POLITECNICO DI TORINO

Facolta di Ingegneria

Corso di Laurea in Ingegneria Informatica

Tesi di Laurea Magistrale

Air quality estimation from satellite acquisitions

Relatori
Prof. Paolo Garza
Dott. Alessandro Farasin

Candidata
Ambra Destino

Anno Accademico 2019-2020

Sommario

L'analisi della qualità dell'aria e l'inquinamento sono temi centrali dei nostri tempi poiché questi problemi hanno un grosso impatto sulla vita dell'uomo. Il monitoraggio delle concentrazioni degli inquinanti permette di studiare quali sono le condizioni favorevoli alla variazione delle stesse e un opportuno meccanismo di previsione permette di attuare misure di protezione e prevenzione da parte degli organi competenti. In questo lavoro sono stati usati dati sugli inquinanti ottenuti tramite la missione Copernicus Sentinel-5p e dati meteorologici. Gli inquinanti presi in considerazione sono: O3, CO, NO2, SO2, HCHO. Sono state implementate ed analizzate diverse tecniche di Machine Learning per creare dei modelli che permettano di effettuare predizioni sulle concentrazioni di un set di inquinanti. I modelli utilizzati sono: Random Forest, Multi-Layer Perceptron e Convolutional Long Short-Term Memory. Per l'analisi sono state considerate dieci città europee: Atene, Barcellona, Bruxelles, Ginevra, Milano, Norimberga, Parigi, Roma, Torino, Zurigo; ed un periodo temporale definito dall'intervallo 5/12/2018- 19/09/2019.

Indice

Capitolo 1

Introduzione

Il progetto presentato in questa tesi si propone di studiare e analizzare fonti di dati eterogenee per costruire modelli di Machine Learning e Deep Learning per prevedere la qualità dell'aria di alcune zone di interesse. In particolare, sono stati usati i dati ottenuti dal satellite Sentinel 5P dell'Agenzia Spaziale Europea.

Il capitolo che segue è strutturato in due parti: la prima presenta una panoramica sul problema dell'inquinamento e sugli enti che operano per la sua risoluzione; la seconda fornisce una breve introduzione al programma Copernicus, all'Agenzia spaziale Europea e al progetto Sentinel.

1.1 Inquinamento

> *"L'inquinamento è l'introduzione diretta o indiretta in un ambiente di sostanze o anche di energia capaci di trasformare gli equilibri naturali producendo anche effetti sulla salute umana. Alcune di queste trasformazioni sono irreversibili nel medio o nel lungo periodo". Attualmente il termine inquinamento viene utilizzato come sinonimo di "ambiente sporco". [1]*

L'inquinamento atmosferico è un problema globale che non conosce confini. L'Organizzazione Mondiale della Sanità (OMS) ha affermato che l'inquinamento atmosferico provoca 4,2 milioni di morti premature ogni anno [2]. Inoltre, l'Agenzia Europea dell'Ambiente [3] sostiene che il 90% della popolazione delle grandi città europee è esposto a fattori inquinanti pericolosi.
Governi e Organizzazioni hanno la possibilità di pianificare strategie, per mitigare e limitare le conseguenze di tutto ciò, grazie allo studio delle condizioni attuali di inquinamento. Tra i maggiori organi mondiali che si occupano di questa problematica, vi sono: l'Organizzazione delle Nazioni Unite (ONU), la US Environmental Protection Agency (EPA) e l'Agenzia Europea dell'Ambiente (AEA). L'ONU ha stabilito degli 'Obiettivi di sviluppo sostenibile'. Essi costituiscono una lista di 17 obbiettivi. Gli argomenti trattati sono tutti differenti tra di loro ma allo stesso tempo correlati. Vengono prese in considerazione diverse problematiche: lo sviluppo economico e sociale, la fame, la povertà, la salute,

l'istruzione, l'uguaglianza di genere, l'acqua, i servizi igenico-sanitari, l'energia, l'urbanizzazione, l'ambiente e il cambiamento climatico, l'uguaglianza sociale. Il tredicesimo obiettivo [4], *Climate action* ,tratta la problematica del cambiamento climatico, con i conseguenti problemi del riscaldamento globale e dell'innalzamento dei mari, e mira a una riduzione delle emisioni di CO_2 pari al 45%.

La US Environmental Protection Agency (EPA) [5], Agenzia per la protezione dell'ambiente degli Stati Uniti, ha come compito quello di stabilire delle norme per un insieme di sei inquinanti comuni, che possono causare danni all'ambiente e alla salute. In questa lista sono presenti: Particolato (PM), Piombo (Pb), Ozono (a livello del suolo, O_3), Monossido di carbonio (CO), Anidride solforosa (SO_2), Biossido di azoto (NO_2).[6]

In Europa, l'Agenzia Europea dell'Ambiente (AEA) si occupa di documentare e valutare la tendenza dell'inquinamento, le misure e le politiche applicate ad essa connesse. L'Agenzia Europea dell'Ambiente afferma che:

"in Europa, le emissioni di molti inquinanti atmosferici sono diminuite in modo sostanziale negli ultimi decenni, determinando una migliore qualità dell'aria. Le concentrazioni di inquinanti sono tuttavia ancora troppo elevate e i problemi legati alla qualità dell'aria persistono."

A dimostrazione di ciò, in Figura 1.1, 1.2 e 1.3 vengono riportati alcuni grafici[7] che mostrano l'andamento delle emissioni di alcuni inquinanti, nei grandi impianti di combustione, nel corso del tempo. È possibile notare che nel 2017, rispetto all'anno 2014, le emissioni di SO_2 e NO_2 sono notevolmente diminuite.

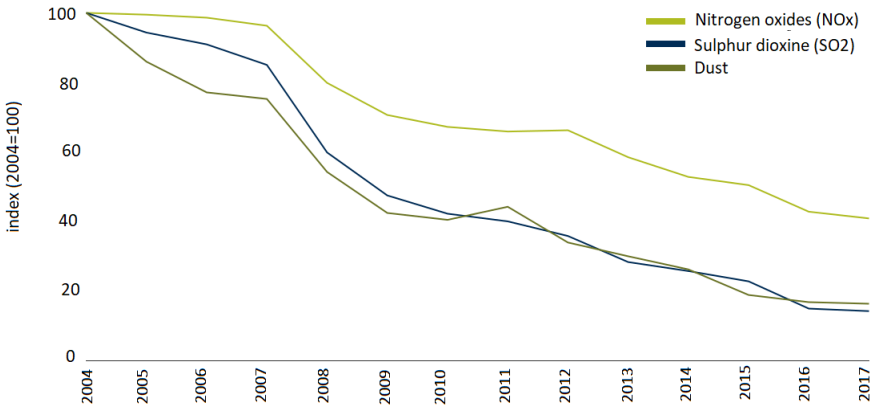

Figura 1.1: Emissioni indicizzate di anidride solforosa, ossidi di azoto e polvere da grandi impianti di combustione nell'Unione Europea

Nonostante le norme e misure cautelari atte a limitare e prevenire le alte concentrazioni, in particolar modo nelle città, i limiti vengono spesso superati mettendo a grave rischio la salute.[8]

> "L'inquinamento atmosferico sta danneggiando la salute umana e gli ecosistemi. Larghe fasce della popolazione non vivono in un ambiente sano, in base alle norme attuali. Per imboccare un cammino sostenibile, l'Europa dovrà essere ambiziosa e andare oltre la legislazione attuale." (Hans Bruyninckx, direttore esecutivo dell'AEA).

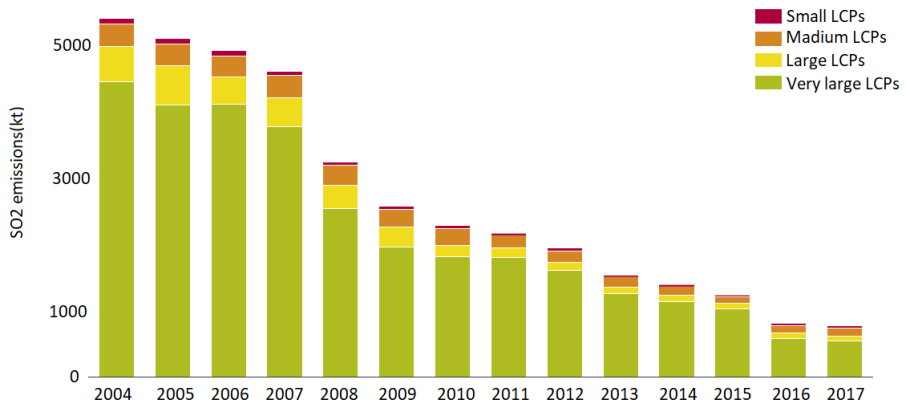

Figura 1.2: Emissioni di SO_2 da grandi impianti di combustione nell'Unione Europea, per classe di capacità

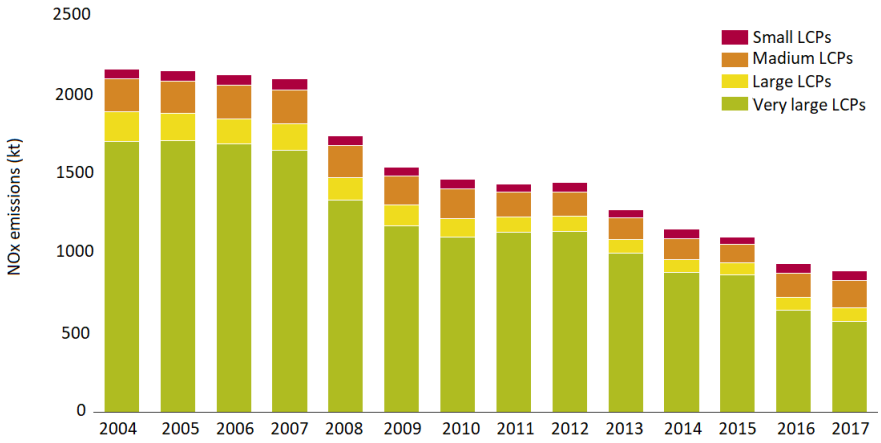

Figura 1.3: Emissioni di NO_2 da grandi impianti di combustione nell'Unione Europea, per classe di capacità

A proposito di questo problema, la Commissione Europea, nel 2013, ha adottato la proposta "Aria pulita", un'insieme di misure favorevoli alla diminuzione dell'inquinamento atmosferico. [9]

Di seguito vengono riportati alcuni tra gli inquinanti, citati nella proposta "aria pulita", che attualmente causano maggiore preoccupazione.

Ozono (O_3): l'ozono presente nell'atmosfera può essere classificato come "buono" o "cattivo" a seconda di dove si trova. L'ozono presente nella stratosfera, "buono", protegge la Terra da radiazioni ultraviolette del sole, le quali permettono la scissione degli atomi di O_2. Mentre quello presente nella troposfera, quindi a contatto con il suolo, è causa di malattie respiratorie. L'ozono troposferico è il risultato di reazioni chimiche tra ossidi di

azoto (NO_x) e composti organici volatili, favorite dall'irraggiamento solare e dalle temperature elevate. Inoltre i venti permettono il trasporto e l'accumulo di questo inquinate, che quindi potrà raggiungere anche zone rurali.

Monossido di carbonio (CO): è un gas incolore e inodore, che inalato in grosse quantità diventa nocivo per la salute. Esso è creato durante il processo di combustione , un esempio può essere la combustione di combustibili fossili effettuata da automobili, velivoli o macchinari oppure , all'interno delle abitazioni: camini, forni, stufe. La presenza di CO nell'aria riduce la percentuale di ossigeno che può essere inalata, creando problemi all'organismo. Un' alta concentrazione di CO, nell'aria inalata da un essere umano, provoca una bassa ossigenazione del flusso sanguigno e in condizioni estreme può provocare vertigini, perdita di sensi e morte. Situazioni che non richiedono un alto livello di concentrazione in soggetti che presentano già problemi cardiaci e respiratori.

Anidride solforosa (SO_2): fa parte di un gruppo di componenti più ampio: gli ossidi di zolfo gassosi SO_x. SO_2 viene utilizzato come indicatore per l'intero gruppo poiché è il componente che crea maggiori preoccupazioni data la maggiore concentrazione. La concentrazione di SO_3 è minore perché la reazione che la genera è molto più lenta rispetto a quella che ha come risultato SO_2. [$S+O_2 \rightarrow SO_2$, $2SO_2+ O_2 \rightarrow 2SO_3$] inoltre SO_3 ha un'alta reattività con l'acqua (H_2O) la cui razione porta alla formazione di acido solforico H2SO4, che presenta consistenza liquida o oleosa a seconda della temperatura ambientale. Le fonti di creazione di questi inquinanti sono comuni all'intero gruppo, tra le più importanti possiamo annoverare: la combustione di combustibili fossili effettuata nelle industrie e nelle centrali elettriche e le eruzioni vulcaniche. L'esposizione ad alte concentrazioni di SO_x può provocare problemi respiratori. L'aspetto che maggiormente spaventa, considerando l'impatto sulla salute umana, è la propensione degli ossidi di zolfo a reagire con altri composti presenti nell'atmosfera che portano alla creazione di piccole particelle, le quali contribuiscono all'inquinamento da particolato (PM). Le piccole particelle possono essere inalate e ciò permette la penetrazione nelle vie respiratorie fino a giungere ai polmoni. Oppure possono depositarsi su statue e monumenti provocando un danneggiamento della pietra e danneggiando beni culturali. Gli effetti della presenza di SO_x nell'atmosfera non sono importanti solo per gli esseri umani , ma anche per le piante e per interi ecosistemi, essendo la causa principale delle piogge acide.

Biossido di azoto (NO_2) L'intero gruppo degli ossidi di azoto (NO_x) , di cui fa parte il biossido di azoto, è altamente reattivo. Essi vengono prodotti , principalmente, durante la combustione. Il biossido di azoto non viene immesso direttamente nell'atmosfera, ma è il prodotto di reazioni chimiche che avvengono nell'atmosfera. Le reazioni dei biossidi di azoto assieme ad altre sostanze presenti nell'atmosfera porta alla formazione di particolato e ozono. Il biossido di azoto è quattro volte più tossico del monossido di azoto per la salute umana.

Gli strumenti messi a disposizione per il monitoraggio e lo studio dell'inquinamento sono molteplici e differenti tra di loro. Principalmente possiamo distinguerli in due grosse macro categorie: strumenti a terra, che permettono di avere un informazione più dettagliata su un area più ristretta; tecnologie satellitari, che forniscono una visione di insieme,

più amplia, con minor dettaglio. Per quanto riguarda la seconda categoria, un ruolo importante è ricoperto dall'Agenzia Spaziale Europea (ESA)[10] e dalla NASA[11]. I quali si sono occupati di molteplici missioni satellitari dedicate al monitoraggio dell'atmosfera terrestre. Basti ricordare: Envisat da parte dell'ESA e Aura della NASA. Nel 2020 partirà la missione Seninel-5, mentre la missione precursore Copernicus Sentinel-5P è stata lanciata nel 2017.

1.2 Copernicus

Copernicus è il programma di osservazione terrestre coordinato dalla Commissione Europea, in collaborazione con gli stati membri. Si occupa di monitorare il pianeta fornendo

Figura 1.4: Logo Copernicus

servizi, liberi e gratuiti, di informazione basati sia su osservazioni satellitari ma anche da misuratori terrestri, aerei, marittimi. Collaborano a questo progetto anche: Agenzia spaziale Europea ESA, organizzazione Europea per l'esercizio dei satelliti meteorologici (EUMETSAT), il centro europeo per le previsioni meteorologiche a medio termine (CEPMMT), Mercator Ocean, le agenzie dell'UE.

1.3 Agenzia Spaziale Europea - ESA

Figura 1.5: Logo Agenzia Spaziale Europea

L'Agenzia spaziale Europea ha come obiettivo quello di sviluppare le capacità spaziali europee, portando avanti programmi per lo studio della terra, della sua atmosfera, sul sistema solare e sull'universo in generale [12].

Gli stati che fanno parte dell'ESA sono 20: Austria, Belgio, Danimarca, Finlandia, Francia, Germania, Grecia, Irlanda, Italia, Lussemburgo, Norvegia, Paesi Bassi, Polonia, Portogallo, Romania, Regno Unito, Repubblica Ceca, Spagna, Svezia e Svizzera. L'unione di questi, permette, sia a livello finanziario che intellettuale di raggiungere risultati, che singolarmente ogni stato non avrebbe potuto raggiungere. Ai 20 stati si aggiungono anche

Estonia, Slovenia, Ungheria e Canada che hanno stretto alcuni accordi di cooperazione. Sono in atto anche collaborazioni con agenzie extra-europee atte a condividere e rafforzare le conoscenze ottenute tramite la ricerca.

Sebbene abbia grossi legami con l'UE, l'ESA svolge un'attività completamente indipendente da quest'ultima, come testimonia la mancanza di alcuni Paesi europei all'interno degli stati cooperanti e la presenza di Paesi extra-europei.

1.4 Progetto Sentinel

ESA e la Commissione Europea del progetto GMES (Global Monitoring for Environment and Security) stanno sviluppando un una serie di missioni finalizzate all'osservazione della Terra. L'intero programma prende il nome di SENTINEL. L'obiettivo di SENTINEL è quello apportare una miglioria tecnologica rispetto alle vecchie missioni utilizzate per l'osservazione terrestre, come ad esempio European Remote Sensing (ERS) [13] , la quale non è più attiva, o altre in procinto di terminare il loro ciclo di vita. Come detto in precedenza, il progetto SENTINEL prevede numerose missioni, le quali si concentreranno su aspetti differenti.

Ogni missione di Sentinel è basata su una costellazione di una coppia di satelliti, costantemente in movimento su orbite diverse, per soddisfare i requisiti di revisione e copertura, ottenendo set di dati affidabili per i servizi Copernicus. Le missioni Sentinel sono riportate di seguito:

- **Sentinel-1**: è una missione di imaging-radar ad apertura sintetica in banda C per il monitoraggio terrestre e oceanico indipendente dalle condizioni meteorologiche. È composta da due satelliti, in orbita polare, che operano sia di giorno che di notte. Il primo satellite è stato lanciato nell'Aprile del 2014;

- **Sentinel-2**: come la precedente è costituita da una famiglia di due satelliti in orbita polare. Ha come obiettivo il monitoraggio della superficie terrestre , delle zone costiere e dei cambiamenti nella vegetazione. I satelliti sono dotati di uno strumento multi-spettrale , che fornisce tredici bande spettrali differenti per ogni acquisizione. Il primo satellite è stato lanciato a Giugno 2015;

- **Sentinel-3**: questa missione ha come lo studio della topografia della superficie del mare, le temperature della superficie terrestre e marittima, il colore della terra e dell'oceano, con la finalità di monitorare i cambiamenti sia del clima che dell'ambiente. È una missione multi-strumento e , a differenza delle precedenti, è costituita da tre satelliti in orbita polare. Il primo satellite è stato inviato a febbraio 2016;

- **Sentinel-4**: ha come missione quella di monitorare la composizione atmosferica al fine di ottenere un analisi della qualità dell'aria. Lo strumento SENTINEL-4 UVN è uno spettrometro a luce Ultraviolet-Visible-Near-Infrared (UVN) che sarà montato a bordo dei satelliti Meteosat di terza generazione, gestiti da EUMETSAT. I dati forniti verranno utilizzati per supportare il monitoraggio e le previsioni in Europa;

- **Sentinel-5**: dedicato al monitoraggio della qualità dell'aria. Lo strumento Sentinel-5 UVNS è uno spettrometro che sarà montato sui satelliti MetOp di seconda generazione. Ha come obiettivo il monitoraggio continuo dell'atmosfera terrestre, fornendo dati globali, su larga scala. Sarà lanciato nel 2021.

- **Sentinel-5P**: è una missione satellitare precursore che mira a colmare le lacune dovute al ritiro del satellite Envisat e la missione Aura della NASA e il lancio di Sentinel-5. Ha come obiettivo la misurazione di alcuni parametri atmosferici con risoluzione spazio-temporale elevata, riguardanti: qualità dell'aria , ozono, radiazioni UV. Il satellite è stato lanciato a Ottobre 2017.

- **Sentinel-6**: l'ultima missione di questo programma, verrà lanciato nel Novembre 2020. Esso trasporterà un altimetro radar per misurare l'altezza globale della superficie del mare. Prevede la collaborazione di Europa e Stati Uniti.

1.4.1 Sentinel-5P

Sentinel-5P é la missione Copernicus precursore di Sentinel-5. La prima missione di copernicus che ha come obiettivo quello di monitorare l'atmosfera terrestre eseguendo misurazioni atmosferiche con alta risoluzione spazio-temporale. Questo progetto nasce dalla collaborazione di ESA con: la Commissione Europea, l'Ufficio spaziale Olandese.

Il satellite è stato lanciato il 13 Ottobre del 2017 dal cosmodromo di Plesetsk in Russia. La fase di accelerazione si è conclusa il 5 Marzo 2019, da allora svolge operazioni di routine.

Durante la fase di accelerazione si è così impostato il piano di fornitura e rilascio dei dati:

- Dopo 8 mesi dal lancio: Livello 1B; colonne totali di ozono, biossido di azoto, monossido di carbonio; cloud e aerosol (10/07/2018)

- Dopo 10 mesi: totale colonne di ozono, formaldeide, anidride solforosa (17/10/2018)

- Dopo 12 mesi: colonne totali di ozono troposferico, colonne di metano (1/03/2019)

- 30/09/2019: Aerosol Layer Height

- Rilascio di Ozone Profile previsto per gli inizi del 2020

La piattaforma meccanica ha una struttura esagonale che ospita tutti i sotto sistemi che costituiscono il satellite. I principali sottosistemi si occupano di: (i) Propulsione, (ii) Energia elettrica (EPS), (iii) Controllo termico (TCS), comprende dei riscaldatori utili a mantenere l'ambiente termico sulla piattaforma, (iv) Sistema di controllo di orbita e assetto (AOCS) composto da : ricevitore GPS, ruote di reazione , un sensore di Terra, magnetometro, (v) Gestione dati (DHS), comprende il computer di bordo, (vi) Comunicazione di banda S (TT&C), (vii) Gestione della trasmissione dei dati del payload (PDHT).

Sentinel-5P utilizza un'orbita con inclinazione, 98,7° circa. L'orbita è calcolata come la distanza angolare del piano orbitale rispetto all'equatore [14]. L'orbita è quasi polare, calcolata in maniera tale che la superficie considerata sia sempre illuminata con la stessa angolatura dai raggi solari. Il ciclo orbitale dura 16 giorni per un totale di 227 orbite, 14

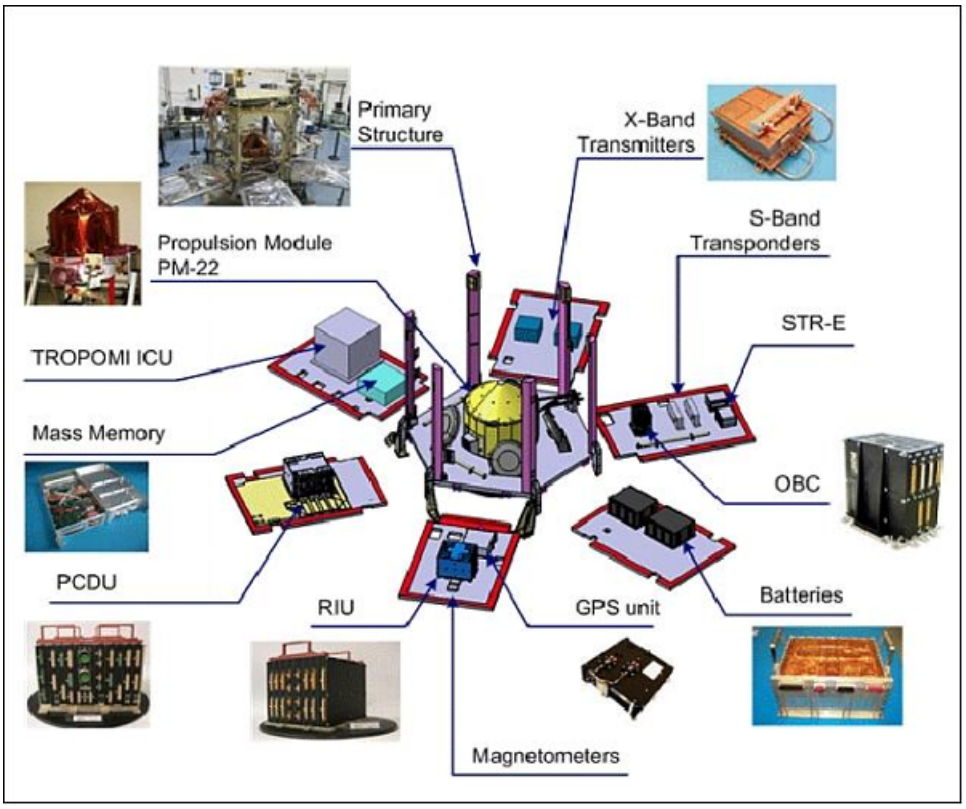

Figura 1.6: SENTINEL-5P: Elementi della piattaforma

giornaliere. Con il termine ciclo si intende il tempo necessario affinché il satellite ricopra l'intera superficie terrestre e ritorni sullo stesso punto di partenza. L'altitudine dell'orbita è pari, circa, a 824 km. L'orbita di Sentinel-5P è calcolata in maniera tale che venga seguito da Suomi-NPP, satellite della NASA della missione NPP iniziata nel 2011, in maniera tale che la zona osservata da Sentinel-5P resti all'interno della zona esaminata da Suomi-NPP, come mostrato in Figura 1.7.

Grazie all'orbita stabilita e alla grandezza dello swath, 108° (circa 2600km a terra), fornisce copertura giornaliera completa della superficie di misurazione di luminosità e riflettanza per latitudini maggiori di 7° e minori di -7°, e una copertura del 95% nella fascia con latitudini comprese del range -7° e 7°.

Il satellite trasporta lo "TROPOspheric Monitoring Instrument" (TROPOMI) che combina punti di forza di strumenti già utilizzati per l'osservazione dell'atmosfera, come: SCIAMACHY (ENVISAT), OMI (AURA) e colma le lacune presenti tra essi e le missioni future di Copernicus Sentinel-4 e Sentinel-5 tramite l'utilizzo di tecnologie avanzate che permettono di ottenere prestazioni non ottenibili tramite gli strumenti già presenti nello spazio.

Le caratteristiche principali dello strumento sono [15]:

- Tipo: spettrometro per imaging a griglia passiva

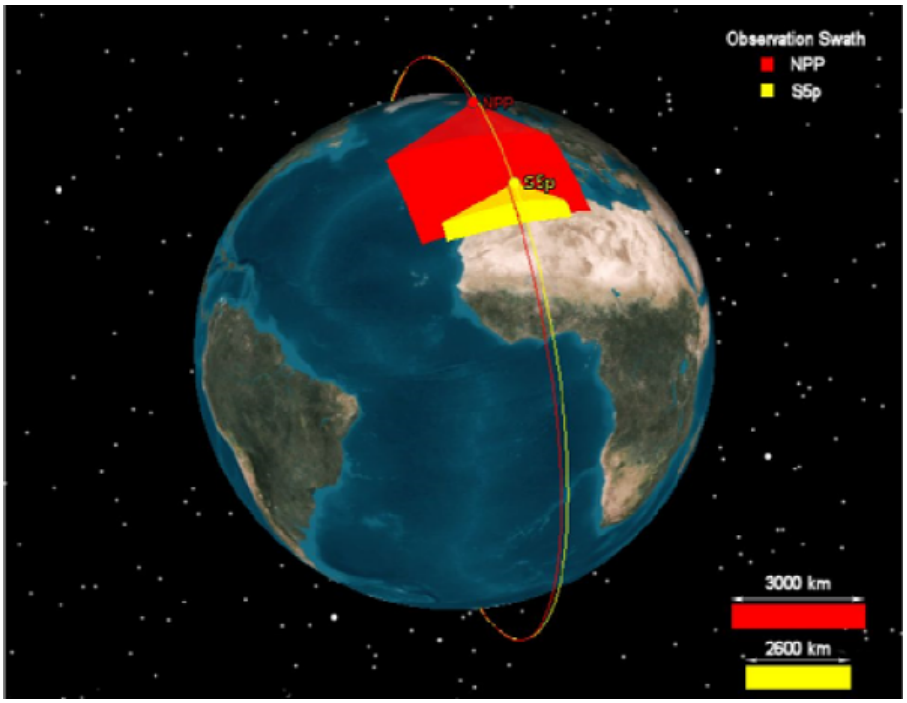

Figura 1.7: Sentinel-5P: Orbita

- Configurazione: broom staring (non-scanning) in nadir viewing
- Larghezza swath: 2.600 km
- Campionamento spaziale: 7x7 km 2
- Spettrale: 4 spettrometri, ciascuno suddiviso elettronicamente in due bande (2 in UV, 2 in VIS, 2 in NIR, 2 in SWIR)
- Precisione radiometrica (assoluta): dall'1,6% (SWIR) all'1,9% (UV) della riflettanza spettrale terrestre misurata.
- Massa complessiva: 204,3 kg esclusa ICU (16,7 kg)
- Dimensioni: 1,40 x 0,65 x 0,75 m
- Durata del progetto: 7 anni
- Assorbimento medio: 155 W
- Volume di dati generato: 139 Gbit per orbita completa.

I prodotti forniti da TROPOMI possono essere suddivisi in tre livelli: (i) Livello 0, (ii) Livello 1B, (iii) Livello 2, i quali rappresentano tre livelli di elaborazione dei dati, dove livello 0 rappresenta il dato grezzo acquisito dallo strumento. All'interno di questo lavoro sono stati considerati solamente i dati del livello 2, il quale include:

- Colonna totale di ozono, anidride solforosa, biossido di azoto, monossido di carbonio, formaldeide e metano.
- Colonna troposferica di ozono
- Profili verticali di ozono
- Informazioni su nuvole e aerosol

Inoltre, vengono fornite tre elaborazioni dei dati:

- NRT quasi in tempo reale
- OFFL offline
- Reprocessing

Nel primo caso i dati sono disponibili entro tre ore dal rilevamento, questo servizio è offerto per la maggior parte dei prodotti. Nel secondo caso le tempistiche differiscono a seconda del prodotto considerato: 12 ore per il livello 1B, fino a cinque giorni per i prodotti del livello 2. La terza elaborazione viene eseguita solo in caso sia ritenuto necessario apportare aggiornamenti al prodotto.

Come esplicitato dal sito ufficiale dell'ESA [16], la missione Sentinel-5P, oltre a fornire dati che permettano di studiare la concentrazione di sostanze inquinanti nell'atmosfera, "contribuirà anche a servizi come il monitoraggio delle ceneri vulcaniche per la sicurezza aerea e per servizi che avvertono di alti livelli di radiazioni UV che possono causare danni alla pelle. Inoltre, gli scienziati useranno anche i dati per migliorare la nostra conoscenza di importanti processi nell'atmosfera legati al clima e alla formazione di buchi nello strato di ozono."

1.5 Struttura documento

La struttura del documento è la seguente:

- **Stato dell'arte**: In questo capitolo viene svolta la descrizione dei lavori, presenti in letteratura, correlati con il lavoro proposto da questo progetto di tesi e l'introduzione al Machine Learning e alle tecniche utilizzare nel progetto.
- **Dataset e Tools**: Dopo una descrizione delle sorgenti di dati e dei servizi utilizzati per l'acquisizione degli stessi è riportata la spiegazione del processo di acquisizione dei dati. Inoltre, è fornita anche una panoramica sulle librerie e i frameworks utilizzati durante l'analisi.
- **Metodologia**: Dopo una definizione del problema si passa al cuore del progetto tramite la descrizione dell'intero processo che, partendo dai dati grezzi acquisiti, porta alla creazione dei dataset utilizzati durante lo sviluppo dei modelli. A proposito di questi ultimi, per ogni modello è brevemente descritto il processo di preparazione dei dati. Nella seconda parte del capitolo sono presenti alcuni approfondimenti su: tecniche utilizzate durante l'analisi, come GridSearch, per la ricerca degli iperparametri, o CrossValidation ,per la validazione dei modelli; e le metriche di valutazione delle performance.

- **Esperimenti**: Il presente capitolo presenta tutti i risultati ottenuti durante il lavoro svolto. I risultati sono suddivisi in due macro categorie: 'predizione inquinanti dal giorno corrente' e 'predizione inquinanti da finestra tempo'. Quest'ultime sono a loro volta divise in due categorie: la casistica che prende in considerazione tutti gli inquinanti e quella che considera il singolo inquinante.
  Nella seconda parte di questo capitolo viene riportata una sezione comparativa di tutti i risultati
- **Conclusioni e sviluppi futuri**

Capitolo 2

Stato dell'arte

Questo capitolo è suddiviso in due parti: la prima parte descrive alcuni lavori, presenti in letteratura, correlati con questo progetto; la seconda parte introduce l'ambito del machine learning e descrive , in maniera teorica, i modelli di predizione utilizzati nell'ambito del progetto di tesi.

2.1 Lavori correlati

L'analisi della qualità dell'aria è un problema che spesso è stato affrontato negli ultimi anni. Molti sono i lavori che, tramite algoritmi di machine learning e deep learning, cercano di studiare il fenomeno per creare modelli che permettano di effettuare previsioni per il futuro. La maggior parte di questi progetti utilizzano dati di sensori posti a terra, quindi hanno un campo applicativo differente dal progetto riportato in questa tesi [17][18]. La principale differenza sta nel fatto che i sensori a terra considerano aree ristrette e soprattutto che i valori riportati vanno a rappresentare solamente una fascia ristretta della troposfera. Tutto ciò rende non comparabili questi dati con i dati satellitari, i quali considerano l'intera colonna atmosferica, o in casi particolari, l'intera colonna troposferica.
Per quanto riguarda la parte minoritaria di lavori che analizzano dati satellitari, questi ottengono dati da tecnologie differenti e quindi spesso ci si ritrova nella condizione in cui le granularità osservate, sia a livello spaziale che temporale, non sono comparabili con quelle ottenute dal satellite Sentinel-5P.
Di seguito verrà riportato un lavoro, che nonostante l'utilizzo di tecnologie differenti, considera le stesse informazioni utilizzate in questo progetto: inquinanti e condizioni meteorologiche. [19]

Il paper *"Hybridization of air quality forecasting models using machine learning and clustering: an original approach to detect pollutant peak"* [19] è un lavoro pubblicato a cura di: Wani Tamas, Gilles Notton, Christophe Paoli, Marie-Laure Nivet, Cyril Voyant. Questo lavoro si propone di presentare un approccio innovativo, dato dalla combinazione di reti neurali e meccanismi di clustering, per la detection dei picchi di inquinanti. Sono stati sviluppati dei modelli predittivi, tramite l'utilizzo di tecniche di Machine Learning, applicati a concentrazioni orarie di ozono (O3), biossido si azoto (NO_2) e particolato

(PM_{10}). Come primo modello è stato utilizzato un MultiLayer-Percepton (MLP). Successivamente è stato sperimentato l'utilizzo di questo modello applicato in combinazione con meccanismi di cluster come: K-Means e Hierarchical clustering. I meccanismi di clustering sono stati utilizzati al fine di suddividere il dataset di partenza in sotto-dataset per poi applicare il modello MLP a ognuno di essi. I dati considerati sono: dati di inquinanti (O_3, NO_2, PM_{10}) e dati meteorologici (temperatura, pressione, vento, umidità, precipitazioni, nuvolosità). In Figura 2.1 viene riportato il flow chart che rappresenta la struttura dell'algoritmo.

2.2 Machine learning

Questo capitolo introduce un breve cenno teorico agli algoritmi di machine learning utilizzati durante il lavoro di tesi. Nel capitolo successivo, sara' trattata in maniera piu' esaustiva l'implementazione e personalizzazione apportata appositamente per questo lavoro di tesi.

Il machine learning è una branca dell'informatica in stretto collegamento con l'intelligenza artificiale, basata sul presupposto che, attraverso i dati, le macchine possano acquisire conoscenza identificando dei modelli autonomamente. Infatti, il machine learning nasce proprio dalla teoria che un computer possa imparare tramite il riconoscimento di schemi tra dati. Si potrebbe definire il machine learning come l'insieme dei meccanismi che permettono a una macchina intelligente di migliorare le proprie capacità e prestazioni nel tempo. La macchina impara a svolgere determinati compiti migliorando, con l'esperienza, le proprie capacità. Volgendo lo sguardo al passato, è possibile ritrovare le origini del machine learning già negli anni cinquanta, quando alcuni matematici e statistici ipotizzavano di poter utilizzare i metodi probabilistici per realizzare macchine in grado di calcolare la probabilità di accadimento di un determinato evento e conseguentemente prendere decisioni. Uno dei grandi nomi del tempo, e che rimase nella storia per la costruzione della macchina a lui omonima, è Alan Turing. Nonostante, in questo periodo, si sviluppò un grosso fervore sull'argomento molto era anche lo scetticismo dimostrato a riguardo. Quest'ultimo, accompagnato dalla mancanza di investimenti, fu una delle cause per cui l'ambito di ricerca fu accantonato per alcuni anni.
A partire dagli anni Ottanta e maggiormente negli anni Novanta, grazie a innovazioni tecniche legate all'ambito statistico, le attività di ricerca si intensificarono fino a giungere al tempo presente in cui il machine learning è uno dei temi più importanti.

È possibile dividere i sistemi di machine learning in tre categorie:

- **Supervised Learning** È una tecnica che permette di addestrare il modello tramite dati etichettati. Nella fase di training viene creata una funzione di mappatura input-output basata sulle coppie (inut, output) fornite. Questa funzione sarà poi utilizzata nella fase di test per assegnare un output a dati, di input, non ancora etichettati.
  A seconda del dominio dell'output possiamo distinguere due casistiche: se il dominio dell'output è finito ed è costituito da valori discreti, allora il modello attuerà una classificazione; se il dominio di output è continuo esso attuerà una regressione.

- **Unsupervised Learning** Ha come obiettivo l'estrazione di informazioni a partire da dati non etichettati e la cui struttura non è conosciuta a priori. Esempi di algoritmi

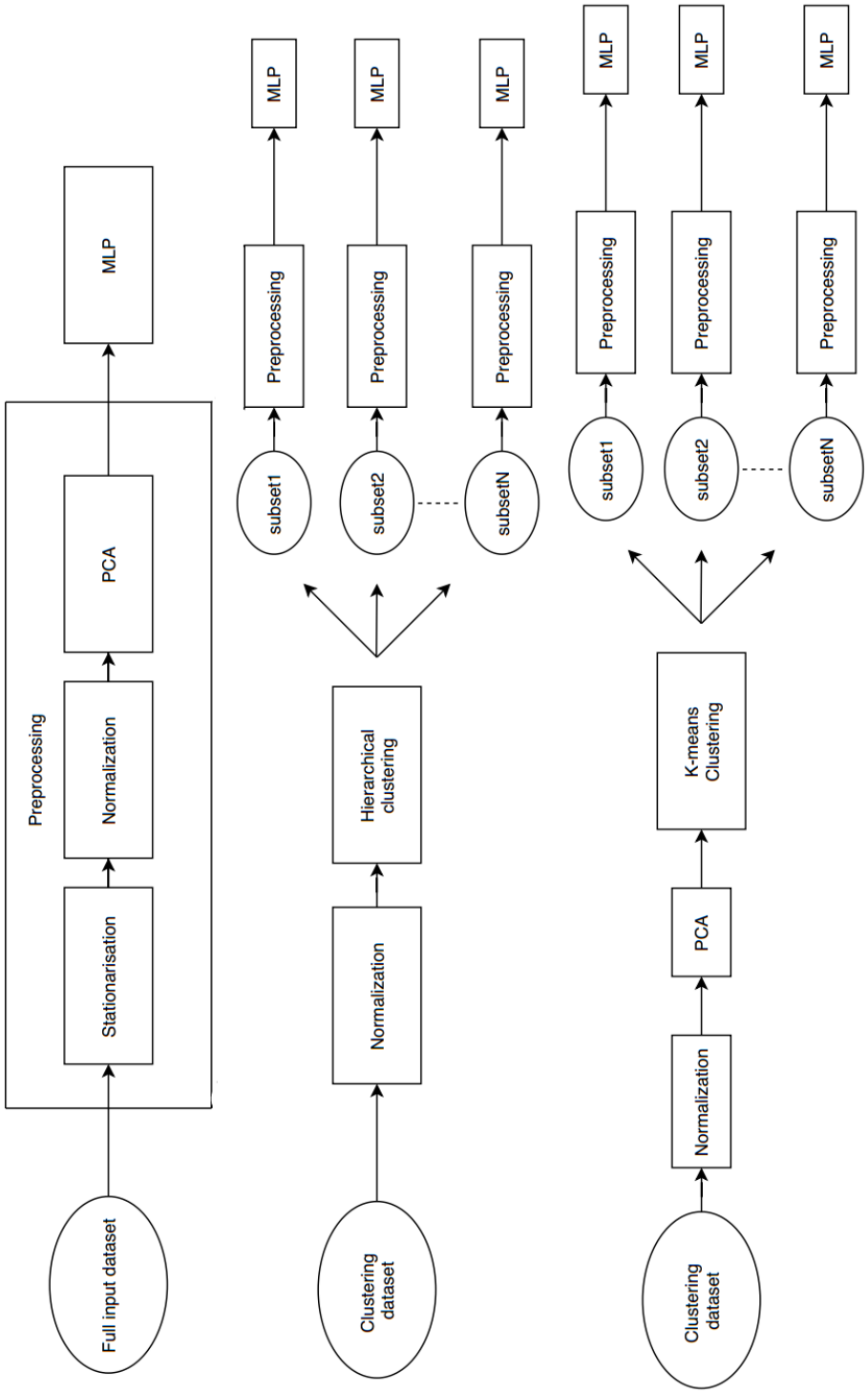

Figura 2.1: Flow chart

non supervisionati sono: algoritmi di clustering e algoritmi per la riduzione della dimensionalità (PCA, t-SNE).

- **Reinforcement learning** Prevede un Sistema (agente) che parta da uno stato iniziale. Successivamente viene applicata un'azione. Sul risultato di questa azione vengono calcolate delle metriche per valutarne le performance. Questa valutazione viene inviata come feedback al sistema, il quale modifica il suo stato.

Questa tesi si focalizza nell'ambito del Supervised learning per la regressione. Di seguito sono introdotti gli approcci ai quali ci si e' ispirati per fornire soluzioni al problema trattato.

2.2.1 Decision Tree

Il Decision Tree [20] è un algoritmo per l'apprendimento supervisionato basato su decisioni strutturate in maniera sequenziale e gerarchica.

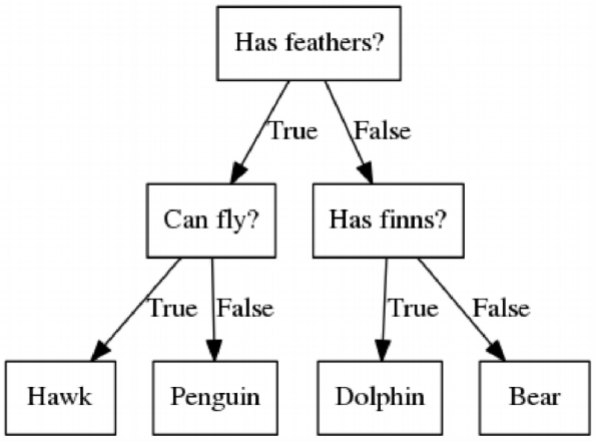

Figura 2.2: Esempio di DecisionTree

Nella struttura dell'albero ogni nodo raffigurato, che non sia un nodo foglia, rappresenta una feature. I rami di uscita di un nodo rappresentano una divisione del dataset di partenza basata sul valore della feature analizzata in quel dato nodo. Gli alberi sono caratterizzati dal numero di rami uscenti dai nodi. Sono definiti alberi binari gli alberi con due rami uscenti per ogni nodo. I nodi dell'albero si differenziano in nodi radice e nodi intermedi, spiegati in precedenza, e nodi foglia. I nodi foglia rappresentano l'output assegnato.

L'utilizzo di questo meccanismo di apprendimento prevede la suddivisione del lavoro in due fasi:

1. La **fase di training**: in cui viene creato l'albero.

   L'approccio utilizzato in questa fase è di tipo: Gready, l'attributo di split è scelto localmente e non globalmente; Top Down, si parte dal nodo radice e si procede fino a giungere ai nodi foglia. In questa fase è importante valutare i seguenti parametri:

- l'ordine di valutazione delle features
- valore di soglia per ogni feature

La scelta dell'attributo locale, su cui effettuare lo split, viene fatta in maniera tale da raggiungere il prima possibile la soluzione finale, in cui sono azzerate il numero di tuple con una certa etichetta oppure sottolinea la maggioranza per una data etichetta (distribuzione omogenea su un valore). Per la valutazione dell'attributo possono essere utilizzate diverse tecniche, tra cui:

- Entropia
  E(t)=$\sum j(P(j|t)logP(j|t))$
- Misclassification Error
  E(t)=$1 - maxP(i|t)$
  P(i|t): frequenza della classe j al nodo t
  L'obiettivo risulta quindi minimizzare F.
- GiniIndex
  Metodo più utilizzato
  G(t)=$1 - \sum_j p(j|t)^2$
  J: classi possibili
  P(j|t): frequenza di j al nodo t
  Casi limite:
  - G=0 tutti gli elementi appartengono alla stessa classe, caso di interesse.
  - G=1-1/ (numero classi) distribuzione uniforme delle classi

  Guadagno = M_0 – $M_{1,2}$
  M_0: GiniIndex nodo padre
  $M_{1,2}$ = Ginnisplit= $\sum_{i=1}^{k} \frac{ni}{n} G(i)$
  ni: numero record della partizione i-esima
  n: numero record nodo padre
  Si effettua il calcolo per tutti gli attributi rimanenti, e si sceglie quello con guadagno maggiore. GiniIndex può essere usato anche per scegliere la soglia: si considerano diverse soglie, di ognuna di essa si calcola G e si considera la soglia con G più piccolo.

2. La **fase di test**, per ogni istanza del test set, si percorre l'albero costruito in fase di training fino a giungere al nodo foglia che rappresenta il valore di predizione attribuito a quell'istanza.

Finora è stato illustrato il funzionamento del Decision Tree applicato a problemi di classificazione.
Per quanto riguarda la regressione, il funzionamento generale è il medesimo, bisogna tener conto di una differenze nella creazione dell'albero, cioè nella fase di train ad ogni nodo foglia viene attribuito il valore medio delle label di tutte le istanze ad esso appartenenti anziché un etichetta di classe. Un albero decisionale suddivide le funzioni di input in diverse aree e assegna un valore di previsione a ciascuna area.

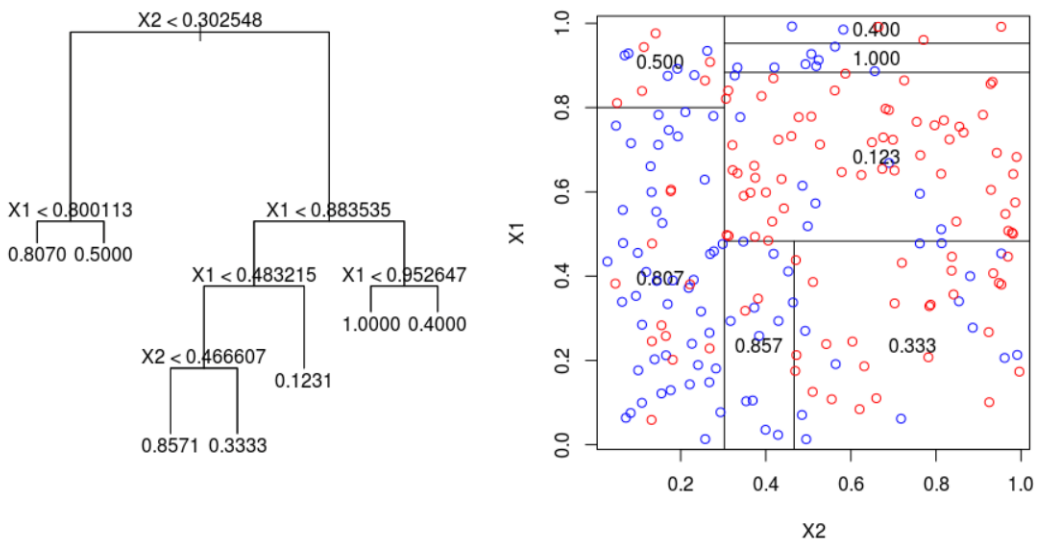

Figura 2.3: Esempio di DecisionTree Regressor

2.2.2 Random Forest

Un *Esemble model* è la combinazione di più algoritmi di machine learning che permette di avere una predizione più accurata rispetto ai modelli considerati singolarmente.
Esistono due tipologie di Esemble model: Boosting e Bagging.

Bagging (Bootstrap aggregating) Il bootstrap è una tecnica di ricampionamento, essa viene utilizzata per ricampionare, più volte, i dati di training. Dato un insieme di training di n unità, esso seleziona n unità a caso con ripetizione, ottenendo molti dataset, tutti di dimensione n da utilizzare per la stima del modello.

Il Random Forest è un algoritmo di apprendimento supervisionato utilizzato sia per la classificazione che per la regressione. L'algoritmo Random Forest permette di superare alcuni limiti presenti nell'utilizzo dell'albero decisionale. Infatti, gli alberi con elevata profondità: sono molto sensibili ai dati di training; hanno elevato costo computazionale e alto rischio di overfitting. Il modello Random Forest è un Esemble model, prevede la combinazione di più alberi decisionali in un singolo modello. Random Forest può essere classificato come una tecnica di **bagging** (Bootstrap aggregating), poiché gli alberi decisionali che lo compongono vengono eseguiti in parallelo senza alcuna dipendenza e interazione tra essi.

Il funzionamento è il seguente: (i) generazione di diversi dataset con bootstrap; (ii) stima di un albero per ciascun dataset; (iii) applicazione di tecniche di regressione e classificazione per le previsioni. Per garantire la de-correlazione degli alberi, durante la costruzione del modello bisogna considerare alcuni fattori:

- Bisogna garantire che il modello non prediliga nessuna features e che attribuisca a ciascuna feature lo stesso peso. Per questo motivo nella scelta della feature successiva

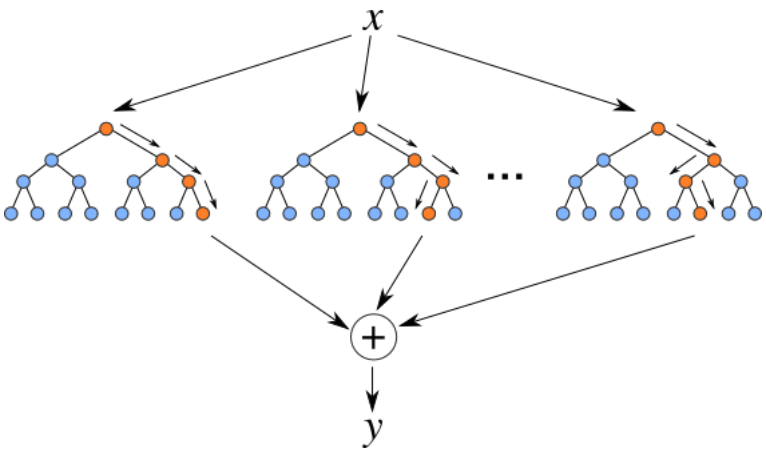

Figura 2.4: Random Forest

viene considerato solamente un sotto insieme casuale di features, in maniera tale che le features più importanti non vengano scelte sempre per prime.

- Ogni albero estrae un campione casuale dal set di dati originale permettendo di aggiungere un ulteriore elemento di casualità che impedisca il sovra-adattamento.

Durante l'esecuzione il modello combina i risultati ottenuti dai singoli alberi decisionali, producendo un singolo output.

2.2.3 Neural Networks

Le reti neurali sono dei modelli complessi composti da neuroni interconnessi. I primi lavori risalgono agli anni Ottanta e prendono ispirazione da un cervello umano composto da neuroni (unità elementari) e sinapsi (collegamenti tra unità).

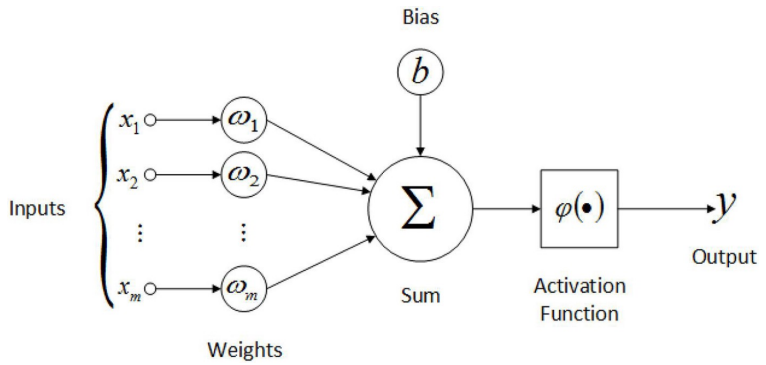

Figura 2.5: Perceptron

Neurone Il neurone è l'elemento base delle reti neurali. Il neurone è un modello non lineare. Per il suo funzionamento ha bisogno di:

- x vettore di input;

- w vettore dei pesi, i dati di input vengono moltiplicati con il vettore dei pesi;
- b bias, offset che permette di modificare il potenziale di azione dell'input;
- f funzione di attivazione (es: funzione gradino, funzione sinusoidale).

Tutto ciò permette di ottenere l'output.

Nelle reti neurali, i neuroni sono organizzati in livelli. I livelli intermedi vengono chiamati *Hidden Layer*. I dati di input, di un neurone intermedio, sono ottenuti dall'output generato dai neuroni del livello precedenti. Le prestazioni del modello dipendono dal numero di nodi e dal numero di livelli. Durante la fase di train bias e pesi vengono modificati in maniera tale da minimizzare la funzione di perdita l=l(y,$\tilde{y}$). La funzione di perdita è utilizzata per la valutazione delle performance.

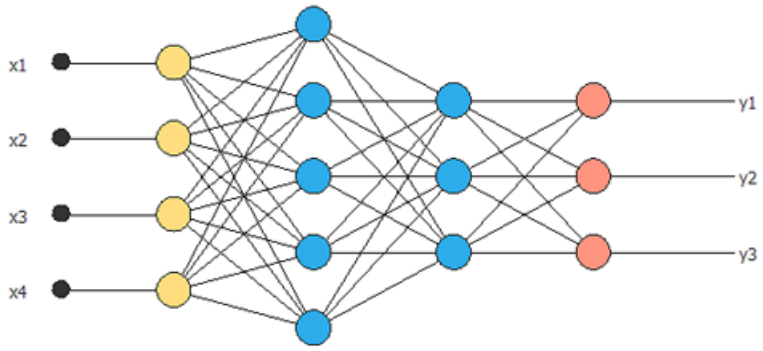

Figura 2.6: Neural Network

CNN Le reti neurali convoluzionali, o ConvNet (CNN) [21] sono un algoritmo di deep learning utilizzate specialmente nel campo della computer vision. La stuttura delle reti neurali convoluzionali rispetta quella delle generiche reti neurali, la differenza sta nel fatto che ogni layer ospita una kernel di dimensione $n * m$. Il kernel viene moltiplicato per una porzione dell'input ottenendo in output una matrice delle stesse dimensioni, nel caso in cui si utilizzi una tecnica di zero padding, o più piccola, data dalla somma di tutti gli elementi moltiplicati. L'output viene generato facendo scorrere il kernel sull'intera immagine di input, ottenendo una feature map.
A differenza delle reti fully connected, in cui i neuroni sono collegati a tutti i neuroni del livello precedente, in questo caso ogni neurone è collegato solamente a una sotto porzione del volume di input.

La rete convoluzionale è caratterizzata da diversi parametri: padding, stride, kernel size, dilatation. A livello architetturale, tre dei layer più comuni sono: (i) livello convoluzionale: applica alle immagini di input dei filtri, i quali vanno ad attivare feature differenti delle immagini (ii) livello pooling: esegue un sotto-campionamento non lineare riducendo il numero di parametri che la rete deve apprendere. (iii) Rectified Linear Unit (ReLU): permette di accelerare l'addestramento e renderlo più efficiente. Per fare ciò mappa a zero

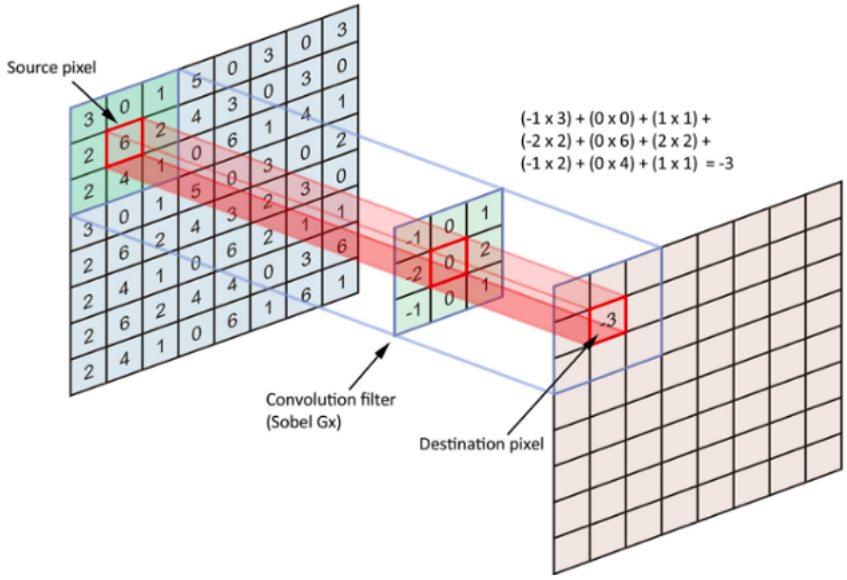

Figura 2.7: Convoluzione

i valori negativi mentre vengono lasciati invariati i valori positivi. Per questo motivo può essere chiamato anche livello di attivazione.

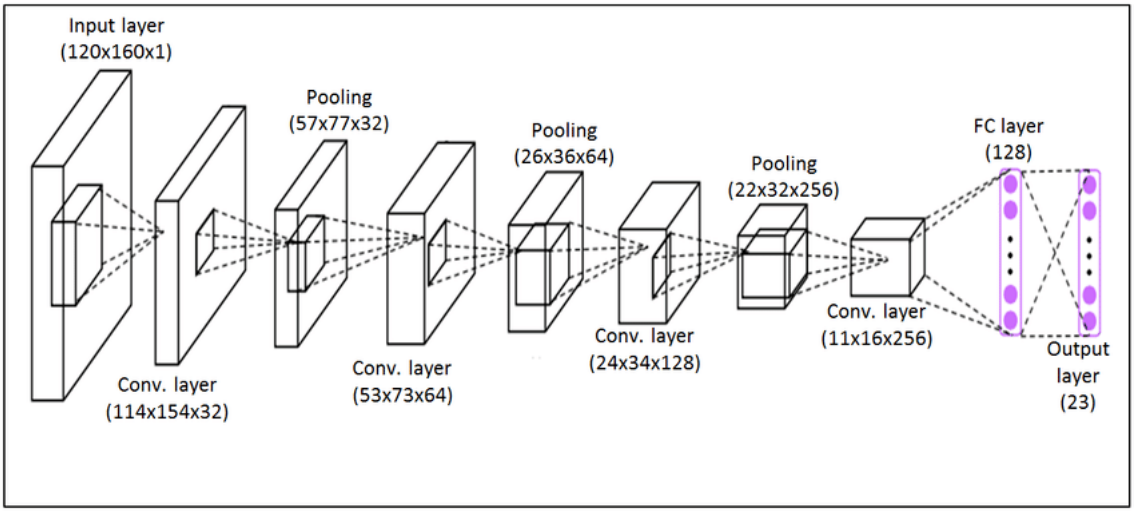

Figura 2.8: Rete convoluzionale

Long Short Term Memory - LSTM Le reti LSTM vengono utilizzate per classificare, elaborare e fare previsioni su serie temporali. Esse sono basate sul meccanismo di back-propagation.

Strutturalmente è possibile considerare tre elementi fondamentali:

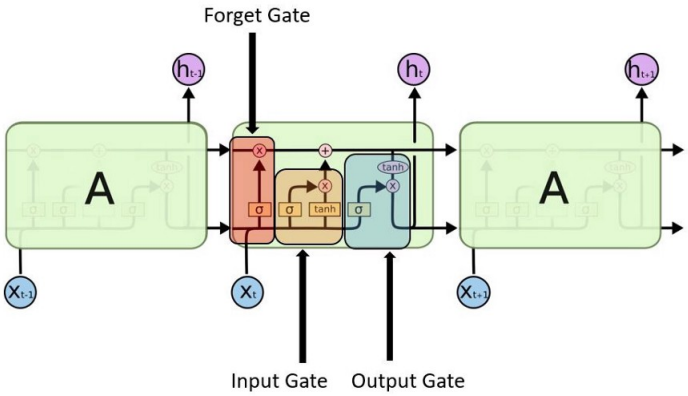

Figura 2.9: Struttura LSTM

- Input Gate: una funzione *sigmoid* assegna un valore 0 o 1 ai valori di input, in maniera tale da decidere quali valori vengono fatti passare. La funzione *tanh* attribuisce a ogni valore di input un peso compreso nel range [-1,1].
- Forget Gate: definisce quali dettagli devono essere considerati utilizzando una funzione *sigmoid* applicata sui dati di input e lo stato precedente. Come output si ha un valore compreso tra 0 (scartato) e 1 (preso).
- Output Gate: anche in questo caso viene utilizzata una funzione *sigmoid* per capile quali sono i valori ammessi, i quali vengono pesati attraverso una funzione *tanh*.

Capitolo 3

Sorgenti dati e Tools

Il capitolo che segue si compone di due parti. Nella prima parte verranno illustrati tutti i servizi, che permettono l'acquisizione dei dati e la visualizzazione degli stessi, utili per lo sviluppo del progetto. Nella seconda parte viene riportata una panoramica sulle librerie e i frameworks utilizzati.

3.1 Sorgente dati e servizi

3.1.1 Copernicus Hub e acquisizione dati

Copernicus Open Access Hub [22] fornisce, come riportato dal sito web, accesso ai prodotti delle missioni Sentinel-1, Sentinel-2, Sentinel-3, Sentinel-5P in maniera gratuita e accessibile a tutti. Per la ricerca e il download dei dati è possibile utilizzare due strumenti differenti messi a disposizione. Il primo, Open Hub, fornisce un interfaccia grafica in cui è possibile, tramite un menù apposito, selezionare i prodotti di interesse impostando alcuni filtri. Per effettuare una ricerca è possibile specificare le seguenti caratteristiche:

- area di interesse: è possibile inserire, nell'apposito campo, l'oggetto POLYGON che specifica i vertici del poligono dell'aria che si vuole esaminare. Viene fornito anche uno strumento grafico che permette di selezionare, direttamente sulla mappa, l'aria di studio.
- tempo di rilevamento o 'ingestion': permette di selezionare , tramite l'inserimento di una data di inizio e una di fine, una finestra temporale più o meno ampia che può far riferimento al tempo di rilevamento, cioè acquisizione del dato, o 'ingestion', cioè il tempo in cui il dato è stato reso disponibile dall'Hub.
- missione: come detto in precedenza, Copernicus Hub mette a disposizione i dati di diverse missioni Sentinel
- piattaforma
- tipo di prodotto
- livello di processo

- altre informazioni aggiuntive differenti da missione a missione.

Nel menù è possibile selezionare anche in che ordine visualizzare i risultati selezionati. Una volta effettuata la richiesta, viene visualizzate una lista di tutti i prodotti che rispettano i filtri impostati. Per ogni prodotto c'è la possibilità di: visionare sulla mappa l'aria di copertura; visionare il contenuto dello stesso, tramite lettura del file json; effettuare il downlowd.
Il secondo strumento è Hub API, basato sulla tecnologia REST, che permette di ottenere prestazioni migliori. Anche in questo caso è possibile specificare i filtri, gli stessi utilizzati nell'interfaccia grafica, al momento della richiesta.

3.1.2 Google Eart Engine

Google Eart Engine [23] è una piattaforma informatica che raccoglie dati satellitari di diverse piattaforme, fornendo un catalogo di immagini satellitari e set di dati geo-spaziali con copertura globale. L'archivio di dati pubblici contiene informazioni degli ultimi trenta anni per un volume di circa venti Petabyte. Google Earth non è solo un'archivio, ma una piattaforma in cui è possibile eseguire analisi geo-spaziali servendosi delle infrastrutture messe a disposizione da Google. Sono fornite molte vie di accesso e interazione con la piattaforma. Viene messa a disposizione un'API che permette l'accesso a tutte le risorse e all'utilizzo del cloud di Google. È stato sviluppato anche un IDE, *Code Editor*, per la scrittura e l'esecuzione di script [24] per lo sviluppo semplificato di algoritmi con accesso all'archivio. Esiste anche un'app web, *Explorer*, per l'accesso al catalogo dei dati e l'esecuzione di analisi semplici.
Riassumendo, Google Earth Engine fornisce, non solamente, un archivio molto ricco, capace di coprire e permettere analisi di diverso tipo, ma anche strumenti che permettano di accedere ai dati, manipolarli e analizzarli in maniera semplice e soprattutto efficace.

3.1.3 API Dark Sky

L'API Dark Sky [25] è un servizio che fornisce informazioni meteorologiche. Tramite l'API Dark Sky è possibile reperire informazioni su osservazioni del passato, sulle condizioni meteorologiche attuali e anche previsioni meteorologiche. È possibile selezionare una delle seguenti granularità temporali, a seconda del prodotto di interesse:

- minuto, per le previsioni fino a un'ora successiva al momento della richiesta;
- ora o giorno, per le previsioni fino a una settimana dopo il momento della richiesta
- ora , per le osservazioni storiche fino a dieci anni precedenti.

L'API Dark Sky è basata su informazioni recepite da una vasta gamma di fonti. [26] Alcune di esse sono:

- Centro meteorologico Canadese e NCEP degli USA , fornisce copertura globale ;
- "gfs": Sistema di previsioni USA NOAA, con copertura globale;
- Ufficio meteorologico tedesco, anche esso fornisce copertura globale;
- Integrated Surface Database' NOAA.

3.1.4 Visan

VISAN[27] è uno strumento di visualizzazione e analisi di dati atmosferici. Esso permette la lettura di prodotti ottenuti da diversi strumenti di missioni come: Sentinel-5P, Aeolus, GOME-2, IASI, OMI ecc. Ma anche di strumenti che forniscono dati a terra come NDACC e EVDC. Le funzioni principali fornite da questo strumento sono: la possibilità di manipolare questi dati tramite funzioni matematiche e soprattutto la possibilità di poter creare grafici 2D e worldmap.
VISAN usufruisce di due software: CODA e HARP, i quali sono usati per l'acquisizione dei dati dei prodotti atmosferici. L'applicazione Visan è completamente open source basata su pacchetti Open Source esistenti come Python, wxPython, VTK, CODA, HARP.
Questa applicazione fornisce anche un'interfaccia di comando, la quale, tramite l'utilizzo del linguaggi Python, permette l'implementazione di script. Per agevolare questo aspetto al momento dell'installazione vengono già scaricate alcune librerie Python utili per l'analisi come Numpy.
CODA e HARP offrono due modi complementari di accesso e gestione dei dati di prodotto. Mentre CODA fornisce l'accesso ai dati mantenendo invariata la struttura dei dati memorizzati nel prodotto, HARP esegue una conversione dei dati del prodotto in una struttura dati semplificata.

3.1.5 Panoply

Panoply[28] è un'applicazione multipiattaforma che funziona su Windows, Linux, Mac, sviluppata dal personale GISS.
Essa permette di lavorare su file come netCDF, HDF, GRIB. Tramite l'utilizzo di Panoply è possibile: tracciare array di latitudine-longitudine, latitudine-verticale, longitudine-verticale o tempo-latitudine da variabili multidimensionali; combinare due matrici in un grafico; tracciare i dati geo-referenziati su una mappa; utilizzare diverse tabelle di colori per le rappresentazioni grafiche; esportare i grafici ottenuti con diverse estensioni.

3.1.6 Acquisizione dati

L' obiettivo della tesi è quelli di utilizzare i dati ottenuti tramite la missione di Copernicus, Sentinel-5P, per effettuare uno studio sulla qualità dell'aria e creare dei modelli che permettano di effettuare previsioni sulla concentrazione degli inquinanti nell'atmosfera.
Sentinel-5P propone una vasta gamma di prodotti che vanno a caratterizzare la composizione atmosferica.
Tra questi si è scelto di utilizzare i seguenti componenti: (i) Anidride solforosa (SO_2); (ii) Ozono (O_3); (iii) Monossido di carbonio (CO); (iv) Formaldeide (HCHO); (v) Biossido di azoto (NO_2).
Tutti i prodotti riportati nell'elenco sono risultati del livello 2 di elaborazione dati, di cui si è parlato nella sezione 1.4.1. Inoltre, è presa in considerazione la possibilità di arricchire l'intero set di dati aggiungendo ulteriori features. Per fare ciò è stato necessario approfondire lo studio dell'*inquinamento* per comprendere quali fattori potessero essere correlati con la sua manifestazione e le sue variazioni.
Dopo un'attenta ricerca è stato deciso di utilizzare dati meteorologici. Per l'ottenimento di questi ultimi è stato utilizzato il servizio gratuito fornito da Dark Sky.

La correlazione presente tra le informazioni sugli inquinanti e le informazioni meteorologiche, supportata scientificamente [29], è stata confermata da un analisi effettuata all'inizio del progetto di tesi, di cui successivamente verranno riportati i dettagli.

Prima di iniziare con l'acquisizione dei dati è stata scelta arbitrariamente come area di studio l'intera superficie ricoperta dalla città di Milano.
Successivamente l'analisi è stata ampliata considerando un set di dieci città europee: Atene, Barcellona, Bruxelles, Ginevra, Milano, Norimberga, Parigi, Roma, Torino, Zurigo.
Tutti i dettagli riportati successivamente fanno riferimento all'analisi su Milano. Si tenga presente che tutto il processo di acquisizione è stato ripetuto per ogni singola città con le medesime modalità.

Figura 3.1: Città esaminate

Come periodo di riferimento è stata considerata la finestra temporale più amplia possibile che fosse coperta dai dati Sentinel. Come già citato nel capitolo precedente, la missione Sentinel-5P è una missione recente, e i dati disponibili forniscono una copertura completa di tutti i prodotti solamente a partire dal giorno 5 Dicembre 2018. Quest'ultima data è stata scelta come inizio del periodo di analisi il cui termine è stato stabilito, per esigenze di studio, alla data 19 Settembre 2019.

Ogni richiesta effettuata tramite l'API Sentinel fornisce come risultato un file netCDF-4.

La libreria netCDF-4 è stata creata a partire dalle librerie netCDF-3 e HDF-5. Essa permette all'utente di utilizzare API accessibili tramite linguaggi di programmazione, come python, per leggere i dati. Ogni file ha un nome che segue una convezione ben precisa, illustrata nella *Tabella 3.1.*

Numero caratteri	Significato
3	Nome missione
4	Tipo elaborazione dati
10	Identificativo prodotto
15	Tempo di inizio acquisizione (UTC)
15	Tempo fine acquisizione (UTC)
5	Numero orbita
2	Numero collezione
6	Numero versione processore
15	Tempo di processamento del prodotto (UTC)
2	Estensione file (.nc)

Tabella 3.1: Convenzione per i nomi dei prodotti Sentinel

Le acquisizioni sentinel presentano granularità differenti a seconda del prodotto considerato. Le acquisizioni di inquinanti come NO_2, SO_2 e HCHO possono essere rappresentate con delle patch di grandezza 3,5x7 Km^2; altri inquinanti come CO e O_3 con patch di grandezza 7x7 Km^2.

DarkSky, il servizio di acquisizione dati meteorologici, fornisce dati con granularità differenti a seconda della feature considerata. Per gli attributi: *Temperatura, Pressione* e *Umidità* la granularità fornita è di 1/120 di grado , circa 0.75 Km; per tutti gli altri 1/4 di grado , circa 20 Km.
Poiché il servizio gratuito fornisce un numero di richieste limitate è stato effettuato uno studio sulla variabilità dei valori, in maniera tale da effettuare un numero di richieste limitate ma che allo stesso tempo fornissero una copertura completa, e soprattutto, minimizzando la perdita di informazioni.

Per fare ciò è stata considerata l'area di Milano e un periodo temporale ristretto, sette giorni (20/6/2019 - 26/6/2019). Utilizzando questi estremi temporali e spaziali sono state effettuate le richieste al servizio DarkSky con granularità più piccola possibile.
Dopo un analisi di dominio delle singole features e uno studio di correlazione tra di esse, è stato effettuato un processo di aggregazione spaziale delle features e un conseguente studio di grandezze statistiche, di cui, di seguito, vengono riportati maggiori dettagli.
L'utilizzo di uno step di 1/120 di grado ha permesso di creare, sull'area di Milano, una griglia con 725 punti di acquisizione.
I dati sono stati divisi, prima di tutto, in 256 partizioni , poi in: 64, 16, 4 e 1.
Di ogni partizione creata è stata calcolata:

- *Deviazione standard* La deviazione standard è un indice di dispersione statistico di una distribuzione di valori

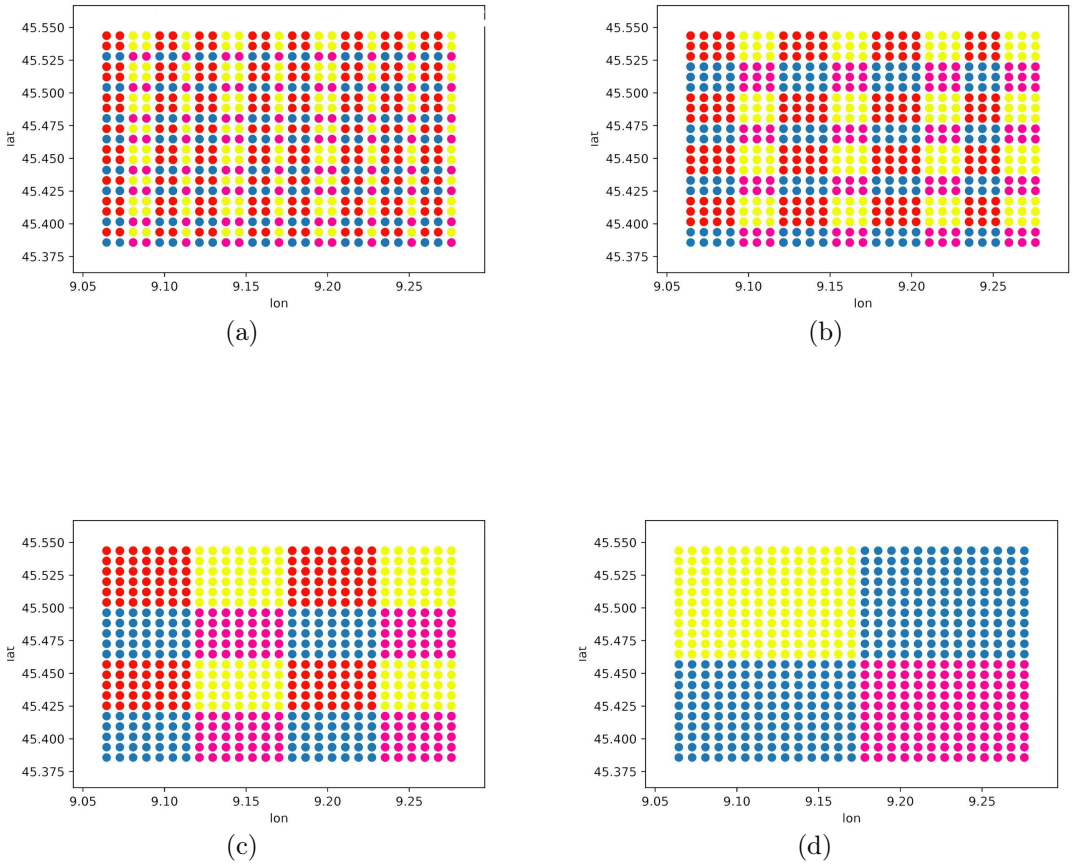

Figura 3.2: Divisione in (a) 256 partizioni, (b) 64 partizioni, (c) 16 partizioni, (d) partizioni.

$$\sigma = \sqrt[2]{\left(\frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}\right)}$$

- *Variabilità*
- *Errore standard della media*

Alla conclusione di questo studio è stato notato che, con una suddivisione in 64 partizioni, all'interno della singola partizione, i dati forniti erano molto simili tra di loro. Per questo motivo, sull'aria di Milano, è stata creata una griglia di dimensioni 8x8 , utilizzata per le acquisizioni del servizio DarkSky. Di seguito vengono riportate le illustrazioni delle prime partizioni create, in maniera tale da rendere più chiaro il criterio utilizzato, e della griglia finale.

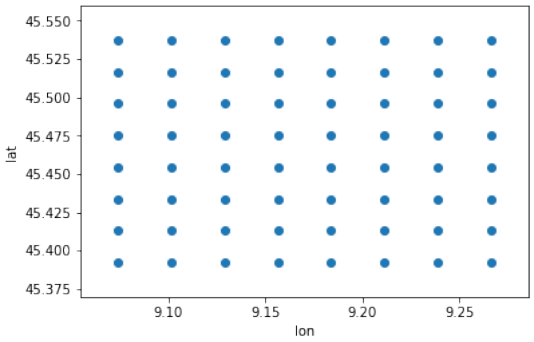

Figura 3.3: Griglia di acquisizione del servizio DarkSky su Milano

3.2 Librerie e Frameworks

3.2.1 Python

Python [30] è un linguaggio di programmazione ad alto livello. Python fu ideato da Guido van Rossum nei primi anni '90 e poi distribuito in maniera open-source, approvata dalla OSI. Diventa uno dei linguaggi più diffusi grazie alla semplicità e la flessibilità che lo caratterizzano. Supporta sia la programmazione a oggetti che la programmazione funzionale. Alcune delle caratteristiche principali sono:

- variabili senza tipo
- indentazione per l'uso di specifiche
- overloading di operatori e funzioni
- innumerevoli librerie standard
- sintassi avanzata.

Il suo utilizzo è largamente diffuso in ambito: Data Analysis,Web Development, Machine Learning, Software Testing ecc.

3.2.2 scikit-learn

scikit-learn [32] è una libreria open-source per il linguaggio di programmazione Python. Permette l'implemetazione di numerosi algoritmi di Machine Learning: Decision Tree, Random Forest, Support Vector Machines ecc. Fornisce metodi per l'implemetazione di: Classificazione, Regressione, Clusteering, riduzione della dimensionalità ecc.
Coopera, principalmente, con altre due librerie numeriche : NumPy e Scipy.

3.2.3 Tensorflow e Keras

Tensorflow [33] è una piattaforma open-source utile per la computazione numerica con alte performance. Tensorflow è stata sviluppata da ricercatori del team Google Brain.

Può funzionare su numerosi tipi di CPU e GPU, grazie al supporto di linguaggi come CUDA e OpenCI.
Inoltre, Google ha sviluppato il processore ASIC dedicato a questo linguaggio, chiamato Tensor Processing Unit (TPU).
È una libreria molto complessa da utilizzare, per questa ragione sono stati sviluppati dei framework per il suo utilizzo in maniera più semplice ed efficiente.
Keras [34] è un API scritta in Python in grado di funzionare su TensorFlow, CNTK e Theano. Essa consente la prototipazione semplice e veloce, supporta reti convoluzionali e ricorrenti, funziona su CPU e GPU.

3.2.4 Neupy

NeuPy [35] è una libreria Python per reti neurali e Deep Learning. Neupy utilizza Tensorflow come backend computazionale per la costruzione di modelli ad apprendimento profondo.

3.2.5 Sentinel Sat

Sentinel Sat [36] è una libreria Python che fornisce API per l'accesso a Copernicus Open Access Hub. Essa permette: l'accesso ai metadati del prodotti Sentinel-1, Sentinel-2, Sentinel-3 e Sentinel-5P; metodi per il download degli stessi specificando i filtri di ricerca in una query.

Capitolo 4

Metodologia

Questo capitolo descrive lo sviluppo del progetto di tesi. Nella prima parte verrà definito l'intero processo che , partendo dal dato grezzo, tramite le fasi di pre e post processing, porta alla creazione dei dataset utilizzati per l'analisi. Nella seconda parte, invece, verrà illustrato come è stato sviluppato ogni singolo modello e come sono stati modellati i dati di input, in maniera tale da ottenere le prestazioni migliori.

4.1 Definizione del problema

L'obiettivo di questo lavoro è quello di analizzare la qualità dell'aria, o meglio studiare la presenza degli inquinanti nell'atmosfera, e creare un modello in grado di predire, in maniera esaustiva, le concentrazioni degli inquinanti. L'intero progetto nasce dalla curiosità di capire quali sono le possibili potenzialità della nuova missione Sentinel-5p, la quale permette di superare il limite spaziale, imposto dall'installazione di sensori a terra, fornendo una copertura globale.

4.2 Data Processing

4.2.1 Pre-processing

Data cleaning Primo processo fondamentale per la creazione di un set di dati affidabile è quello definito *Data Cleaning*. In questa fase dell'analisi si procede con una sorta di filtraggio delle features, in maniera tale da mantenere solamente quelle strettamente necessarie per l'analisi che si vuole effettuare.

Dati Sentinel. Di ogni singolo prodotto Sentinel è stato ritenuto necessario salvare le seguenti informazione:

- latitudine e longitudine di riferimento della patch, punto centrale;
- latitudini e longitudini della bounding box che definisce la patch;
- valore dell'inquinante;
- indicatore di qualità del prodotto.

Un'ulteriore fase di filtraggio è stata applicata ai dati ottenuti dal satellite Sentinel-5P , grazie all'utilizzo di un indicatore fornito con gli stessi, che ne attesta precisione e affidabilità. Questo valore di affidabilità, rappresentato all'interno dei metadati con l'identificativo *qa_ value*, è rappresentato tramite un numero compreso nel range [0, 1], dove il valore uno rappresenta una precisione del dato associato pari al 100% e il valore zero indica la mancanza del dato. Come consigliato dal manuale [39], sono stati considerati solamente i dati con qa_value > 0.5.

Dati meteorologici. Sono state prese in considerazione le seguenti features: *cloudCover*, *dewPoint*, *humidity*, *latitude*, *longitude*, *ozone*, *precipIntensity*, precipProbability, pressure, *temperature*, *time*, *uvIndex*, *visibility*, *windBearing*, *windGust*, *windSpeed*. Alcune di queste features sono state modificate in maniera tale da essere facilmente manipolate. L'attributo *'Time'*, rappresentato tramite timestamp, è stato convertito ottenendo le informazioni : *'Data'* nel formato 'gg/mm/aaaa' e *Hour* nel formato 'hh'; l'attributo *'Temperature'* è stato convertito in gradi *Celsius*.

Interpolazione dati Sentinel Nonostante il servizio garantisca una copertura giornaliera globale, i dati Sentinel presentano delle mancanze. Talvolta sono presenti a dati mancanti o a dati di pessima qualità che non possono essere presi in considerazione. Conseguentemente, le serie temporali che rappresentano l'andamento dei singoli inquinanti presentano delle interruzioni. Questo si verifica anche per finestre temporali di larghezza superiore al singolo giorno. Al fine di limitare l'incidenza di queste mancanze nell'analisi finale, è stato applicato un metodo di interpolazione. L'interpolazione è un processo che cerca di stimare un valore mancante di una serie tra due valori dati della stessa. I metodi di interpolazione comportano l'inserimento di un errore intrinseco nei dati stimati. Per limitare ciò sono state attuate alcune attenzioni particolari: si è scelto di utilizzare una funzione di interpolazione *cubica* in maniera tale che la funzione interpolante riuscisse a seguire al meglio la curva dei dati originali; inoltre la funzione di interpolazione è stata applicata solamente se la finestra temporale di dati mancanti aveva una durata massima di tre giorni.

Mapping L'aria considerata è stata suddivisa in settori. La grandezza di questi settori è pari alla dimensione della patch sentinel più piccola, pari a: 3,5x3,5 Km^2. È stata utilizzata la formula di *Haversine* [40] per il calcolo delle distanze, poichè tutte le grandezze, nei dati sentinel, vengono fornite in gradi e in questo caso era necessaria una conversione in km.
Al termine di questa elaborazione, ogni città presenta un numero di settori differenti a seconda della grandezza dell'aria considerata, tutte le informazioni vengono riportate nella Tabella 4.1. Poiché la copertura Sentinel non è uniforme, cioè esistono settori che si intersecano con più patch sentinel e settori che non sono intersecati da queste ultime, e questa copertura varia a seconda dell'inquinante considerato, si è agito nel seguente modo:
(i) ai settori che si intersecano con più prodotti sentinel, per un dato inquinante, è stato assegnato il valore medio degli stessi;
(ii) ai settori che non presentano nessun intersezione è stato assegnato, di ogni inquinante, il valore della patch più vicina.
Per quanto riguarda i valori meteo, è stata effettuata una aggregazione giornaliera dei

Città	Numero settori
Atene	2
Barcellona	16
Bruxelles	12
Ginevra	2
Milano	25
Norimberga	28
Parigi	12
Roma	36
Torino	12
Zurigo	12

Tabella 4.1: Suddivisione delle città di analisi in settori

dati appartenenti a ogni settore. Di ogni feature è stata calcolato: valore massimo, valore minimo e media.

Correlation matrix L'analisi di correlazione è un metodo per valutare statisticamente la relazione reciproca tra due variabili. Per il calcolo della correlazione viene utilizzato un indicatore. In statistica esisto vari coefficienti, in questa analisi è stato utilizzato il *Coefficiente di Pearson.*

Il coefficiente di correlazione di Pearson è una misura che rappresenta la correlazione tra due variabili X e Y. Esso può assumere un valore compreso tra -1 e 1, dove: 1 indica una correlazione lineare positiva; -1 una correlazione lineare negativa; 0 mancanza di correlazione.

$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_x \sigma_Y}$

Dove:

- $\boldsymbol{cov(X,Y)}$ è la covarianza
- $\boldsymbol{\sigma_X}$ è la deviazione standard di X
- $\boldsymbol{\sigma_Y}$ è la deviazione standard di Y

È stata creata una matrice di correlazione. Essa è una matrice simmetrica che presenta su entrambi gli assi le features di cui si vuole studiare la correlazione. Ogni cella(i,j) rappresenta l'indice di correlazione tra l'attributo i-esimo e j-esimo. Da notare che se i e j coincidono allora l'indice di correlazione è pari a 1.
Particolare importanza è stata data allo studio delle correlazione di ogni singolo inquinante con gli altri inquinanti e le caratteristiche meteorologiche. Risultati di notevole interesse:

- $\rho = -0.7$ tra temperatura al giorno t e concentrazione di NO_2 dello stesso giorno
- $\rho = 0.55$ tra indicatore di umidità al giorno t e concentrazione di NO_2 dello stesso giorno

- $\rho = -0.5$ tra indicatore di umidità al giorno t e concentrazione di SO_2 dello stesso giorno
- $\rho = 0.6$ tra indicatore di presenza di nubi al giorno t e concentrazione di CO dello stesso giorno

Gli altri indicatori riportano un livello di correlazione inferiore, ma comunque degno di nota per essere considerato tra le features dei modelli.

Dalla correlation matrix (Fig.4.1) si può evincere che ogni inquinante è fortemente correlato con gli altri, sia al giorno corrente, sia nei giorni precedenti. Inoltre, si apprezzano elevate autocorrelazioni degli inquinanti, rispetto ai giorni precedenti.

4.2.2 Dataset

Dopo aver applicato la funzione di Mapping, che permette di creare un collegamento geografico tra i dati degli inquinanti e quelli delle condizioni meteorologiche, sono stati creati creati tre tipologie di dataset differenti.

Il dataset ottenuto dalla funzione di mapping, che potremmo considerare come dataset di base per la creazione dei dataset utilizzati nel processo di modellamento, è costituito dalle seguenti informazioni:

- Data: giorno di riferimento t
- Settore
- Inquinanti (t)
- Informazioni meteorologiche (t).

Il dataset contiene informazioni per 287 date differenti e 157 settori, per una shape finale pari a (45059, 59).
Il dataset sopra descritto è stato utilizzato per la creazione dei dataset utili ai modelli spiegati nel capitolo successivo.

Dataset completo La prima tipologia di dataset creato per i modelli presenta la seguente struttura, bisogna considerare che è stato creato un dataset differente per ogni inquinante:

- Data (t)
- Settore
- Inquinanti: SO_2(t), NO_2(t), O_3(t), CO(t), HCHO(t)
- Informazioni meteorologiche al tempo (t)
- Inquinante , di interesse, al tempo (t+1)

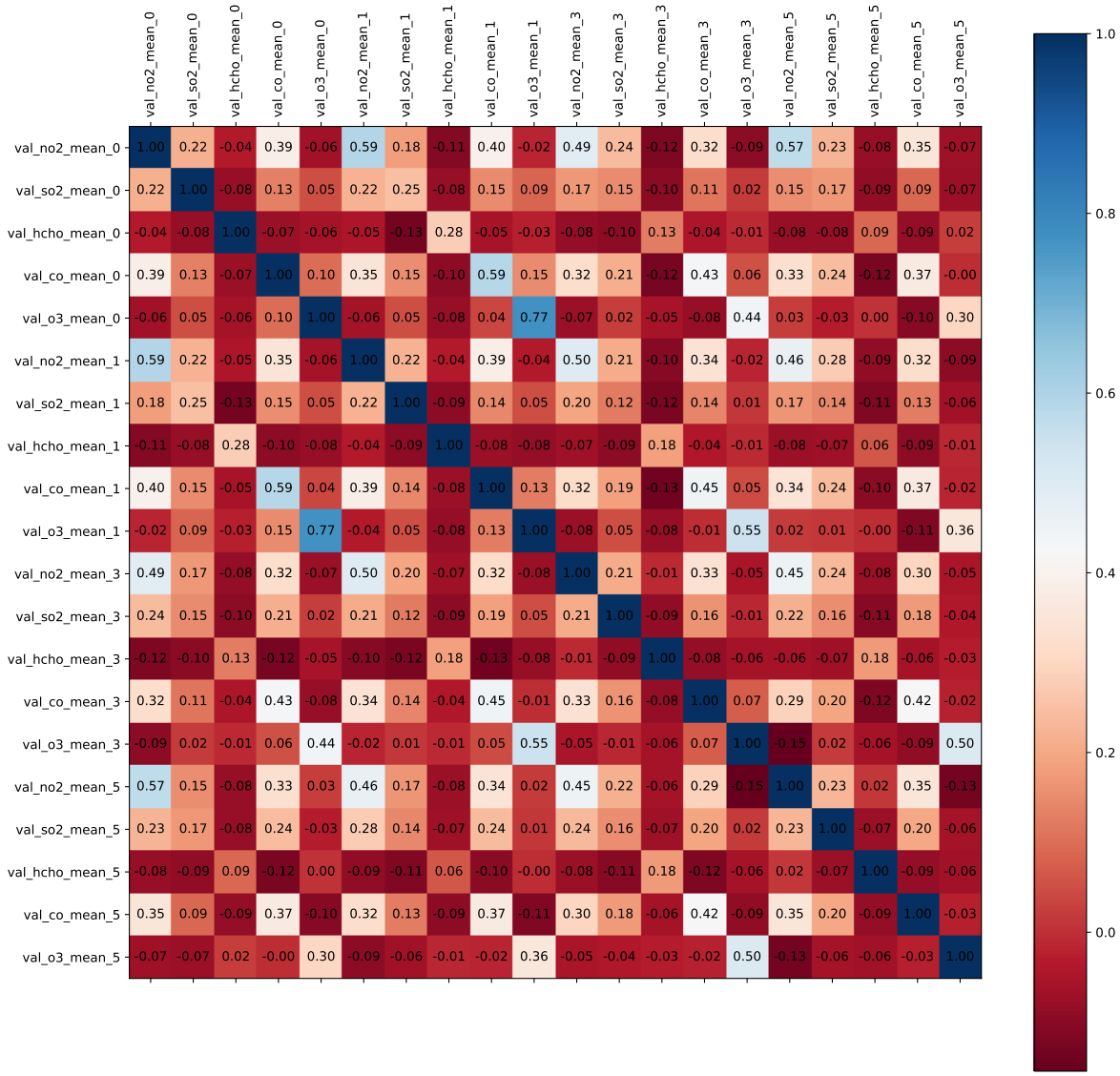

Figura 4.1: Correlation matrix tra inquinanti considerando una finestra temporale di 5 giorni

	O3	CO	NO2	SO2	HCHO
Shape	(22137, 63)	(22061, 63)	(22137, 63)	(21976, 63)	(21945, 63)
Numero giorni	141	137	141	127	132
Numero settori	157	157	157	157	157

Tabella 4.2: Informazioni riguardanti i dataset completi

Dataset con singolo inquinante Il dataset sopra citato è stato modificato in maniera tale da scremare le informazioni sugli inquinanti al tempo t , lasciando solamente quelle riguardanti l'inquinante di interesse. La struttura è la seguente:

- Data (t)
- Settore
- *Inquinante , di interesse, al tempo (t)*
- Informazioni meteorologiche al tempo (t)
- Inquinante, di interesse, al tempo (t+1)

Dataset con finestra temporale I dataset spiegati in precedenza sono stati ulteriormente modificati, ottenendo tre tipologie differenti di dataset:

- **Dataset con delta 1** presenta tutte le informazioni , meteo e inquinanti, del giorno *t* e ha come label l'informazione sull'inquinante scelto al tempo *t+1*
- **Dataset con delta 3** presenta tutte le informazioni , meteo e inquinanti, dei giorni *t, t-1, t-2* e ha come label l'informazione sull'inquinante scelto al tempo *t+1*
- **Dataset con delta 5** presenta tutte le informazioni , meteo e inquinanti, del giorno *t, t-1, t-2, t-3, t-4* e ha come label l'informazione sull'inquinante scelto al tempo *t+1*.

È stata applicata una fase di pulizia dei dataset con finestra in maniera tale che tutti e tre i dataset contenessero le informazioni per lo stesso set di giorni *t*.

	O3	CO	NO2	SO2	HCHO
Shape	(11775, *)	(11763, *)	(11775, *)	(11698, *)	(11710, *)
Numero giorni	75	74	75	69	71
Numero settori	157	157	157	157	157

Tabella 4.3: Informazioni riguardanti i dataset con finestra temporale
(*)Il numero delle features dipende dal delta considerato (delta1:63, delta3:171, delta5:279)

Dataset con finestra temporale con singolo inquinante La struttura dei dataset è simile ai *Dataset con finestra temporale* con la differenza che anziché avere informazioni su tutti gli inquinanti, vengono prese in considerazione solamente le informazioni riguardanti l'inquinante di interesse.

4.3 Preparazione dati per i modelli

Partendo dai dataset costruiti, descritti nella sezione precedente, è stata applicata una fase di pre-processing al fine di modellare i dati di input in maniera ottimale per ogni modello. Qui di seguito ne verranno riportati i dettagli.

4.3.1 Baseline

Prima di attuare una modellizzazione utilizzando le tecniche di Machine Learning e Deep Learning, riportate nei paragrafi successivi, sono state applicate due baseline molto semplici.
La prima baseline attribuisce, ad ogni istanza del dataset, un valore di predizione di un dato inquinante in maniera randomica, calcolando un valore randomico compreso tra -*std* e +*std*. Dove *std* rappresenta la deviazione standard delle *labels*.
La seconda baseline attribuisce, ad ogni istanza del dataset, come valore di predizione, il valore dell'inquinante al giorno precedente.

4.3.2 Hybridization algorithm

Nel capitolo 2.1 è presente una breve introduzione del lavoro svolto dalla *Taiwan Association for Aerosol Research* [37], associazione senza scopo di lucro fondata nel 1992 e che ha come obiettivo quello di unire scienza e ingegneria e promuovere la ricerca e lo sviluppo in ambito tecnologico.
Poiché di questo lavoro era disponibile solamente la descrizione ma non il codice, quest'ultimo è stato implementato al fine di riadattarlo ai dati disponibili e poter comparare le prestazioni con i modelli successivamente implementati durante il percorso di tesi.
Questo lavoro presenta l'applicazione di tre modelli differenti, che di seguito verranno nominati come: fMLP, hMLP, kMLP.

fMLP é il primo modello, il quale utilizza l'intero dataset. Per questo algoritmo è stato utilizzato il Dataset completo , descritto nel paragrafo 4.2.2. La fase di pre-processing di questo algoritmo verrà applicata anche nei modelli successivi. Essa è costituita da tre step:

1. *Stationarisation*: alla serie temporale, vengono sottratte la media annuale e quella mensile.

2. Normalization: tutti le features vengono riportate in un range di valori [-1, 1], in maniera tale che nessuna prevalga sulle altre.

3. *PCA*: questa tecnica permette di incrementare la precisione del modello predittivo [38] e ridurre la dimensionalità del dataset. In questa fase è stata applicata una riduzione delle componenti, tale da garantire l'85% di varianza cumulativa.

hMLP prevede l'utilizzo della tecnica di clustering gerarchico. Infatti, il dataset di partenza è stato prima normalizzato e successivamente sottoposto a clustering gerarchico con numero di cluster N variabile da 2 a 5. Seguendo l'etichetatura di cluster il dataset è stato diviso in N sotto dataset, ai quali è stata applicata la fase di preprocessing descritta per il modello fMLP.

kMLP prevede l'applicazione della tecnica di clustering k-means. Il dataset di partenza è stato prima elaborato applicando la normalizzazione e la PCA e successivamente sottoposto a clustering, anche in questo caso, con N, numero di cluster, definito tra 2 e 5. Come nel caso precedente, il dataset è stato suddiviso in sotto dataset ai quali è stata applicata la fase di preprocessing definita nel modello fMLP.

In tutti i modelli, i dati sono stati ordinati per data e l'intero dataset è stato diviso in cinque parti, ordinate temporalmente, così organizzate: tre parti per il train set, una parte per il validation set e una per il test set.

4.3.3 Random Forest e Multilayer Perceptron

I due modelli, Random Forest e Multilayar Perseptron, sono stati applicati a dataset differenti creando diverse casistiche di analisi. Come descritto nella sezione 4.2.2, sono state create ben quattro tipologie di dataset , differenziati tra loro dalle informazioni sugli inquinanti e/o dalla finestra temporale considerata. L'analisi effettuata su i differenti dataset può mettere in luce possibili correlazioni tra : informazioni correnti e informazioni che fanno riferimento a un tempo passato oppure correlazioni presenti tra inquinanti differenti, anche in tempi diversi.

L'intero dataset è stato ordinato per data, imponendo una divisione temporale al momento della separazione in training, validation e test set.

Ognuno di questi dataset, prima di essere dato in input al modello scelto, è stato normalizzato. La normalizzazione è una trasformazione del dataset che permette di riproporzionare il dominio di ogni features in un dominio comune. Solitamente tutte le features vengono mappate in un range [0, 1] oppure [-1, 1]. Questa operazione permette di assegnare, durante l'analisi, lo stesso peso ad ogni attributo, evitando che features che presentano valori molto grandi possano influenzare maggiormente il modello rispetto a features con valori minori.
Al fine di evitare overflow di calcolo, prima di applicare il Multilayer Percepron le *labels* sono state moltiplicate per una potenza negativa di dieci e poi riportate al valore iniziale prima della valutazione del modello.

4.3.4 Convolutional Long Short-Term Memory

Per questo algoritmo è stato utilizzato il Dataset completo , descritto nel paragrafo 4.2.2. Il dataset è stato ordinato in maniera tale da avere un ordine spaziale e temporale, infatti

è stato applicato un sort per: *città, zona, tempo*. Successivamente il dataset è stato normalizzato, in maniera tale da riportare tutte le feature nel dominio [0, 1].
Il dataset è stato riorganizzato tramite *reshape* riportando l'intero dataset da dimensione $(\sum_i^N n\_zone\_city_i * timestamp, n\_features)$ a $(\sum_i^N n\_zone\_city_i, timestamp, n\_features)$.

Dove:
N: numero città
n_zone_city$_i$: numero di zone nella città i-esima
timestamp: numero di giorni
n_features: numero di features del dataset.

In questa maniera vengono create serie temporali distinte per ogni zona di ogni città, le quali vengono passate in input alla rete. Inoltre, nella divisione del dataset di partenza in training, validation e test set le serie temporali vengono suddivise in maniera tale da garantire che tutte le serie appartenenti alla stessa città vengano assegnate allo stesso sotto dataset.
Durante la costruzione del modello si è scelto di operare nel seguente modo:
(I) prima è stato creato un modello **LSTM**, costituito da: *LSTM Layer* e *Dense Layer*;
(III) poi sono stati combinati una Convolutional neural network (CNN) con un LSTM ottenendo un modello denominato **convLSTM**, il quale ha la seguente struttura: *Conv1d layer*, *Maxpooling1D layer*, *LSTM layer*, *Dense layer*,*Dropout Layer*, *Dense layer*.

Metodo di arresto Nell'apprendimento automatico viene utilizzata la tecnica *Early stopping* per interrompere l'addestramento del modello e far in modo che esso non si adatti troppo ai dati di training. Il modello viene addestrato utilizzano il training set, e ad ogni epoca viene calcolato l'RMSE usando il validation set. Quando l'RMSE del validation set resta pressoché costante, cioè con una variazione inferiore o uguale a $\varepsilon = 0.01$, per cinque iterazioni consecutive, l'apprendimendo viene interrotto. In alcuni casi potrebbe anche non verificarsi questa situazione, poichè l'RMSE potrebbe oscillare con variazioni maggiori di ε, in tal caso è stata prefissata una soglia massima di numero di epoche pari a 200.

4.4 Grid Search

Gli **iperparametri** sono dei parametri caratteristici del modello i quali non possono essere stimati dai dati, poiché essi devono essere impostati prima del processo di apprendimento. Il *GridSearch* è una tecnica utile per trovare gli iperparametri ottimali, affinché il modello abbia delle prestazioni migliori. Dopo aver specificato , per ogni iperparametro, un insieme finito di valori, il GridSearch si occupa di testare tutte le combinazioni di questi valori su un set di addestramento. Ogni modello costruito viene valutato in base a una funzione di perdita specificata. Al termine della ricerca viene riportato un vettore contenente la combinazione di iperparametri che ha ottenuto la valutazione migliore.

4.5 Validazione modelli

Per la validazione dei modelli è stata utilizzata la tecnica del **Cross-Validation**. Essa è una tecnica statistica che prevede la suddivisione del dataset totale in K parti, di uguali dimensioni. In maniera ciclica, uno dei K subset viene utilizzato come validation set e i restanti *K-1* subset vengono utilizzati per l'addestramento del modello.
Il modello viene allenato K volte, in questo modo si supera il problema dell'overfitting. (Figura 4.2)

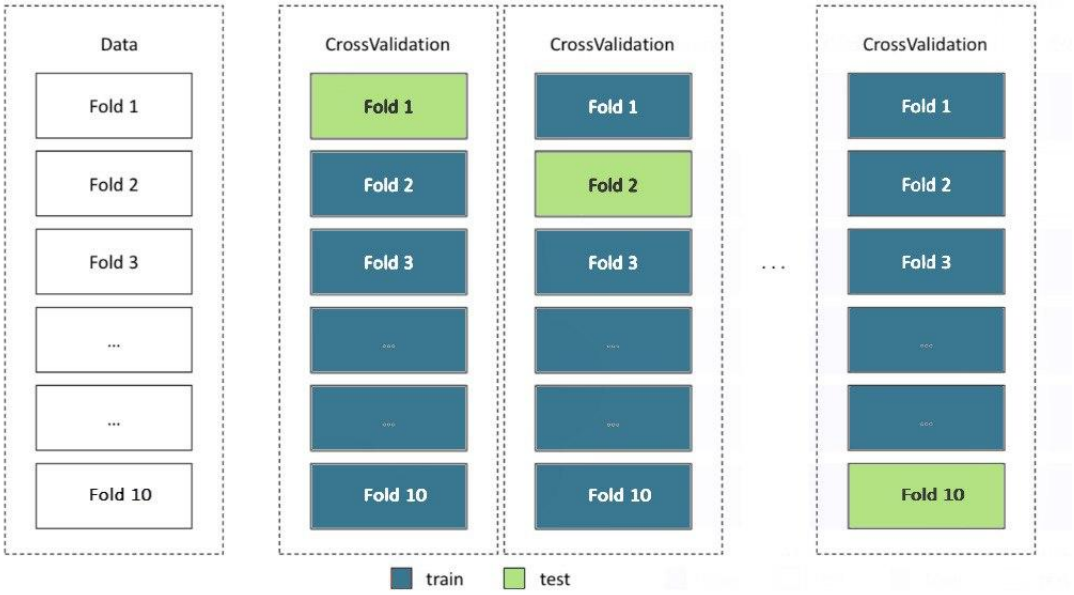

Figura 4.2: Cross Validation

I dati contenuti nel dataset utilizzato per lo svolgimento della tesi presentano una distribuzione spaziale, oltre che temporale. Per questo motivo non è stato possibile utilizzare tecniche di cross validation standard ma è stato implementato un algoritmo apposito. L'algoritmo sviluppato effettua la divisione del dataset di partenza con numero di fold K pari a 10. Inoltre, occorre verificare che durante lo split Train-Validation-Test non ci sia sovrapposizione temporale dei dataset. Poichè ogni istanza del dataset di partenza va a considerare una finestra temporale di durata 1, 3, 5 giorni, per assicurarsi che non si crei sovrapposizione, di cui sopra, tra i tre sottodataset viene lasciato un *gap* di durata pari a quella della finestra temporale (Figura 4.3).
Il metodo precedente è stato utilizzato nel caso in cui il dataset in analisi venisse suddiviso

in Train, Validation, Test set utilizzando una divisione temporale.

Figura 4.3: Divisione applicata al dataset con finestra temporale

Nei casi in cui è stata applicata una divisione spaziale il metodo di validazione è stato applicato nel seguente modo: ogni città (10) rappresenta un fold; ad ogni iterazione è stata utilizzata una città per il test set, una per il validation set e le restanti per il training set.

4.6 Metriche di valutazione delle performance

I modelli, sopra citati, sono tutti modelli predittivi, cioè hanno come obiettivo quello di assegnare, ad ogni dato di input del modello, un valore continuo di output.

4.6.1 RMSE

Mean Squared Error (MSE) , in statistica, indica la differenza quadratica media tra i valori attesi e i valori stimati. Esso viene calcolato come la somma della varianza e del quadrato del bias dello stimatore.
$MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2]$
$MSE(\hat{\theta}) = Var(\hat{\theta}) + (Bias(\hat{\theta}, \theta))^2$
È possibile rappresentare l'indicatore anche nel seguente modo:

$MSE(\hat{\theta}) = \frac{\sum_i^N (x_i - \hat{x}_i)}{n}$

Root Mean Squared Error (RMSE) è la deviazione standard dei residui. I residui sono le distanze calcolate tra il valore predetto e il valore reale.
$RMSE = MSE^2$
RMSE è definito in un range [0, inf], più piccolo è in valore migliore sarà il risultato.

4.6.2 MAE

Mean Absolute Error (MAE) misura la media degli errori in una serie di previsioni. viene calcolato come la media dei valori assoluti delle differenze tra valore previsto e valore reale.
$MAE = \frac{\sum_i^N (|y_i - \hat{y}_i|)}{n}$
MAE è definito in un range [0, inf], più piccolo è in valore migliore sarà il risultato.

4.6.3 R2

R-squared (R^2) è una metrica di valutazione dei modelli di regressione. Essa assume valori nel range [0, 100], dove 0 rappresenta un modello non in grado di seguire le variazioni della variabile dipendente, al contrario, il valore 100 rappresenta un modello in grado di seguire tutte le variazioni.

Capitolo 5

Esperimenti

In questo capitolo viene descritta l'implementazione dei modelli e vengono analizzati i risultati ottenuti al fine di valutarne le performance e stabilire il modello che meglio riesce a predire le concentrazioni degli inquinanti d'analisi nell'atmosfera.
Il capitolo seguente è strutturato in tre parti: (i) la prima parte riporta una panoramica sulle baseline applicate e le loro prestazioni; (ii) nella seconda viene effettuata un'analisi predittiva utilizzando le informazioni del giorno corrente per ottenere quelle del giorno successivo; (iii) nella terza parte viene effettuata l'analisi prendendo in considerazione le informazioni di una finestra temporale, di durata variabile, per predire la concentrazione degli inquinanti al giorno successivo. Inoltre, in ogni test case vengono presentati due casi di analisi differenti: nel primo, si considera il dataset completo con le informazioni su tutti gli inquinanti; nel secondo caso si prendono in considerazione solamente le informazioni dell'inquinante esaminato.
Ogni caso di studio presenta i risultati ottenuti da ogni modello e un paragrafo conclusivo in cui essi vengono comparati.

Per far si che i risultati, riportati in questo capitolo, siano più facilmente comprensibili vengono esplicitati i domini dei singoli inquinanti:

- **O_3** min: 6,73E+18 max: 1,28E+19 std: 6,10E+17
- **CO** min: 1,11E+18 max: 3,02E+18 std: 2,24E+17
- **NO_2** min: 6,17E+13 max: 5,72E+16 std: 4,33E+15
- **SO_2** min: 3,08E+12 max: 5,77E+17 std: 4,12E+16
- **HCHO** min: 5,13E+12 max: 7,86E+16 std: 6,95E+15

5.1 Baseline

Prima di passare ai risultati ottenuti dai modelli sviluppati, di seguito (Tabelle: 5.1, 5.2, 5.3) sono riportati i risultati ottenuti dall'applicazione delle baseline. Si noti che dell'algoritmo Hybridization algorithm, per ogni inquinante, viene riportato solamente il caso migliore e che nelle tabelle, per ogni inquinante viene evidenziato, in grassetto, il caso migliore.

inquinante	**RMSE**(E+17)	**MAE**(E+17)	**R2**
O_3	9,57	7,67	-0,46
NO_2	0,05	0,04	-0,39
CO	2,70	2,17	-0,45
SO_2	0,48	0,33	-0,36
HCHO	0,08	0,06	-0,35

Tabella 5.1: Metriche di valutazione della Baseline randomica

inquinante	**RMSE**(E+17)	**MAE**(E+17)	**R2**
O_3	7,77	5,23	0,36
NO_2	0,05	0,03	0,13
CO	1,82	1,38	0,38
SO_2	0,71	0,39	-0,25
HCHO	0,08	0,06	-0,44

Tabella 5.2: Metriche di valutazione della Baseline Day Before

inquinante	**Id**	**RMSE**(E+17)	**MAE**(E+17)	**R2**
O_3	k2	13,10	10,30	0,014
NO_2	h4	0,08	0,05	0,005
CO	h3	3,76	2,80	0,002
SO_2	k4	0,32	0,20	0,253
HCHO	h5	0,04	0,03	0,274

Tabella 5.3: Metriche di valutazione della Baseline 'Hybridization algorithm'

Dall'osservazione delle tre baseline (Tabelle: 5.1, 5.2, 5.3) è possibile già distinguere un comportamento differente dei vari inquinanti. Infatti è possibile osservare che O3, NO2 e CO le prestazioni miglior sono ottenute utilizzando la baseline *DayBefore* il che potrebbe sottolineare una bassa variabilità temporale e un alta correlazione tra giorno precedente e attuale. Mentre, SO2 e HCHO ottengono prestazioni migliori con il modello Hybridization algorithm.

5.2 Test case: predizione inquinanti dal giorno corrente

5.2.1 Caso: tutti gli inquinanti

Random Forest

I parametri utilizzati per il processo di training sono riportati nella Tabella A.1

Inquinante	**RMSE**(E+17)	**MAE**(E+17)	**R2**
O_3	5,28	3,82	0,56
NO_2	0,04	0,02	0,31
CO	1,68	1,31	0,44
SO_2	0,39	0,24	0,10
HCHO	0,07	0,05	0,12

Tabella 5.4: Metriche di valutazione RandomForest

Multi Layer Perceptron

I parametri utilizzati per il processo di training sono riportati nella Tabella A.2

inquinante	**RMSE**(E+17)	**MAE**(E+17)	**R2**
O_3	4,85	3,45	0,63
NO_2	0,04	0,02	0,36
CO	1,62	1,26	0,47
SO_2	0,39	0,24	0,10
HCHO	0,07	0,05	0,11

Tabella 5.5: Metriche di valutazione MultiLayerPerceptron

LSTM

I parametri utilizzati per il processo di training sono riportati nella Tabella A.3

inquinante	**RMSE**(E+17)	**MAE**(E+17)	**R2**
O_3	4,56	3,21	0,67
NO_2	0,02	0,01	0,81
CO	1,54	1,18	0,52
SO_2	0,37	0,23	0,47
HCHO	0,06	0,05	0,25

Tabella 5.6: Metriche di valutazione LSTM

ConvLSTM

I parametri utilizzati per il processo di training sono riportati nella Tabella A.3

inquinante	**RMSE**(E+17)	**MAE**(E+17)	**R2**
O_3	3,75	2,65	0,78
NO_2	0,02	0,01	0,84
CO	1,58	1,21	0,50
SO_2	0,29	0,18	0,51
HCHO	0,04	0,03	0,66

Tabella 5.7: Metriche di valutazione convLSTM

Riassumendo

Modello	**O_3(E+17)**		**NO_2(E+17)**		**CO(E+17)**		**SO_2(E+17)**		**HCHO(E+17)**	
	Id	**RMSE**	**Id**	**RMSE**	**Id**	**RMSE**	**Id**	**RMSE**	**Id**	**RMSE**
BL Day Before		7,77		0,05		1,82		0,71		0,08
Hybrid Alg	k2	13,10	h4	0,08	h3	3,76	k4	0,32	h5	0,04
RF		5,28		0,04		1,68		0,39		0,07
MLP		4,85		0,04		1,62		0,39		0,07
LSTM		4,56		0,02		**1,54**		0,37		0,06
convLSTM		**3,75**		**0,02**		1,58		**0,29**		**0,04**

Tabella 5.8: Tabella riassuntiva TestCase: predizione inquinanti dal giorno corrente - case: tutti gli inquinanti

Dall'osservazione dei risultati ottenuti da questo primo caso di studio (Tabella 5.8) è possibile notare che, considerando ogni inquinante come casistica separata dalle altre, i primi due modelli, Random Forest e MultiLayer Perceptron, hanno risultati simili, mentre l'LSTM e convLSTM ottengono prestazioni differenti e migliori degli altri. Da questi primi risultati ottenuti è possibile notare un comportamento differente tra gli inquinanti. Mentre l'LSTM riesce a effettuare predizioni migliori, anche se di poco, per il Monossido di Carbonio, convLSTM ottiene prestazioni migliori sugli altri quattro inquinanti. Un'ulteriore differenza è possibile notarla nelle prestazioni ottenute dai modelli RandomForest e MultiLayer Perceptron, i quali riescono a ottenere prestazioni migliori rispetto alla Base Line solamente con tre inquinanti su cinque.

5.2.2 Caso: singolo inquinante

Random Forest

I parametri utilizzati per il processo di training sono riportati nella Tabella A.4

inquinante	RMSE(E+17)	MAE(E+17)	R2
O_3	5,20	3,75	0,57
NO_2	0,04	0,02	0,30
CO	1,66	1,29	0,45
SO_2	0,39	0,24	0,11
HCHO	0,07	0,05	0,12

Tabella 5.9: Metriche di valutazione RandomForest - Case: singolo inquinante

Multi Layer Perceptron

I parametri utilizzati per il processo di training sono riportati nella Tabella A.5

inquinante	RMSE(E+17)	MAE(E+17)	R2
O_3	4,87	3,45	0,62
NO_2	0,04	0,02	0,34
CO	1,63	1,26	0,47
SO_2	0,39	0,24	0,09
HCHO	0,07	0,05	0,09

Tabella 5.10: Metriche di valutazione MultiLayerPerceptron - Case: singolo inquinante

LSTM

I parametri utilizzati per il processo di training sono riportati nella Tabella A.6

inquinante	RMSE(E+17)	MAE(E+17)	R2
O_3	4,47	3,13	0,68
NO_2	0,03	0,02	0,41
CO	1,54	1,18	0,53
SO_2	0,39	0,24	0,10
HCHO	0,06	0,05	0,16

Tabella 5.11: Metriche di valutazione LSTM - Case: singolo inquinante

convLSTM

I parametri utilizzati per il processo di training sono riportati nella Tabella A.6

inquinante	**RMSE(E+17)**	**MAE(E+17)**	**R2**
O_3	3,91	2,74	0,76
NO_2	0,02	0,01	0,84
CO	1,29	0,96	0,67
SO_2	0,31	0,19	0,45
HCHO	0,05	0,03	0,54

Tabella 5.12: Metriche di valutazione convLSTM - Case: singolo inquinante

Riassumendo

Modello	**O_3(E+17) RMSE**	**NO_2(E+17) RMSE**	**CO(E+17) RMSE**	**SO_2(E+17) RMSE**	**HCHO(E+17) RMSE**
RF	5,20	0,04	1,66	0,39	0,07
MLP	4,87	0,04	1,63	0,39	0,07
LSTM	4,47	0,03	1,54	0,39	0,06
convLSTM	**3,91**	**0,02**	**1,29**	**0,31**	**0,05**

Tabella 5.13: Tabella riassuntiva TestCase: predizione inquinanti dal giorno corrente - case: singolo inquinante

Come già descritto nel capitolo precedente, questo caso di analisi è molto simile al precedente, la differenza consiste in una modifica del dataset in maniera tale da mantenere solamente le informazioni dell'inquinante di interesse e non quelle di tutti gli inquinanti. Confrontando le prestazioni dei modelli (Tabella 5.13) è possibile notare che in questo caso, qualsiasi sia l'inquinante in esame, le prestazioni migliori vengono ottenute con il modello convLSTM.

Con riferimento ai risultati complessivi mostrati in Tabella 5.8 e 5.13, il modello convLSTM performa complessivamente meglio considerando il contributo anche di altri inquinanti. Tale risultato è supporto dell'analisi dati riportata in Sezione 4.2.1, nella quale si apprezzano sensibili correlazioni tra inquinanti diversi

5.3 Test case: predizione inquinanti da finestra temporale

In questa sezione vengono riportati solo i valori dell'RMSE, metrica principale di analisi, nelle Tabelle contenute nell' Appendice B vengono riportati tutti i dettagli.

5.3.1 Caso: tutti gli inquinanti

Random Forest

I parametri utilizzati per il processo di training sono riportati nella Tabella A.7

Inquinante	Delta1 **RMSE**(E+17)	Delta3 **RMSE**(E+17)	Delta5 **RMSE**(E+17)
O_3	**4,38**	4,77	4,78
NO_2	0,04	**0,03**	0,04
CO	**1,74**	2,04	2,08
SO_2	**0,388**	0,40	0,41
HCHO	**0,07**	0,07	0,07

Tabella 5.14: Metriche di valutazione RandomForest

Multi Layer Perceptron

I parametri utilizzati per il processo di training sono riportati nella Tabella A.8

Inquinante	Delta1 **RMSE**(E+17)	Delta3 **RMSE**(E+17)	Delta5 **RMSE**(E+17)
O_3	**3,96**	4,22	4,92
NO_2	0,03	0,04	**0,03**
CO	**1,77**	2,17	2,27
SO_2	0,51	0,42	**0,41**
HCHO	**0,07**	0,07	0,07

Tabella 5.15: Metriche di valutazione MultiLayerPerceptron

LSTM

I parametri utilizzati per il processo di training sono riportati nella Tabella A.9

Inquinante	Delta1 **RMSE**(E+17)	Delta3 **RMSE**(E+17)	Delta5 **RMSE**(E+17)
O_2	3,88	3,82	**3,70**
NO_2	**0,02**	0,02	0,02
CO	1,52	1,51	**1,46**
SO_2	0,39	**0,36**	0,38
HCHO	**0,06**	0,07	0,07

Tabella 5.16: Metriche di valutazione LSTM

convLSTM

I parametri utilizzati per il processo di training sono riportati nella Tabella A.9

Inquinante	Delta1 **RMSE**(E+17)	Delta3 **RMSE**(E+17)	Delta5 **RMSE**(E+17)
O_2	**2,94**	3,09	3,02
NO_2	0,02	**0,01**	0,01
CO	**1,84**	1,92	1,84
SO_2	**0,28**	0,30	0,33
HCHO	**0,04**	0,04	0,05

Tabella 5.17: Metriche di valutazione convLSTM

Riassumendo

Nella Tabella 5.18, per ogni coppia Modello-Inquinante, viene riportato solamente il risultato migliore rispetto alle tre casistiche: Delta1, Delta3 e Delta5 (Colonna 'Id').

Come specificato nel capitolo precedente, dopo aver analizzato diversi modelli e aver confrontato le prestazioni ottenute, è stato ritenuto interessante capire se ampliare la finestra temporale di analisi portasse delle migliorie.

Dall'osservazione della tabella riassuntiva 5.18 è possibile dedurre che alcuni inquinanti non presentano correlazioni con le concentrazioni degli inquinanti dei giorni precedenti, o almeno queste non sono notevolmente importanti. Infatti, si ottengono prestazioni migliori, principalmente, con modelli che considerano un delta temporale pari a un giorno, anziché un delta più ampio. Un comportamento differente viene presentato dal Biossido di Azoto il quale presenta risultati migliori con finestre temporali più ampie. Inoltre è possibile notare una conferma delle prestazioni dei modelli: nel caso 'tutti gli inquinanti' il convLSTM è il modello che performa meglio tranne per quanto riguarda il Monossido di carbonio che ottiene prestazioni migliori con LSTM.

Modello	O_3(E+17)		NO_2(E+17)		CO(E+17)		SO_2(E+17)		HCHO(E+17)	
	Id	**RMSE**	**Id**	**RMSE**	**Id**	**RMSE**	**Id**	**RMSE**	**Id**	**RMSE**
RF	d1	4,38	d3	0,04	d1	1,74	d1	0,39	d1	0,07
MLP	d1	3,96	d5	0,03	d1	1,77	d5	0,51	d1	0,07
LSTM	d5	3,70	d1	0,02	d5	**1,46**	d3	0,36	d1	0,06
convLSTM	d1	**2,94**	d3	**0,01**	d1	1,84	d1	**0,28**	d1	**0.04**

Tabella 5.18: Tabella riassuntiva TestCase: predizione inquinanti da finestra temporale - case: tutti gli inquinanti

5.3.2 Caso: singolo inquinante

Random Forest

I parametri utilizzati per il processo di training sono riportati nella Tabella A.10

Inquinante	Delta1 **RMSE**(E+17)	Delta3 **RMSE**(E+17)	Delta5 **RMSE**(E+17)
O_3	**4,35**	4,72	4,60
NO_2	0,04	**0,03**	0,03
CO	**1,79**	1,97	2,04
SO_2	**0,38**	0,40	0,40
HCHO	**0,07**	0,07	0,07

Tabella 5.19: Metriche di valutazione RandomForest, case: singolo inquinante, finestra temporale

Multi Layer Perceptron

I parametri utilizzati per il processo di training sono riportati nella Tabella A.11

Inquinante	**Delta1** **RMSE**(E+17)	**Delta3** **RMSE**(E+17)	**Delta5** **RMSE**(E+17)
O_3	4,28	**4,14**	4,68
NO_2	**0,03**	0,04	0,04
CO	**1,79**	1,88	1,90
SO_2	0,43	0,43	**0,41**
HCHO	**0,07**	0,07	0,07

Tabella 5.20: Metriche di valutazione RandomForest, case: singolo inquinante, finestra temporale

LSTM

I parametri utilizzati per il processo di training sono riportati nella Tabella A.12

Inquinante	Delta1 **RMSE**(E+17)	Delta3 **RMSE**(E+17)	Delta5 **RMSE**(E+17)
O2	3,71	3,65	**3,46**
NO_2	**0,02**	0,02	0,03
CO	1,50	**1,49**	1,50
SO_2	0,40	0,37	**0,36**
HCHO	0,07	0,06	**0,06**

Tabella 5.21: Metriche di valutazione LSTM

convLSTM

I parametri utilizzati per il processo di training sono riportati nella Tabella A.12

Inquinante	Delta1 **RMSE**(E+17)	Delta3 **RMSE**(E+17)	Delta5 **RMSE**(E+17)
O_2	3,21	3,12	**3,07**
NO_2	**0,02**	0,02	0,02
CO	1,23	1,33	**1,19**
SO_2	**0,29**	0,30	0,38
HCHO	0,04	0,04	**0,04**

Tabella 5.22: Metriche di valutazione convLSTM

Riassumendo

Nella Tabella 5.23, per ogni coppia Modello-Inquinante, viene riportato solamente il risultato migliore rispetto alle tre casistiche: Delta1, Delta3 e Delta5 (Colonna 'Id').
A differenza del TestCase precedente, è possibile notare un numero maggiore di casistiche che ottengono prestazioni migliori con un delta maggiore di un giorno. Si potrebbe ipotizzare che la mancanza di informazioni sugli altri inquinanti faccia sì che il modello abbia necessità di utilizzare e informazioni dei giorni precedenti per lavorare al meglio.

Modello	O_3(E+17)		NO_2(E+17)		CO(E+17)		SO_2(E+17)		HCHO(E+17)	
	Id	**RMSE**	**Id**	**RMSE**	**Id**	**RMSE**	**Id**	**RMSE**	**Id**	**RMSE**
RF	d1	4,35	d3	0,03	d1	1,79	d1	0,38	d1	0,07
MLP	d3	4,14	d1	0,03	d1	1,79	d5	0,41	d1	0,07
LSTM	d5	3,46	d1	0,02	d3	1,49	d5	0,36	d5	0,06
convLSTM	d5	**3,07**	d1	**0,02**	d5	**1,19**	d1	**0,29**	d5	**0,04**

Tabella 5.23: Tabella riassuntiva TestCase: predizione inquinanti da finestra temporale - case: singolo inquinante

5.4 Risultati

L'intero progetto è stato basato sull'obbiettivo di creare un modello predittivo, tramite tecniche di Machine Learning e Deep Learning. Il lavoro di modellizzazione è partito dall'analisi e dalla conseguente implementazione di un modello ibrido già presente in letteratura e da ulteriori due baseline.
Successivamente sono stati implementati altri modelli (RandomForest, MultiLayer Perceptron, LSTM e convLSTM) al fine di comparare le loro prestazioni e comprendere quale risultasse migliore considerando il set di dati da analizzare. Come mostrato nella tabella 5.8 il modello convLSTM permette di ottenere prestazioni migliori per gli inquinanti: Ozono (O_2), Biossido di Azoto (NO_2), Anidride Solforosa (SO_2) e Formaldeide (HCHO); mentre per quanto riguarda il Monossido di Carbonio (CO) le migliori prestazioni sono ottenute con l'LSTM.

Dopo aver raggiunto l'obbiettivo primario è nata la curiosità di capire se per la predizione dell'inquinante X_i fosse necessario avere informazioni riguardanti gli altri inquinanti $X_{j!=i}$. Dalla Tabella 5.13 è possibile notare come la modifica del dataset abbia modificato l'andamento reciproco dei modelli, ottenendo le migliori prestazioni, comunemente con il modello convLSTM. Paragonando i risultati con quelli ottenuti nella casistica precedente è possibile notare come solamente nel caso del Monossido di carbonio si siano ottenute delle migliorie; le prestazioni restano le medesime nel caso del Biossido di Azoto e peggiorano in tutti gli altri casi.

Un'ulteriore analisi svolta è stata quella di analizzare modelli che considerassero finestre temporali di durata differente, in maniera tale da capire se esiste o meno una correlazione temporale tra i dati analizzati. Anche questo secondo TestCase (sezione 5.3) è stato svolto considerando tutti gli inquinanti e successivamente il singolo inquinante di interesse. Dai risultati ottenuti, dalla casistica *'tutti gli inquinanti'* (Tabella 5.18) è possibile evincere una scarsa correlazione nel tempo delle features analizzate, questo perché, nella maggior parte dei casi, si sono ottenuti risultati migliori con una finestra temporale di un singolo giorno.
Mentre, nella casistica *'singolo inquinante'* (Tabella 5.18), per alcuni inquinanti, è stata

esplicitata la necessità di una finestra temporale più amplia per ottenere prestazioni migliori. Per quanto riguarda la comparazione dei modelli, i risultati sono i medesimi del TestCase precedente: il modello convLSTM ottiene risultati migliori in tutti i casi e con tutti gli inquinanti, tranne nel caso del CO:'tutti gli inquinanti' in cui l'LSTM riesce ad ottenere risultati migliori.

Effettuando una comparazione che considera tutti i casi analiazzati, è possibile concludere che: il modello che permette di effettuare predizioni più accurate è il convLSTM e per quattro inquinanti su cinque, il caso migliore è quello che considera le informazioni di tutti gli inquinati. Fa eccezione il Monossido di carbonio che a seconda del caso di analisi ottiene prestazioni migliori con LSTM o convLSTM, e che considerando la totalità dei casi analizzati, la casistica 'singolo inquinante' fornisce risultati migliori.

Capitolo 6

Conclusioni e sviluppi futuri

In questo capito verranno presentati possibili sviluppi futuri per la manipolazione dei dati di inquinanti al fine di effettuare delle predizioni. Inoltre, verranno presentate anche delle considerazioni sul lavoro svolto nel percorso di tesi.

Questo lavoro è stato concepito per analizzare i dati atmosferici ed ottenere delle predizioni sulla concentrazione degli inquinanti nell'atmosfera. Per fare ciò sono stati utilizzati i dati ottenuti dalla nuova missione Copernicus Sentinel-5p. Inoltre il lavoro è stato svolto progettando ed analizzando vari modelli di machine learning e deep learning e soprattutto applicando configurazioni differenti al database da analizzare, cercando di ottenere da essi le migliori prestazioni.

6.1 Sviluppi futuri

Uno dei possibili sviluppi futuri potrebbe essere quello di aggiungere nuove features che permettano di arricchire il dataset ottenendo un modello che analizzi contemporaneamente diverse informazioni, ottenute anche da fonti di diversa natura, e che permetta di migliorare le prestazioni ottenute durante il processo di predizione. Possibili informazioni, di cui ci si potrebbe servire, potrebbero essere: dati topologici, che permettano di capire come la conformazione territoriale incida sulla propagazione degli inquinanti; dati riguardanti fonti di emissioni degli inquinanti, come ad esempio la presenza nel territorio di zone industriali e la caratterizzazione delle stesse che permetta di capire su quale inquinante incidano le emissioni.
Un ulteriore lavoro futuro, che è possibile sviluppare a partire da questo progetto, è quello di convertire il problema analizzato da problema di regressione a problema di classificazione. In questo caso sarebbe opportuno analizzare l'andamento di ogni singolo inquinante per suddividere il dominio di azione in n classi, tramite la scelta di opportuni valori di soglia.
Le n classi dovrebbero essere comparabili con gli indici di qualità dell'aria, come ad esempio il CiteairII Air Quality Index (CAQI).
Il CAQI è un indice di qualità dell'aria usato in Europa a partire dal 2012, rappresentato tramite un numero da 1 a 100. I valori ad esso associati vengono divisi in sotto insiemi,

ad ognuno di essi è associato un nome qualitativo('molto basso', 'basso', 'medio', 'alto', 'molto alto') che descrive il livello di inquinamento ed anche un colore (verde acceso, verde chiaro, giallo, arancione, rosso) che permette di rendere più veloce ed efficace la trasmissione dell'informazione al pubblico.

La scelta delle soglie potrebbe essere effettuata tramite la comparazione, per un periodo di tempo limitato, dei valori sentinel con valori di sensori a terra. Questo processo permetterebbe di comprendere la correlazione , se esistente, tra la concentrazione degli inquinanti presenti nei primi strati della troposfera ,captati da sensori a terra, i quali incidono sulla salute dell'uomo, e quella dell'intera colonna troposferica e stratosferica.

6.2 Considerazioni finali

L'acquisizione satellitare di dati sull'inquinamento permette di superare il problema dell'installazione locale di sensori, la quale impone grossi limiti fisici dovuti soprattutto alla conformazione del territorio. Inoltre, l'analisi di queste informazioni tramite algoritmi di Machine Learning e Deep Learning fornisce un'importante strumento per lo studio del fenomeno e l'attuazione di possibili politiche che permettano di limitare il problema dell'inquinamento e non solo.
Lo studio effettuato in questa tesi dimostra che la missione Sentinel-5P ha grosse potenzialità poiché fornisce uno sguardo di insieme all'intero globo terrestre. Alcuni degli avvenimenti recenti hanno permesso di mettere in luce questo aspetto come ad esempio: l'incendio sviluppatosi nella foresta amazzonica durante gennaio 2019; gli incendi in Australia durante i mesi scorsi; l'eruzione dell'Etna a Settembre 2019; la diminuzione dell'inquinamento, nell'intera provincia di Hubei , in Cina, e nel nord Italia in seguito alle misure di quarantena per il COVID-19.
Un limite presentato dalla missione Sentinel-5P consiste nel fornire informazioni riguardanti l'intera colonna atmosferica che ingloba sia lo strato troposferico che stratosferico, ciò rende i dati forniti non comparabili con dati ottenuti da sensori a terra.

Appendice A

Parametri dei modelli

A.1 Test case: predizione inquinanti dal giorno corrente

A.1.1 Caso: tutti gli inquinanti

Random Forest

	Valori per singolo modello				
Parametri	**O_3**	**CO**	**NO_2**	**SO_2**	**HCHO**
bootstrap	True	True	False	False	True
min samples leaf	3	2	3	3	1
n estimators	20	20	15	20	15
min samples split	2	2	2	2	2
random state	42	42	42	42	42
max features	sqrt	sqrt	sqrt	sqrt	sqrt
max depth	5	10	10	5	5

Tabella A.1: Parametri RandomForest

Multi Layer Perceptron

	Valori per singolo modello				
Parametri	**O_3**	**CO**	**NO_2**	**SO_2**	**HCHO**
alpha	0.0005	0.0005	5e-05	0.0005	0.0005
activation	identity	logistic	logistic	logistic	relu
solver	lbfgs	lbfgs	adam	sgd	sgd
learning rate	constant	constant	constant	adaptive	adaptive
hidden layer sizes	(50,)	(1,)	(50,)	(50,)	(50,)

Tabella A.2: Parametri MultiLayerPerceptron

LSTM e convLSTM

Parametri	**Valore**
Batch Size	1
Optimizer	adam
Loss function	mse

Tabella A.3: Parametri LSTM e convLSTM

A.1.2 Caso: singolo inquinante

Random Forest

	Valori singolo modello				
Parametri	**O_3**	**CO**	**NO_2**	**SO_2**	**HCHO**
bootstrap	False	True	True	True	True
min samples leaf	1	1	3	1	2
n estimators	15	20	20	10	20
min samples split	2	2	2	5	2
random state	42	42	42	42	42
max features	sqrt	sqrt	sqrt	sqrt	sqrt
max depth	5	10	10	5	10

Tabella A.4: Parametri RandomForest

Multi Layer Perceptron

	Valori singolo modello				
Parametri	**O_3**	**CO**	**NO_2**	**SO_2**	**HCHO**
alpha	5e-05	5e-05	0.0005	0.0005	0.0005
activation	identity	logistic	relu	logistic	logistic
solver	lbfgs	lbfgs	sgd	sgd	adam
learning rate	constant	adaptive	adaptive	adaptive	constant
hidden layer sizes	(50,)	(1,)	(50,)	(50,)	(1,)

Tabella A.5: Parametri MultiLayerPerceptron

LSTM e convLSTM

Parametri	**Valore**
Batch Size	1
Optimizer	adam
Loss function	mse

Tabella A.6: Parametri LSTM e convLSTM

A.2 Test case: predizione inquinanti da finestra temporale

A.2.1 Caso: tutti gli inquinanti

Random Forest

Inquinante	Parametri	Valore Delta1	Valore Delta3	Valore Delta 5
O_3	bootstrap	True	False	False
	min samples leaf	1	1	1
	n estimators	20	15	20
	min samples split	2	5	5
	random state	42	42	42
	max features	sqrt	sqrt	sqrt
	max depth	10	15	15
CO	bootstrap	False	False	True
	min samples leaf	2	2	1
	n estimators	20	20	15
	min samples split	5	5	5
	random state	42	42	42
	max features	sqrt	sqrt	sqrt
	max depth	10	15	10
NO_2	bootstrap	True	True	False
	min samples leaf	2	1	2
	n estimators	20	20	20
	min samples split	2	4	2
	random state	42	42	42
	max features	sqrt	sqrt	sqrt
	max depth	5	10	10
SO_2	bootstrap	False	True	True
	min samples leaf	3	3	2
	n estimators	20	10	20
	min samples split	2	2	5
	random state	42	42	42
	max features	sqrt	sqrt	sqrt
	max depth	3	3	3
HCHO	bootstrap	False	True	False
	min samples leaf	1	1	2
	n estimators	20	20	20
	min samples split	4	5	5
	random state	42	42	42
	max features	sqrt	sqrt	sqrt
	max depth	10	15	10

Tabella A.7: Parametri RandomForest

Multi Layer Perceptron

Inquinante	Parametri	Valore Delta1	Valore Delta3	Valore Delta5
O_3	alpha	5e-05	0.0005	0.0005
	activation	logistic	identity	identity
	solver	lbfgs	lbfgs	lbfgs
	learning rate	adaptive	constant	adaptive
	hidden layer sizes	(20,)	(20,)	(1,)
CO	alpha	0.0005	5e-05	0.0005
	activation	identity	logistic	logistic
	solver	lbfgs	lbfgs	adam
	learning rate	constant	adaptive	constant
	hidden layer sizes	(1,)	(1,)	(20,)
NO_2	alpha	5e-05	0.0005	0.0005
	activation	identity	identity	identity
	solver	adam	sgd	sgd
	learning rate	constant	constant	adaptive
	hidden layer sizes	(1,)	(20,)	(20,)
SO_2	alpha	5e-05	5e-05	0.0005
	activation	logistic	logistic	logistic
	solver	lbfgs	lbfgs	sgd
	learning rate	constant	constant	adaptive
	hidden layer sizes	(1,)	(1,)	(1,)
HCHO	alpha	5e-05	0.0005	0.0005
	activation	logistic	relu	relu
	solver	adam	sgd	sgd
	learning rate	constant	constant	adaptive
	hidden layer sizes	(20,)	(20,)	(20,)

Tabella A.8: Parametri MultiLayerPerceptron

LSTM e convLSTM

Parametri	**Valore**
Batch Size	1
Optimizer	adam
Loss function	mse

Tabella A.9: Parametri LSTM e convLSTM

A.2.2 Caso: singolo inquinante

Random Forest

Inquinante	Parametri	Valore Delta1	Valore Delta3	Valore Delta 5
O_3	bootstrap	False	False	False
	min samples leaf	3	1	1
	n estimators	20 15	20	
	min samples split	2	4	5
	random state	42	42	42
	max features	sqrt	sqrt	sqrt
	max depth	15	15	15
CO	bootstrap	False	True	True
	min samples leaf	1	3	1
	n estimators	20	20	20
	min samples split	5	2	4
	random state	42	42	42
	max features	sqrt	sqrt	sqrt
	max depth	15	15	10
NO_2	bootstrap	False	False	True
	min samples leaf	3	2	1
	n estimators	20	10	20
	min samples split	2	2	2
	random state	42	42	42
	max features	sqrt	sqrt	sqrt
	max depth	10	5	15
SO_2	bootstrap	False	False	True
	min samples leaf	1	3	2
	n estimators	20	10	20
	min samples split	2	2	5
	random state	42	42	42
	max features	sqrt	sqrt	sqrt
	max depth	3	3	3
HCHO	bootstrap	True	False	True
	min samples leaf	3	1	1
	n estimators	20	20	20
	min samples split	2	2	4
	random state	42	42	42
	max features	sqrt	sqrt	sqrt
	max depth	15	10	10

Tabella A.10: Parametri RandomForest, case: singolo inquinante, finestra temporale

Multi Layer Perceptron

Inquinante	Parametri	Valore Delta1	Valore Delta3	Valore Delta5
O_3	alpha	0.0005	5e-05	0.0005
	activation	identity	identity	identity
	solver	lbfgs	lbfgs	lbfgs
	learning rate	adaptive	adaptive	adaptive
	hidden layer sizes	(1,)	(1,)	(1,)
CO	alpha	5e-05	5e-05	0.005
	activation	identity	identity	identity
	solver	lbfgs	lbfgs	lbfgs
	learning rate	adaptive	adaptive	constant
	hidden layer sizes	(1,)	(1,)	(1,)
NO_2	alpha	0.0005	0.0005	0.0005
	activation	logistic	identity	relu
	solver	adam	sgd	sgd
	learning rate	adaptive	adaptive	adaptive
	hidden layer sizes	(20,)	(20,)	(20,)
SO_2	alpha	5e-05	5e-05	5e-05
	activation	logistic	logistic	logistic
	solver	lbfgs	lbfgs	sgd
	learning rate	adaptive	adaptive	constant
	hidden layer sizes	(1,)	(1,)	(1,)
HCHO	alpha	5e-05	5e-05	5e-05
	activation	logistic	relu	logistic
	solver	adam	sgd	sgd
	learning rate	adaptive	constant	adaptive
	hidden layer sizes	(20,)	(20,)	(20,)

Tabella A.11: Parametri MultiLayerPerceptron, case: singolo inquinante, finestra temporale

LSTM e convLSTM

Parametri	**Valore**
Batch Size	1
Optimizer	adam
Loss function	mse

Tabella A.12: Parametri LSTM e convLSTM

Appendice B

Metriche

B.1 Test case: predizione inquinanti da finestra temporale

B.1.1 Caso: tutti gli inquinanti

Random Forest

	delta1			delta3			delta5		
Inquinante	**RMSE**[1]	**MAE**[1]	**R2**	**RMSE**[1]	**MAE**[1]	**R2**	**RMSE**[1]	**MAE**[1]	**R2**
O_3	4,38	3,22	0,48	4,77	3,48	0,39	4,78	3,50	0,39
NO_2	0,04	0,02	0,23	0,03	0,02	0,31	0,04	0,02	0,25
CO	1,74	1,35	0,29	2,04	1,57	0,03	2,08	1,60	-0,01
SO_2	0,39	0,24	0,08	0,40	0,25	0,00	0,41	0,26	-0,05
HCHO	0,07	0,05	0,10	0,07	0,06	0,08	0,07	0,05	0,10

Tabella B.1: Metriche di valutazione RandomForest

Multi Layer Perceptron

	delta1			delta3			delta5		
Inquinante	**RMSE**[1]	**MAE**[1]	**R2**	**RMSE**[1]	**MAE**[1]	**R2**	**RMSE**[1]	**MAE**[1]	**R2**
O_3	3,96	2,86	0,58	4,22	3,05	0,52	4,92	3,56	0,35
NO_2	0,03	0,02	0,29	0,04	0,02	0,22	0,03	0,02	0,29
CO	1,77	1,38	0,27	2,17	1,66	-0,10	2,27	1,75	-0,21
SO_2	0,51	0,30	-0,57	0,43	0,25	-0,13	0,41	0.25	-0,04
HCHO	0,07	0,05	0,08	0,07	0,05	0,07	0,07	0,05	0,03

Tabella B.2: Metriche di valutazione MultiLayerPerceptron

[1]I valori di RMSE e MAE sono divisi per (E+17)

LSTM

	delta1			delta3			delta5		
Inquinante	**RMSE**[1]	**MAE**[1]	**R2**	**RMSE**[1]	**MAE**[1]	**R2**	**RMSE**[1]	**MAE**[1]	**R2**
O_3	3,88	2,75	0,60	3,82	2,63	0,61	3,70	2,64	0,63
NO_2	0,02	0,01	0,81	0,02	0,01	0,74	0,02	0,01	0,70
CO	1,52	1,18	0,46	1,51	1,17	0,47	1,46	1,12	0,50
SO_2	0,39	0,24	0,09	0,36	0,21	0,21	0,38	0,24	0,09
HCHO	0,06	0,05	0,32	0,07	0,05	0,21	0,07	0,05	0,21

Tabella B.3: Metriche di valutazione LSTM

convLSTM

	delta1			delta3			delta5		
Inquinante	**RMSE**[1]	**MAE**[1]	**R2**	**RMSE**[1]	**MAE**[1]	**R2**	**RMSE**[1]	**MAE**[1]	**R2**
O_3	2,94	2,06	0,77	3,09	2,17	0,74	3,02	2,13	0,75
NO_2	0,02	0,01	0,46	0,01	0,01	0,50	0,01	0,01	0,51
CO	1,84	1,15	0,79	1,92	1,18	0,78	1,84	1,22	0,80
SO_2	0,28	0,17	0,50	0,30	0,19	0,44	0,33	0,20	0,35
HCHO	0,04	0,03	0,67	0,04	0,03	0,69	0,05	0,03	0,61

Tabella B.4: Metriche di valutazione convLSTM

B.1.2 Caso: singolo inquinante

Random Forest

	delta1			delta3			delta5		
Inquinante	**RMSE**[1]	**MAE**[1]	**R2**	**RMSE**[1]	**MAE**[1]	**R2**	**RMSE**[1]	**MAE**[1]	**R2**
O_3	4,35	3,18	0,49	4,72	3,47	0,40	4,60	3,39	0,43
NO_2	0,04	0,02	0,21	0,04	0,02	0,31	0,03	0,02	0,30
CO	1,79	1,39	0,25	1,97	1,52	0,09	2,04	1,58	0,02
SO_2	0,38	0,23	0,10	0,40	0,24	0,04	0,40	0,25	0,02
HCHO	0,07	0,05	0,08	0,07	0,05	0,10	0,07	0,05	0,08

Tabella B.5: Metriche di valutazione RandomForest

Multi Layer Perceptron

	delta1			delta3			delta5		
Inquinante	**RMSE**[1]	**MAE**[1]	**R2**	**RMSE**[1]	**MAE**[1]	**R2**	**RMSE**[1]	**MAE**[1]	**R2**
O_3	4,28	3,19	0,51	4,14	3,01	0,54	4,68	3,47	0,41
NO_2	0,03	0,02	0,33	0,04	0,02	0,18	0,04	0,02	0,15
CO	1,79	1,39	0,25	1,88	1,48	0,17	1,90	1,49	0,15
SO_2	0,43	0,26	-0,16	0,43	0,26	-0,14	0,41	0,25	-0,01
HCHO	0,07	0,05	0,08	0,07	0,05	0,06	0,07	0,05	0,08

Tabella B.6: Metriche di valutazione MultiLayerPerceptron

LSTM

	delta1			delta3			delta5		
Inquinante	**RMSE**[1]	**MAE**[1]	**R2**	**RMSE**[1]	**MAE**[1]	**R2**	**RMSE**[1]	**MAE**[1]	**R2**
O_3	3,71	2,56	0,63	3,65	2,61	0,64	3,46	2,43	0,68
NO_2	0,02	0,01	0,75	0,02	0,01	0,79	0,03	0,02	0,60
CO	1,50	1,16	0,47	1,49	1,15	0,48	1,50	1,16	0,47
SO_2	0,40	0,26	-0,01	0,37	0,22	0,16	0,36	0,21	0,22
HCHO	0,07	0,05	0,16	0,06	0,05	0,31	0,06	0,04	0,34

Tabella B.7: Metriche di valutazione LSTM

convLSTM

	delta1			delta3			delta5		
Inquinante	**RMSE**[1]	**MAE**[1]	**R2**	**RMSE**[1]	**MAE**[1]	**R2**	**RMSE**[1]	**MAE**[1]	**R2**
O_3	3,21	2,26	0,72	3,12	2,17	0,74	3,07	2,14	0,75
NO_2	0,02	0,01	0,82	0,02	0,01	0,71	0,02	0,01	0,82
CO	1,23	0,94	0,65	1,33	0,98	0,59	1,19	0,91	0,67
SO_2	0,29	0,18	0,50	0,30	0,19	0,45	0,38	0,23	0,14
HCHO	0,04	0,03	0,64	0,04	0,03	0,64	0,04	0,03	0,66

Tabella B.8: Metriche di valutazione convLSTM

[1]I valori di RMSE e MAE sono divisi per (E+17)

Bibliografia

[1] Treccani-inquinamento
http://www.treccani.it/enciclopedia/inquinamento

[2] World Health Organization
https://www.who.int/news-room/detail/02-05-2018-9-out-of-10-people-worldwide-breathe-polluted-air-but-more-countries-are-taking-action

[3] Agenzia Europea per l'Ambiente
https://www.eea.europa.eu/it/themes/air/intro

[4] ONU, Sustainable Development Goals, Goal 13
https://www.undp.org/content/undp/en/home/sustainable-development-goals/goal-13-climate-action.html

[5] United States Environmental Protection Agency
https://www.epa.gov/

[6] United States Environmental Protection Agency: Criteria Air Pollutants
https://www.epa.gov/criteria-air-pollutants

[7] EEA-Emissions of air pollutants from large combustion plants in Europe.
https://www.eea.europa.eu/data-and-maps/indicators/emissions-of-air-pollutants-from-16/assessmenttab-used-in-publications

[8] Lega Ambiente: Emergenza Smog
https://www.legambiente.it/emergenza-smog-i-nuovi-dati-di-malaria-il-report-di-legambiente-sullinquinamento-atmosferico-in-citta/

[9] Consiglio dell'Unione Europea: il pacchetto "Aria pulita"
https://www.consilium.europa.eu/it/policies/clean-air/

[10] European Space Agency
http://www.esa.int/

[11] National Aeronautics and Space Administration
https://www.nasa.gov/

[12] Ansa.it- Agenzia spaziale Europea
https://www.ansa.it/scienza/notizie/rubriche/eccellenze/esa.html

[13] European Remote Sensing (ERS) mission
https://earth.esa.int/web/guest/missions/esa-operational-eo-missions/ers

[14] Copernicus-Sentinel-5P: Orbita
https://sentinels.copernicus.eu/web/sentinel/missions/sentinel-5p/orbit

[15] Copernicus : s5p instrumental payload
https://sentinels.copernicus.eu/web/sentinel/missions/sentinel-5p/instrumental-payload

[16] ESA: Introduccing Sentinel 5P
https://www.esa.int/Applications/Observing$_t$he$_E$arth/Copernicus/Sentinel − 5P/Introducing$_S$entinel − 5P
[17] K. B. Shaban, A. Kadri, E. Rezk *Urban Air Pollution Monitoring System With Forecasting Models.* (IEEE Sens. J.) [*vol. 16, no. 8, pp. 2598-2606*] (2016).
[18] Peng, H., Lima, A.R., Teakles, A. et al. *Evaluating hourly air quality forecasting in Canada with nonlinear updatable machine learning methods.* [*Air Qual Atmos Health 10, 195–211*] (2017).
https://doi.org/10.1007/s11869-016-0414-3
[19] Wani Tamas, Gilles Notton, Christophe Paoli, Marie-Laure Nivet, Cyril Voyant *Hybridization of air quality forecasting models using machine learning and clustering: an original approach to detect pollutant peaks. Aerosol and Air Quality Research.* [*Aerosol ans Air Quality Research, 16: 405-416, 2016*]
[20] Science Direct: decision tree
https://www.sciencedirect.com/topics/computer-science/decision-trees
[21] Convolutional Neural Network
https://it.mathworks.com/solutions/deep-learning/convolutional-neural-network.html
[22] Copernicus Open Access Hub
https://scihub.copernicus.eu/
[23] google earth engine.
https://earthengine.google.com/
[24] Google earth engine code editor.
https://explorer.earthengine.google.com/workspace
[25] API Dark Sky
https://darksky.net/dev
[26] API Dark Sky: sources
https://darksky.net/dev/docs/sources
[27] Visan documentation
https://stcorp.github.io/visan/
[28] Panoply documentation
https://www.giss.nasa.gov/tools/panoply/
[29] 2005 H.K. Elminir *Dependence of urban air pollutants on meteorology* [*Sci. Total Environ., 350 (2005), pp. 225-237*]
[30] python
https://www.python.it/
[31] scikit-learn
https://scikit-learn.org/stable/
[32] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *"Scikit-learn: Machine learning in python"*. [*vol. 12, no. Oct, pp. 2825–2830, 2011*].
[33] Tensorflow
https://www.tensorflow.org/
[34] Keras
https://keras.io/
[35] NeuPy
http://neupy.com/pages/home.html

[36] Sentinel sat.
`http://sentinelsat.readthedocs.io/en/stable/api.html`
[37] Taiwan Association for Aerosol Research
`http://www.taar.org.tw/about/detail/4`
[38] S.I.V., Sousa, F.G., Martins, M.C.M., Alvim-Ferraz, M.C., Pereira *Multiple linear regression and artificial neural networks based on principal components to predict ozone concentrations* (2007) [*Environ. Modell. Software 22:97-103*]
[39] Sentinel-5P: Products and Algorithms https://sentinels.copernicus.eu/web/sentinel/technical-guides/sentinel-5p/products-algorithms
[40] Haversine formula https://www.geeksforgeeks.org/haversine-formula-to-find-distance-between-two-points-on-a-sphere/
[41] Sentinel hub.
`https://www.sentinel-hub.com/`