UNIVERSITÀ
DEGLI STUDI
DI BRESCIA

DIPARTIMENTO DI INGEGNERIA
DELL'INFORMAZIONE

Corso di Laurea in Ingegneria Informatica

Relazione Finale

Tecniche data-driven per la previsione degli inquinanti in atmosfera

Relatore:

Chiar.mo Prof. Claudio Carnevale

Laureando:

Federico Dagani
Matricola n. 732028

ANNO ACCADEMICO 2022/2023

Indice

Introduzione

L'analisi degli inquinanti atmosferici è diventata una priorità cruciale per gli scienziati, i governi e le organizzazioni ambientali in tutto il mondo. Questa analisi non solo ci fornisce dati preziosi sulla composizione chimica dell'atmosfera, ma anche sulle fonti ed i processi che contribuiscono all'inquinamento.
Comprendere l'origine e la diffusione degli inquinanti atmosferici è fondamentale per sviluppare politiche efficaci di controllo dell'inquinamento, per proteggere la salute pubblica e per mitigare gli impatti del cambiamento climatico.

L'inquinamento atmosferico è composto da una serie di sostanze nocive e particelle che vengono rilasciate nell'atmosfera a causa dell'attività umana, industriale ed ambientale. Queste sostanze comprendono gas come il biossido di azoto (NO2), il biossido di zolfo (SO2), l'anidride carbonica (CO2) e molte altre particelle sospese nell'aria.

L'accumulo di tali inquinanti ha conseguenze significative sulla qualità dell'aria che respiriamo, sulla salute umana, sulla biodiversità e sul cambiamento climatico.

Nel contesto di questa sfida globale, la mia tesi si concentra sulla predizione dei livelli di biossido di azoto (NO2) nell'atmosfera. Il biossido di azoto è un inquinante atmosferico significativo, noto per le sue gravi implicazioni sulla salute umana e per il suo contributo ai cambiamenti climatici.

Attraverso l'analisi dei dati e l'utilizzo di modelli predittivi, cercheremo di comprendere meglio i fattori che influenzano la concentrazione di NO2 nell'aria, contribuendo così ad una comprensione più ampia del cambiamento climatico e all'identificazione di strategie per mitigarne gli effetti.
La nostra ricerca dimostra che l'analisi dell'inquinamento atmosferico, come il NO2, non solo è rilevante per la salute pubblica, ma è anche fondamentale per affrontare la crisi climatica che sta plasmando il nostro pianeta ed il nostro futuro.

Nel corso della mia ricerca, ho adottato un approccio data-driven per affrontare la sfida della predizione dei livelli di biossido di azoto (NO2) nell'atmosfera.
Questo ha permesso di sfruttare l'enorme quantità di dati disponibili per comprende-

re meglio le dinamiche complesse che influenzano la concentrazione di NO2 nell'aria.

Per l'implementazione ho utilizzato il linguaggio di programmazione Python insieme ad una serie di strumenti e librerie specializzate per l'analisi dei dati e la costruzione di modelli predittivi.
L'uso di Python come linguaggio principale mi ha fornito la flessibilità necessaria per eseguire analisi approfondite sui dati e sviluppare modelli avanzati.

In particolare, ho utilizzato algoritmi di machine learning e tecniche di data science per idetificare e validare i modelli predittivi. Questi modelli sfruttano un vasto insieme di dati storici sulle concentrazioni di NO2, insieme a variabili correlate come la temperatura al suolo, emissioni del traffico, la velocità del vento e la radiazione solare.

Capitolo 1

Inquinanti atmosferici

In questo capitolo viene analizzato il problema dell'inquinamento dell'aria. Particolare attenzione viene posta sull'inquinante NO2 (biossido di azoto), alla sua formazione ed ai suoi effetti sulla persona e sull'ecosistema. Infine vengono brevemente riportate le normative in vigore in tema di inquinamento atmosferico.

1.1 L'atmosfera

L'atmosfera è uno strato di gas che circonda la Terra ed è fondamentale per la vita sul nostro pianeta. Essa è costituita da una miscela di gas, particelle e vapore acqueo che si estende per centinaia di chilometri sopra la superficie terrestre.
L'atmosfera svolge diverse funzionalità cruciali, tra cui la protezione dalla radiazione solare dannosa, la regolazione delle temperature sulla Terra e la fornitura di ossigeno necessario per la respirazione.

I principali componenti [6] chimici dell'atmosfera sono:

- **Azoto (N2)**: L'azoto è il componente più abbondante dell'atmosfera, costituisce circa il 78% del volume complessivo. Esso è un gas inerte, il che significa che non reagisce facilmente con altre sostanze chimiche. L'azoto è essenziale per la vita, in quanto costituisce una parte fondamentale delle molecole delle proteine e del DNA.

- **Ossigeno (O2)**: L'ossigeno rappresenta circa il 21% dell'atmosfera. È vitale per la respirazione della maggior parte degli esseri viventi, poiché è utilizzato nei processi di combustione cellulare per generare energia.

- **Argon (Ar)**: L'argon è un gas nobile che costituisce circa l'1% dell'atmosfera. È chimicamente inerte e non partecipa a reazioni chimiche significative nell'atmosfera.

- **Anidride carbonica (CO2)**: L'anidride carbonica è un gas poco presente nell'atmosfera (circa 0.0415%), ma svolge un ruolo cruciale nell'effetto serra e nel ciclo del carbonio. Aumenti significativi della concentrazione di CO2 nell'atmosfera sono responsabili del riscaldamento globale e del cambiamento climatico.

- **Vapore acqueo (H2O)**: Il vapore acqueo è il gas responsabile delle precipitazioni. La sua concentrazione varia notevolmente in base alla regione e alle condizioni meteorologiche.

- **Gas nobili**: Oltre all'argon, l'atmosfera contiene tracce di altri gas nobili come il neon, il cripto e lo xeno. Questi gas sono chimicamente inattivi e si trovano in quantità molto ridotte.

- **Inquinanti atmosferici**: Infine l'atmosfera può contenere anche una varietà di inquinanti atmosferici, tra cui il biossido di azoto (NO2), il biossido di zolfo (SO2), il monossido di carbonio (CO) e particelle sospese. Questi inquinanti sono spesso il risultato di attività industriali e possono avere effetti negativi sulla salute umana e sull'ambiente.

L'atmosfera è uno dei sistemi più complessi e dinamici del nostro pianeta, e la comprensione dei suoi componenti chimici è fondamentale per comprendere i fenomeni meteorologici, il cambiamento climatico e l'impatto dell'attività umana sull'ambiente.

1.1.1 I principali inquinanti [5]

Gli inquinanti atmosferici sono sostanze chimiche e/o particelle che vengono rilasciate nell'atmosfera a seguito di varie attività umane, industriali ed ambientali.

Esaminiamo alcuni dei principali inquinanti atmosferici:

- **Biossido di Azoto (NO2)**: Il biossido di azoto è un inquinante gassoso comunemente prodotto da veicoli a motore, centrali elettriche e altre fonti industriali. È una causa importante dell'inquinamento atmosferico urbano ed è associato a problemi respiratori e cardiaci negli esseri umani. Inoltre, contribuisce alla formazione dell'ozono troposferico, un altro inquinante atmosferico dannoso.

- **Biossido di Zolfo (SO2)**: Il biossido di zolfo è prodotto principalmente dalla combustione di combustibili fossili contenenti zolfo, come il carbone e il petrolio. Esso può causare problemi respiratori e contribuire all'acidificazione del suolo e delle acque superficiali.

- **Monossido di Carbonio (CO)**: Il monossido di carbonio è un gas incolore e inodore prodotto dalla combustione incompleta di combustibili fossili. L'esposizione al CO può portare a intossicazioni e persino alla morte se avviene in ambienti chiusi.

- **Particelle Sospese (PM2.5 e PM10)**: Le particelle sospese sono piccole particelle solide o liquide nell'aria, con diametri inferiori a 2,5 o 10 μm. Queste particelle possono penetrare in profondità all'interno dei polmoni umani e causare una serie di problemi alla salute, tra cui malattie cardiovascolari e respiratorie. Le fonti principali di PM includono la combustione di carburanti fossili, l'industria e la polvere da strada.

- **Ozono Troposferico (O3)**: L'ozono troposferico, diverso dall'ozono stratosferico che ci protegge dai raggi UV dannosi, è un inquinante atmosferico che si forma attraverso delle reazioni chimiche tra il biossido di azoto (NO2) e i composti organici volatili (COV) in presenza di luce solare. L'ozono troposferico è irritante per le vie respiratorie e può causare problemi respiratori.

- **Composti Organici Volatili (COV)**: I COV sono una classe di composti chimici che evaporano facilmente in aria. Essi sono emessi da diverse fonti, tra cui veicoli a motore, industrie, vernici e solventi. I COV possono contribuire alla formazione di ozono troposferico e possono essere tossici per la salute umana.

- **Metalli Pesanti**: Metalli come il piombo, il mercurio e il cadmio possono essere emessi nell'aria attraverso processi industriali e dalla combustione di combustibili fossili. Questi metalli possono accumularsi negli ecosistemi e avere effetti dannosi sulla salute umana.

Il controllo e la riduzione degli inquinanti atmosferici sono di fondamentale importanza per preservare la qualità dell'aria, proteggere la salute umana, mitigare il cambiamento climatico e preservare la biodiversità.

Gli sforzi per monitorare, regolamentare e ridurre le emissioni di inquinanti atmosferici sono essenziali per garantire un ambiente più sano e sostenibile per le generazioni future.

1.2 Il Biossido di Azoto

Il caso studio preso in esame prende in considerazione le concentrazioni di Biossido di Azoto rilevate dall' ARPA nella zona di Milano, si procede quindi ad un breve excursus sulle dinamiche della sua formazione e dei suoi effetti.

1.2.1 La formazione del Biossido di Azoto

Il biossido di azoto (NO2) si forma principalmente attraverso processi di combustione che coinvolgono azoto (N2) e ossigeno (O2) presenti nell'atmosfera. Questa formazione è il risultato di reazioni chimiche complesse che avvengono ad alte temperature, come quelle prodotte nei motori a combustione interna, nelle centrali elettriche e in altri processi industriali.

Ecco un'analisi dettagliata della formazione del biossido di azoto:

- **Combustione**: La principale fonte di NO2 è la combustione di combustibili fossili, come benzina, diesel, carbone e gas naturale. Durante la combustione, gli atomi di azoto (N2) e di ossigeno (O2) reagiscono per formare ossido di azoto (NO).

- **Formazione dell'ossido di azoto (NO)**: In un primo momento l'azoto molecolare (N2) presente nell'aria viene diviso in due atomi di azoto (N). Questo processo richiede elevate temperature ($> 1200°$) e una notevole quantità di energia, che è fornita dalla combustione. Gli atomi di azoto (N) reagiscono quindi con l'ossigeno (O2) per formare ossido di azoto (NO):
  $N2 + O2 \rightarrow 2NO$

- **Formazione del biossido di azoto (NO2)**: Successivamente l'ossido di azoto (NO) può reagire nuovamente con l'ossigeno (O2) atmosferico per formare biossido di azoto (NO2) attraverso una reazione di ossidazione:
  $2NO + O2 \rightarrow 2NO2$

- **Equilibrio tra NO e NO2**: È importante notare che l'equilibrio tra NO e NO2 dipende dalle condizioni locali, compresa la temperatura, la pressione e la presenza di altri inquinanti atmosferici. A temperature più elevate e in presenza di sufficiente ossigeno, la formazione di NO2 può essere favorita, mentre a temperature più basse il NO può essere più stabile.

- **Emissioni antropiche e naturali**: Le emissioni antropiche, cioè quelle causate dalle attività umane, sono la principale fonte di biossido di azoto nell'atmosfera. Tuttavia, esistono anche fonti naturali di NO2, come le eruzioni vulcaniche ed i fulmini, che possono contribuire alla concentrazione di NO2 nell'aria.

- **Ruolo nella formazione dell'ozono**: Il NO2 è coinvolto anche nella formazione dell'ozono troposferico (O3), un altro inquinante atmosferico. Il NO2 può reagire con il radicale idrossile ($OH\cdot$) nell'atmosfera, formando ossigeno atomico (O) e nitrato ($NO3\cdot$), che è un precursore dell'ozono troposferico:
  $NO2 + OH\cdot \rightarrow NO3\cdot$
  $NO3\cdot + O2 \rightarrow NO2 + O3$

In sintesi, il biossido di azoto (NO2) si forma principalmente attraverso reazioni chimiche durante la combustione di combustibili fossili, della quale siamo i maggiori responsabili.

1.2.2 Effetti sulla salute

- **Problemi respiratori**: L'esposizione prolungata al biossido di azoto (NO2) può causare una serie di problemi respiratori, tra cui tosse, respiro sibilante, bronchite e aumento della suscettibilità alle infezioni respiratorie come il raffreddore e la polmonite. Le persone con asma sono particolarmente vulnerabili agli effetti negativi del NO2, poiché può innescare o peggiorare gli attacchi asmatici.

- **Irritazione delle vie aeree**: Il NO2 è noto per irritare le vie respiratorie superiori ed inferiori. Questo può causare dolore al petto e difficoltà nella respirazione.

- **Effetti cardiovascolari**: Ricerche recenti suggeriscono che l'esposizione al NO2 può aumentare il rischio di malattie cardiovascolari, inclusi problemi cardiaci e ictus.

1.2.3 Effetti sull'ambiente

- **Formazione dell'Ozono Troposferico**: Il NO2 è uno dei precursori dell'ozono troposferico (O3). Quando reagisce con il radicale idrossile (OH·) nell'atmosfera, può contribuire alla formazione dell'ozono troposferico. Questo ozono non è benefico come quello presente nella stratosfera, poiché è un inquinante che irrita le vie respiratorie e contribuisce all'inquinamento atmosferico estivo.

- **Acidificazione**: Il biossido di azoto può contribuire all'acidificazione del suolo e delle acque superficiali. Quando il NO2 si dissolve nell'acqua, forma acido nitrico (HNO3), che può abbassare il pH dell'acqua, danneggiare gli ecosistemi acquatici e influenzare negativamente la flora e la fauna.

- **Eutrofizzazione**: Il NO2 può anche contribuire all'eutrofizzazione, ovvero l'arricchimento delle acque con sostanze nutritive, come il nitrato. Ciò può portare alla crescita eccessiva di alghe e all'alterazione degli ecosistemi acquatici, causando la morte di pesci e altri organismi acquatici.

- **Impatti sulla vegetazione**: L'esposizione prolungata al NO2 può danneggiare le piante ed i raccolti. Il NO2 può interferire con la fotosintesi e danneggiare le foglie, riducendo la produttività agricola e influenzando la salute delle foreste.

1.2.4 Il quadro normativo

Livello Nazionale

In Italia, la normativa sulla qualità dell'aria [1], che comprende quella relativa al biossido di azoto (NO2), è regolamentata principalmente dalla Legge Nazionale n. 155 del 31 luglio 1995, nota come "Disposizioni in materia di tutela dell'aria". Questa legge è stata successivamente modificata ed integrata da una serie di decreti ministeriali e regionali.

Ecco alcuni punti chiave delle normative italiane relative al NO2:

- **Standard di qualità dell'aria**: In base alla normativa italiana, gli standard di qualità dell'aria per il NO2 devono essere rispettati al fine di proteggere la salute umana e quella dell'ambiente. Il decreto legislativo 155/2010, ad esempio, stabilisce i limiti di esposizione media annuale e di esposizione oraria per il NO2. Il limite medio annuale è di 40 $\mu g/m^3$, mentre il limite orario è di 200 $\mu g/m^3$, non superabile per più di 18 volte in un anno civile.

- **Piani di azione**: Quando i limiti di qualità dell'aria non vengono rispettati in determinate aree, le autorità locali sono tenute a sviluppare piani di azione per migliorare la qualità dell'aria. Questi piani possono includere misure come il controllo delle emissioni veicolari, l'implementazione di zone a traffico limitato e la promozione dei trasporti pubblici.

- **Monitoraggio dell'inquinamento atmosferico**: In Italia, esiste una rete di monitoraggio dell'aria gestita dalle autorità locali e regionali. Questa rete raccoglie dati sulla concentrazione di NO2 e altri inquinanti atmosferici in molte città italiane. I dati vengono utilizzati per valutare la conformità alle normative e per adottare misure correttive quando necessario.

- **Divieti e limitazioni di circolazione**: Alcune città italiane hanno introdotto misure di controllo del traffico per ridurre le emissioni di NO2 e migliorare la qualità dell'aria. Queste misure possono includere zone a traffico limitato (ZTL), restrizioni sull'accesso dei veicoli più inquinanti ed incentivi per l'uso di veicoli a basse emissioni.

- **Programmi di sensibilizzazione**: In aggiunta alle misure di controllo delle emissioni, molte città italiane promuovono programmi di sensibilizzazione pubblica sull'inquinamento dell'aria e sull'importanza di adottare comportamenti sostenibili per ridurre l'impatto ambientale.

Le normative sulla qualità dell'aria in Italia sono mirate a proteggere la salute pubblica e ambientale, le autorità locali lavorano costantemente per migliorare la qualità dell'aria nelle aree in cui i livelli di NO2 e altri inquinanti superano i limiti stabiliti.

Livello Mondiale

A livello mondiale [2] il WHO (World Healt Organization) ha pubblicato delle linee guida riguardanti l'inquinamento atmosferico.

Queste sono frutto di un costante sviluppo inziato nel 1987 fino ad arrivare alla pubblicazione nel 2005, ovvero quella precedente all'ultima del 2021.

All'interno di queste linee guida troviamo delle regolamentazioni riguardanti i più comuni inqunanti atmosferici: PM, O3, NO2 e SO2.

In figura si può notare i valori aggiornati dal 2005 al 2021:

Pollutant	Averaging Time	2005 AQGs	2021 AQGs
$PM_{2.5}$, $\mu g/m^3$	Annual	10	5
	24-hour[a]	25	15
PM_{10}, $\mu g/m^3$	Annual	20	15
	24-hour[a]	50	45
O_3, $\mu g/m^3$	Peak season[b]	-	60
	8-hour[a]	100	100
NO_2, $\mu g/m^3$	Annual	40	10
	24-hour[a]	-	25
SO_2, $\mu g/m^3$	24-hour[a]	20	40
CO, mg/m^3	24-hour[a]	-	4

Figura 1.1: WHO guidelines

Capitolo 2

Modelli matematici

In questo capitolo si vedranno i modelli utilizzati all'interno del mio caso studio, facendo un introduzione iniziale per poi visualizzare le formulazioni teoriche.
Essi mettono i dati al centro del processo decisionale, andandoli ad elaborare, cercando pattern che guideranno le scelete dell'algoritmo.

2.1 Serie storiche

Il primo elemento che analizziamo sono le serie storiche, ovvero possiamo definire la serie storica y(t) come la registrazione cronologica (non necessariamente con campionamento uniforme) di osservazioni sperimentali di una variabile.

Essendo interessati al comportamento macroscopico del sistema, si andrà a trascurare quello interno del fenomeno in analisi, quindi saremo interessati maggiormente su "cosa accade" puttosto del "modo in cui accade".

In altre parole, si tratta il fenomeno come una black-box: si conosce il segnale in ingresso (dati in input) e si osserva il segnale in uscita (dati in output) senza conoscere ciò che succede all'interno del sistema stesso.

Qindi possiamo indicare con Y il fenomeno da osservare ed Y_t ogni sua osservazione, cosicchè si va ad ottenere la serie storica $Y_t = Y_1, Y_2, Y_3, \ldots, Y_T$ delle osservazioni.

In seguito alla costruzione è possibile analizzarne le caratteristiche e metterle in relazione, per far ciò è possibile rappresentarla nel seguente modo:

$$Y(t) = \tau(t) + S(t) + r(t) \tag{2.1}$$

2.1.1 Analisi delle componenti di una serie storica

- **trend $\tau(t)$**: descrive il comportamento medio rispetto ad una scala temporale. Si tratta della variazione a lungo termine della serie, la quale viene rappresentata con una curva lineare, quadratica o esponenziale (solitamente di grado non elevato). In sostanza permette di analizzare l'evoluzione strutturale del fenomeno studiato.

- **stagionalità $S(t)$**: rappresenta le variazioni che si riscontrano con analoga intensità a distanza temporale costante nella serie. Le serie che presentano stagionalità sono anche dette periodiche, poichè in esse si può osservare la ripetizione ciclica di eventi stagionali in un periodo fisso.

- **residuo $r(t)$**: è composto dagli elementi che non risultano compresi nel trend o nella stagionalità. É una sequenza formata da valori tutti indipendenti tra loro, quindi priva di informazione.

Per concludere questa introduzione alle serie storiche è utile definire il concetto di dinamica di una serie, ovvero l'evoluzione del movimento dei corpi a causa di fattori esterni e/o circostanze che lo influenzano, questo in termini fisici generali.
Analogamente, si nota una certa dipendenza tra i valori delle grandezze esaminate in istanti di tempo diversi.
Da qui, si può definire la dinamica di una serie storica come l'evoluzione delle grandezze misurate in istanti di tempo consecuitivi, tenendo conto del fatto che il valore rilevato in un certo istante dipende sempre dai valori rilevati in uno o più istanti precedenti.

2.2 Modelli Data-Driven

Definiamo modello di un sistema la descrizione di un oggetto, del dispositivo, del fenomeno e/o della variazione nel tempo e/o nello spazio delle grandezze fisiche che lo caratterizzano.

Il suo scopo principale è la trasmissione di informazioni relative al sistema stesso, descrivendo tramite opportune equazioni la sua evoluzione nel tempo. L'utilizzo di un modello consente di osservare i valori ottenuti in uscita dal sistema (output) a fronte di determinati segnali in ingresso (input) senza coinvolgere il sistema reale.

Tra i vari modelli realizzabili, distinguiamo i modelli matematici, ovvero modelli astratti che descrivono l'evoluzione del sistema mediante opportune equazioni.
Sostanzialmente questi modelli analizzano i dati a loro disposizione e creano dei legami matematici, che stabiliranno il funzionamento e comportamento del modello stesso.

Questi possono essere divisi in due categorie:

- Modelli **deterministici**:le variabili in ingresso assumono valori fissi e le elaborazioni del sistema non tengono conto di eventuali incertezze associate a tali variabili.
- Modelli **stocastici**: le variabili in ingresso sono sottoposte a variazioni (casuali e non), quindi i risultati possono essere espressi in termini di probabilità.

I modelli stocastici risultano in generale strutturalmente più complessi rispetto a quelli deterministici,in quanto tengono conto della variabilità (anche casuale) dei dati di input. Tuttavia, questa stessa caratteristica rende i modelli stocastici più versatili, affidabili e aderenti alla realtà.

2.2.1 modelli AR

I modelli autoregressivi (AR) fanno parte della famiglia dei modelli stocastici: si tratta di modelli ampiamente utilizzati nell'analisi delle serie storiche, in quanto mettono in relazione il valore presente di una variabile con i suoi valori ritardati.
In questo modo è possibile tenere conto della possibile dipendenza statistica tra osservazioni della stessa grandezza corrispondenti a istanti diversi.

I modelli autoregressivi assumono, in generale, la forma seguente:

$$y(t) = \alpha_1 \cdot y(t-1) + \alpha_2 \cdot y(t-2) + \ldots + \alpha_n \cdot y(t-n) + \alpha_0 \tag{2.2}$$

Nella quale posso evidenziare diversi elementi:

i parametri $\alpha_1, \alpha_2, \ldots, \alpha_n$ sono i coefficienti di regressione lineare

n è l'ordine della parte autoregressiva, ovvero il numero di valori passati presi in considerazione

α_0 è il termine di depolarizzazione, ovvero il termine di errore

Nei modelli autoregressivi più semplici, dove la variabile y_t è legata al suo stesso valore nel solo istante precedente alla rilevazione si parla di modelli autoregressivi di ordine 1, o AR(1), e sono espressi dalla seguente relazione lineare:

$$y_t = \alpha_1 \cdot y(t-1) + \alpha_0 \tag{2.3}$$

dove il parametro α_1 costituisce il coefficiente di regressione lineare mentre α_0 rappresenta il termine di errore.

2.2.2 Modelli ARX

Per arriavre ai modelli ARX possiamo estrendere semplicemente il modello autoregressivo, andando a considerare sia dei suoi valori in istanti di tempo precedenti, sia degli ingressi esogeni: si tratta del modello autoregressivo con ingressi (ARX).
Quindi definiamo gli ingressi di tipo esogeno come gli ingressi che provengono dall'esterno, andando ad influenzare la grandezza presa in esame, ma rimanendo estranei ad essa.
Per questo tipo di modello vengono quindi definiti due tipi di ingressi distinti:

- definiamo gli **ingressi esogeni** quelli provenienti dall'esterno: si tratta di grandezze esterne che influenzano la grandezza in esame, ma estranee ad essa.
- definiamo invece gli **ingressi endogeni** quelli della grandezza di interesse in istanti di tempo precedenti.

Un modello autoregressivo con ingressi ha (in generale) la forma:

$$y(t) = \alpha_1 \cdot y(t-1) + \ldots + \alpha_n \cdot y(t-n) + \alpha_0 + \beta_0 \cdot u(t-k) + \beta_1 \cdot u(t-1-k) + \ldots + \beta_m \cdot u(t-m-k) \tag{2.4}$$

Nella quale posso evidenziare diversi elementi:

n è l'ordine della parte autoregressiva, degli ingressi endogeni

m è l'ordine della parte esogena

k è il ritardo della parte esogena

α_0 è il termine di depolarizzazione

Nella formula appena vista (Eq. 2.4), abbiamo un unico ingresso esogeno u, però possiamo avere modelli influenzati da più sorgenti esogene diverse, in questo caso m e k saranno 2 grandezze vettoriali.

In generale, un modello ARX può essere scritto in forma matriciale come:

$$y(t) = M(t) \times \theta \tag{2.5}$$

dove:

$$M(t) = [y(t-1), \ldots, y(t-n), u(t-k), \ldots, u(t-m-k)] \tag{2.6}$$

$$\theta = \begin{bmatrix} \alpha_1 \\ \vdots \\ \alpha_n \\ \beta_0 \\ \vdots \\ \beta_m \end{bmatrix} \tag{2.7}$$

Dunque andando ad utilizzare la famiglia dei modelli ARX per l'identificazione black-box, a fronte delle misure $(\bar{u}, \bar{y})$, si avrà che la generica uscita al tempo t della famiglia (fissati n e m) sarà:

$$\hat{y}_\theta(t) = M(t) \times \theta \tag{2.8}$$

Ora quello che dovremo fare sarà andare a stimare i valori di θ affinchè la distanza $D(\hat{y}_\theta, \bar{y})$ sia minimizzata.

Nei modelli ARX la tecnica utilizzata per la prezione dei valori del vettore colonna θ è quella della regressione lineare, ovvero si andrà ad adattare un modello lineare con coefficienti :

$$\theta = [\alpha_1, \ldots, \alpha_n, \beta_{10}, \ldots, \beta_{1m}, \beta_{20}, \ldots, \beta_{2m}, \beta_{30}, \ldots, \beta_{3m}, \beta_{40}, \ldots, \beta_{4m}] \tag{2.9}$$

utilizzando la "classica" metrica dei minimi quadrati.

Nota: Per completezza sono andato ad inserire nell'equazione 2.9 tutti i coefficienti presi in esame nel mio caso studio specifico (quindi n coefficienti per la parte autoregressiva e $m \cdot 4$ coefficienti per la parte esogena, questo perchè nel mio caso studio utilizzo 4 ingressi di tipologia diversa).

Dunque si cercherà di minimizzare la somma delle differenze al quadrato tra i target osservati nel set di dati e i target previsti dall'approssimazione lineare, secondo la formula seguente:

$$\min D(\hat{y}_\theta, \bar{y}) = \min \sum_t (\hat{y}_\theta(t) - \bar{y}(t))^2 \tag{2.10}$$

Graficamente andiamo ad ottenere un elemento lineare che approssima i dati come nella seguente figura:

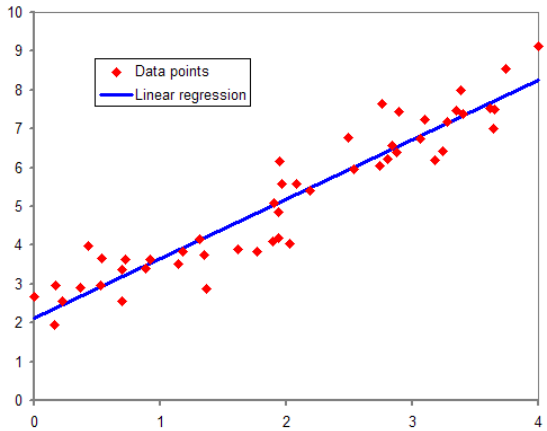


Figura 2.1: Esempio del funzionamento di una regressione lineare

2.2.3 Decision Tree Regressor [4]

Per poter parlare del Decision Tree Regressor bisogna prima introdurre il Decision Tree Learning, ovvero un approccio di apprendimento supervisionato largamente utilizzato in diversi ambiti (statistica, data mining, machine learning, ...) che permette di creare un albero decisionale utilizzando un set di dati di addestramento, questa struttura ad albero sarà poi utilizzata per prendere decisioni basate su delle regole.

Quindi si va ad ottenere una struttura come in Figura:

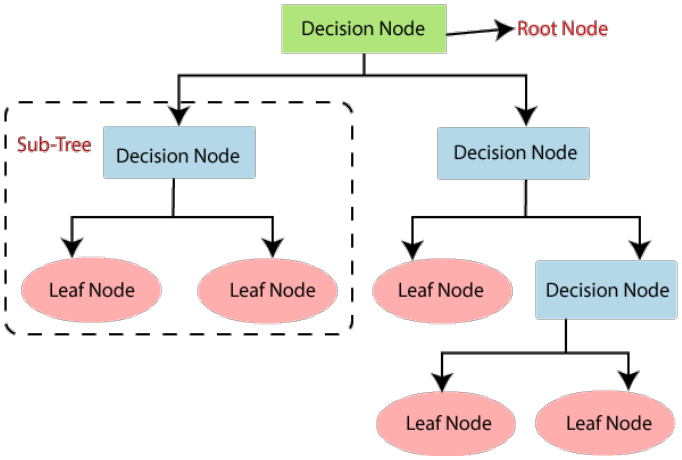


Figura 2.2: Struttura di un albero decisionale [3]

Un Decision Tree Regressor è un tipo specifico di modello di albero decisionale utilizzato per problemi di regressione, quindi viene addestrato per prevedere valori numerici o continui, mentre una albero decisionale classico va a stimare delle classi o categorie.

Innanzitutto bisogna evidenziare il motivo per la quale si decide di utilizzare una tecnica diversa dalla Regressione Lineare (essendo la più semplice e classica da implementare).

Come possiamo notare in Figura 2.3 è possibile approssimare a sufficienza il grafico a sinistra (a), mentre risulterà difficile fare altrettanto con il grafico a destra (b), data la sua distribuzione "poco lineare".

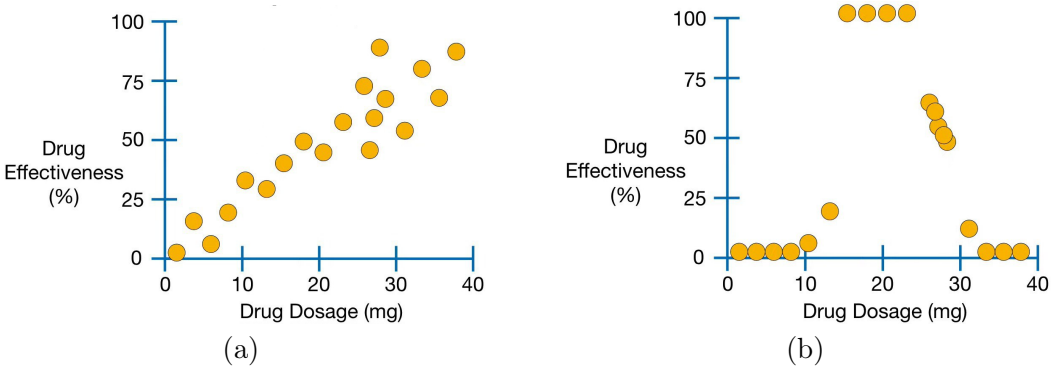


Figura 2.3: Diverse distribuzioni di dati

Nota: In questo esempio utilizzerò un set di dati monodimensionale, per facilitarne la comprensione a livello grafico, infine mostrerò l'estensione al caso multidimensionale.

Il risultato che vorrò ottenere sarà un albero che segue la seguente logica:

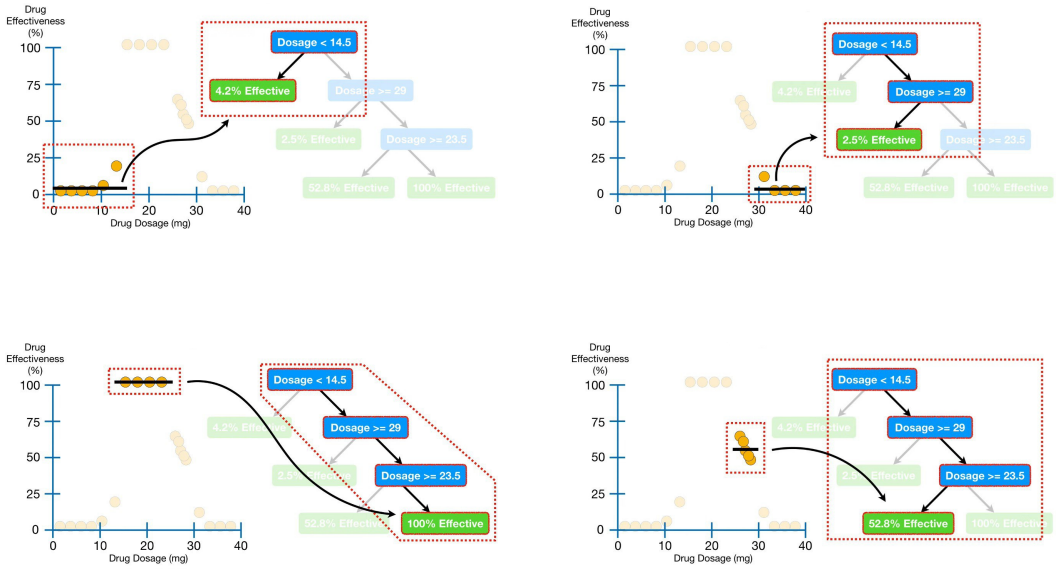


Figura 2.4: Corrispondenza grafica tra il grafico e le componenti dell'albero

Inizialmente si procede selezionando i primi 2 elementi e calcolandone la media (in questo caso ho Dosage = 3) come in Figura:

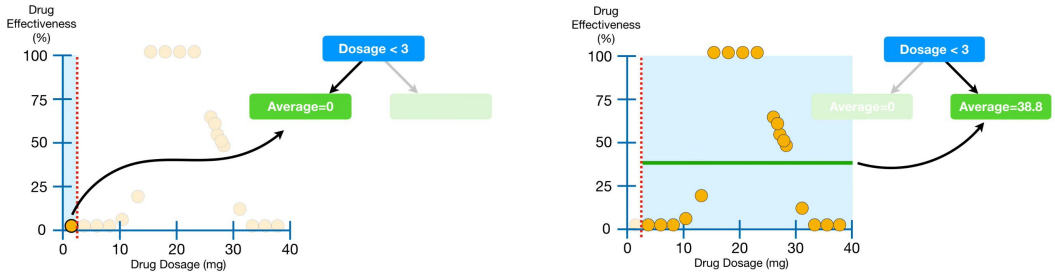


Figura 2.5: Prima iterazione per stabilire il root node

Successivamente si calcola la somma dei quadrati residui (RSS) rispetto alla media:

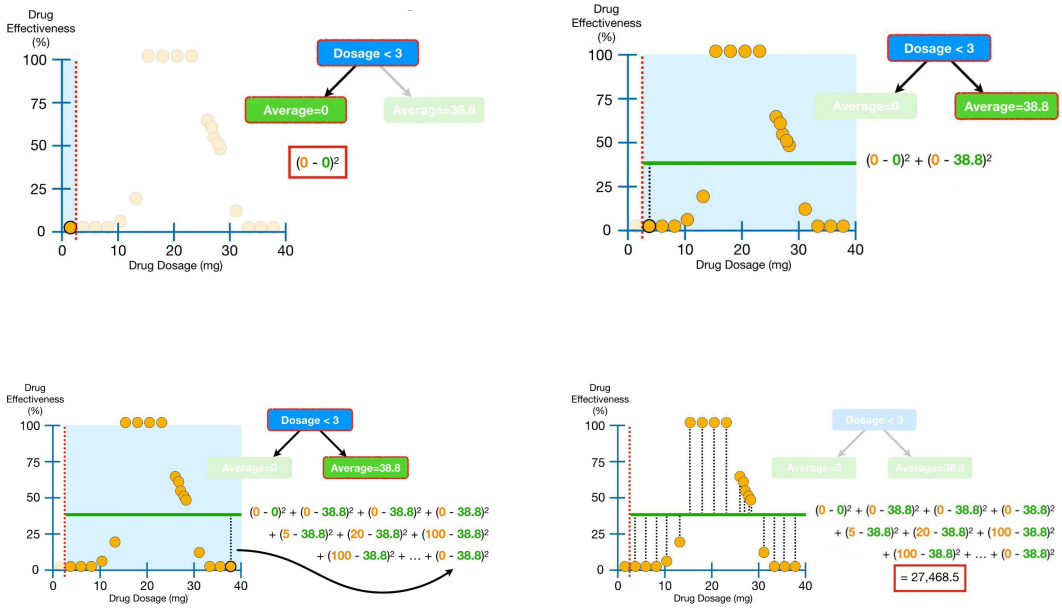


Figura 2.6: Passaggi per il calcolo dell'RSS iniziale (Dosage = 3)

Il valore dell'RSS trovato verrà confrontato con tutti i valori di RSS calcolati utilizzando valori della media diversi, graficamente otteniamo il plot in Figura:

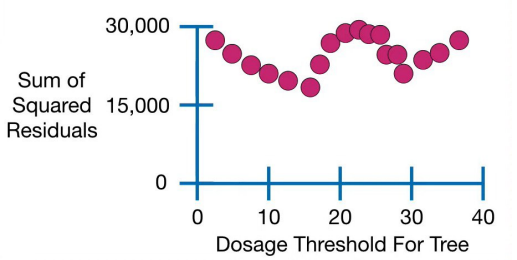


Figura 2.7: RSS inizialmente calcolati

Ora che abbiamo i valori dei diversi RSS andiamo a scegliere il valore minore (ovvero quello corrispondente a Dosage = 14.5) trovando il nodo radice.

Dopo iniziamo ad espandere i due diversi rami dell'albero, quindi come in precenza andrò a calcolare i diversi valori delle medie per poi calcolre i diversi RSS, selezionando infine il valore minore:

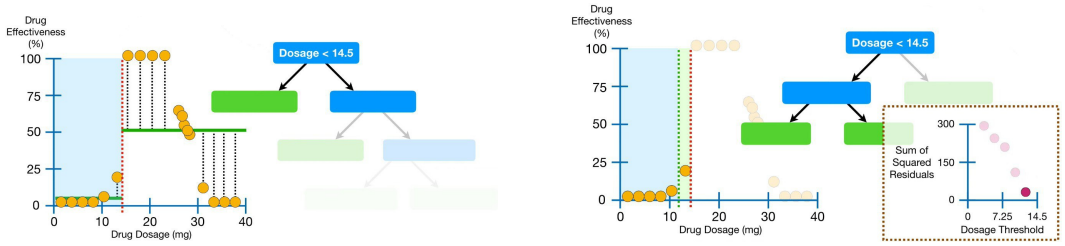


Figura 2.8: Espansione del lato sinistro dell'albero

Così facendo si rischia di espandere eccessivamente l'albero e andare in overfit (come in Figura 2.9 lato sinistro).

Quindi bisogna definire un limite di valori sotto la quale l'albero non può essere espanso ulteriormente (cosicchè si va ad ottenere il risultato in Figura 2.9 lato destro).

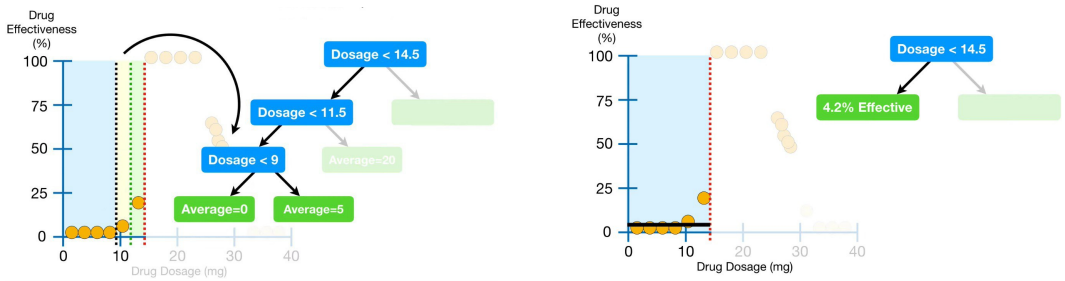


Figura 2.9: Overfit e risultato corretto

Similmente andremo a calcolare anche la parte restante dell'albero, sempre tenendo in considerazione il limite minimo di elementi da considerare.

Infine nel caso multidimensionale basterà eseguire il calcolo inziale dei diversi RSS e scegliere quello minore per ogni features, come nell'esempio in Figura:

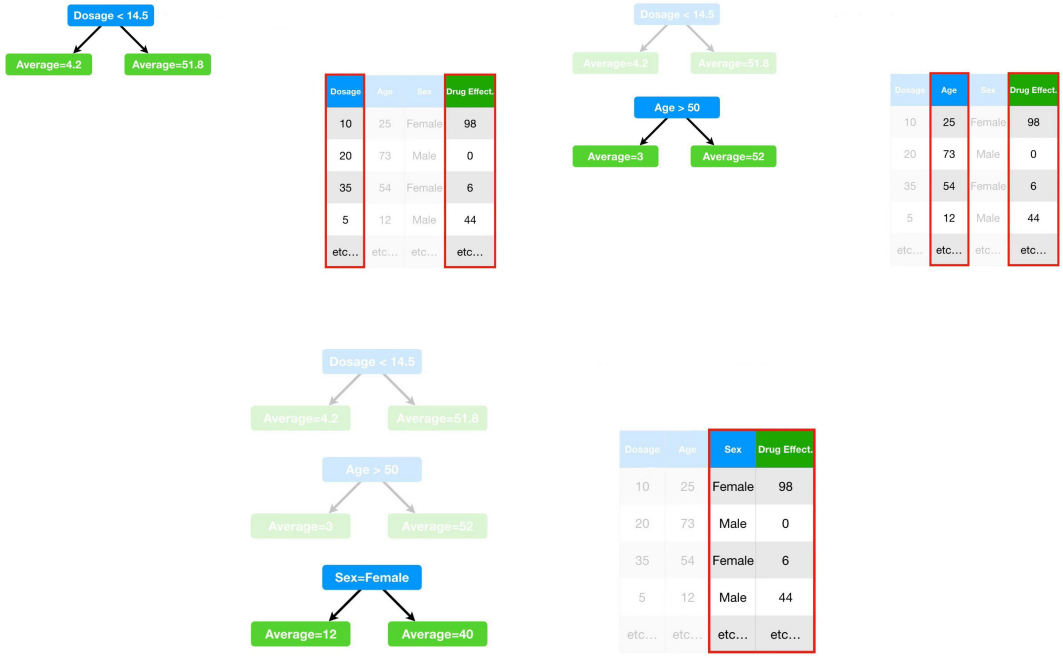


Figura 2.10: Step per la scelta del nodo radice

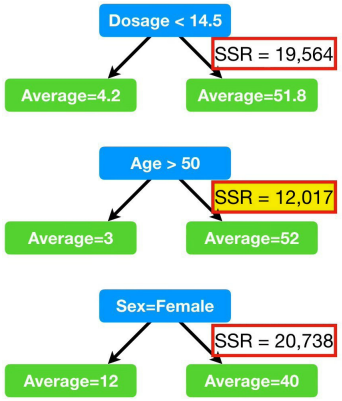


Figura 2.11: Decisione dell'RSS migliore

Infine basterà ripetere iterativamente questo processo per ottenere l'albero finale:

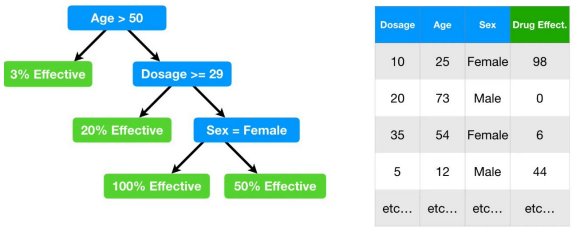


Figura 2.12: Albero finale

Capitolo 3

Caso studio

In questo capitolo analizzerò i diversi aspetti del caso studio preso in esame nel mio lavoro di tesi.

I diversi aspetti trattati saranno rispettivamente:

- Dati di input/output
- Suddivisione del dataset
- Trattamento dei dati non validi
- Prestazioni in validazione

3.1 Dati di input/output

Il dataset in input corrisponde a osservazioni giornarliere eseguite dal 01 Gen 2014 al 31 Dic 2019 (per un totale di 2191 giorni, essendo il 2016 un anno bisestile).

Queste osservazioni giornaliere sono suddividise in 2 gruppi come in Figura:

	temperatura	velocità vento	radiazione solare	emissioni
0	0.892308	93.069231	0.000000	43.658640
1	3.500000	86.715385	0.016667	66.949758
2	3.361538	98.138462	0.200000	66.949758
3	4.846154	94.946154	0.200000	52.284212
4	4.723077	98.484615	0.166667	43.658640
...	...	...	...	...
2186	3.866109	84.574542	57.042712	54.149256
2187	3.713009	87.693241	55.308594	44.411578
2188	2.937712	96.403378	43.130056	37.411999
2189	3.472094	90.442144	58.676518	54.149256
2190	2.593199	88.833948	67.534259	54.149256

Figura 3.1: Dataset in input - parte esogena

	concentrazione max
0	[106.7688]
1	[82.3099]
2	[61.6159]
3	[63.3935]
4	[64.6059]
...	...
2186	[82.5421]
2187	[86.8299]
2188	[72.5528]
2189	[55.7206]
2190	[74.3971]

Figura 3.2: Dataset in input - parte autoregressiva

3.2 Suddivisione del dataset

La suddivisione del dataset in un set di identificazione (addestramento) e un set di validazione è una pratica fondamentale nell'ambito delle tecniche data-driven.

Questo perchè serve a capire quanto bene il modello sia in grado di generalizzare al di fuori dei dati di addestramento.

Addestrando un modello utilizzando l'intero dataset, si potrebbero ottenere prestazioni eccellenti sui dati di addestramento ma non si sarebbe in grado di valutare in modo affidabile il comportamento con nuovi dati, il che è l'obiettivo finale.

Generalmente si deve suddividere il dataset in un 80% per l'identificazione e il restante 20% per l validazione, nel mio caso per l'addestramento son stati utilizzati i dati dal 01 Gen 2014 al 31 Dic 2018, di conseguenza per la validazione sono stati utilizzati i dati relativi all'anno 2019.

3.3 Trattamento dei dati non validi

Trattare i dati non validi è un passo essenziale nella gestione dei dati che verranno processati da un modello.

Questo perchè possono compromettere l'affidabilità dei risultati, portando a conclusioni errate o distorte.

Infatti i dati non validi possono causare errori durante il processo di analisi e l'elaborazione dei dati, creando errori nei modelli di identificazione e/o validazione. Di conseguenza questi errori si propagheranno nelle stime dei valori successivi.

Infine la presenza di dati non validi può compromettere la coerenza e la consistenza dei dati. Ciò può rendere difficile o impossibile condurre analisi statistiche e confronti significativi tra i dati.

Per affrontare i dati non validi da un punto di vista sintattico, solitamente è necessario eseguire azioni come la correzione degli errori di formattazione, l'individuazione e la rimozione dei dati duplicati o mancanti, la gestione degli outliers (valori anomali) e altro ancora.

Nel mio lavoro di tesi questa fase è stata realizzata tramite la funzione in Figura:

```
def clean (y,X):

    S=np.sum(X,axis=1)
    bad=np.isnan(S)
    X=np.delete(X,bad,axis=0)
    y=np.delete(y,bad)
    bad=np.isnan(y)
    X=np.delete(X,bad,axis=0)
    y=np.delete(y,bad)
    return (y,X)
```

Figura 3.3: funzione per la pulizia del dataset

3.4 Prestazioni in validazione

In quest'ultima sezione mostrerò le prestazioni ottenute in fase di validazione dalle tecniche data-driven utilizzate (ovvero Linear Regression e Decision Tree Regressor).

Ogni tecnica è stata applicata a 30 modelli, ottenuti facendo variare il grado della parte autoregressiva (na $\in [1, 5]$) e quello della parte esogena (nb $\in [0, 5]$).

Dunque per ogni tecnica mostrerò rispettivamente i risulati riguardanti 3 parametri:

- **Errore Medio Normalizzato (NME)**: É un valore normalizzato rispetto alla media che indica la sovrastima o sottostima del modello.

$$NME = \frac{\sum_{i=1}^{N} e_i}{\sum_{i=1}^{N} \bar{y}}$$

- **Errore Medio Assoluto Normalizzato (NMAE)**: É un valore normalizzato rispetto alla media, più "robusto" del NMAE perchè non ho compensazione di segno.

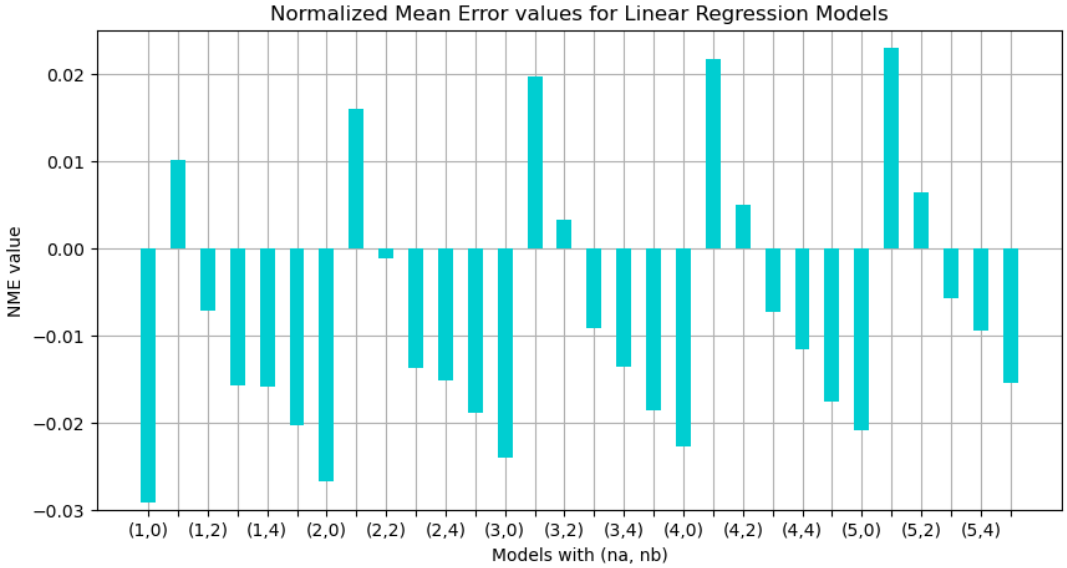
$$NMAE = \frac{\sum_{i=1}^{N} |e_i|}{\sum_{i=1}^{N} \bar{y}}$$

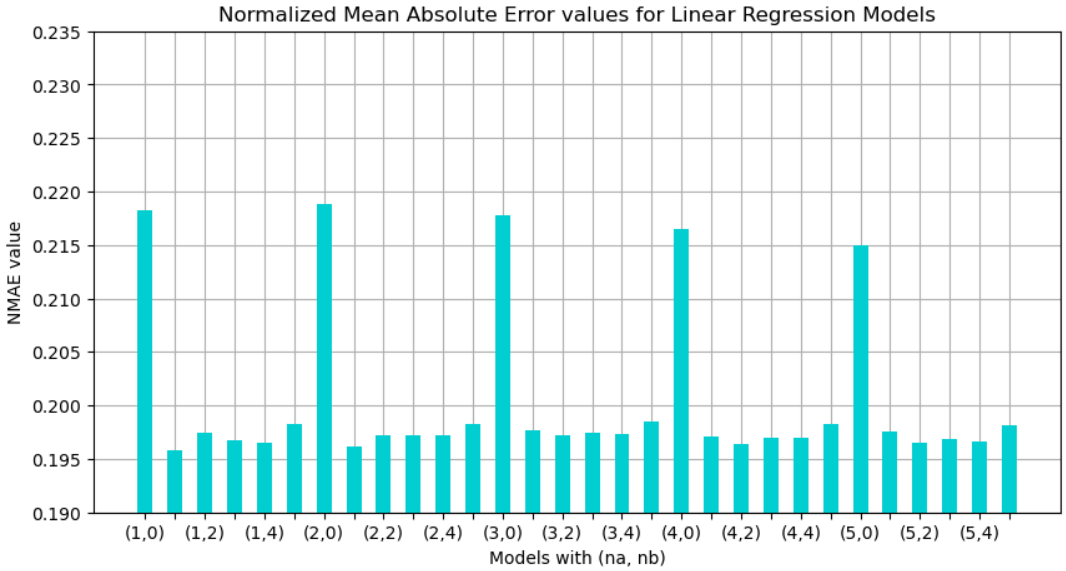
- **Correlazione**: Indica la capacità del modello di un andamento simile a quello della serie (non considera eventuali traslazioni).

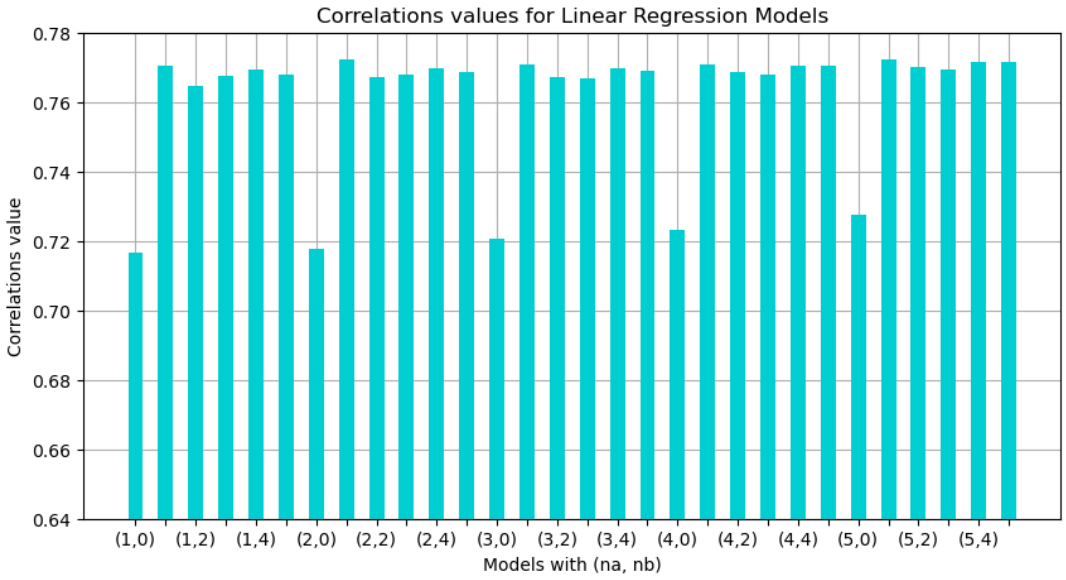
$$Correlazione = \frac{\sum_{i=1}^{N}(y_i - u_y) \cdot (\hat{y}_i - u_{\hat{y}})}{\sqrt{\sum_{i=1}^{N}(y_i - u_y)^2} \cdot \sqrt{\sum_{i=1}^{N}(\hat{y}_i - u_{\hat{y}})^2}}$$

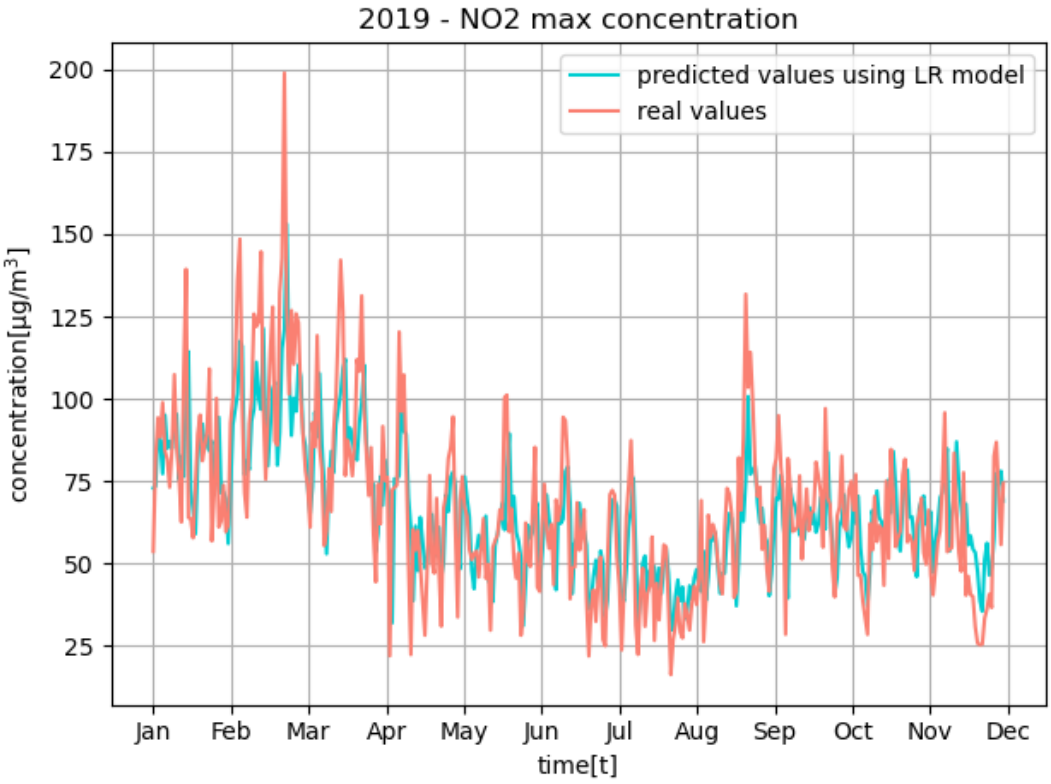
Infine si trova il plot avente la migliore predizione, ovvero quello con il valore di NMAE più basso.

3.4.1 Prestazioni Linear Regression

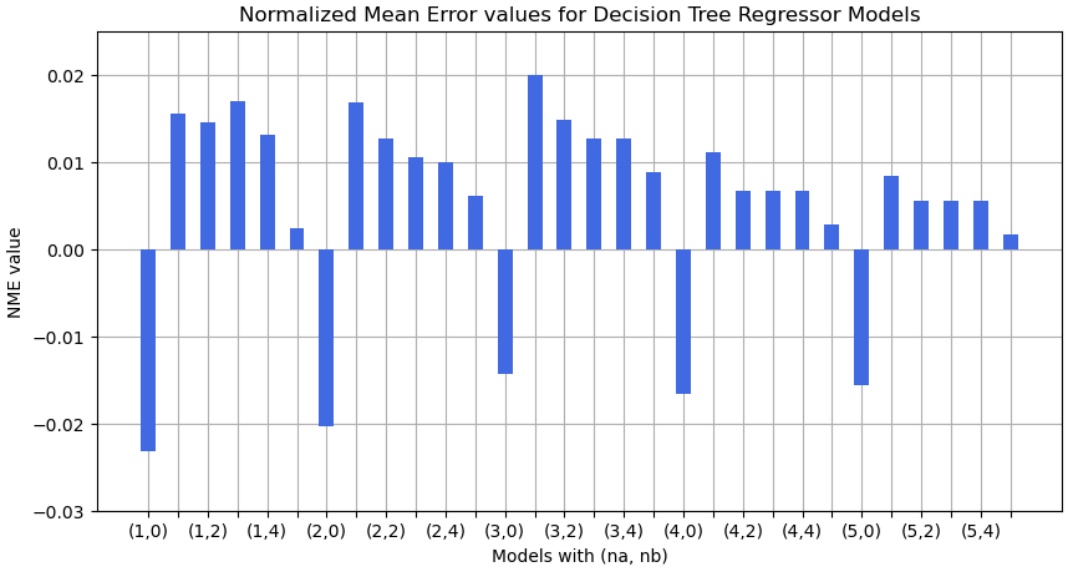


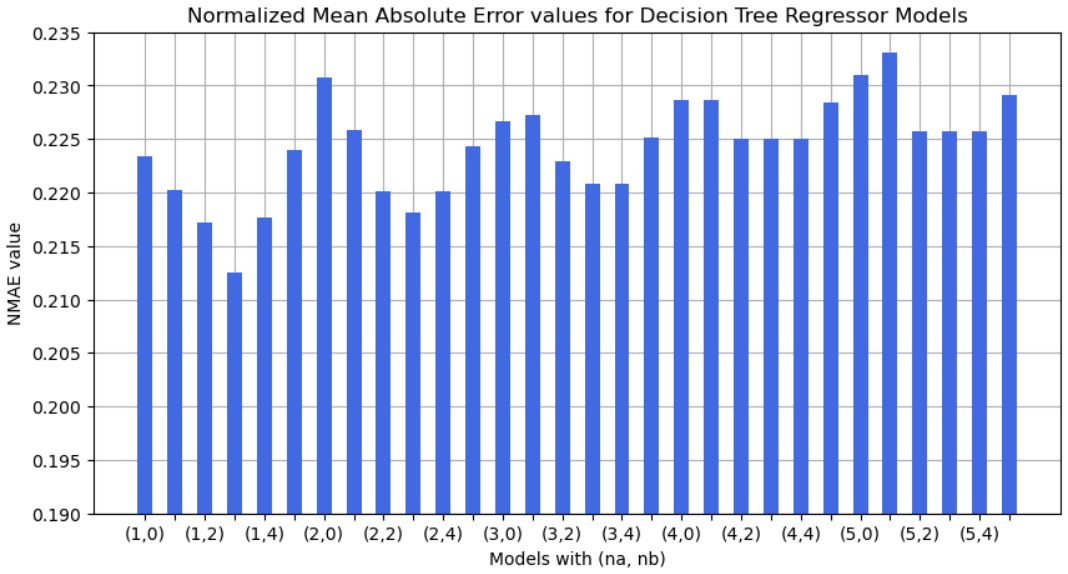


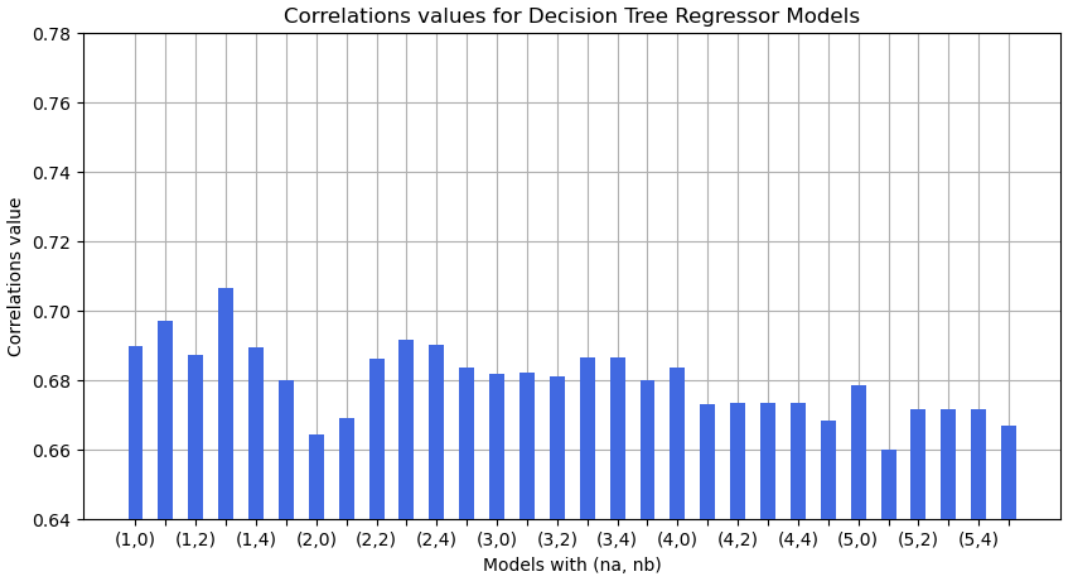

Correlations values for Linear Regression Models
Correlations value
0.78
0.76
0.74
0.72
0.70
0.68
0.66
0.64
(1,0) (1,2) (1,4) (2,0) (2,2) (2,4) (3,0) (3,2) (3,4) (4,0) (4,2) (4,4) (5,0) (5,2) (5,4)
Models with (na, nb)

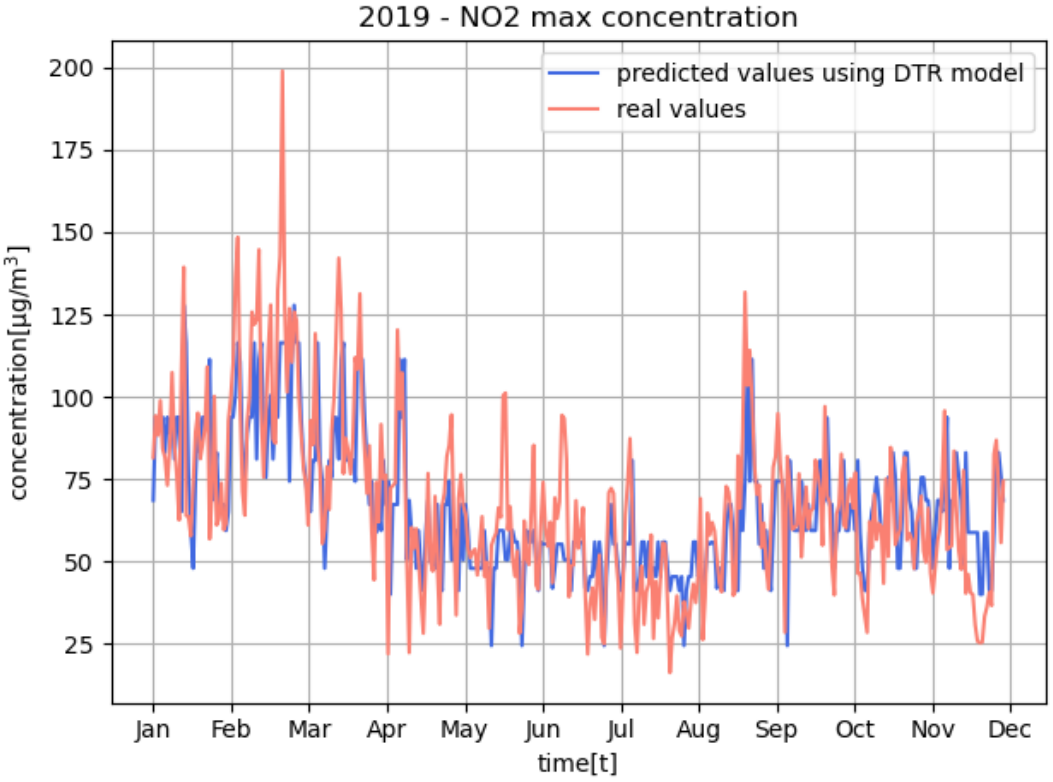

2019 - NO2 max concentration
predicted values using LR model
real values
concentration[µg/m³]
200
175
150
125
100
75
50
25
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
time[t]

3.4.2 Prestazioni Decision Tree Regressor






Correlations values for Decision Tree Regressor Models
Correlations value
0.78
0.76
0.74
0.72
0.70
0.68
0.66
0.64
(1,0) (1,2) (1,4) (2,0) (2,2) (2,4) (3,0) (3,2) (3,4) (4,0) (4,2) (4,4) (5,0) (5,2) (5,4)
Models with (na, nb)


2019 - NO2 max concentration
predicted values using DTR model
real values
concentration[µg/m³]
200
175
150
125
100
75
50
25
Jan Feb Mar Apr May Jun Jul Aug Sep Oct Nov Dec
time[t]

Conclusioni

Essendo in un momento storico decisivo per lo sviluppo della Terra e del suo ecosistema, vista la crescente importanza riguardante lo studio ,e di conseguenza la prevenzione, degli elementi dannosi per essa, la comunità scientifica sta aumentando la ricerca nell'ambito dei modelli previsionali in grado di andare oltre la semplice constatazione della situazione attuale.

Questi modelli costituiscono le fondamenta sulla quale si basano le decisioni prenti dai diversi enti in tutto il mondo.

Dunque l'implementazione di questi modelli predittivi si deve basare su una corretta identificazione di essi, i quali partendo da dati storici misurati all'interno di una rete di monitoraggio, saranno in grado di prevedere e descrivere la dinamica non lineare che mette in relazione la formazione e rimozione degli inquinanti nel breve termine.

In questo lavoro di tesi mi son basato su un modello ARX (autoregressivo con ingressi esogeni) che permette di ottenere un'elevata flessibilità ed adattabilità in molteplici casi.

Nello specifico mi son focalizzato sulla predizione del livello di inquinamento del Biossido di Azoto (NO2) nell'aria, un gas che produce effetti dannosi all'uomo e all'ambiente. Il dataset conteneva i valori delle concentrazioni massime di NO2, la radiazione solare, le emissioni, la velocità del vento e la temperatura relative a 5 anni (dal 01 Gen 2014 al 31 Dic 2019) provenineti dall'ARPA nel comune di Milano.

Con i modelli ARX generati ho utilizzato due tecniche data-driven per il fit del modello. In primis il Linear Regression Model che utilizza la minimizzazione della somma dei quadrati minimi, tecnica molto generica ma altrettanto efficace. Poi ho utilizzato il Decision Tree Regressor che permette di creare un albero decisionale per stabilire la predizione.

L'utilizzo di entrambi i modelli ha portato a buoni risultati con previsioni del livello di NO2 soddisfacenti.

Bibliografia

[1] Ministero dell Ambiente e della Tutela del Territorio e del Mare. Qualità dell'aria ambiente: Biossido di azoto (no2).

[2] World Healt Organization. What are the who air quality guidelines? 2021.

[3] scikit Learn. Decision tree regressor illustration, link: https://scikit-learn.org/stable/modules/generated/sklearn.tree.decisiontreeregressor.html.

[4] Josh Starmer. Regression trees, link: https://youtu.be/g9c66tuylz4?si=1da8inqq7q-ck2nn. *StatQuest*.

[5] ARPA Veneto. Inquinanti amosferici link: http://www.arpa.veneto.it/temi-ambientali/aria/a-proposito-di-ozono. 12 - 2022.

[6] Wikipedia. Atmosphere of earth link: https://en.wikipedia.org/wiki/atmosphere-of-earth.

Ringraziamenti

Ringrazio in primis il Prof. Claudio Carnevale, per l'aiuto durante la stesura del lavoro di tesi, dalla scelta dell'argomento ad ogni accorgimento suggerito.

Un sentito grazie a tutta la mia famiglia, per il costante supporto nel mio percorso di studio, soprattutto nei momenti più difficili.

Inoltre ringrazio tutti i miei amici conosciuti in università, con i quali ho condiviso esperienze e costruito un bellissimo rapporto che porterò sempre con me. Siete stati fondamentali nel raggiungimento di questo traguardo, è anche grazie a voi se ora son arrivato fin qui.

Grazie a tutti coloro che ho incontrato lungo questo (difficile) percorso, ho conosciuto parecchie persone per bene, sarò sempre grato di questo con ognuno di voi.

Concludo questi ringraziamenti sottolineando la mia gratitudine verso quest'esperienza di "soli" 3 anni, i quali credo abbiano avuto un impatto sostanziale nello sviluppo della mia persona, Grazie ancora.