UNIVERSITÀ
DEGLI STUDI DI BARI
ALDO MORO

Dipartimento di
Informatica

Tesi di Laurea in
SISTEMI AD AGENTI

Modello Predittivo per lo Studio delle Relazioni tra Qualità dell'Aria e Sintomatologia di Soggetti Allergici

Relatore:
Prof.ssa Berardina De Carolis

Laureando:
Luigi Daddario
Mat. 685195

Anno Accademico 2021-2022

"If a machine is expected to be infallible, it cannot also be intelligent."

- Alan Turing

Indice

Capitolo 1

Introduzione

L'inquinamento atmosferico rappresenta una preoccupazione crescente per le comunità in tutto il mondo, in quanto gli inquinanti presenti nell'aria possono avere effetti negativi sulla salute umana. In particolare, le persone con allergie possono essere particolarmente sensibili alla qualità dell'aria, poiché gli inquinanti atmosferici possono irritare le vie respiratorie e aumentare la produzione di muco, rendendo le allergie più gravi e le reazioni più intense.

La necessità di affrontare questo problema ha portato alla ricerca di modelli predittivi per lo studio delle relazioni tra qualità dell'aria e sintomatologia di soggetti allergici. Questi modelli possono essere utilizzati per comprendere come l'inquinamento atmosferico influisce sulle allergie e per sviluppare strategie di prevenzione e intervento.

Le ricerche hanno dimostrato che gli inquinanti atmosferici come il particolato fine (PM2,5), l'ozono troposferico (O3) e gli ossidi di azoto (NOx) possono aumentare il rischio di sviluppare allergie e asma. Inoltre, l'esposizione a lungo termine all'inquinamento atmosferico può aggravare i sintomi delle allergie esistenti e aumentare la gravità delle reazioni allergiche.

Pertanto, la creazione di un modello predittivo per lo studio delle relazioni tra

qualità dell'aria e sintomatologia di soggetti allergici può aiutare a identificare i fattori di rischio associati all'inquinamento atmosferico e alle allergie, e a sviluppare interventi efficaci per prevenire e trattare le allergie.

Inoltre, è importante che i governi e le comunità lavorino insieme per ridurre l'inquinamento atmosferico e migliorare la qualità dell'aria. Ciò può includere l'adozione di politiche ambientali più rigorose, la promozione di tecnologie pulite e il miglioramento dei sistemi di trasporto pubblico.

1.1 Problema affrontato

Il presente studio si propone di analizzare le conseguenze dell'inquinamento atmosferico sulla sintomatologia di soggetti allergici, esaminando le correlazioni, covarianze e dipendenze tra le concentrazioni degli inquinanti nell'aria e i sintomi manifestati in specifiche situazioni.

Iniziamo definendo il concetto di qualità dell'aria e il relativo indice di valutazione. Esistono diversi indici per valutare la qualità dell'aria, ma per il nostro studio utilizzeremo l'Indice di Qualità dell'Aria (IQA), il quale viene calcolato come il massimo valore tra gli indici di ogni inquinante presente nell'aria, in accordo con la seguente formula:

$$IQA = max(Ix; \forall Ix) \tag{1.1}$$

dove Ix rappresenta l'indice di un singolo inquinante. L'indice Ix viene calcolato come rapporto tra la concentrazione dell'inquinante e il valore massimo consentito nell'aria (Imax), moltiplicato per 100, ovvero:

$$Ix = \frac{Dx}{Imax} \times 100 \tag{1.2}$$

Dove Dx rappresenta la concentrazione dell'inquinante e Imax la concentrazione massima consentita di tale inquinante nell'aria in un'ora.

Con le concentrazioni di tutti gli inquinanti misurate in una certa zona di una città in un dato intervallo di tempo, è possibile calcolare l'indice IQA secondo la formula (1.1). Esiste un certo range di valori dell'IQA che definisce il concetto di "buona qualità dell'aria" o "cattiva qualità dell'aria", come riportato nella Tabella sotto.

valore IQA	Qualità
≤ 50	Buona
$\geq 50 \leq 100$	Accettabile
$\geq 100 \leq 150$	Mediocre
$\geq 150 \leq 200$	Scadente
≥ 200	pessima

Tabella 1.1: Valori di riferimento qualità dell'aria.

Nella nostra analisi, consideriamo un dataset contenente le misurazioni orarie delle concentrazioni degli inquinanti in una specifica zona e utilizziamo l'indice IQA per valutare la qualità dell'aria in ogni ora. Successivamente, analizziamo la relazione tra le misurazioni dell'IQA e i sintomi di soggetti allergici in quella stessa zona e periodo di tempo.

1.2 Stato dell'arte

La mia tesi si basa su una ricerca bibliografica che ha identificato numerose fonti scientifiche che giustificano le fondamenta del mio lavoro. In particolare, ho raccolto una serie di rilevanze che dimostrano l'importanza di studiare la correlazione tra la

qualità dell'aria e i sintomi delle allergie, e come tali studi siano utili per comprendere gli effetti dell'inquinamento atmosferico sulla salute umana.

Inoltre, ho potuto constatare come l'interesse verso questa tematica sia molto forte nell'ambito dell'analisi dei dati, il che ha reso valide fin da subito le mie ipotesi. Infatti, ho potuto consultare alcune pubblicazioni scientifiche che hanno esplorato l'utilizzo di modelli matematici e algoritmi di machine learning per prevedere la qualità dell'aria e studiare le sue conseguenze sulla salute umana.

Questi studi dimostrano come l'analisi dei dati sia un valido strumento per comprendere le dipendenze e le correlazioni tra le concentrazioni degli inquinanti nell'aria e i sintomi delle allergie, e come ciò possa fornire preziose informazioni per adottare misure di mitigazione dell'inquinamento atmosferico e proteggere la salute umana.

1.3 Relazioni e correlazioni tra sintomatologie e qualità dell'aria

Quando si parla di qualità dell'aria, è importante distinguere tra due possibili scenari di misurazione: indoor e outdoor. Nel mio studio, ho considerato come scenario di riferimento i casi di misurazioni outdoor, soprattutto durante periodi dell'anno particolarmente problematici per i soggetti allergici, come la primavera. Tuttavia, non ho trascurato i riferimenti ai sintomi che altre ricerche hanno evidenziato in casi di qualità dell'aria precaria in ambienti chiusi.

Per quanto riguarda i sintomi di riferimento, ho scelto quelli più comuni in caso di attacchi allergici più o meno intensi, poiché sono i più semplici da riconoscere, nonostante la loro natura pressoché generica.

L'esposizione a particelle fini e altre sostanze inquinanti presenti nell'aria può aumentare l'infiammazione delle vie respiratorie, che può portare ad un aggravamento

dei sintomi dell'asma e delle allergie respiratorie come la rinite allergica. Inoltre, l'inquinamento dell'aria può anche indebolire il sistema immunitario e rendere gli individui più suscettibili alle allergie e alle malattie respiratorie.

È importante sottolineare che esistono studi che dimostrano una correlazione tra la concentrazione di polline nell'aria e l'incidenza di sintomi allergici. In particolare, l'aumento dei livelli di polline può aumentare il rischio di rinite allergica e asma allergico.

Correlare i sintomi e la qualità dell'aria è fondamentale per evidenziare gli effetti a breve termine che l'inquinamento atmosferico può avere sulla salute umana. Inoltre, queste correlazioni ci permettono di identificare gli inquinanti che più concorrono alla sintomatologia e di avere dati sui quali basare eventuali misure per ridurre le concentrazioni di tali inquinanti o supportare ulteriori studi scientifici.

1.4 Struttura della tesi

La tesi è strutturata in tre capitoli:

- Nel secondo capitolo, **"Il ruolo della Data Science e Machine Learning"**, viene illustrato come l'applicazione di queste discipline possa essere utile per risolvere problemi simili a quello affrontato nel progetto. In particolare, vengono descritti i principali problemi e algoritmi dell'apprendimento supervisionato e non supervisionato che sono stati utilizzati.

- Nel terzo capitolo, intitolato **"Il modello predittivo"**, viene presentato il progetto realizzato, focalizzandosi in particolare sul modello predittivo adottato. Sono inoltre descritte le varie fasi del progetto, la verifica delle performance e le sperimentazioni effettuate.

- Nel quarto e ultimo capitolo, intitolato **"Conclusione e sviluppi futuri"**, sono state tratte le conclusioni a seguito dell'analisi delle performance del modello e delle correlazioni evidenziate. Inoltre, sono state indicate alcune possibili direzioni di sviluppo future per il progetto.

Capitolo 2

Il ruolo della Data Science e Machine Learning

In questo capitolo vedremo come la Data Science possa essere utile ad evidenziare in maniera *intelligente* i pattern che sussistono tra i dati ed in particolare tra le misurazioni delle concentrazioni orarie degli inquinanti. Successivamente vedremo come correlare questi dati con i sintomi dei soggetti che vogliamo analizzare.

2.1 Approccio generale

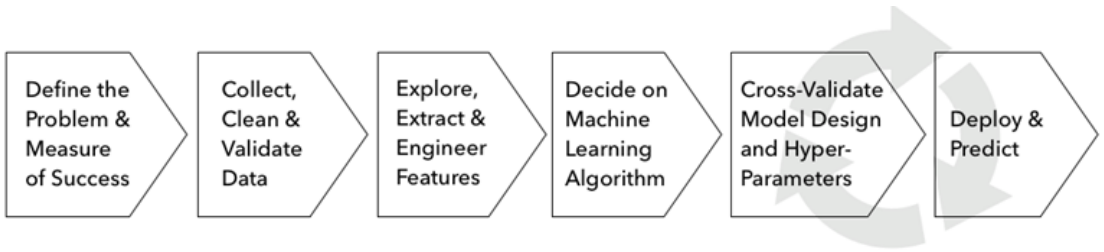


Figura 2.1: Risoluzione di un problema in Data Science / Machine Learning

In generale, l'approccio ad un problema di analisi dei dati o di Intelligenza artificiale è molto diverso da quello che solitamente si applica a problemi dell'Informatica

classica; è anche vero che alcuni problemi dell'informatica classica possono essere risolti con metodi e algoritmi di IA, ma sarebbe computazionalmente improponibile risolvere problemi di intelligenza artificiale con metodi dell'Informatica classica.

In realtà però dobbiamo fare chiarezza su cosa sia la Data Science e perchè è strettamente correlata all'Intelligenza Artificiale. Per Data Science ci si riferisce all'applicazione di tecniche e metodi di IA ai fine di analisi dei dati. Conseguentemente possiamo intuire come, anche l'apprendimento automatico possa essere relazionato alle due scienze menzionate prima.

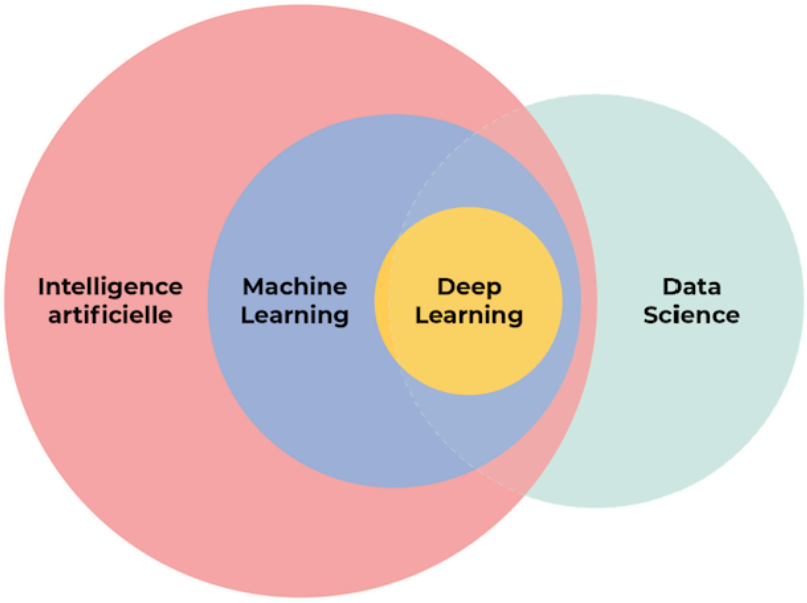


Figura 2.2: Inclusione insiemistica tra le varie scienze dell'IA

2.2 Data collection

Dopo aver definito il problema, che per questa Tesi è chiaro essere lo studio delle regressioni tra sintomatologie allergiche e concentrazioni degli inquinanti nell'aria, lo step successivo è sicuramente quello di raccogliere i dati. Nel mio caso la raccolta dati in realtà è divisa in due parti, questo perchè lo scopo è stato fin da subito

avere due dataset da *fondere*, quello relativo alle misurazioni degli inquinanti e quello relativo ai sintomi percepiti dagli individui. Raccogliere i dati è un'operazione abbastanza complessa, che richiede attenzione e criterio. Questo perchè, in generale, chiaramente dipende dal tipo di dati che dobbbiamo raccogliere e può richiedere diverse tecnologie o diversi strumenti.

Per poter raccogliere le misurazioni degli inquinanti è necessario possedere degli strumenti, come ad esempio dei sensori, che lo facciano al posto nostro. Nel mio caso ho scelto di utilizzare dei dataset che vedremo più avanti e che settimanalmente vengono aggiornati dall'ente che si occupa di raccogliere e convertire in file CSV.

Nonostante questa metodologia non sia scevra da problemi, è chiaro che la raccolta dati in questo modo ha una rilevanza molto alta, perchè di fatto si tratta di prendere delle misurazioni con dei sensori ad alta precisione. Al contrario, in generale, raccogliere dati può essere costoso e molto più complesso perchè, ad esempio gli utenti possono non essere disposti ad essere intervistati o possono non essere troppo precisi.

Nel mio caso, per poter raccogliere i dati che mi servivano, quelli relativi ai sintomi percepiti dall'utente in certi periodo dell'anno, mi sono affidato alle tecniche di Interviste e focus group. Dopo aver scelto un certo numero di utenti che potenzialmente si vuole raggiungere, si sceglie di somministrare agli stessi dei questionari, che ci possano aiutare, con domande specifiche, a scoprire ciò di cui abbiamo bisogno. A volte però abbiamo bisogno di metodologie più efficaci, con una confidenza molto più alta e che possano essere il più accurati possibile, questo è il motivo per il quale ho scelto di condurre, dopo aver fatto compilare un questionario, delle interviste individuali, che mi hanno aiutato a capire con quale probabilità le informazioni fornite da quello specifico individuo, sarebbero state utili e affidabili.

In alcuni casi però gli utenti intervistati, facendo parte dello stesso nucleo familiare, avevano delle relazioni tra loro, ed è per questo che ho condotto dei focus

group per potermi focalizzare meglio su alcune questioni e su alcuni valori che erano stati dichiarati nei questionari.

Un focus group consiste in una intervista di gruppo, durante la quale si sceglie di focalizzarsi su alcuni specifici argomenti che possono risultare rilevanti in ambito della raccolta dati. Queste interviste vengono svolte tra i utenti simili tra loro, o che comunque possano avere dei punti in comune, come nel mio caso.

2.3 Apprendimento supervisionato

L'apprendimento è la capacità di sfruttare l'esperienza per poter migliorare il comportamento. Ciò significa estendere le abilità e migliorare l'accuratezza e l'efficienza.

L'apprendimento supervisionato è una tecnica di apprendimento automatico che, dato un insieme di esempi, coppie di input-output, consente di predire l'output per nuovi casi di cui si conosce solo l'input. Possiamo avere diversi approcci per l'apprendimento supervisionato, ciò dipende dalla natura delle features di input o dal dominio del problema. Possiamo scegliere una singola ipotesi che si possa ben adattare agli esempi o possiamo fare delle predizioni direttamente a partire dagli esempi, è il caso dei metodi non parametrici. In alternativa possiamo selezionare il sottoinsieme di spazio delle ipotesi coerenti con gli esempi oppure fare delle predizioni in base alle distribuzioni di probabilità a posteriori delle ipotesi condizionate sugli esempi.

2.3.1 Definizione di problema di apprendimento

Definiamo un problema di apprendimento in questo modo: Data una conoscenza di fondo (background knowledge) e dei dati (esperienza pregressa), lo scopo è quello di creare una rappresentazione interna, che sarà parte di una base di conoscenza e che possa essere utilizzata per decisioni future.

2.3.2 Definizione formale di apprendimento supervisionato

Definizione 2.3.1 (Supervised Learning)**.** Dato un training set di N esempi nella forma $(x1, y1), ..., (xN, yN)$ t.c x_i è il vettore delle feature dell'i-esimo esempio di N e y è y_i è il target (es. la classe), un algoritmo di apprendimento definisce una funzione g: $X \to Y$, dove X è lo spazio di input e Y lo spazio di output.

La definizione di apprendimento supervisionato è molto interessante perchè ci fa capire fin da subito la differenza di approccio ai problemi dell'informatica classica, rispetto a quelli di IA. Un Informatico è abituato a partire da un problema e a risolverlo producendo un algoritmo, che di fatto costituisce una funzione che risolve il problema stesso. Con un problema di apprendimento il procedimento è esattamente l'opposto, noi partiamo da dei dati che assumiamo essere affidabili e l'algoritmo produce una funzione che ne stabilisce la relazione. [2]

2.3.3 Algoritmi di Regressione

La regressione in Machine Learning, Data Science e in statistica in generale è un problema che consiste, a partire da esempi senza etichetta, nella previsione di valori reali.

Quindi, dati degli esempi lo scopo è quello di predire il valore di $y \in \mathbf{R}$ La stima che viene fatta, in particolare è sulla stima tra le relazioni che sussistono tra le variabili. Mentre la classificazione identifica a quale categoria appartiene una osservazione, un modello di regressione ne stima il valore. L'algoritmo viene anche chiamato **regressore**.

Come funziona la regressione

Un dataset o Training Set è composto da N esempi. Ogni esempio è suddiviso in alcune caratteristiche, dette feature ed eventualmente nel caso dell'apprendimento

supervisionato, cioè quello che stiamo trattando, anche da una variabile target.

x1	x2	x3	x4	y
1.34	43.1	121.3	32.1	91.0
1.34	43.1	121.3	32.1	91.0
1.34	43.1	121.3	32.1	91.0
1.34	43.1	121.3	32.1	91.0
1.34	43.1	121.3	32.1	91.0

Tabella 2.1: Esempio di training set.

Cerchiamo di capire meglio cosa significa: l'insieme delle caratteristiche rappresentano tutte le informazioni che ci servono per poter descrivere un certo oggetto, che vogliamo classificare o sul quale vogliamo un regressore. Se stiamo trattando la qualità dell'aria, avremo bisogno di capire quali siano tutte le caratteristiche che contribuiscono alla definizione della qualità dell'aria. Se invece trattiamo un oggetto meno complesso come un cane, ci dovremo chiedere quali siano le caratteristiche che definiscono un cane. Ad esempio, l'altezza al garrese, il colore del manto, alcune caratteristiche della dentatura e cosi via.

Conoscendo tutte queste caratteristiche, sarà poi più semplice rappresentare in uno spazio vettoriale tutti questi esempi, per i quali potremo definire una funzione $y = f(x)$ che ne definirà le relazioni. Questo è quello che realizza un regressore.

Regressione lineare

Il modello di regressione di base per eccellenza è sicuramente la regressione lineare. La regressione lineare è un modello statistico, che, se vogliamo rappresenta una metodologia di stima del valore atteso, condizionando lo stesso da una variabile dipendente Y dati i valori di altre variabili indipendenti $X1, X2...Xk$ in modo tale che risulti: $E[Y|X1, ...Xk]$

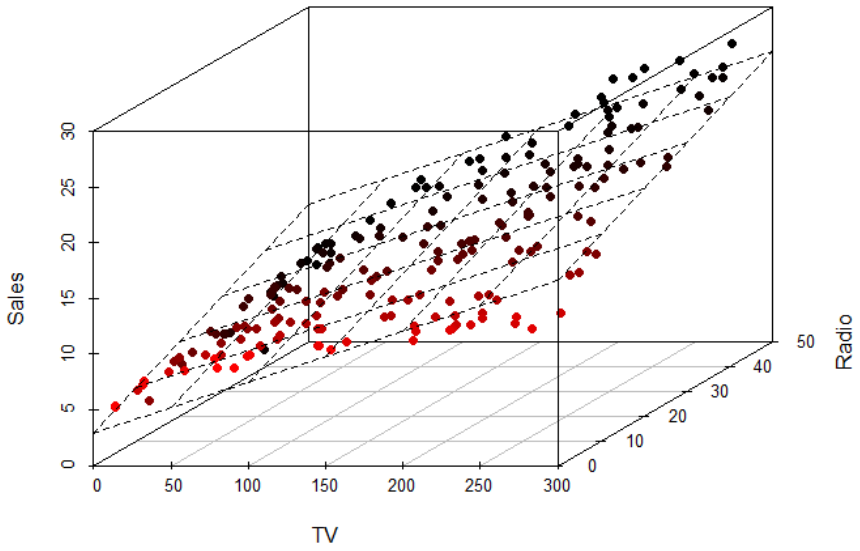


Figura 2.3: Esempio di regressione multidimensionale

Il modello di regressione lineare è allora cosi definito: $yi = \beta 0 + \beta 1 Xi + ui$

dove i varia tra le osservazioni, Yi è la variabile dipendente, Xi è la variabile indipendente o regressore, $\beta 0 + \beta 1$ è la retta di regressione o funzione di regressione della popolazione, $\beta 0$ è l'intercetta della retta di regressione, $\beta 1$ è il coefficiente angolare e ui è l'errore statistico.

Per ogni osservazione si cerca di determinare in maniera formale, una relazione lineare tra la variabile Y e k variabili deterministiche.

Un esempio tipico è quello che lega la i consumi con il reddito, scegliendo una funzione di regressione che ne spieghi le relazioni.

$C = f(Y)$ è una generica relazione che determina il consumo

$C = a + bY$ è la relazione lineare, dove a è l'intercetta e b è la pendenza della retta.

2.3.4 Regressione lineare in Machine Learning

La regressione lineare in machine learning rappresenta l'adattamento di una funzione lineare ad un training set con feature numeriche.

Avremo quindi un vettore $X_1, \ldots, X_n$ di feature e un obiettivo Y. Il modello sarà rappresentato da una funzione lineare delle feature di input.

$$Yw(e) = w_0 + w_1 X_1(e) + \ldots + w_n X_n(e) = \sum_{i=0}^{n} w_i * X_i(e)$$

[3]

2.3.5 Train set e Test Set

Una procedura per la valutazione del modello di apprendimento consiste nel dividere il dataset in due (o più) parti, questo per poter testare il modello e cercare, appunto, di validarlo.

Il train/test splitting consiste nel dividere il dataset in due parti, una parte dedicata al train del modello, che lo stesso utilizzerà per poter apprendere i dati, e una parte, quella di test, che sarà utilizzata per testare gli stessi. Il modello quindi apprenderà dai dati di training e sarà poi testato su dati che non ha mai visto, i dati di test.

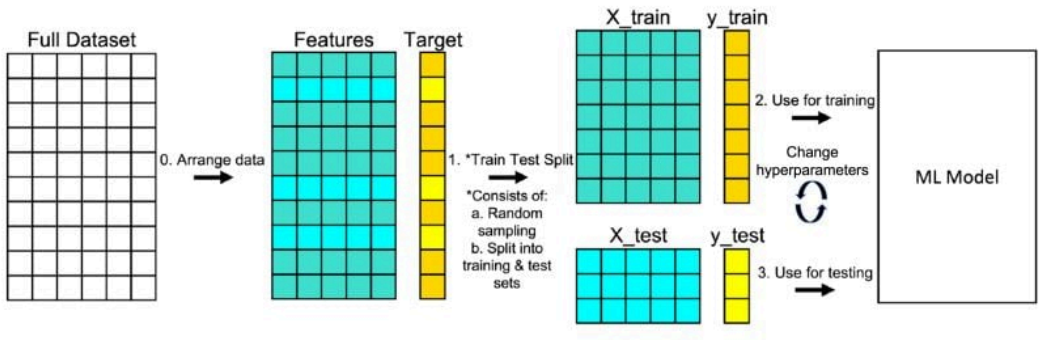


Figura 2.4: Train/Test splitting

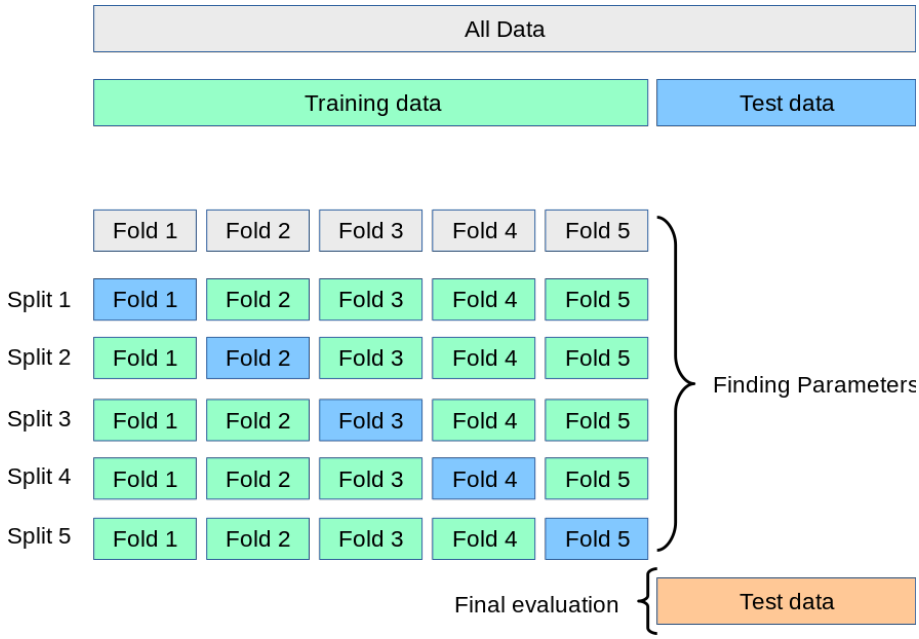


Figura 2.5: Cross validation

Una tecnica statistica molto utilizzata è la cosidetta cross-validation. Quando abbiamo un numero molto elevato di campioni, la convalida incrociata ci consente di suddividere l'insieme dei dati in k parti uguali. Ad ogni passo una delle k parti costituirà la parte dei dati di convalida del modello e le restanti saranno l'insieme di addestramento.

Allenando il modello con ognuna delle k parti, si eviterà, tendenzialmente il problema del sovradattamento, che chiariremo nel prossimo paragrafo. Altra motivazione per la quale dovremmo utilizzare questa metodologia di convalida è relativa al fatto che evita anche il campionamento asimmetrico, che tipicamente è presente nella suddivisione dei dati in due parti. La convalida incrociata è meglio conosciuta come k-fold cross validation.

2.3.6 Problemi di Underfitting e Overfitting

In generale, in Matematica o Informatica, con induzione o ragionamento induttivo ci si riferisce all'apprendimento di concetti generali da semplici esempi specifici.

Il primo principio della Matematica che ci viene in mente a riguardo è sicuramente il principio di induzione. Questo è diverso dalla deduzione, che è esattamente il contrario: apprendimento di concetti specifici da regole generali.

La generalizzazione è la modalità con la quale un modello di apprendimento automatico sia in grado di applicare concetti appresi in fase di training, su dati che non ha visto durante l'allenamento.

Trovare una funzione che definisca precisamente le relazioni tra le feature e i target spesso non è molto semplice, di conseguenza a volte capita che il modello possa avere dei comportamenti anomali, diversi rispetto a quelli che ci si aspetterebbe. Questo potrebbe essere il caso di overfitting o underfitting. Il caso di underfitting è un caso poco trattato, poichè di fatto è una problematica che il più delle volte si risolve utilizzando algoritmi di apprendimento più efficaci, migliorando la qualità dei dati o ancora aumentandone la quantità. L'underfitting quindi si riferisce ad un modello che non è in grado di comprendere bene i dati di training ne tantomeno i dati di test.

Differente invece è il caso di overfitting. Quest'ultimo si riferisce ad un modello che non punta tanto alla generalizzazione quanto alla memorizzazione dei pattern dei dati, conseguentemente diventa molto bravo con i dati di training ma molto con i dati di test.

Una delle procedure per poter evitare l'overfitting è la procedura di convalida citata prima, la k-fold cross-validation.

Come vediamo da queste rappresentazioni, l'overfitting è un chiaro caso di memorizzazione dei dati, che lo porta si ad essere "molto bravo" con i dati di training,

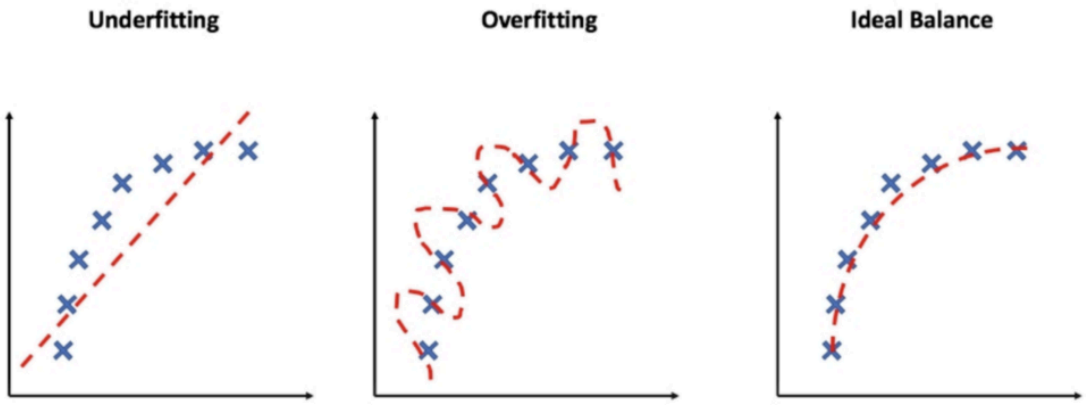


Figura 2.6: Underfitting/Overfitting e ideal

ma a non avere ben chiara la rappresentazione, nello spazio, degli esempi di test, o comunque di dati che non ha mai visto, non riuscirà mai a definirne una funzione precisa.

2.3.7 Modelli neurali

I modelli neurali sono di totale ispirazione dai neuroni cerebrali, tuttavia differiscono per semplicità e numerosità. Questi modelli sono molto utilizzati per problemi di **basso livello** ma per i quali abbiamo tanti dati disponibili.

Si parte dall'ipotesi di poter replicare i meccanismi del cervello per poterne simulare il funzionamento.

Senza allontanarci troppo dalla generalità, questo modello matematico è formato da un gruppo di interconnessioni di informazioni, formate da *neuroni artificiali*, che vengono elaborate con uno specifico approccio di calcolo.

Una rete neurale riceve dei segnali sullo strato di input, i cui nodi sono tutti collegati a tantissimi nodi interni, dipendentemente dagli iperparametri scelti. Ogni singolo nodo elabora il segnale ricevuto e trasmette il risultato ai nodi successivi.

La rete neurale è un modello che può essere considerato composito, poichè di

fatto la funzione che determina questo modello $f(x)$ può essere definita come la composizione funzionale di altre funzioni $G(X)$. In alcuni casi potremmo considerare la funzione definita da una rete neurale come la composizione di k funzioni predefinite:

$$f(x) = k(\sum_i Wi * gi(x))$$

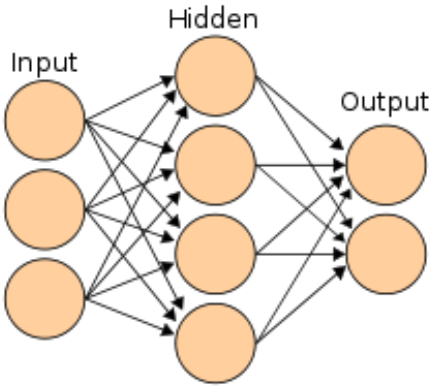


Figura 2.7: Rete Neurale semplice

LSTM

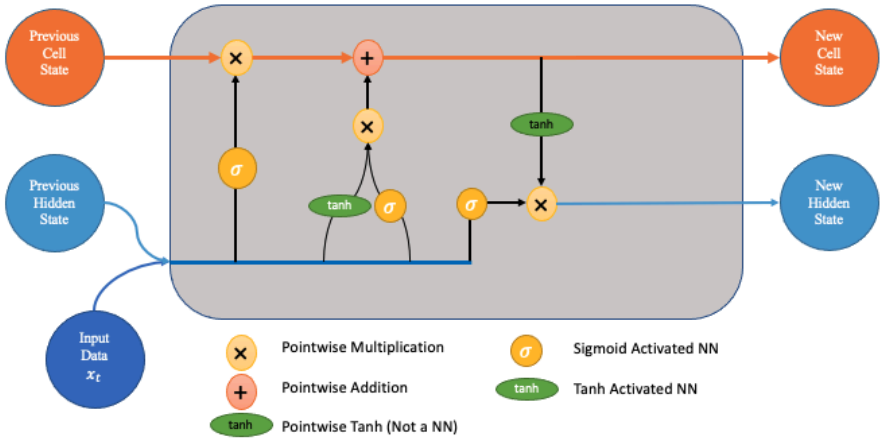


Figura 2.8: LSTM

Le Long short term memory neural networks sono progettate per poter permettere ad una rete neurale di ricordare ciò di cui ha bisogno per poter mantenere il contesto, ma consente anche di dimenticare cose che non sono più applicabili.

Per quanto questa introduzione possa essere poco formale, cerchiamo di capire il background: supponiamo di avere un problema da risolvere. Questo problema ha una soluzione che tendenzialmente varia nel tempo, poichè dipende dallo stesso, ma anche da certe situazioni e certe caratteristiche che sono il tempo riesce a mostrare. In questo caso le LSTM svolgono un ruolo fondamentale, perchè riescono a capire le relazioni degli avvenimenti, in funzione del tempo. Questo ha un potere molto forte. Le LSTM hanno connessioni di feedback diverse dalle rete neurali tradizionali, ciò consente di poter elaborare appunto, serie temporali, ma, conservando, nel tempo, informazioni utili.

Cosa vuol dire? Possiamo considerare delle serie temporali, durante le quali le condizioni e le relazioni dei dati, cambiano, grazie a tanti fattori. Un esempio sono le previsioni del tempo. Nelle previsioni meteo gli esperti analizzano l'ambiente atmosferico in funzione del tempo, questo perchè risulta molto utili poter fare delle previsioni per i giorni successivi a quelli considerati.

Una rete LSTM è, di fatto, una rete neurale ricorrente. Vediamo quali sono le caratteristiche di una RNN e le differenze con le reti neurali LSTM. Una RNN ha un input, che viene elaborato e produce un risultato. La ricorrenza risiede nel fatto che tutte le volte che, a partire dal passaggio successivo al primo, la rete riceve un input, avrà oltre questo, informazioni riguardanti l'output precedente.

Una RNN è soggetta ad un problema chiamato: problema della dipendenza a lungo termine, ciò significa che più andiamo avanti nel tempo, più saranno le informazioni ricevute dalla rete e meno la stessa sarà disposta ad imparare cose nuove.

L'equivalenza mostrata sopra ci consente di capire molto l'importanza della ci-

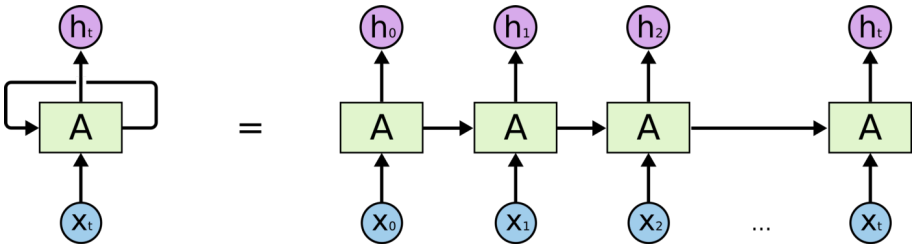


Figura 2.9: Esempio di Rete neurale ricorrente

clicità all'interno di una rete neurale ricorrente. L'idea è proprio quella di poter apprendere dal passato.

Torniamo però sul problema della dipendenza introdotto poco fa. LSTM offre una soluzione a questo problema, aggiungendo uno stato interno al nodo RNN, di conseguenza quando riceverà l'input corrente e l'output del passaggio precedente, salverà in questo stato alcune informazioni e ne preleverà altre, in modo tale da poter ragionare sul contesto, senza perdere le informazioni di tutti gli input precedenti.

Di fatto, lo stato della rete è una cella divisa in tre parti:

- Forget Gate
- Input Gate
- Output gate

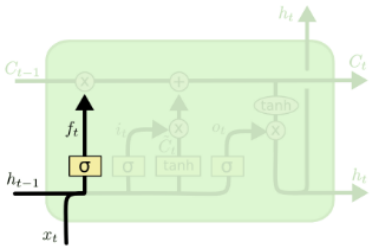
Il **Forget Gate** ci dice quali sono le informazioni che possono essere dimenticate, ovvero quelle informazioni che non ha più senso ricordare, poichè non contestualmente rilevanti. **L'input** gate ci dice quali informazioni sarebbe il caso di aggiungere o aggiornare sullo spazio di archiviazione. Infine il gate di **output** ci dice, in una particolare istanza, quali dovrebbero essere le informazioni da dover "mostrare".

In realtà però è importante specificare che, il tipo di informazioni che vogliamo ricordare sono sempre relative al contesto. Se stiamo cercando di predire delle parole mancanti in una frase, potrebbe ad esempio essere utile dover ricordare che il

soggetto della frase è di sesso maschile, poichè, considerando questa informazione si presume che il resto della frase possa dipendere da questo. E, in caso di cambiamento del soggetto in frasi successive a quella considerata, sarà rilevante modificare le informazioni salvate nella cella LSTM.

Vediamo adesso la procedura di **apprendimento** di una LSTM.

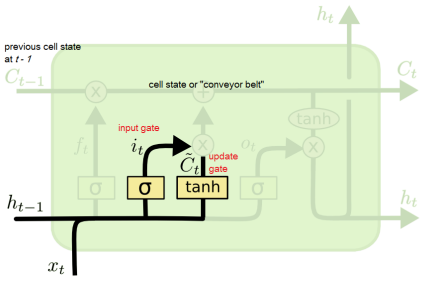
Il primo step è decidere quali informazioni dobbiamo eliminare dalla cella di stato. Questa decisione è affidata al **sigmoid layer**, anche chiamato "forget gate layer". Esso prende $Ht-1$ e Xt e ritorna un booleano, 0 o 1. Esegue questo per ogni numero nella cella $Ct-1$. 1 significa "mantieni" mentre zero significa "puoi ignorare questa informazione". La funzione definita:



$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \; + \; b_f\right)$$

Figura 2.10: Forget gate

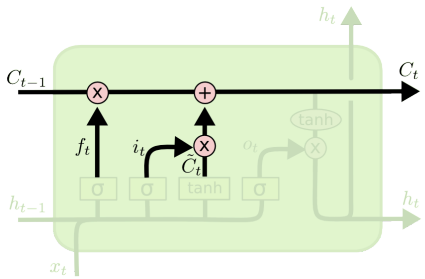
Il secondo step prevede di scegliere quale informazione dobbiamo salvare nella cell state. Questa operazione è divisa in due parti: nella prima parte l'input gate layer decide quali valori aggiornare, successivamente il **tanh layer** realizza un vettore di valori Ct che sono i candidati a dover essere aggiunti.



$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \; + \; b_i\right)$$
$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \; + \; b_C)$$

Figura 2.11: Tanh gate

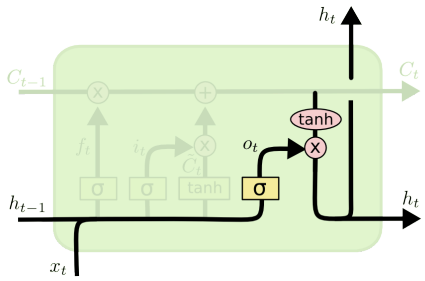
Successivamente dovremo combinare questi due per poter creare uno stato di update. Come facciamo? Dobbiamo moltiplicare lo stato precedente ft dimenticando ciò che abbiamo deciso di dimenticare, per poi aggiungere $it * Ct$. Questi sono esattamente i **valori candidati**. (fig. 2.12)



$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

Figura 2.12: Tanh gate

Sono alla fine possiamo decidere quale sarà il nostro output (Fig. 2.13) . Prima di tutto eseguiamo un sigmoid layer, che avrà il compto di decidere quale parte della cella sarà l'output, dopodichè diamo il valore dello stato della cella in input a tanh, moltiplicandolo per l'output del gate sigmoide.[4][5]



$$o_t = \sigma\left(W_o\ [h_{t-1}, x_t]\ +\ b_o\right)$$
$$h_t = o_t * \tanh\left(C_t\right)$$

Figura 2.13: output

Metriche per la valutazione degli algoritmi di Regressione

Le metriche dei modelli di regressione in Data Science e Machine Learning si basato sempre sui residui, ovvero sulla differenza tra le previsioni del modello e le risposte corrette. Ad esempio, nell'ipotesi di aver diviso il dataset in train e test set, effet-

tueremo il test sui dati di test, controllando la distanza tra il valore predetto e il valore effettivo.

Tra le metriche più utilizzate ne citiamo due:

- RMSE (RootMean Squared Error)
- MAE (Mean Absolute Error

RMSE è la radice quadrata della media degli errori al quadrato. L'errore è nullo quando MSE=0. Viene definito come segue:

$$RMSE = 1/n * \sum_{i=1}{}_n (Yi - Pi)^2$$

MAE è la media delle differenze in valore assolute tra previsioni e target. Viene definito come:

$$MAE = 1/n * \sum_{i=1}{}_n |Yi - Pi|$$

Qual'è la differenza tra queste due metriche? Esperienza e contesto ci aiutano a utilizzarle nella maniera corretta. Dal punto di vista matematico però la RMSE indica una maggior sensibilità per i valori anomali. MAE invece fornisce una misura di distanza effettiva, poichè, includendo il valore assoluto non dà spazio ad eventuali negatività.

2.4 Apprendimento non supervisionato

L'apprendimento non supervisionato è una forma di machine learning che non richiede un insieme di dati etichettati per formare un modello. Invece, l'algoritmo cerca di scoprire strutture o relazioni nei dati che sono presenti. L'obiettivo dell'ap-

prendimento non supervisionato è quello di comprendere i dati e di rappresentare il loro significato sotto forma di modelli o schemi.

2.4.1 Clustering

Il clustering è uno dei metodi più comuni di apprendimento non supervisionato. Il suo scopo è quello di raggruppare o suddividere i dati in cluster o gruppi omogenei. Ciò significa che i dati all'interno di un cluster sono simili tra loro, mentre i dati tra i cluster sono diversi. Il clustering viene utilizzato in molte applicazioni, tra cui la segmentazione del mercato, la scoperta di regole di associazione e l'analisi di cluster di dati.

Algoritmi di clustering

Gli algoritmi di clustering più comuni sono: K-Means, Hierarchical Clustering, DBSCAN, e Gaussian Mixture Model.

K-Means è un algoritmo di clustering iterativo che divide i dati in k cluster predefiniti. L'algoritmo funziona assegnando ogni punto di dati a un cluster e quindi riducendo la distanza media tra i punti di dati e il centroide del cluster.

L'Hierarchical Clustering è un algoritmo che crea una gerarchia di cluster partendo dai singoli punti di dati fino a formare cluster più grandi. Esistono due tipi di Hierarchical Clustering: Agglomerativo e Divisivo.

DBSCAN è un algoritmo di clustering basato sulla densità che identifica i cluster come regioni di alta densità di punti di dati circondati da regioni di bassa densità.

Il Gaussian Mixture Model è un algoritmo di clustering che utilizza una distribuzione gaussiana per modellare i cluster. Questo algoritmo è più flessibile rispetto a K-Means perché i cluster possono avere forme diverse da una semplice sfera.

2.4.2 K-Means

K-Means è un algoritmo di clustering che divide i dati in k cluster predefiniti. Il funzionamento di K-Means è molto semplice: inizialmente, vengono scelti casualmente k centroidi come rappresentanti dei cluster, e quindi ogni punto di dati viene assegnato al cluster il cui centroide è più vicino. Questo processo di assegnamento e riduzione della distanza viene ripetuto finché i centroidi non convergono o fino a quando il numero massimo di iterazioni è stato raggiunto.

Tuttavia, K-Means presenta alcuni problemi comuni. In primo luogo, l'algoritmo è sensibile alla scelta dei centroidi iniziali, il che significa che potrebbero essere necessarie più iterazioni per trovare la soluzione ottimale. Inoltre, K-Means presuppone che i cluster siano sfere e che i punti di dati siano uniformemente distribuiti all'interno di queste sfere, il che può non essere sempre il caso. Infine, K-Means può essere influenzato da outlier o punti di dati anomali che possono influenzare significativamente la posizione dei centroidi.

2.4.3 Metodo di Elbow

Uno dei problemi più comuni di K-Means è determinare il numero ottimale di cluster. Questo problema può essere risolto utilizzando il metodo di "Elbow". Il metodo di Elbow consiste nel calcolare la somma quadratica delle distanze tra i punti di dati e il centroide del loro cluster (conosciuta anche come "inertia"), per una serie di valori di k, e quindi disegnare un grafico di questi valori. Il punto in cui la somma quadratica delle distanze inizia a rallentare significativamente indica il numero ottimale di cluster.

In sintesi, il metodo di Elbow è uno strumento utile per determinare il numero ottimale di cluster in un dataset quando si utilizza K-Means, ma è importante tenere

Figura 2.14: Esempio di clustering effettuato tramite K-Means

presente che potrebbe non essere sempre una soluzione definitiva e che potrebbero essere necessarie ulteriori valutazioni per determinare il numero di cluster ottimali.

2.5 Pre-processing dei dati

Il pre-processing dei dati è uno step fondamentale in Data Science, questo perchè è importantissimo che i dati vengano trattati nella maniera corretta in modo tale da influire positivamente sulle performance dei modelli che andremo ad utilizzare.

In questa sede non ci concentreremo su tutte le tecniche, ma solo su quelle che sono state utilizzate nel progetto e, sulle basi teoriche dalle quali derivano.

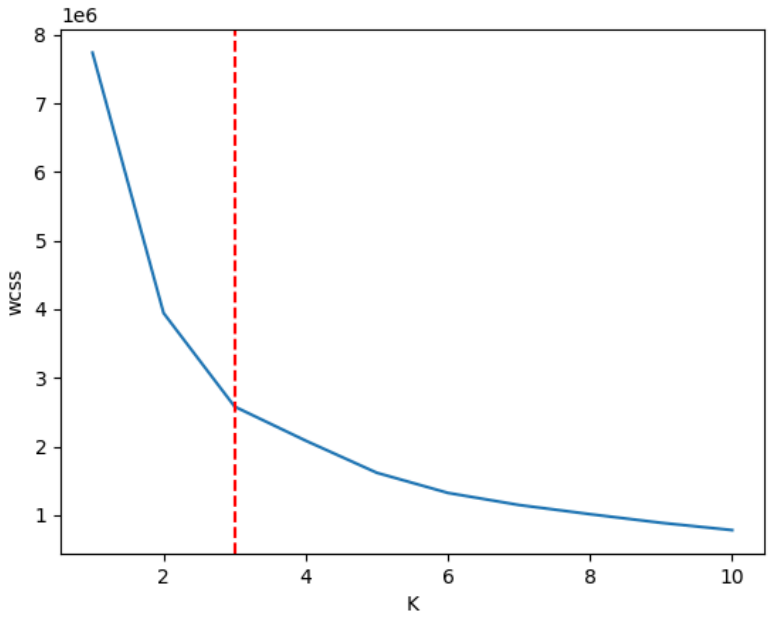


Figura 2.15: Metodo di Elbow per determinare il numero ottimale di cluster con K-Means

2.5.1 Scalare i dati

Scalare i dati è una tecnica molto importante in ambito di analisi dei dati, questo consente di normalizzare il dominio delle features che stiamo utilizzando per far si che i modelli possano lavorare con dati che condividono lo stesso dominio. Cosa significa questo? Significa che se ci sono variabili di input con valori molto grandi e relazionati ad altre variabili di input, esse possono dominare e si può tradurre in una maggior attenzione sulle stesse da parte degli algoritmi di Machine learning.

Gli algoritmi che generalmente sono soggetti a questo sono gli algoritmi come la regressione lineare, la regressione logistica, le reti neurali (incluse le LSTM) o ancora, algoritmi che utilizzano misure di distanza, come KNN. Un possibile approccio allo scaling dei dati richiede di calcolare media e deviazione standard di ogni variabile, utilizzando questi valori per poter avere una media pari a zero e una deviazione standard pari a uno, avendo una "normale standard", come distribuzione di proba-

bilità. Questo processo viene chiamato standardizzazione ed è chiaramente molto utile quando le variabili in input hanno una distribuzione di probabilità Gaussiana.

Vediamo il calcolo della standardizzazione:

$$value = \frac{(value–mean)}{\sigma} \tag{2.1}$$

A volte le variabili di input possono avere degli outlier. Questi valori ai lati della distribuzione possono avere una probabilità di occorrenza più bassa. Gli outliers possono causare una distorsione della distribuzione di probabilità e lo scaling risulterà più complesso perchè chiaramente saranno compromesse media e deviazione standard.

Un possibile approccio in presenza di outliers può essere quello di calcolare la standardizzazione ignorando gli stessi nel calcolo di media e deviazione standard, dopodichè utilizzando i valori calcolati per scalare la variabile.

Questo è chiamato Robust data scaling, che è la metologia che ho utilizzato nel progetto. Possiamo realizzare questo calcolando la mediana (50mo percentile), il 25 e il 75mo percentile. I valori di ciascuna variabile da scalare hanno la loro media sottratta e sono divise per l'interquartile (IQR), cioè la differenza p75 - p25.

$$value = \frac{(value–median)}{IQR} \tag{2.2}$$

La variabile risultante ha media e mediana pari a 0 e σ pari a 1. La distorsione risulta nulla ma la relazione tra gli outliers (che sono comunque presenti) e le altre variabili, sussiste.

Per visualizzare graficamente l'effetto dello scaling dei dati, vedendo l'immagine sotto capiamo perfettamente come possa essere utile. Dapprima sicuramente ci rendiamo conto di quanto possa essere utile per algoritmi come KNN, tuttavia risulta in ogni caso efficace proprio perchè come anticipato prima, fa si che gli algoritmi e

i modelli si concentrino in maniera equa su tutti i dati, senza preferenze di sorta.

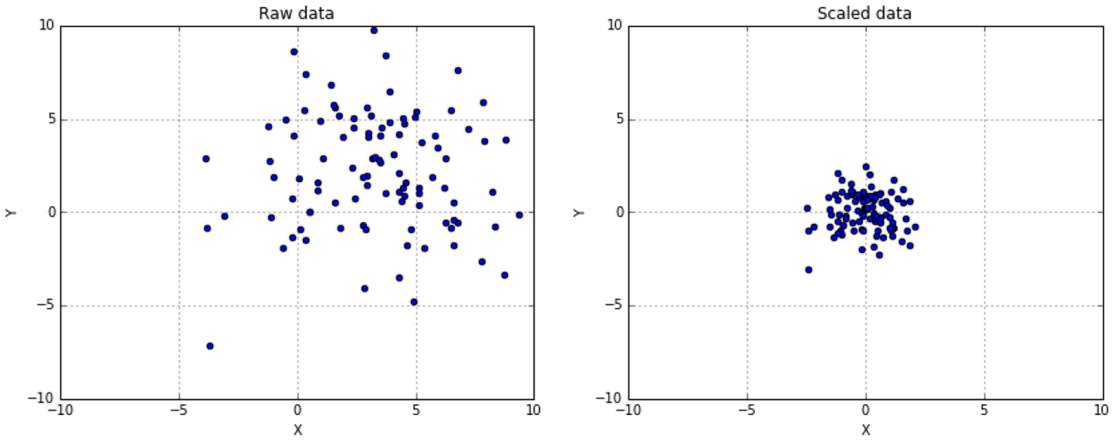


Figura 2.16: Prima e dopo lo scaling

2.5.2 Dati mancanti

Il problema dei dati mancanti è una sfida comune nell'Intelligenza Artificiale e nella Data Science. I dati mancanti possono verificarsi per una serie di motivi, come la raccolta inadeguata di dati, la perdita di informazioni durante il trasferimento o la mancanta registrazione di dati. Il problema con i dati mancanti è che possono influire negativamente sulla qualità delle analisi e dei modelli.

In molti casi, i dati mancanti possono essere semplicemente ignorati, ma questo può portare a una perdita di informazioni importanti e a una riduzione della precisione delle analisi. In alternativa, i valori mancanti possono essere sostituiti con un valore medio o con la moda, ma questo può anche distorcere i risultati delle analisi.

Per affrontare il problema dei dati mancanti, ci sono diverse tecniche di imputazione dei dati, come la media, la moda, la regressione lineare, la regressione logistica e la modellizzazione basata su processi. La scelta della tecnica più adatta dipende dalle caratteristiche specifiche del dataset e dai requisiti del progetto.

2.5.3 Undersampling e oversampling

L'undersampling e l'oversampling sono tecniche di pre-elaborazione dei dati ampiamente utilizzate nel machine learning per gestire problemi di sbilanciamento delle classi. In molti casi, i dati a disposizione possono presentare una distribuzione asimmetrica, con una classe di minoranza che rappresenta una piccola percentuale del dataset. Questa situazione può causare problemi per gli algoritmi di apprendimento automatico, che tendono a sovra-stimare la classe di maggioranza e sotto-stimare quella di minoranza.

L'undersampling prevede la riduzione della classe di maggioranza eliminando alcune delle sue istanze. In questo modo, si cerca di bilanciare la distribuzione delle classi e di ridurre l'impatto della classe di maggioranza sul risultato dell'algoritmo di apprendimento. L'oversampling, d'altra parte, prevede la replicazione della classe di minoranza per aumentarne il numero di istanze. Questa tecnica consente di rendere più equilibrata la distribuzione delle classi e di migliorare la capacità dell'algoritmo di apprendimento di rilevare le istanze della classe di minoranza.

Entrambe le tecniche presentano vantaggi e svantaggi e la scelta tra undersampling e oversampling dipende dal tipo di dati a disposizione e dal problema specifico che si vuole risolvere. Inoltre, esistono anche tecniche più sofisticate come SMOTE (Synthetic Minority Over-sampling Technique) che generano nuove istanze della classe di minoranza attraverso la creazione di istanze sintetiche basate su quelle esistenti.

2.5.4 SMOTE

SMOTE (Synthetic Minority Over-sampling Technique) è una tecnica di oversampling molto utilizzata nel machine learning per gestire problemi di sbilanciamento delle classi. SMOTE crea nuove istanze della classe di minoranza attraverso la

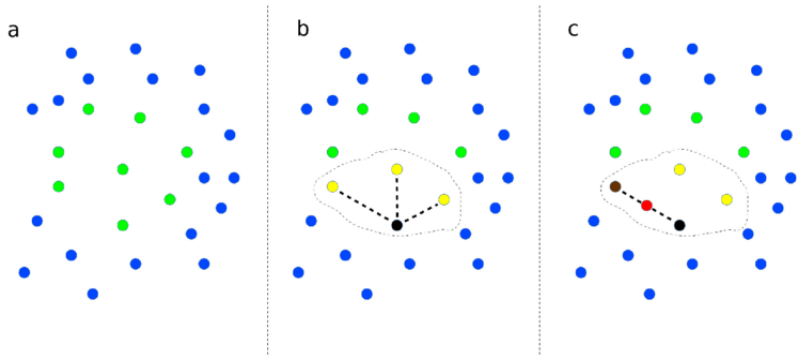


Figura 2.17: Rappresentazione grafica di SMOTE

creazione di istanze sintetiche basate su quelle esistenti. Questa tecnica prevede la selezione di una istanza di minoranza a caso e la creazione di nuove istanze attraverso l'interpolazione lineare tra questa istanza e alcune delle sue istanze vicine. In pratica, per ogni istanza di minoranza, si selezionano i suoi k-NN (k-nearest neighbors) e si creano nuove istanze lungo la linea che congiunge l'istanza originale con i suoi k-NN.

SMOTE consente di aumentare il numero di istanze della classe di minoranza senza dover replicare le istanze esistenti, migliorando così la capacità dell'algoritmo di apprendimento di rilevare le istanze della classe di minoranza. Inoltre, SMOTE può essere combinato con tecniche di undersampling per creare dataset bilanciati e gestire in modo efficace problemi di sbilanciamento delle classi.

Capitolo 3

Il modello predittivo

In questo lavoro di tesi, mi sono concentrato sulle misurazioni della qualità dell'aria nella mia città al fine di identificare eventuali relazioni e/o correlazioni tra sintomi di allergie e le concentrazioni di inquinanti. Questo è particolarmente importante poiché l'inquinamento atmosferico può aumentare la suscettibilità alle allergie e aggravare i sintomi delle persone già allergiche.

Per raggiungere il mio obiettivo, ho utilizzato un modello di apprendimento neurale, in particolare un modello LSTM, per elaborare e analizzare i dati sulle concentrazioni di inquinanti nell'aria e sui sintomi riportati dai soggetti allergici. L'obiettivo del mio studio è quello di identificare eventuali correlazioni tra le concentrazioni di inquinanti e la sintomatologia allergica al fine di comprendere meglio gli effetti dell'inquinamento atmosferico sulla salute umana.

Questo lavoro può portare a nuove strategie di prevenzione e trattamento per le persone che soffrono di allergie e di altre patologie correlate all'inquinamento atmosferico. Inoltre, questo studio potrebbe essere utilizzato come base per futuri studi sulla qualità dell'aria e sulla salute umana.

3.1 Scelta del dataset

La scelta del dataset è stata una parte fondamentale e complessa del lavoro di tesi, poiché è essenziale disporre di dati di qualità per addestrare modelli di Machine Learning. Tuttavia, la raccolta di dati di qualità può rivelarsi una sfida, in quanto non sempre le misurazioni raccolte sono indicative e affidabili, nonostante l'uso di semplici sensori.

Ho cercato di raccogliere un grande quantitativo di dati, al fine di filtrarli e selezionare quelli più rilevanti per la mia ricerca. Ho individuato alcuni inquinanti che sembrano avere un impatto significativo sulla sintomatologia allergica e ho scelto di concentrarmi su di essi.

Tuttavia, durante l'analisi dei dati, mi sono reso conto che alcuni file presentavano dati mancanti e, purtroppo, ho dovuto escluderli dalle misurazioni. Questa scelta ha comportato una perdita di informazioni e una riduzione delle dimensioni del dataset, ma era necessaria per garantire la qualità dei dati utilizzati.

Inoltre, ho deciso di utilizzare solo una delle zone raccolte per avviare l'apprendimento del modello predittivo, in modo da poter utilizzare le altre per migliorare le performance del modello. Questo ha richiesto una selezione accurata della zona, basata sulla disponibilità di dati di alta qualità e sulla rilevanza degli inquinanti misurati.

3.2 Analisi dei dati

Dopo aver effettuato la scelta del dataset, ho proceduto ad analizzare approfonditamente i contenuti dello stesso, impiegando tecniche di statistica e data visualization al fine di ottenere una comprensione approfondita della natura dei dati raccolti.

In primo luogo, ho eseguito un'analisi temporale delle misurazioni dei vari inquinanti e della qualità dell'aria, mediante l'utilizzo di grafici appositamente elaborati. Successivamente, ho confrontato i dati raccolti e ho commentato i risultati ottenuti, attraverso una procedura di brainstorming, volto ad individuare eventuali correlazioni tra i dati. In particolare, ho potuto identificare inquinanti come il PM10, che presentavano una mediana superiore rispetto agli altri inquinanti, e che quindi potrebbero avere un maggiore impatto sulla qualità dell'aria.

Tale procedura di analisi dei dati ha rappresentato una fase fondamentale del mio lavoro di tesi, poiché mi ha permesso di acquisire una conoscenza più approfondita dei dati raccolti, individuando eventuali problematiche e criticità, che successivamente ho dovuto affrontare per giungere ad una elaborazione dei dati di maggiore affidabilità.

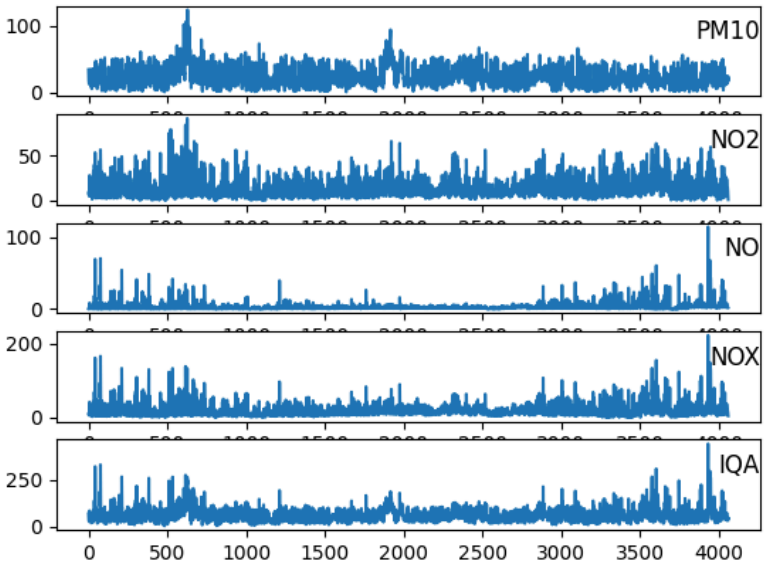


Figura 3.1: Misurazioni nel tempo

Clusters

Nella mia tesi ho approfondito l'uso del clustering come tecnica di apprendimento non supervisionato per dividere il dataset in gruppi, al fine di identificare le varie qualità dell'aria. Secondo le linee guida del Ministero della Salute, sarebbero stati necessari 5 gruppi, ma utilizzando il K-means, il risultato è stato la definizione di soli tre cluster ben definiti. Questo risultato è stato interessante perché ha permesso di evidenziare come la qualità dell'aria annuale della città non sia fortunatamente troppo scarsa, e che i dati del dataset presentino pattern ben definiti solo per alcune range di qualità, in particolare quelli più buoni.

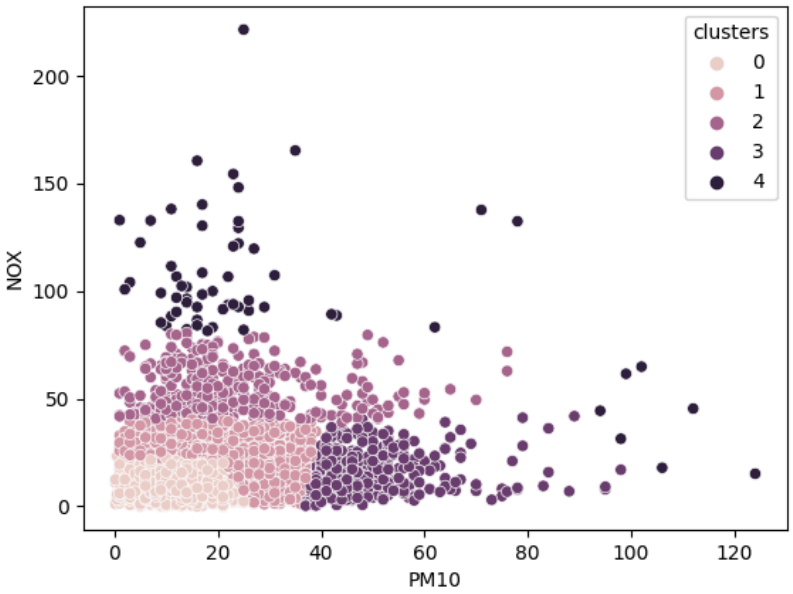


Figura 3.2: Clustering con K-Means pre-oversampling

Durante la fase di pre-processing dei dati, ho compreso l'importanza di approfondire la questione relativa alla distribuzione dei dati in funzione della qualità dell'aria. Se avessi utilizzato questi dati per addestrare un modello predittivo, questo avrebbe ottenuto risultati migliori solo per i range di qualità dell'aria relativamente buoni,

risultando poco efficace per dati relativi a qualità dell'aria scarsa. Pertanto, ho considerato fondamentale studiare le distribuzioni di alcuni inquinanti x in funzione di altri inquinanti y, al fine di comprendere meglio la distribuzione complessiva dei dati.

Successivamente, ho utilizzato l'algoritmo K-means per suddividere il dataset in 5 clusters, tuttavia ho ritenuto necessario confermare questa scelta analizzando i dati in modo più approfondito. Per fare ciò, ho applicato il metodo di Elbow per determinare il numero ideale di clusters. I risultati ottenuti hanno confermato l'ipotesi precedente, rivelando che il valore di K ideale sarebbe pari a 3.

Oversampling

Per confermare la teoria esposta in precedenza, ho deciso di utilizzare il metodo di oversampling SMOTE per generare dati relativi ai gruppi che sembravano essere una minoranza. Come si può notare, dopo aver effettuato il processo di oversampling, i clusters risultano essere molto più definiti rispetto a prima. È importante sottolineare che non mancano problemi di rumore o outliers, tuttavia ciò era prevedibile dal momento che, fondamentalmente, ho creato dei dati partendo da una minoranza.

Distribuzioni e correlazioni

Successivamente ho visualizzato le distribuzioni di ogni inquinante in funzione degli altri, questo ha contribuito ad alcune scelte fatte successivamente, nel progetto.

Successivamente, mi sono focalizzato sulla ricerca di correlazioni tra gli inquinanti nel dataset. L'obiettivo della mia tesi è di studiare le regressioni e le correlazioni tra gli inquinanti e la sintomatologia di soggetti allergici. Tuttavia, prima di esplorare tali correlazioni, è stato importante valutare se ci fossero correlazioni tra gli inquinanti stessi.

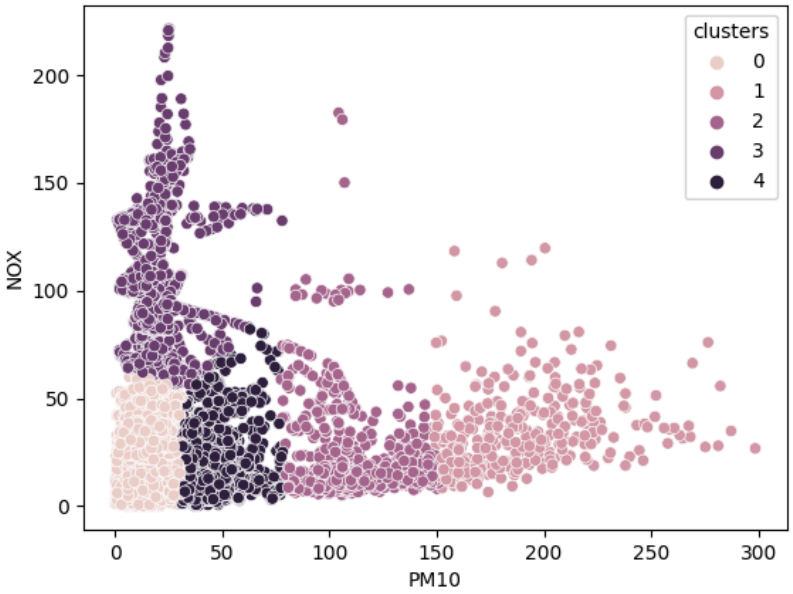


Figura 3.3: Clustering con K-Means post-oversampling

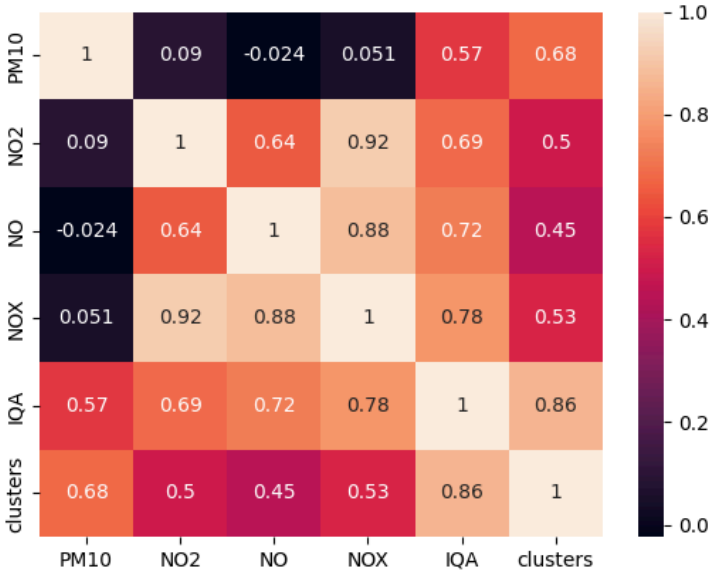


Figura 3.4: Correlazioni

Utilizzando il coefficiente di correlazione di Pearson (r), ho verificato che, ad esempio, la correlazione tra NOX e IQA è pari a 0.72, mentre quella tra PM10 e IQA è pari a 0.57 . Inoltre, ho notato che NO, NO2 e NOX sono fortemente correlati tra loro, mentre sembra che nessuno dei tre inquinanti sia correlato con il PM10. In particolare, nel caso di NO e NOX, sembra che questi siano inversamente correlati con il PM10. Tuttavia, va sottolineato che tutti questi inquinanti hanno un forte impatto sulla qualità dell'aria e che l'assenza di una correlazione tra due inquinanti non implica necessariamente che uno dei due non abbia effetto sull'altro.

3.3 Raccolta dati

La prima fase del mio lavoro è stata dedicata alla ricerca di sintomi specifici per soggetti allergici, al fine di correlarli con i dati appena visualizzati. Questa fase, seppur delicata, è stata una delle più importanti in quanto mi ha permesso di trattare i sintomi da un punto di vista statistico e matematico, fornendomi così un altro aspetto su cui lavorare.

I sintomi sono la manifestazione di uno stato patologico e, per alcune patologie o situazioni, è dimostrato empiricamente che una certa patologia possa portare a una determinata tipologia di sintomi. Spesso i sintomi possono essere facilmente dimostrabili e considerati come appartenenti al range di valori prodotti da una certa situazione patologica.

Nel caso delle allergie, sono noti alcuni sintomi comuni che possono dipendere dalla tipologia di allergia e dagli allergeni o pollini coinvolti. Tuttavia, è importante sottolineare che questi sintomi possono variare da individuo a individuo, e che situazioni di stress o ansia possono portare ad un aumento della sintomatologia, rendendola meno affidabile nell'analisi dei dati. Nonostante ciò, in media, è possibile identificare i seguenti sintomi come comuni:

- Sonnolenza
- Asma
- Gola Irritata
- Occhi gonfi
- Lacrimazione
- Starnuti
- Prurito
- Difficoltà respiratorie
- Tosse

La raccolta dei dati è stata effettuata attraverso tecniche di intervista e di focus group, poiché la stagionalità dei pollini più frequenti non era compatibile con il periodo di scrittura di questa tesi. È stato selezionato un campione di soggetti allergici a cui è stato somministrato un breve questionario, con l'obiettivo di valutare la frequenza dei sintomi sopracitati in un periodo di circa 120 giorni scelto dai partecipanti stessi.

Assumendo che ogni individuo abbia scelto il periodo dell'anno più problematico dal punto di vista dei sintomi, l'informazione sulla frequenza f risulta utile per comprendere la distribuzione dei sintomi durante il periodo considerato. Inoltre, considerando anche le informazioni sulle concentrazioni degli inquinanti raccolte durante lo stesso periodo, questa informazione risulta ancora più preziosa per la nostra analisi.

È importante notare che il questionario è stato talvolta somministrato a membri dello stesso nucleo familiare, il che potrebbe portare a considerazioni importanti

dal punto di vista probabilistico. I risultati del questionario sono stati riportati in una tabella simile a quella rappresentata in figura, dove ogni riga rappresenta un individuo, le colonne rappresentano i sintomi e i valori nella tabella corrispondono alla frequenza presunta del sintomo per ciascun individuo durante i 120 giorni considerati. In particolare, abbiamo convenuto che il valore 0 indica la totale assenza di sintomi, mentre il valore 120 indica la presenza del sintomo in tutti i giorni del periodo considerato.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Parm	PM10	NO2	NO	NOX	IQA	starnuti	tosse	asma	difficolta_	mal_di_te	sonnolenz	occhi_gon	lacrimazio	gola_irrita	prurito
2	1/3/21 12.	6.35	3.28	0.07	3.85	13	0	0	0	0	0	0	0	0	0	0
3	1/3/21 6.0	13.68	4.97	0.52	6.24	27	0	13	0	0	0	0	0	0	0	0
4	1/3/21 8.0	11.97	19.66	3.33	25.23	24	30	0	8	8	0	0	0	0	0	0
5	1/3/21 10.	17.83	18.45	4.11	25.21	36	30	13	8	0	0	0	0	0	24	43
6	1/3/21 12.	16.61	10.24	2.09	13.9	33	30	0	0	0	0	0	0	0	24	0
7	1/3/21 2.0	47.62	12.75	3.45	18.5	95	30	13	0	0	0	0	0	0	0	43
8	1/3/21 4.0	23.69	13.61	4.07	20.31	47	0	13	0	0	0	0	0	0	24	0
9	1/3/21 6.0	16.36	18.52	2.89	23.41	33	30	0	0	0	0	0	0	0	0	0
10	1/3/21 8.0	19.05	27.54	1.94	30.98	38	0	13	8	8	18	18	0	0	0	43
11	1/3/21 10.	20.76	19.62	0.72	21.18	42	30	0	8	0	18	0	0	0	0	0
12	2/3/21 12.	24.42	6.57	0.05	7.11	49	0	13	8	0	18	18	0	0	24	43
13	2/3/21 4.0	25.15	3.33	0.07	3.89	50	30	0	8	0	18	0	0	0	0	43
14	2/3/21 6.0	17.09	10.86	0.77	12.52	34	30	0	8	0	0	18	0	0	0	43
15	2/3/21 8.0	14.41	53	31.58	101.88	51	0	13	8	0	0	18	0	0	0	43
16	2/3/21 10.	21.49	37.81	18.09	66.01	43	0	13	8	0	0	18	0	0	0	43
17	2/3/21 12.	27.59	16.8	5.14	25.15	55	0	0	8	0	18	18	0	0	24	0
18	2/3/21 2.0	30.04	10.78	2.91	15.71	60	0	0	8	0	18	18	0	0	0	0
19	2/3/21 4.0	55.19	12.65	2.95	17.64	110	0	13	8	0	18	18	0	0	24	0
20	2/3/21 6.0	33.94	34.66	8.9	48.76	68	0	13	8	0	18	18	0	0	0	43
21	2/3/21 8.0	25.64	66.21	15.84	90.95	51	0	13	8	0	18	18	0	0	0	0
22	2/3/21 10.	30.04	40.18	5.42	48.95	60	0	13	8	0	18	0	0	0	24	43
23	3/3/21 12.	39.56	12.41	0.24	13.23	79	0	13	8	0	18	0	0	0	0	0
24	3/3/21 2.0	42.49	5.68	0.1	6.3	85	0	13	8	0	18	18	0	0	24	0
25	3/3/21 4.0	38.1	5.37	0.53	6.65	76	0	13	8	0	0	18	0	0	0	0

Figura 3.5: Dataset concentrazioni più sintomi

Dato il dataset, ho deciso di riassumere i risultati distribuendoli in modo probabilistico tra i vari step temporali delle misurazioni, ottenendo la probabilità $P(s)$ per ogni sintomo. Successivamente, ho calcolato la probabilità condizionata $P(s|m)$ per ogni riga, ovvero la probabilità che quel sintomo fosse presente o meno in una determinata frazione di tempo considerata, conoscendo la sua media.

Per implementare questa idea, ho utilizzato Python e scikit-learn [6] per costruire un classificatore Bayesiano. In particolare, per ogni tempo t, il classificatore ha assegnato una classe Cr al sintomo, dove $Cr \in P, NP$ a seconda che la probabilità $P(s|m)$ fosse maggiore o minore di una certa soglia.

Il risultato finale è stato un nuovo dataset contenente le misurazioni, i sintomi e le relative frequenze medie nel caso in cui il classificatore avesse scelto la classe P, e il valore 0 nel caso in cui avesse scelto la classe NP. Sebbene questo sia stato un metodo semplice per valutare le interviste e il questionario somministrati, potrebbe essere migliorato ad esempio utilizzando un sistema di apprendimento basato sulla conoscenza della probabilità al tempo $t-1$. Tuttavia, dal mio punto di vista, questa metodologia potrebbe essere computazionalmente impraticabile e sarebbe preferibile adottarne altre più semplici per la raccolta dei dati.

3.4 LSTM

Dopo aver analizzato i dati raccolti, ho deciso di testare le mie ipotesi costruendo un modello predittivo in grado di stimare i valori dell'IQA, delle concentrazioni degli inquinanti e delle probabilità di presenza o assenza di sintomi. La scelta è ricaduta su una rete neurale a memoria a lungo termine (LSTM), in quanto ideale per affrontare problemi che coinvolgono serie temporali. Tuttavia, le serie temporali erano presenti solo a livello concettuale e ho dovuto sviluppare una procedura significativa per trasformare il dataset in un formato adatto a un modello di apprendimento supervisionato. Di seguito, viene presentato il codice sviluppato per questa operazione:

```
1
2 # conversione serie temporali in SL
3 def series_to_supervised(data, n_in=1, n_out=1, dropnan=True↩
    ):
4     n_vars = 1 if type(data) is list else data.shape[1]
5     df = pd.DataFrame(data)
```

```
6       cols, names = list(), list()
7       # input (t-n, ... t-1)
8       for i in range(n_in, 0, -1):
9           cols.append(df.shift(i))
10          names += [('var%d(t-%d)' % (j + 1, i))
11          for j in range(n_vars)]
12      # time (t, t+1, ... t+n)
13      for i in range(0, n_out):
14          cols.append(df.shift(-i))
15          if i == 0:
16              names += [('var%d(t)' % (j + 1))
17              for j in range(n_vars)]
18          else:
19              names += [('var%d(t+%d)' % (j + 1, i))
20              for j in range(n_vars)]
21      # Concatena
22      agg = pd.concat(cols, axis=1)
23      agg.columns = names
24      # toglie righe con val NaN
25      if dropnan:
26          agg.dropna(inplace=True)
27      return agg
```

Le reti LSTM rappresentano un approccio multivariato al problema di previsione, in quanto consentono di produrre più parametri in output. Nel mio lavoro di tesi, ho scelto di focalizzarmi su un solo output, ma la metodologia può essere facilmente estesa per prevedere più valori per ogni livello.

Per ottenere ciò, ho utilizzato la proprietà *Dense* della rete. Questa proprietà

indica che il livello di neuroni è completamente connesso ai neuroni del livello successivo, implementando così le proprietà della LSTM descritte nei capitoli precedenti. Sarebbe interessante valutare le performance del modello in caso di variazione del parametro della proprietà.

Passando agli iperparametri della rete neurale, cerchiamo di analizzarli nel dettaglio, motivando la scelta di ognuno di essi.

```
1
2  model = Sequential()
3  model.add(LSTM(50, input_shape=(train_X.shape[1], train_X.↩
       shape[2])))
4  model.add(Dense(1))
5  model.compile(loss='mae', optimizer='adam')
6
7
8  history = model.fit(train_X, train_y, epochs=100,
9  batch_size=72, validation_data=(test_X, test_y),
10 verbose=2,
11                      shuffle=False)
```

Listing 3.1: "Configurazione della LSTM"

Il codice Python scritto rappresenta la configurazione della LSTM. In particolare, il modello è stato costruito con il metodo *Sequential*(), il quale prevede una pila di livelli, ognuno dei quali contiene esattamente un tensore in input e un tensore in output. Per il nostro lavoro di tesi, abbiamo scelto un modello con un singolo output, ma il modello sequenziale ci permette di estendere facilmente la previsione a più valori per ogni livello, grazie alla proprietà *Dense*. Un livello Dense è un livello

completamente connesso, dove tutti i neuroni del livello N sono connessi a tutti i neuroni del livello N+1, e implementa le proprietà della LSTM esposte nei capitoli precedenti.

Per quanto riguarda gli iperparametri scelti per la rete neurale, abbiamo cercato di darne una motivazione ad ognuno di essi. Per esempio, l'iperparametro *Dense* ha valore 1 perché non abbiamo previsto soluzioni alternative per questo lavoro di tesi. Invece, per il metodo *compile*, abbiamo scelto la metrica di errore "MAE" come funzione di loss e l'ottimizzatore "Adam".

3.5 Sperimentazioni

Le sperimentazioni rappresentano una fase fondamentale in ogni lavoro di ricerca e sviluppo di un modello predittivo. In particolare, per la costruzione del modello predittivo è necessario effettuare una serie di sperimentazioni che permettano di valutare l'efficacia del modello stesso nella previsione dei sintomi dei soggetti allergici in base alle concentrazioni degli inquinanti nell'aria. Le sperimentazioni consentono di valutare la bontà del modello, fornendo così informazioni importanti sulla sua capacità di generalizzare a dati non visti in precedenza. In questo paragrafo, saranno presentati i dettagli delle sperimentazioni condotte, le tecniche utilizzate per la valutazione del modello e i risultati ottenuti.

3.5.1 Predizione

Dopo aver scalato i dati del dataset utilizzando il Robust Scaler di scikit-learn, ho effettuato il fitting della rete neurale utilizzando i dati di una specifica area della città metropolitana di Bari. Una volta addestrato il modello, ho testato le sue prestazioni utilizzando un set di dati proveniente da un'altra zona della città, ottenendo un RMSE di 6.811. Questo ha sollevato il sospetto che il modello potesse

avere difficoltà a generalizzare su dati diversi da quelli utilizzati per l'addestramento e la convalida iniziali.

Per capire meglio la situazione, ho deciso di espandere il set di dati di test con dati provenienti da entrambe le zone, in modo da valutare se il modello fosse in grado di generalizzare su un range più ampio di pattern o condizioni. Tuttavia, è importante considerare la possibilità che l'espansione del set di dati possa introdurre bias e quindi essere sicuri che questi siano stati correttamente selezionati e preparati per l'uso nella valutazione del modello.

Dopo aver aggiunto i dati della seconda zona al set di dati originale, ho ottenuto un miglioramento significativo delle prestazioni del modello, con un RMSE di 0.726. Ci sono diverse spiegazioni possibili per questo comportamento della rete.

Potrebbe essere il caso che la rete abbia imparato bene a rappresentare le relazioni tra inquinanti e qualità dell'aria della prima zona, ma non sia riuscita a generalizzare bene sulla seconda zona, a causa di differenze tra le due zone come inquinanti differenti, topografia o densità della popolazione.

Al contrario, aggiungendo parte dei dati della seconda zona alla prima, la rete ha avuto la possibilità di apprendere informazioni aggiuntive, migliorando le prestazioni nella seconda zona ma non necessariamente anche nella prima.

Inoltre, l'RMSE è solo una misura della deviazione tra i valori previsti della rete neurale e i valori reali della qualità dell'aria e non ci fornisce alcuna indicazione sulla qualità della rappresentazione dei dati. È possibile che la rete stia fornendo previsioni precise ma sbagliate a causa di una relazione errata appresa tra inquinanti e qualità dell'aria.

Per comprendere meglio il comportamento della rete neurale sarebbe necessario effettuare ulteriori analisi sui dati e sulla sua architettura. Tuttavia, l'aggiunta di dati del set di test provenienti da diverse zone della città metropolitana di Bari ha dimostrato di essere una tecnica efficace per migliorare le prestazioni del modello e

aumentare la sua capacità di generalizzazione su dati diversi da quelli utilizzati per l'addestramento.

3.5.2 Sperimentazione

Una parte della sperimentazione è stata dedicata allo studio delle correlazioni tra gli inquinanti considerati nelle zone utilizzate per il test e i sintomi raccolti dagli utenti con il questionario. Ho utilizzato la tabella realizzata con l'algoritmo descritto al paragrafo 3.3 Raccolta dati.

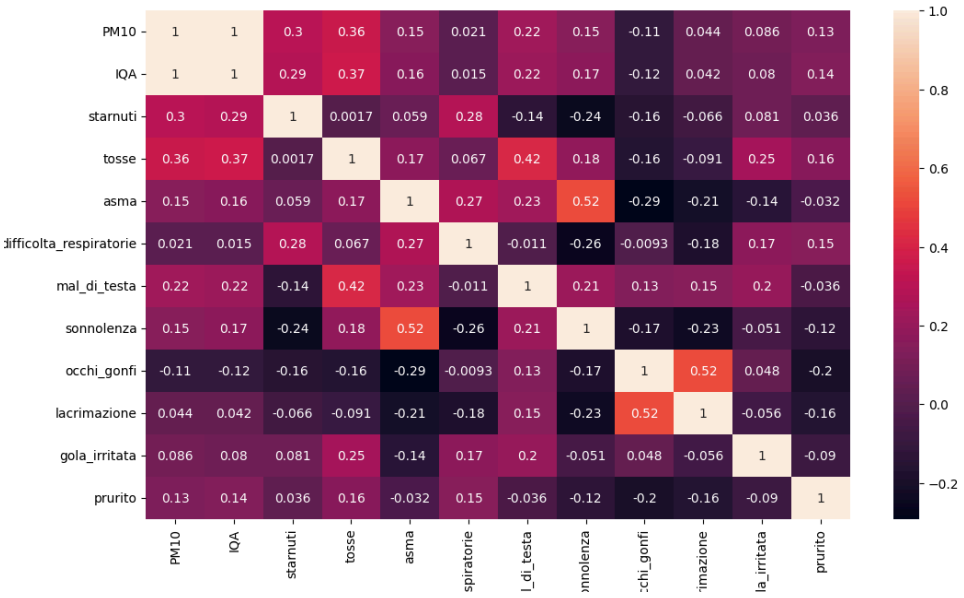


Figura 3.6: Matrice di correlazione dei sintomi raccolti con PM10 e IQA

Osservando la matrice dei dati raccolti è possibile individuare alcune correlazioni che meritano un approfondimento. In particolare, si può notare una correlazione debolmente positiva tra il PM10 e la qualità dell'aria con la tosse. Questo risultato ha un valore statistico e conferma, seppur debolmente, parte degli studi scientifici in materia. Infatti, è stato dimostrato che il PM10 è impattante nelle malattie allergiche respiratorie pediatriche [7].

Tuttavia, è importante considerare che la debolezza della correlazione potrebbe essere dovuta all'età media dei soggetti del questionario e alla quantità limitata di dati disponibili.

Un'altra correlazione debolmente positiva rilevata è quella tra la qualità dell'aria (IQA) e il PM10 con il mal di testa. Anche in questo caso, il valore statistico della correlazione potrebbe essere influenzato dalla qualità e dalla quantità dei campioni analizzati. Tuttavia, questo risultato è coerente con altre ricerche condotte in passato.

In particolare, uno studio pubblicato su "Environmental Health Perspectives" nel 2010 ha esaminato la relazione tra l'esposizione a PM10 e il mal di testa in una popolazione di bambini allergici residenti nella città di Fresno, California. I ricercatori hanno riscontrato che l'aumento dei livelli di PM10 era associato ad un aumento del rischio di mal di testa nei bambini allergici [8].

Un altro studio pubblicato su "The Journal of Allergy and Clinical Immunology" nel 2015 ha esaminato la relazione tra l'esposizione a PM10 e la gravità dei sintomi allergici, tra cui il mal di testa, in una popolazione di pazienti con rinite allergica. I ricercatori hanno riscontrato che l'aumento dei livelli di PM10 era associato ad un aumento della gravità dei sintomi allergici, compreso il mal di testa [9].

La seconda matrice analizzata evidenzia la correlazione tra sonnolenza e inquinanti atmosferici considerati. Tale correlazione è coerente con diversi studi che indicano una relazione tra l'esposizione a inquinanti atmosferici come NO (ossido di azoto), NOx (ossidi di azoto) e NO2 (biossido di azoto) e i sintomi allergici, in particolare l'asma e la rinite allergica.

Ad esempio, uno studio pubblicato su "The Journal of Allergy and Clinical Immunology" nel 2019 ha esaminato la relazione tra l'esposizione agli inquinanti atmosferici e l'asma allergica in una popolazione di bambini in età scolare in Cina. I risultati hanno indicato che l'esposizione a NO2 e PM2.5 (particolato fine) era

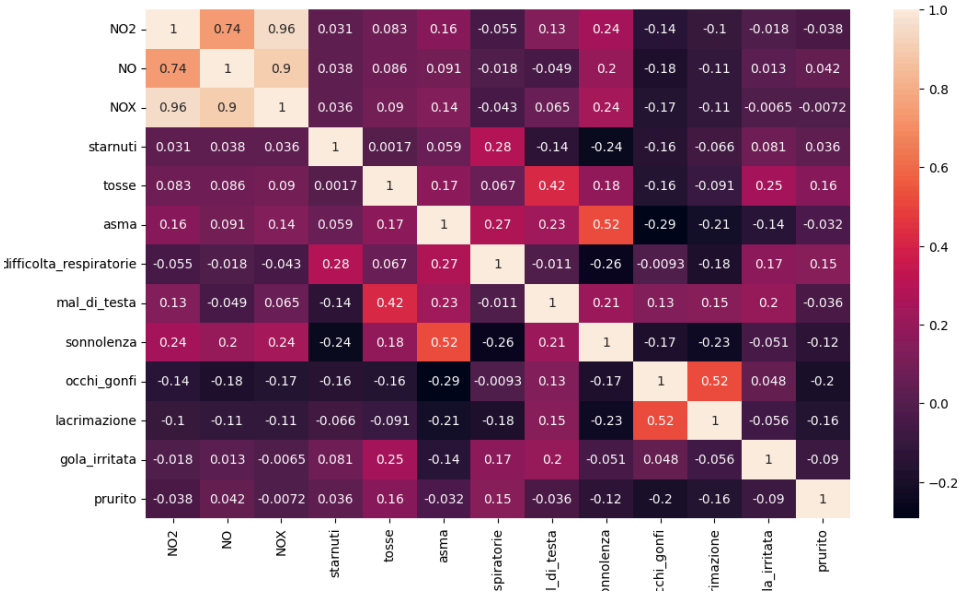


Figura 3.7: Matrice di correlazione dei sintomi raccolti con NO, NO2, NOX

associata ad un aumentato rischio di sintomi asmatici. [10]

In un altro studio pubblicato su "Environmental Health Perspectives" nel 2018, i ricercatori hanno valutato l'effetto dell'esposizione agli inquinanti atmosferici sulla rinite allergica in una popolazione di adulti in Spagna. I risultati hanno indicato che l'esposizione a NO2 e PM2.5 era associata ad un aumento dei sintomi di rinite allergica. [11]

Un terzo studio pubblicato su "Environmental Research" nel 2020 ha esaminato la relazione tra l'esposizione agli inquinanti atmosferici e la prevalenza di allergie respiratorie in una popolazione di adulti in Giappone. I risultati hanno mostrato che l'esposizione a NO2 e PM2.5 era associata ad un aumentato rischio di asma e rinite allergica. [12]

L'insieme di questi studi suggerisce che l'esposizione agli inquinanti atmosferici come NO, NOx e NO2 può aumentare la suscettibilità ai sintomi allergici. Tuttavia, va sottolineato che gli effetti dell'inquinamento atmosferico sulla salute dipendono

da diversi fattori, come la concentrazione degli inquinanti, la durata dell'esposizione e la suscettibilità individuale.

Capitolo 4

Conclusioni e sviluppi futuri

4.1 Conclusioni

Il mio lavoro di tesi mi ha permesso di confermare, seppur in maniera limitata, alcune ipotesi riguardanti l'inquinamento atmosferico e la sintomatologia ad esso correlata, soprattutto nei soggetti allergici. Grazie a semplici interviste, sperimentazioni e uno studio approfondito dei risultati scientifici degli ultimi anni, ho avuto la possibilità di concretizzare queste considerazioni.

È interessante notare come i sintomi tipicamente legati alle manifestazioni allergiche siano simili a quelli attribuiti all'inquinamento atmosferico. Questa osservazione è stata il punto di partenza della mia ricerca. Infatti, la maggior parte dei campioni considerati nel mio studio sono stati soggetti allergici, il che ha portato a una serie di riflessioni che ritengo importante approfondire.

In sintesi, il mio lavoro di tesi ha contribuito ad ampliare la conoscenza sull'influenza dell'inquinamento atmosferico sulla salute umana, in particolare sui soggetti allergici. Spero che i risultati ottenuti possano essere utili per futuri approfondimenti e per sviluppare nuove strategie di prevenzione e di cura delle patologie correlate all'inquinamento.

4.2 Sviluppi futuri

Una possibile soluzione per migliorare la raccolta dei dati potrebbe essere l'utilizzo di sensori portatili indossabili dai soggetti che rilevino in tempo reale i livelli di inquinamento atmosferico e il manifestarsi dei sintomi. Questo approccio potrebbe permettere una raccolta continua e precisa dei dati, riducendo al minimo la possibilità di errori e imprecisioni.

Inoltre, sarebbe utile approfondire le correlazioni positive considerate in maniera più dettagliata, analizzando anche le eventuali interazioni tra i diversi sintomi e l'inquinamento atmosferico. Questo potrebbe fornire ulteriori informazioni sulla patologia e permettere di individuare eventuali fattori di rischio specifici per i soggetti allergici.

Infine, il modello predittivo realizzato potrebbe essere utilizzato per fornire all'utente informazioni personalizzate sulla previsione della sintomatologia, in modo da permettere un intervento tempestivo e mirato. Ad esempio, l'utente potrebbe ricevere una notifica quando i livelli di inquinamento atmosferico sono particolarmente elevati, in modo da poter prendere precauzioni per evitare l'insorgere di sintomi allergici. In questo modo, il modello potrebbe contribuire a migliorare la qualità della vita dei soggetti allergici e a prevenire eventuali complicazioni legate alla patologia.

Bibliografia

[1] Dhairya Kumar (25 Dicembre 2018). *Introduction to Data Preprocessing in Machine Learning (towardsdatascience.com)*

[2] D.Poole,A.Mackworth *Artificial Intelligence: Foundations of Computational Agents*

[3] Wikipedia *Regressione Lineare*

[4] Rian Dolphin *LSTM Networks — A Detailed Explanation*

[5] Colah's blog *Understanding LSTM*

[6] Naive Bayes *Naive Bayes - scikit-learn*

[7] L'impatto dell'inquinamento atmosferico nelle malattie allergiche respiratorie pediatriche *Rivista Italiana di Allergologia e Immunologia Pediatrica, organo ufficiale SIAIP*

[8] Esteban G. Burchard et al. (2010). Environmental Health Perspectives. 118(11): 1431-1435. *Particulate Matter and Childhood Asthma: Recent Advances and Future Directions*

[9] Jin H. Kim et al. (2015). The Journal of Allergy and Clinical Immunology. 135(1): 116-123.e1 *Association between air pollution and allergic rhinitis in Korean children: A population-based study*

[10] Huang, F., Liu, X., Zhang, Y., Liu, Y., Wang, B., Ren, P., Zhang, Y. (2019) *Association between air pollutants and asthma symptoms in pediatric population: A time-series study in Suzhou, China*

[11] Dadvand, P., Nieuwenhuijsen, M. J., Esnaola, M., Forns, J., Basagaña, X., Alvarez-Pedrerol, M., Sunyer, J. (2018) *Air pollution and the prevalence of rinitis among adults in southern Europe: A cross-sectional study. Environmental Health Perspectives*

[12] Nakayama, S. F., Takeuchi, A., Miyake, Y., Tanaka, Y., Kurasawa, M., Araki, A., Shibata, E. (2020). Environmental Research *Associations between exposure to NO2 and PM2. 5, residential greenness, and allergic diseases in Japanese adults: A cross-sectional study.*