Engin...	M	United States
4	1.00 → 100000 - 200000	0.000	100000 - 200000	No	EN	Research Engin...	M	United States
5	1.00 → 100000 - 200000	0.000	100000 - 200000	No	EX	Data Architect	M	United States
6	1.00 → 100000 - 200000	0.000	100000 - 200000	No	EX	Data Architect	M	United States
7	1.00 → 100000 - 200000	0.000	100000 - 200000	No	EN	Research Engin...	M	United States
8	1.00 → 100000 - 200000	0.000	100000 - 200000	No	EN	Research Engin...	M	United States
9	1.00 → 100000 - 200000	0.000	100000 - 200000	No	MI	Data Architect	M	United States
10	1.00 → 100000 - 200000	0.000	100000 - 200000	No	MI	Data Architect	M	United States
11	1.00 → 100000 - 200000	0.000	100000 - 200000	No	MI	MLOps Engineer	M	United States
12	1.00 → 100000 - 200000	0.000	100000 - 200000	No	MI	MLOps Engineer	M	United States
13	1.00 → 100000 - 200000	0.000	100000 - 200000	No	EN	Research Engin...	M	United States
14	1.00 → 100000 - 200000	0.000	100000 - 200000	No	EN	Research Engin...	M	United States
15	1.00 → 100000 - 200000	0.000	100000 - 200000	No	MI	Data Architect	M	United States
16	1.00 → 100000 - 200000	0.000	100000 - 200000	No	MI	Data Architect	M	United States
17	1.00 → 100000 - 200000	0.000	100000 - 200000	No	SE	Cloud Database...	L	United States
18	1.00 → 100000 - 200000	0.000	100000 - 200000	No	SE	Data Infrastruct...	M	United States
19	1.00 → 100000 - 200000	0.000	100000 - 200000	No	SE	Data Infrastruct...	M	United States
20	1.00 → 100000 - 200000	0.000	100000 - 200000	No	MI	ML Engineer	M	United States
21	1.00 → 100000 - 200000	0.000	100000 - 200000	No	MI	ML Engineer	M	United States
22	1.00 → 100000 - 200000	0.000	100000 - 200000	No	EN	Research Scient...	M	United States
23	1.00 → 100000 - 200000	0.000	100000 - 200000	No	EN	Research Scient...	M	United States
24	1.00 → 100000 - 200000	0.000	100000 - 200000	No	SE	Machine Learni...	M	United Kingdom
25	1.00 → 100000 - 200000	0.000	100000 - 200000	No	SE	Machine Learni...	M	United Kingdom
26	1.00 → 100000 - 200000	0.000	100000 - 200000	No	SE	BI Developer	L	United States
27	1.00 → 100000 - 200000	0.000	100000 - 200000	No	SE	BI Analyst	M	United States
28	1.00 → 100000 - 200000	0.000	100000 - 200000	No	SE	BI Analyst	M	United States
29	1.00 → 100000 - 200000	0.000	100000 - 200000	No	EN	Research Scient...	M	United States
30	1.00 → 100000 - 200000	0.000	100000 - 200000	No	EN	Research Scient...	M	United States
31	1.00 → 100000 - 200000	0.000	100000 - 200000	No	SE	Lead Data Scien...	L	United States
32	1.00 → 100000 - 200000	0.000	100000 - 200000	No	EX	Head of Data S...	M	United Kingdom
33	1.00 → 100000 - 200000	0.000	100000 - 200000	No	EX	Head of Data S...	M	United Kingdom
34	1.00 → 100000 - 200000	0.000	100000 - 200000	No	MI	Deep Learning ...	M	United States
35	1.00 → 100000 - 200000	0.000	100000 - 200000	No	MI	Deep Learning ...	M	United States
36	1.00 → 100000 - 200000	0.000	100000 - 200000	No	MI	Data Infrastruct...	M	United States
37	1.00 → 100000 - 200000	0.000	100000 - 200000	No	MI	Data Infrastruct...	M	United States
38	1.00 → 100000 - 200000	0.000	100000 - 200000	No	MI	Data Infrastruct...	M	United States
39	1.00 → 100000 - 200000	0.000	100000 - 200000	No	MI	Data Infrastruct...	M	United States
40	1.00 → 100000 - 200000	0.000	100000 - 200000	No	SE	Director of Data...	M	Canada

Figura 12 Predizioni di un Albero Decisionale sul Dataset Salari Data Science

Nella figura 12 è rappresentata l'output di un albero decisionale che effettua predizioni sul salario (espresso in dollari USA) di professionisti nel settore Data Science. Dai parametri mostrati, si osserva che il modello ha buone prestazioni in termini di precisione (70.3%), richiamo (72.2%) e accuratezza (72.2%), indicando che è capace di classificare con una certa affidabilità i salari annuali con valori che intervallano dai 100.000 ai 200.000 dollari, soprattutto negli Stati Uniti e nel Regno Unito. Il modello sembra essere particolarmente efficace per aziende di medie dimensioni e per individui con diversi livelli di esperienza in particolare con i dipendenti con un livello esperienza senior e media. Il valore AUC di 0.820 (82%) suggerisce che il modello ha una buona capacità di predire correttamente le classi di salari annuali più alte e più basse. Tuttavia, si intuisce che potrebbero esserci alcune difficoltà nella classificazione accurata dei salari più bassi, specialmente al di fuori degli Stati Uniti e del Regno Unito.

L'analisi dei salari annuali dei dipendenti nel settore della Data Science, confrontata con quella dei dipendenti nel settore della Cyber Security, evidenzia l'evoluzione rapida del mondo del lavoro, influenzata dall'emergere dei Big Data e, in particolare, dall'introduzione dell'intelligenza artificiale.

La successiva fase dell'analisi si focalizza sullo studio di vari indicatori economici e finanziari, con l'obiettivo di esaminare le correlazioni tra le serie storiche e creare un modello predittivo Var per il Nasdaq. Quest'ultimo è un importante indice azionario che riflette il valore complessivo delle azioni quotate al NASDAQ Stock Market, uno dei più grandi mercati azionari a livello globale, caratterizzato da una forte presenza di aziende tecnologiche.

Nel corso dello studio, vengono utilizzate due tipologie di correlazioni: quella di Pearson e quella di Spearman. La correlazione di Pearson è impiegata per verificare la presenza di una relazione lineare tra due variabili, i valori di -1 e 1 indicano una perfetta relazione lineare tra le variabili. La correlazione di Spearman è adottata per relazioni monotone, lineari e no, dove i valori di -1 e 1 indicano una perfetta relazione monotona. Tuttavia, è importante sottolineare che le analisi delle correlazioni e della causalità di Granger non implicano necessariamente un legame causale tra le serie storiche esaminate.

Al fine di facilitare il confronto tra le diverse serie storiche, i dati sono stati normalizzati. Questo processo ha comportato il ridimensionamento dei valori all'interno di un intervallo compreso tra -1 e 1, permettendo così un'analisi più agevole e omogenea tra le varie serie temporali.

Di seguito una breve descrizione degli indicatori economici e finanziari utilizzati per valutare lo stato e le tendenze dei mercati finanziari.

- VIX (CBOE Volatility Index): Misura giornaliera della volatilità del mercato azionario USA basata sulle opzioni dell'indice S&P 500. Un valore elevato indica alta volatilità, un valore basso indica bassa volatilità. Fonte
- NASDAQ Composite Index: Indice ponderato per capitalizzazione di mercato con oltre 3.000 azioni, focalizzato maggiormente sul settore tecnologico. Monitorato giornalmente, usa il prezzo di chiusura per l'analisi. Fonte
- EFFR (Effective Federal Funds Rate): Tasso d'interesse stabilito mensilmente dalla Federal Reserve. Influenza i tassi di interesse generali e ha impatti significativi sull'economia e sui mercati finanziari. Fonte
- UMCSENT (University of Michigan Consumer Sentiment Index): Indice mensile che misura la fiducia dei consumatori USA. Un incremento indica fiducia nell'economia, una diminuzione può segnalare incertezza o preoccupazioni. Fonte
- UNRATE (Unemployment Rate): Serie storica mensile del tasso di disoccupazione. Indica la percentuale di persone disoccupate rispetto alla forza lavoro totale, influenzando la spesa dei consumatori e le dinamiche economiche. Fonte

Nell'analisi delle serie storiche, i dati sono stati uniti giorno per giorno con l'utilizzo di SQL lite. Per i dati con frequenza mensile, è stato assegnato lo stesso valore a tutti i giorni del mese; in caso di dati mancanti, le serie storiche sono state interpolate utilizzando il valore più vicino.

L'analisi delle serie è stata condotta su base giornaliera, considerando un intervallo temporale dal 1° agosto 2000 al 30 dicembre 2023.

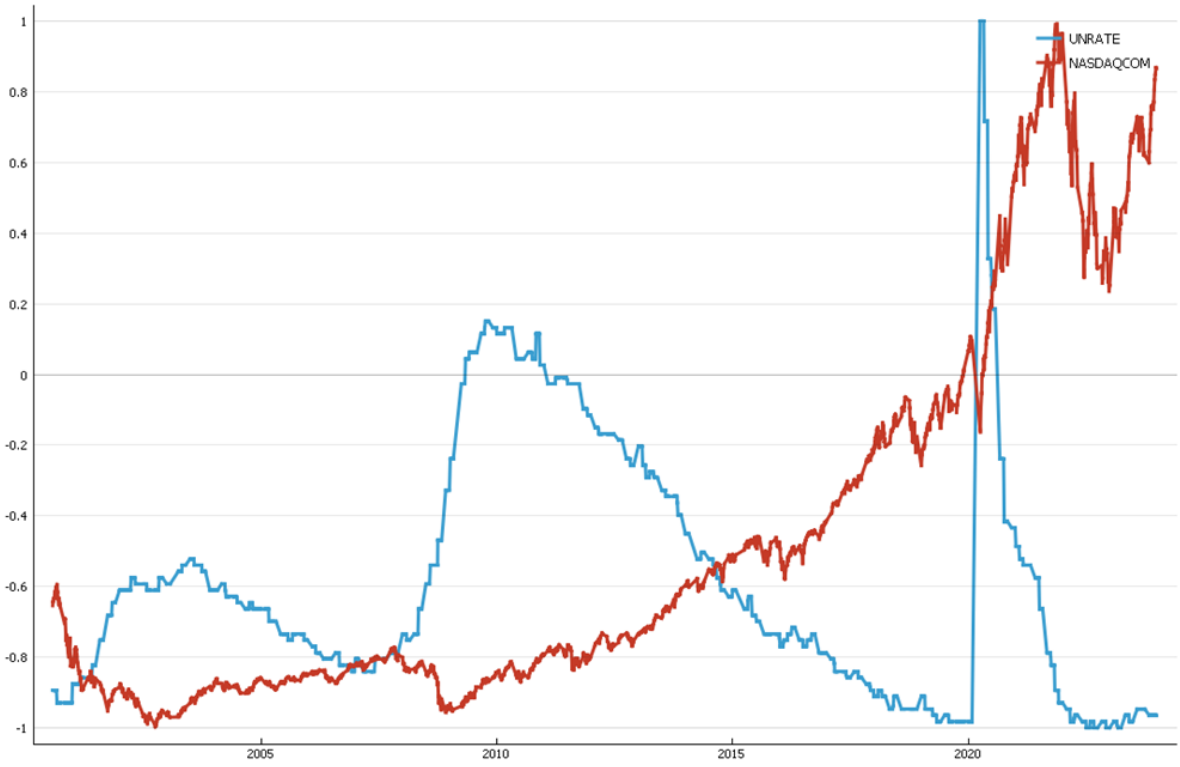


Figura 13 Confronto dell'andamento UNRATE e NASDAQ

Nella Figura 13 viene illustrati gli andamenti di due serie storiche. La correlazione di Pearson presenta un valore negativo di -0.317, mentre la correlazione di Spearman è di -0.435. Sebbene questi valori non indichino una correlazione particolarmente forte, risultano comunque significativi.

Le serie storiche evidenziano l'impatto di tre crisi rilevanti: la bolla dot-com del 2000, la crisi finanziaria del 2008 e la pandemia del 2020. Ognuna di queste crisi ha determinato un incremento del tasso di disoccupazione e un calo del valore dell'S&P 500. Durante la pandemia, il tasso di disoccupazione ha raggiunto il picco massimo, ma è stata la crisi del 2008 a lasciare il segno più profondo sul mercato del lavoro. Inoltre, la serie storica dell'S&P 500 mostra un'impennata seguita da un rapido calo della disoccupazione dopo la pandemia. Tuttavia, intorno al 2022 si registra un marcato declino del NASDAQ, questa dinamica evidenzia che le due serie non sono sempre correlate.

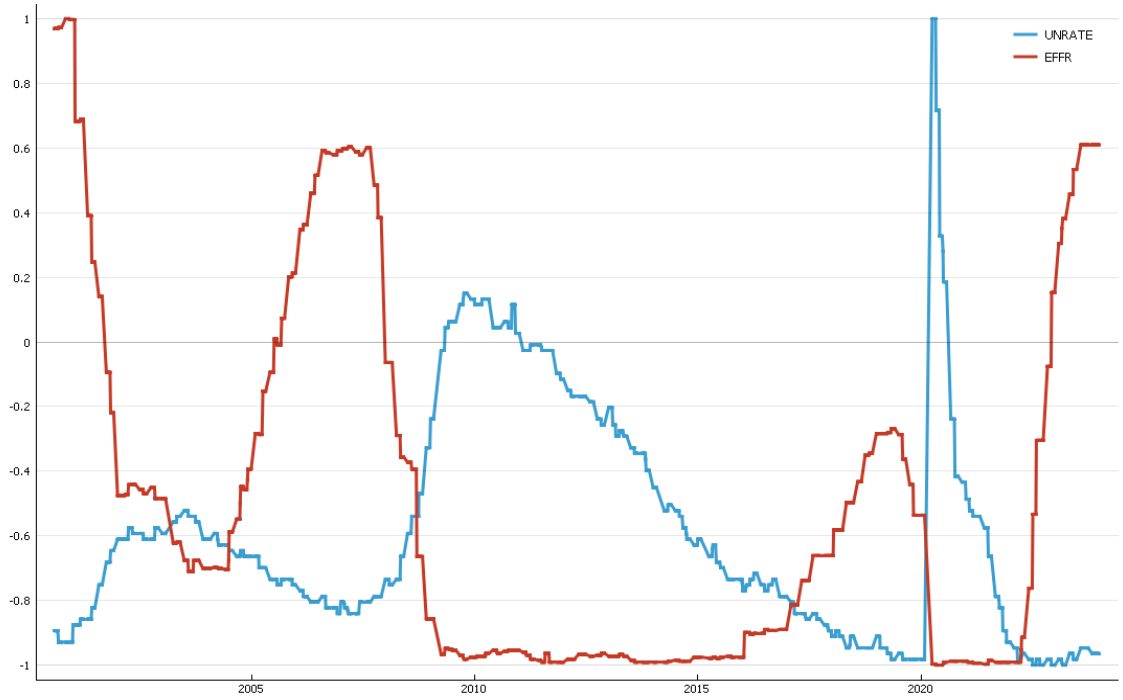


Figura 14 Confronto dell'andamento tra UNRATE e EFFR

La Figura 14 mostra una correlazione inversa tra il tasso di disoccupazione (UNRATE) e il tasso dei fondi federali effettivo (EFFR), con coefficienti di Pearson e Spearman pari a -0.552 e -0.630, rispettivamente, evidenziando una relazione negativa significativa. Tale legame è spesso attribuibile alle politiche economiche adottate in risposta alle variazioni congiunturali: in periodi di crisi economica, il tasso d'interesse tende a essere abbassato per stimolare l'attività economica e contrastare l'aumento della disoccupazione. Quando l'economia inizia a riprendersi, è comune assistere a un innalzamento dei tassi d'interesse come misura di precauzione contro l'inflazione, la quale, se eccessiva, può ostacolare gli investimenti e, in modo paradossale, causare un incremento della disoccupazione.

La forte correlazione tra questi due indicatori riflette il delicato equilibrio che le autorità monetarie cercano di mantenere tra la necessità di stimolare l'economia durante i periodi di recessione e l'importanza di prevenire bolle speculative e successive crisi. Questa dinamica evidenzia il ruolo chiave delle politiche monetarie nel plasmare le condizioni del mercato del lavoro e, più in generale, la salute dell'economia. Attraverso questa analisi, è possibile sottolineare come le decisioni della Federal Reserve abbiano un impatto diretto non solo sui mercati finanziari, ma anche sulla vita quotidiana dei cittadini, influenzando il potere d'acquisto, la capacità di spesa e la sicurezza occupazionale.

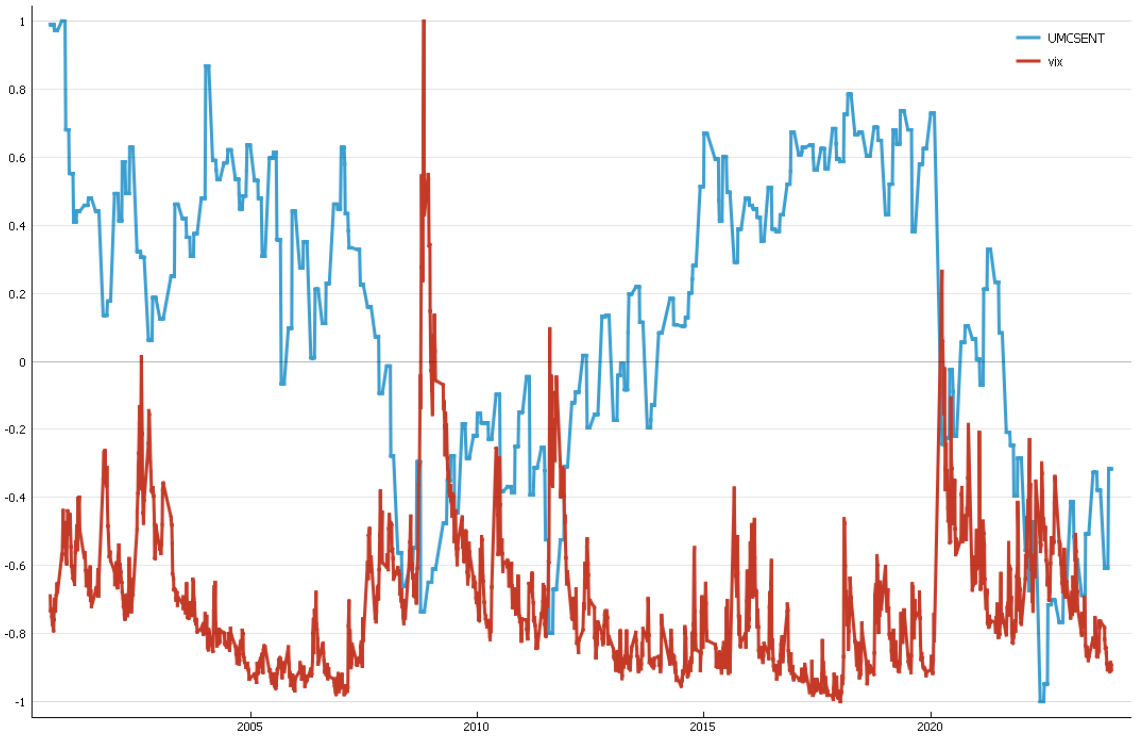


Figura 15 Confronto dell'andamento tra UMSCENT e VIX

La Figura 15 illustra una correlazione inversa tra la fiducia dei consumatori negli Stati Uniti, misurata dall'indice UMCSENT, e la volatilità del mercato azionario, rappresentata dall'indice VIX. Con un coefficiente di correlazione di Pearson pari a -0.460 e di Spearman a -0.462, i dati indicano che un aumento della fiducia dei consumatori tende a coincidere con una riduzione della volatilità del mercato, e viceversa. Un valore elevato del VIX suggerisce un'alta incertezza nel mercato, mentre un UMCSENT alto riflette l'ottimismo dei consumatori riguardo alla situazione economica.

Nella Figura 15, si osserva inoltre che la fiducia dei consumatori non ha toccato il punto più basso durante la crisi finanziaria del 2008, bensì nel 2022. Al contrario, nel 2008, all'apice della crisi finanziaria, la volatilità del mercato ha raggiunto il suo massimo storico, come evidenziato dai picchi significativi del VIX.

Figura 16 Modello predittivo VAR

Il grafico in figura 16 rappresenta un modello VAR (Vector Autoregression). Si tratta di un approccio statistico multivariato utilizzato nell'analisi delle serie storiche, che permette di catturare le interazioni dinamiche tra più variabili temporali. Questo modello si basa sulla premessa che ciascuna variabile di un sistema è una funzione lineare delle proprie passate osservazioni (lags) e delle osservazioni passate delle altre variabili nel sistema. In un modello VAR, ogni equazione del sistema modella la variabile dipendente in funzione dei valori passati di tutte le variabili nel sistema, inclusa sé stessa. Nell'analisi è stato utilizzato un modello VAR con 16 lags su cinque serie temporali: VIX, NASDAQ, EFFR, UMCSENT, UNRATE. Le previsioni del modello sono state fatte a 20 giorni.

Il modello è stato valutato con la funzione model evaluation di orange data mining che testa il modello usando la cross validation. In questa procedura, il modello viene addestrato e testato in modo iterativo su venti sottoinsiemi di dati del dataset, detti 'folds'. Inizialmente, il modello viene addestrato sul primo fold e testato sul primo set di test. Successivamente, il modello viene addestrato cumulativamente sui primi due folds e testato sul secondo set di test. Questo processo viene ripetuto aggiungendo progressivamente un altro fold ai dati di addestramento, e testando su un nuovo set di test, fino a raggiungere un totale di venti folds. In questo processo, ogni fold viene usato come set di test una volta, permettendo l'uso di tutti i dati per l'addestramento e test. Ogni test fa previsioni per venti giorni e le prestazioni del modello sono valutate sia durante l'addestramento che su diversi set di test.

Model Evaluation - Orange

Evaluation Parameters

Number of folds: 20

Forecast steps: 20

	RMSE	MAE	MAPE	POCID	R^2	AIC	BIC
VAR(16,bic,ctt)	0.087	0.053	0.116	34.8	0.788	-12.0	-11.9
VAR(16,bic,ctt) (in-sample)	0.013	0.002	0.008	30.6	0.999	-11.3	-11.2

Figura 17 Modello di valutazione

Il modello di valutazione della figura 17, mostra un modello VAR(16,bic,ctt), con 16 lags, il criterio di informazione bayesiano per la selezione del modello (bic) e la presenza di un trend e termine costante (ctt). Nella figura si può notare un valore del RMSE (Root Mean Square Error) di 0.087. Questo valore suggerisce una discreta accuratezza nelle previsioni dei valori futuri rispetto ai dati osservati. Proseguendo con l'analisi del modello di valutazione, si nota che il MAE (Mean Absolute Error) nel test è di 0.053, indicando che la media degli errori del modello è vicina ai valori reali. Il MAPE (Mean Absolute Percentage Error) è dell'11,6%, fornendo una misura percentuale dell'errore rispetto ai valori osservati. Il POCID (Percentage of Correct Directional Predictions) è del 34,8% nel test, indicando la scarsa capacità del modello di prevedere correttamente la direzione del movimento della serie temporale. Nonostante ciò, la differenza tra addestramento e test è bassa, dimostrando una buona generalizzazione del modello nella previsione della direzione.

Il valore del R^2, o coefficiente di determinazione, è pari a 0.788, indicando che il 78,8% della varianza della variabile dipendente è spiegata dal modello. Per la selezione del modello, vengono utilizzati l'AIC (Akaike Information Criterion) e il BIC (Bayesian Information Criterion), che penalizzano il numero di parametri.

I risultati della cross-validation indicano un possibile overfitting del modello, suggerendo la necessità di migliorare la sua capacità di generalizzazione. I valori del R^2 e il MAPE peggiorano notevolmente dal set di addestramento al set di test. Inoltre, un altro aspetto importante da migliorare è il POCID, per ottimizzare la performance del modello in scenari futuri.

In conclusione, questo studio ha sottolineato l'importanza di monitorare un insieme di indicatori per comprendere appieno le dinamiche del mercato azionario e prevedere le sue fluttuazioni future. La capacità di interpretare correttamente tali dati è fondamentale per gli investitori, analisti e responsabili delle politiche economiche, consentendo loro di prendere decisioni informate in un contesto economico globale in continua evoluzione.