ALMA MATER STUDIORUM – UNIVERSITÀ DI BOLOGNA
CAMPUS DI CESENA

Scuola di Ingegneria e Architettura
Corso di Laurea in Ingegneria e Scienze Informatiche

Monitoraggio e predizione per la qualità dell'aria all'interno del veicolo

Tesi di laurea in
Tecnologie Web

Relatore
(**Prof.**) **Silvia Mirri**

Candidato
Thomas Baldi

Correlatore
(**Prof.**) **Roberto Girau**

3° Sessione di Laurea
Anno Accademico 2020-2021

Ai miei genitori.

Indice

Elenco delle figure

Capitolo 1

Introduzione

Questo progetto presenta una proposta di sviluppo di un monitoraggio della qualità dell'aria che ne rende possibile la predizione all'interno del veicolo tramite l'utilizzo di algoritmi di machine learning.
L'idea di questo progetto è quella di immergersi all'interno del progetto europeo NextPerception che punta allo sviluppo di un modello di intelligenza distribuita per il monitoraggio umano sicuro ed affidabile da poter utilizzare nel settore sanitario e automobilistico, il progetto è molto ampio, vi partecipano 43 partners di 7 paesi, tra cui l'università di Bologna. L'università di Bologna come partner italiano si occupa dell'analisi di diversi aspetti legati alle caratteristiche e agli elementi ambientali che possono incidere sulla guida. Un concetto sul quale porre attenzione, che è una delle basi di questo percorso, è quello di comprendere se e come l'esposizione a questi agenti inquinanti possa influenzare le performance cognitive, in modo da riuscire a preservare la salute del guidatore ed aumentare la sicurezza stradale. Nello specifico l'elaborato di tesi pone le sue basi sugli argomenti di NextPerception focalizzandosi sull'analisi dei dati ambientali all'interno del veicolo come temperatura, pressione, umidità e soprattutto agenti inquinanti. Gli agenti inquinanti in questione sono descritti in modo esaustivo all'interno della letteratura per la loro importanza e soprattutto per gli effetti nocivi che possono causare, sia a lungo termine che a breve termine. I più importanti agenti inquinanti e sostanze, soprattutto in ambito indoor sono il particolato, CO_2 e TVOC. I dati ambientali registrati all'interno e all'esterno del veicolo hanno reso possibile lo svolgimento del progetto, le registrazioni sono state effettuate da un kit di sensori Canarin. Il progetto Canarin è un tema centrale in quanto è stato creato con lo scopo di mettere a disposizione uno strumento per valutare l'esposizione individuale al particolato (PM) in tempo reale e in relazione alla mobilità, descrivendo un approccio innovativo nella progettazione dell'architettura dei sensori specificatamente finalizzato a migliorare l'interpretazione dell'esposizione personale all'inquinamento atmosferico. La disposizione di questi Canarin

ha permesso di effettuare delle registrazioni sull'inquinamento dell'aria all'interno e all'esterno di una Nissan leaf 40kWh, l'automobile in questione ha effettuato le registrazioni durante i mesi invernali eseguendo un percorso nella zona urbana di Bologna, raccogliendo dati contenenti informazioni su agenti inquinanti e sostanze come particolato(PM_1, $PM_{2.5}$, PM_4, PM_{10}), TVOC, CO_2, H2, etanolo e HCHO. Il controllo della qualità dell'aria è sempre stato un tema di centrale importanza soprattutto in ambiti outdoor, nell'ultimo decennio comunque questa analisi sul monitoraggio degli agenti inquinanti si è spostata molto in ambienti indoor dai quali sono stati sviluppati progetti e pubblicazioni che sono esposti nello stato dell'arte di questa tesi. Il controllo della qualità dell'aria all'interno del veicolo è un lavoro molto specifico in quanto è difficile il monitoraggio di queste particelle per controllare il loro impatto sulla salute. Essendo un lavoro molto specifico, l'importante innovazione che questo progetto punta a raggiungere, è quella di riuscire ad utilizzare gli algoritmi di regressione principalmente sulla base dei dati esterni. I lavori presenti nello stato dell'arte mirano alla predizione della qualità dell'aria interna al veicolo considerando solo i valori interni senza prendere in considerazione l'importante correlazione con l'ambiente esterno. Per predire la concentrazione dei singoli agenti inquinanti nell'ambiente interno sono stati utilizzati degli algoritmi di machine learning, un ramo dell'intelligenza artificiale che permette ad un sistema di imparare dai dati piuttosto che attraverso la programmazione esplicita. L'analisi dello stato dell'arte ci ha permesso di individuare quali sono gli algoritmi più performanti in grado di fare predizioni sulla qualità dell'aria, nello specifico nella predizione delle concentrazioni interne per il singolo inquinante. L'elaborato di tesi infatti punta alla previsione delle concentrazioni dei singoli inquinanti all'interno del veicolo basandosi principalmente sulle registrazioni esterne. Prima di fornire i dati ai vari algoritmi di machine learning verrà effettuata un'attenta fase di analisi dei dati e preprocessing con lo scopo di comprenderne la struttura ed il significato focalizzandosi sulla modellazione e sulla scoperta della conoscenza per scopi predittivi piuttosto che descrittivi. Per capire il funzionamento degli algoritmi e dei risultati che ne derivano l'applicazione dei modelli di regressione verrà effettuata gradualmente, partendo da semplici modelli di regressione lineare per poi arrivare a modelli più complessi come il multilayer perceptron. Nello specifico per ogni singolo agente inquinante saranno testati degli algoritmi di regressione lineare, per poi passare a modelli di regressione non lineare tra cui modelli polinomiali (semplice, con aggiunta della regolarizzazione L1, L2 e una combinazione delle due con un modello ElasticNet), regressione con funzioni kernel per aumentare il grado senza introdurre maggiore complessità, alberi di regressione (DecisionTree, RandomForest e l'algortitmo di gradien boosting XGBoost) e per ultimo verrà proposto il modello multilayer perceptron. I metodi di valutazione che saranno utilizzati sono : MAE(Mean Absolute Error), MSE(Mean Squared Error), errore

relativo, RMSE(Root Mean Square Error) e coefficiente di determinazione R^2. Nei prossimi capitoli saranno spiegati i concetti e gli obbiettivi del progetto, evidenziando l'importanza futura che potrà avere questo progetto.

Struttura della tesi

- Capitolo 2 - Stato dell'arte, questo capitolo è estremamente importante in quanto espone i concetti principali sui quali si basa la tesi, si espone il background del progetto analizzando gli agenti inquinanti, prima di analizzare i lavori correlati al nostro progetto, viene proposta una piccola introduzione sul machine learning e gli algoritmi più utilizzati.

- Capitolo 3 - Progetto, in questo capitolo viene esposto l'ambiente in cui si propone di sviluppare l'elaborato di tesi, vengono spiegati gli obbiettivi e gli scopi dell'università di Bologna all'interno del progetto NextPerception. Successivamente saranno descritti i principi di funzionamento, deployment e gli utilizzi all'interno dei vari progetti del kit di sensori Canarin, dal quale sono stati ricavati i dati per lo svolgimento del progetto. Null'ultima parte infine saranno riproposti gli obbiettivi e la spiegazione dei dati raccolti.

- Capitolo 4 - Implementazione Progetto, il penultimo capitolo si offre di presentare l'implementazione del progetto, precisamente è suddivisa in varie fasi, la fase di preprocessing dei dati si occupa di modificare il dataset in modo da strutturarlo adeguatamente per la realizzazione delle previsioni. La fase di analisi fa un'ispezione generale raccogliendo informazioni sui dati utilizzati. La fase dell'applicazione degli algoritmi di regressione e l'ultima fase mostra le considerazioni sui risultati ottenuti.

- Capitolo 5 - Conclusioni: vengono esposte le conclusioni del progetto, considerazioni e progetti futuri.

Capitolo 2

Stato dell'arte

2.1 Qualità dell'aria

La qualità dell'aria è diventata ormai un tema di quotidiana importanza, oggetto di preoccupazione per i principali enti, istituzioni e governi internazionali [14] .
Data l'importanza dell'argomento, in letteratura sono presenti molti articoli e progetti che spiegano dettagliatamente quali sono i principali parametri per la valutazione della qualità dell'aria [39] [19] .
L'inquinamento dell'aria per definizione è qualunque sostanza che possa danneggiare l'uomo, gli animali o la vegetazione, come noto l'inquinamento dell'aria è causa di gravi malattie anche mortali.
I vari tipi di agenti inquinanti differiscono per composizione chimica, proprietà di reazione, persistenza, abilità di essere trasportate e impatto sull'ambiente [17] .
Si possono suddividere in quattro categorie:

- Inquinanti gassosi (e.g. SO_2, NO_x, CO, ozono, composti organici volatili)
- Inquinanti organici persistenti (e.g. diossine).
- Metalli pesanti.
- Particolato.

Il particolato viene indicato come PM, il termine sta ad indicare una miscela di particelle solide e liquide presenti nell'aria, non può essere percepito ad occhio nudo. Comunemente si suddivide in:

- PM_{10}: il numero 10 sta ad indicare la grandezza del diametro della particella, misurata in micron, in questo caso varia dai 10 micron o meno.
- $PM_{2.5}$: diametro di 2.5 micron o meno.

La provenienza del particolato sottile è multipla, può essere emesso direttamente dalle risorse come cantieri, incendi o fabbriche, mentre la maggior parte delle particelle si formano nell'atmosfera come risultato di complesse reazioni chimiche come anidride solforosa e ossidi di azoto provenienti da centrali elettriche, industrie e automobili.
L'inalazione di queste sostanze può causare seri danni, la grandezza delle particelle è direttamente proporzionale all'incidenza sulla salute, infatti particelle con un diametro inferiore a 2.5 micron sono quelle che possono gravare maggiormente [29].
In passato si è sempre considerato l'anidride carbonica (CO_2) come indicatore della qualità dell'aria interna, nell'ultimo decennio ha perso questa funzione anche perché oggi vi incidono molteplici fattori.
L'uso diffuso di nuovi prodotti e materiali ha comportato un aumento delle concentrazioni di inquinanti indoor, in particolare di composti organici volatili (VOC), che inquinano l'aria interna e influiscono sulla salute umana [22], infatti TVOC è considerato come un importante indicatore per la qualità dell'aria interna(IAQ) [20]. I composti organici volatili (VOC) sono prodotti a base di idrocarburi come prodotti petroliferi e solventi organici che possono essere facilmente vaporizzati nell'aria a causa dell'elevata pressione atmosferica.
I VOC possono essere causa di una vasta gamma di effetti che vanno dal disagio sensoriale fino a gravi alterazioni dello stato di salute, alte concentrazioni negli ambienti interni possono causare effetti su numerosi organi o apparati, in particolare il sistema nervoso centrale. E' stato ipotizzato che l'inquinamento indoor da VOC possa costituire un rischio cancerogeno per i soggetti che trascorrono molto tempo in ambienti confinanti, il che indica l'importanza del monitoraggio [10] .

2.2 Effetti cognitivi dell'inquinamento dell'aria

Sono stati pubblicati diversi studi anche dal World Health Organization(WHO) con delle guide, dettagliatamente delle raccomandazioni giornaliere e annuali sull'esposizione del $PM_{2.5}$ che non deve superare 25 e 50 $\mu g/m^3$ in base al tempo di esposizione per limitare gli effetti negativi [26] .
Un concetto sul quale porre attenzione, che è una delle basi di questo percorso, è quello di comprendere se e come l'esposizione al particolato e ad altri agenti inquinanti possano influenzare negativamente le performance cognitive, poiché in questo momento non c'è un'assoluta certezza. Articoli come "Efects of short-term exposure to particulate matter air pollution on cognitive performance" [38] hanno proprio questo obiettivo, effettuare esperimenti all'interno di spazi chiusi con alta concentrazione di PM e utilizzando tecniche di Mini-mental state examination(MMSE), Stroop color e un test selettivo sull'attenzione, cercare di misurare

l'influenza sulle performance cognitive.
L'esposizione della persona al particolato sottile (PM) è superiore nei trasporti motorizzati come automobili, autobus e treni, ma anche nelle località urbane, proprio perché le emissioni dei veicoli è maggiore. In aggiunta al particolato vengono emessi altri inquinanti come idrocarburi (HC), ossidi di azoto (NOx), monossido di carbonio (CO) e anidride solforosa (SO_2).
I processi cognitivi che possono essere influenzati sono: visione spaziale (la capacità di percepire le informazioni visive nell'ambiente), funzioni esecutive, scioltezza verbale, memoria, attenzione e orientamento.
Questo studio suggerisce importanti implicazioni per la capacità cognitiva, la salute mentale umana e la loro dipendenza sull'esposizione al particolato. Evidenzia che i cittadini di città e paesi più inquinati avranno, in media, una peggiore capacità cognitiva rispetto a quelli che vivono in città e paesi nei quali le emissioni di particolato sono inferiori. Pertanto, la riduzione del particolato non solo si tradurrà in un miglioramento della morbilità umana e nel decremento della mortalità, ma contribuirà anche al miglioramento delle prestazioni cognitive.

2.3 Qualità dell'aria all'interno del veicolo

Nell'ultimo decennio, il controllo della qualità dell'aria ha coinvolto anche gli ambienti interni, l'ampliamento della prospettiva è dato dal cambiamento nello stile di vita come conseguenza dell'incremento dell'urbanizzazione [12] .
Si è dimostrato che l'IAQ(Indoor air quality) nelle aree residenziali o negli edifici è significativamente influenzato da vari fattori [31] : (i) qualità dell'aria esterna, (ii) attività umana negli edifici e (iii) materiali da costruzione, attrezzature e mobili.
È noto che le concentrazioni di contaminanti all'aperto e l'ermeticità degli edifici influenzano l'IAQ, a causa della possibilità di trasporto di contaminanti dall'esterno all'interno.
Queste nozioni fanno veramente comprendere quello che è il focus di questo articolo, la qualità dell'aria all'interno del veicolo.
La qualità dell'aria all'interno del veicolo dipende da molti fattori: apertura dei finestrini, utilizzo dell'aria condizionata, tipo di filtri interni, anni e condizioni della macchina e molti altri [24] .
La quantità di agenti inquinanti all'interno è elevata a causa del ricircolo dell'aria che introduce all'interno del veicolo inquinanti emessi dai veicoli circostanti. Poiché i veicoli non sono costruiti per essere a tenuta d'aria, gli inquinanti entrano attraverso prese d'aria ed altre aperture. La concentrazione di inquinanti all'interno dei veicoli potrebbe essere da nove a dodici volte superiore rispetto all'ambiente esterno [16] .

Il microclima all'interno del veicolo è diventato una fonte significativa di esposizione agli agenti inquinanti quali particolato(PM_s), composti organici volatili(VOC_s), composti organici semi-volatili($SVOC_s$), monossido di carbonio, ossido di azoto.
Modalità di ventilazione e portata d'aria, età, ermeticità del veicolo, materiali interni, numero di passeggeri e il livello di inquinamento ambientale esterno al veicolo svolgono un ruolo importante nel determinare la concentrazione di sostanze inquinanti [44] .

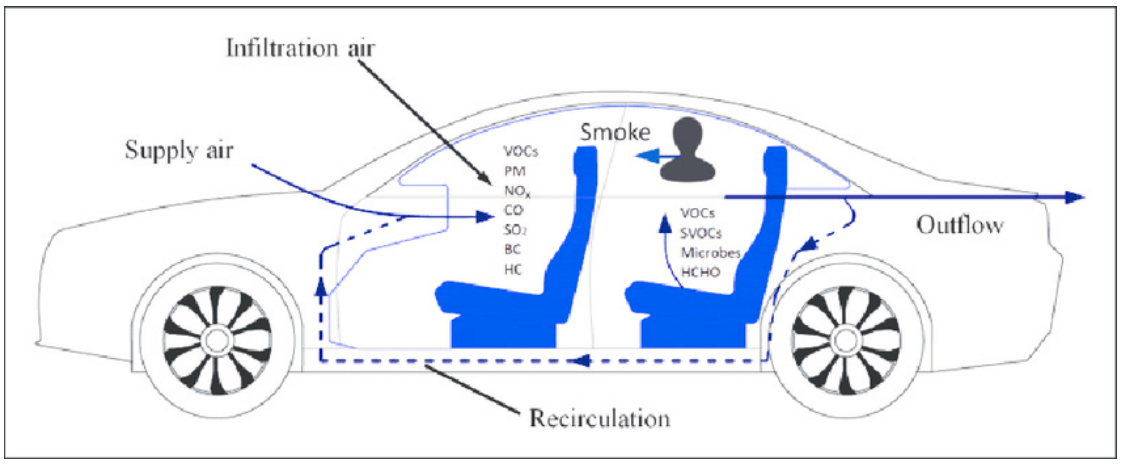


Figura 2.1: Rappresentazione schematica delle fonti di inquinamento all'interno dell'auto e del trasporto [44].

2.4 Indice di qualità dell'aria

Molti progetti e articoli che saranno esposti nella sezione seguente utilizzano indici per rappresentare la qualità dell'aria.
In diversi paesi del mondo, quelli più utilizzati e aggiornati si basano sull'AQI calcolato dall'US EPA(U. S. Environmental Protection Agency), in Europa il progetto CITEAIR mira a supportare le città e le regioni europee nello sviluppo di strumenti e tecnologie per affrontare i problemi di qualità dell'aria. In particolare, viene implementato il Common Air Quality Index (CAQI) con lo scopo di rendere le città comparabili in tutta Europa, questo indicatore, sebbene computazionalmente molto simile all'EPA AQI, non fornisce informazioni dettagliate relative all'incidenza sulla salute [33] .
L'US EPA(U. S. Environmental Protection Agency) ha introdotto il singolo indice inquinante, noto anche come indice di qualità dell'aria (AQI) che funge da indicatore per la misurazione in ambienti mirati. La qualità dell'aria è rappresentata dal valore più alto ottenuto effentuando la misurazione su ogni inquinante presente

[30] .

$$I_p = \frac{I_{Ji} - I_{Lo}}{BP_{Hi} - BP_{Lo}}(C_p - BP_{Lo}) + I_{Lo}$$

I_p = indice per inquinante p
C_p = la concentrazione arrotondata dell'inquinante p
BP_{Hi} = il breakpoint maggiore o uguale a Cp
BP_{Lo} = il breakpoint inferiore o uguale a Cp
I_{Hi} = il valore AQI corrispondente a BPHi
I_{Lo} = il valore AQI corrispondente a BPLo

Di seguito viene mostrata una Tabella che indica i breakPoints per l'AQI.

		BreakPoints					AQI	categoria
O_3 (ppm) 8-hour	O_3(ppm) 1-hour	$PM_{2.5}$	PM_{10}	CO(ppm) 8-hour	SO_2(ppb) 1-hour	NO_2(ppb) 1-hour	AQI	
0.000 - 0.054	-	0.0 -12.0	0-54	0.0-4.4	0-35	0-53	0-50	Buona
0.055 - 0.070	-	12.1-35.4	55-154	4.5 -9.4	36-75	54-100	51-100	Moderata
0.071 - 0.085	0.125 - 0.164	35.5 - 55.4	155-254	9.5 - 12.4	76 - 185	101 - 360	101-150	Dannosa per gruppi sensibili
0.086 - 0.105	0.165-0.204	55-150.4	255-354	12.5 - 15.4	186-304	361-649	151 - 200	Dannosa
0.106 - 0.200	0.205-0.404	150.5 - 250.4	355 - 424	15.5 - 30.4	305 - 604	650 - 1249	201-300	Molto dannosa

Tabella 2.1: Breakpoints per l'AQI(Air Quality Index) [30] .

L'EPA AQI, basato sull'inquinante più elevato, è particolarmente efficace nell'identificare episodi di inquinamento a breve termine, quando le concentrazioni salgono a livelli molto alti. Tuttavia, se lo scopo principale è monitorare l'esposizione a lungo termine o l'impatto dei piani di miglioramento della qualità dell'aria, gli AQI integrati, tenendo conto dell'interazione degli inquinanti e considerando diversi tempi medi, sono più adatti [33] .

La presenza di anidride carbonica e VOC_s hanno un impatto negativo sulle prestazioni cognitive e sulla salute umana, di seguito vengono riportati esempi di breakpoints della CO_2 e VOC_s per l'IAQI [40] .

Index	CO_2 (ppm)	VOC_s	Rating
1	0-400	0-50	Eccellente
2	400-100	51-100	Buona
3	100-1500	101-150	Moderata
4	1500-2000	151-200	Dannosa
5	2000-5000	201-300	Molto Dannosa
6	da 5000	301-500	Grave

Tabella 2.2: Breakpoints dell'IAQI(Indoor Air Quality Index) [40] .

2.5 Monitoraggio dell'aria utilizzando sistemi IoT

Le informazioni precedenti evidenziano l'importanza di monitorare la qualità dell'aria anche in ambienti chiusi come l'interno del veicolo.
L'internet of things(IoT) fornisce concetti e tecnologie, ad esempio WSN(Wireless Sensor Network), reti composte da nodi-sensori comunicanti via wireless, per monitorare l'ambiente circostante [15] .
Negli anni si è registrato un crescente interesse nell'implementazione di piattaforme Internet of Things (Iot) per il monitoraggio dell'inquinamento atmosferico [18], utilizzando vari device connessi tra di loro che si scambiano dati ottimizzando le performance. Le moderne piattaforme di sensori Iot analizzano dati raccolti, eseguono analisi in tempo reale e forniscono notifiche tempestive dello stato dei sensori basate sulle misurazioni, per aumentare l'efficienza gestionale e le risposte. Le tecnologie utilizzate per i dati possono aumentare le performance dei sensori a basso costo, raggiungendo un livello di precisione simile a quelli di strumenti di precisione, riducendo l'investimento iniziale, soprattutto per i grandi sensori [8] .
Sempre in [8] si può trovare un esempio di piattaforma di sensori IoT che include quattro componenti principali : (i) sensori, (ii) Processing Network, (iii) data analysis, (iiii) system monitoring.

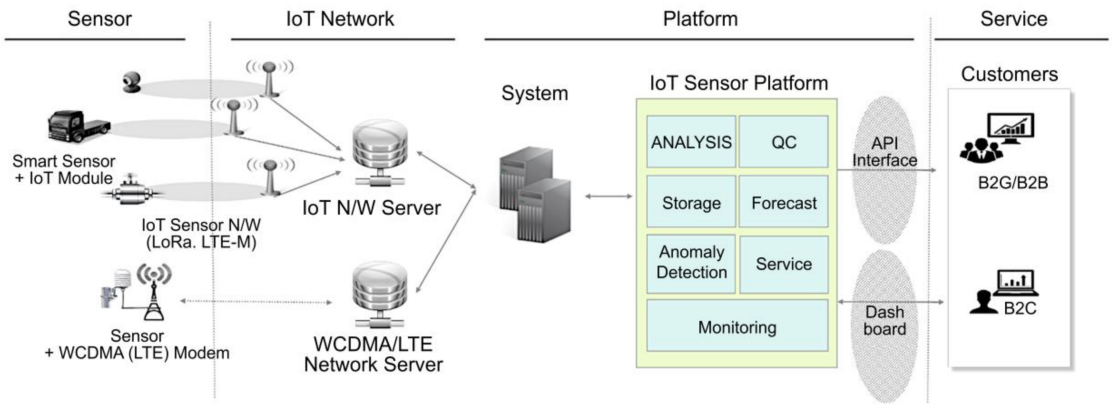


Figura 2.2: Piattaforma di sensori Internet of Things (IoT) [8].

Nuovi sistemi basati su sensori Iot sono continuamente proposti. I Dati raccolti dai sensori possono svolgere un ruolo fondamentale nel gestire e misurare la qualità dell'aria. Una sfida importante riguarda il trattamento delle informazioni, è necessario assicurare che l'analisi dei dati sia efficiente e affidabile. Una falsa interpretazioni dei dati può portare a decisioni errate, il che può diventare molto pericoloso [4]. L'analisi dei dati è un processo di ispezione, trasformazione e modellazione dei dati con l'obiettivo di scoprire informazioni utili, lavori come "In-Cabin Air Quality during Driving and Engine Idling in Air-Conditioned Private Vehicles in Hong Kong " [7] illustrano la qualità dell'aria all'interno della cabina del veicolo in macchine private ad Hong Kong, durante la guida e a motore acceso ma non in movimento.
Utilizzando i seguenti sensori:

- TSI AeroTrak® Handheld Particle Counter Model 9303 - per $PM_{2.5}$ e $PM_{0.3}$.
- Handheld VOC Monitor ppb RAE 3000 - livello TVOC.
- TSI IAQ-Calc™ Indoor Air Quality Meter 7545 - CO e CO_2.
- HOBO temp/RH Logger UX100-003 - Temperatura e RH all'interno della cabina del veicolo.

Tra maggio e settembre 2017 si è svolto un progetto durante il quale sfruttando 51 veicoli è stata misurata la qualità dell'aria mantenendo l'aria condizionata attiva per tutta la durata del progetto. La qualità dell'aria all'interno della cabina è influenzata dall'ambiente interno ed esterno ma anche dall'utilizzo della vettura.
Il tempo di percorrenza medio è stato di circa 30 minuti, le misurazioni di $PM_{2.5}$, $PM_{0.3}$, $TVOC_s$, CO, CO_2, temperatura e umidità relativa (RH) sono stati acquisiti ad intervalli di un minuto.

L'analisi dei dati mostra che per i primi 5 minuti molti dei parametri fluttuano significativamente, in quanto il conducente accende l'aria condizionata, apre e chiude le portiere.
La differenza tra i mezzi e la correlazione tra i diversi parametri, le condizioni della vettura e la percezione del conducente della qualità dell'aria sono stati misurati utilizzando un gruppo indipendente T-test e sfruttando il modello di correlazione Pearson. L'analisi ha preso in considerazione l'età dell'automobile, chilometraggio, utilizzo e analisi della frequenza di pulizia, valori individuali raccolti dal veicoli piuttosto che la percentuale in diversi gruppi sono stati utilizzati nell'analisi.

Informazioni sui risultati ottenuti

- Particolato sottile: la misurazione del $PM_{0.3}$ durante la guida e a motore spento non presenta differenze significative, $PM_{2.5}$ è uno dei parametri più comunemente monitorati in aria esterna in quanto è strettamente correlato al processo di combustione. $PM_{2.5}$ durante la guida ha una misurazione di 40 ± 28 $PtL^{-}1$, valore molto più alto in confronto a 23 ± 19 $PtL^{-}1$ del motore non in movimento. In generale la più grande variazione è registrata durante la guida rispetto al motore acceso ma non in movimento.

- CO: Le concentrazioni medie di CO durante queste condizioni non erano significativamente diverse e all'interno della IAQ le registrazioni vengono identificate come classe eccellente (1,7 ppmv).

- CO_2: Il 96% delle vetture non rispettano le raccomandazioni di CO_2 per essere identificate come classe buona ($\leq$ 1000 ppmv) ed il 90% non supera la stessa identificazione a motore non in movimento con una media di 3096 ppmv.

- TVOC : durante la guida il 76% delle macchine possono essere identificate di classe buona o eccellente, al contrario il 70% dei veicoli non rispetta le statistiche a motore non in movimento.
  La concentrazione media di TVOC a motore non in movimento (1351 ppbv) era significativamente superiore a quello durante la guida (331 ppbv).

I livelli di CO_2 misurati durante il progetto non rappresentano una minaccia diretta per la salute fisica dei pendolari ma queste misurazioni possono causare vertigini e provocare sonnolenza guida (ad esempio $\geq$ 2000 ppm di CO_2), oltre a ridurre il rendimento complessivo del processo decisionale (studi condotti in una camera ambientale controllata con 1000 e 2500 ppm di CO_2). I conducenti devono essere vigili in situazioni diverse e di conseguenza prendere decisioni durante la guida.

L'elevato livello di CO_2 in questo ambiente può costituire un rischio potenziale per i pendolari e per la sicurezza stradale.
Notando le caratteristiche che comportano l'utilizzo dell'HVAC(Heating, Ventilation and Air Conditioning) non possiamo non prenderla in considerazione per il nostro lavoro.

2.6 Machine Learning

Il machine learning è un ramo dell'intelligenza artificiale che permette a un sistema di imparare dai dati piuttosto che attraverso la programmazione esplicita. Gli algoritmi assimilano i dati di addestramento per produrre modelli più precisi basati su tali dati.
Il machine learning quindi ci fornisce metodi generali per estrarre modelli di conoscenza da insiemi di dati, sono cresciuti rapidamente negli ultimi anni nel contesto dell'analisi dei dati e dell'elaborazione che in genere consente alle applicazioni di funzionare in modo intelligente [37] .

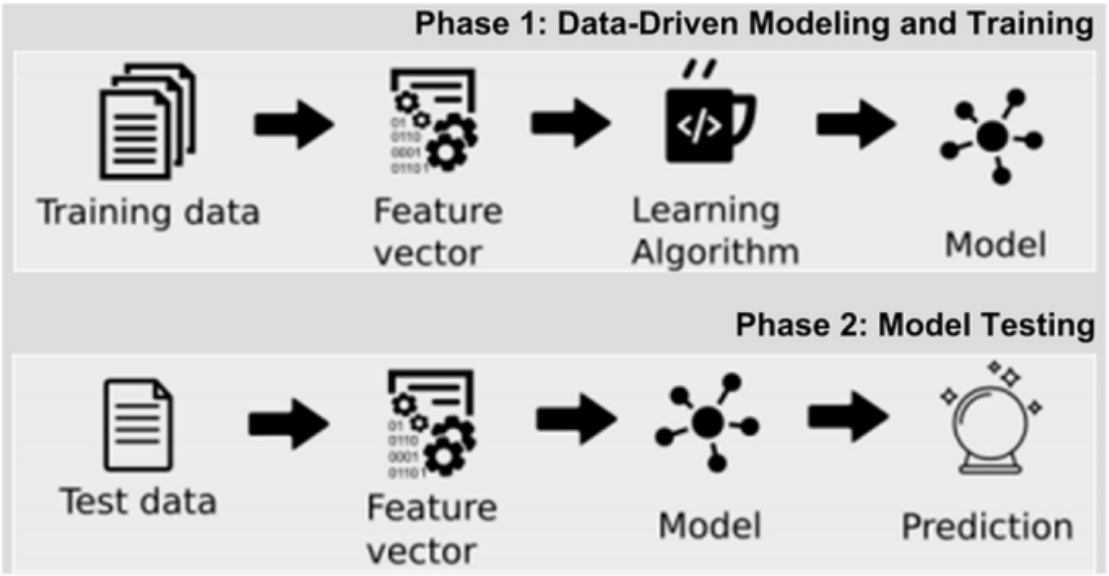


Figura 2.3: Una struttura generale di un modello predittivo basato sul machine learning che considera sia la fase di addestramento che quella di test [36] .

Il Machine Learning si basa su diversi algoritmi per risolvere problemi di dati. Il tipo di algoritmo impiegato dipende dal tipo di problema che si desidera risolvere, il numero di variabili, il tipo di modello che gli si addice meglio e così via.

Negli ultimi anni nell'ambito della previsione dell'inquinamento atmosferico sono state proposte varie soluzioni:

	Problema	Tecniche	Punti di forza
[9]	Tecniche di machine learning per la classificazione della qualità dell'aria	Decision Tree e Naive Bayes Algorithm	91 % di accuratezza per il DecisionTree.
[45]	Sensori per la misurazione dell'AQI a basso costo e utilizzo di machine learning	Random forest	Riduzione del costo
[42]	Tecniche di machine learning per l'AQI in canada	Multilayer neural networks con regressioni non lineari	Gli algoritmi proposti hanno ridotto l'errore
[5]	Machine Learning e modellazione della qualità dell'aria	Randomized Matrix Decompositions e Random forest regression	Il metodo proposto ha diminuito i costi dei sensori

Tabella 2.3: Esempi di progetti che utilizzano algoritmi di machine learning per la predizione della qualità dell'aria.

2.7 Metodi di machine learning per il monitoraggio dell'inquinamento dell'aria

Di seguito vengono accennati gli algoritmi più comunemente utilizzati nella sezione dei related work.

2.7.1 Predizione mediante algoritmi di machine learning

Regressione lineare e polinomiale

La regressione lineare viene utilizzata per predire il valore di una variabile dipendente in base al valore di un'altra variabile indipendente, con più variabili indipendenti si parla di regressione multivariata, con una sola si parla di regressione univariata. Questa forma di analisi stima i coefficienti dell'equazione lineare e implica una o più variabili indipendenti che meglio predicono il valore della variabile dipendente. La regressione lineare corrisponde a una linea retta o a una superficie che minimizza le discrepanze tra i valori di output previsti ed effettivi. Linear regression su Scikit-learn si adatta a un modello lineare con coefficienti w = (w1, ..., wp) per minimizzare la somma residua di quadrati tra gli obiettivi osservati nel dataset e gli obiettivi previsti dall'approssimazione lineare. La regressione polinomiale è una generalizzazione di quella lineare con termini di grado superiore,

in grado di ottenere modelli capaci di descrivere dataset più complessi. Nella regressione polinomiale i parametri e l'algoritmo di regressione rimane il medesimo di quello lineare.

$$\hat{y} = 0_0 + 0_1 X_1 + +0_2 X_2 + ... + +0_n X_n$$

$\hat{y}$—Valore predetto
n— Numero delle features
x_n—nth valore della feature

Support Vector Machine

Support vector machine (SVM) può eseguire la classificazione lineare o non lineare, inoltre, SVM supporta anche applicazioni di regressione lineare e non lineare, noto come SVR. L'obbiettivo di SVM è quello di individuare la separazione lineare ottimale secondo un criterio geometrico tra le istanze delle due classi, è uno dei metodi più utilizzati ed è efficace con training set di piccole dimensioni. La Support vector regression (SVR) usa gli stessi principi delle SVM ma per i problemi di regressione. Il modello prodotto dalla classificazione dei vettori di supporto dipende solo da un sottoinsieme dei dati di addestramento, la funzione di costo per costruire il modello non si preoccupa dei punti di addestramento che si trovano oltre il margine. Analogamente, il modello prodotto da Support Vector Regression dipende solo da un sottoinsieme dei dati di allenamento, la funzione di costo ignora i campioni la cui previsione è vicina al loro obiettivo.

Multilayer Perceptron

Le reti neurali sono composte da gruppi di neuroni artificiali organizzati in livelli. Tipicamente sono presenti: un livello di input, un livello di output e uno o più livelli intermedi o nascosti (hidden). MLP è il modello più comunemente usato nella rete neurale feed-forward dove le connessioni collegano i neuroni di un livello con i neuroni di un livello successivo. Se la rete neurale deve risolvere un problema di regressione, l'output desiderato è il valore corretto della variabile dipendente, in corrispondenza del valore della variabile indipendente fornita in input.

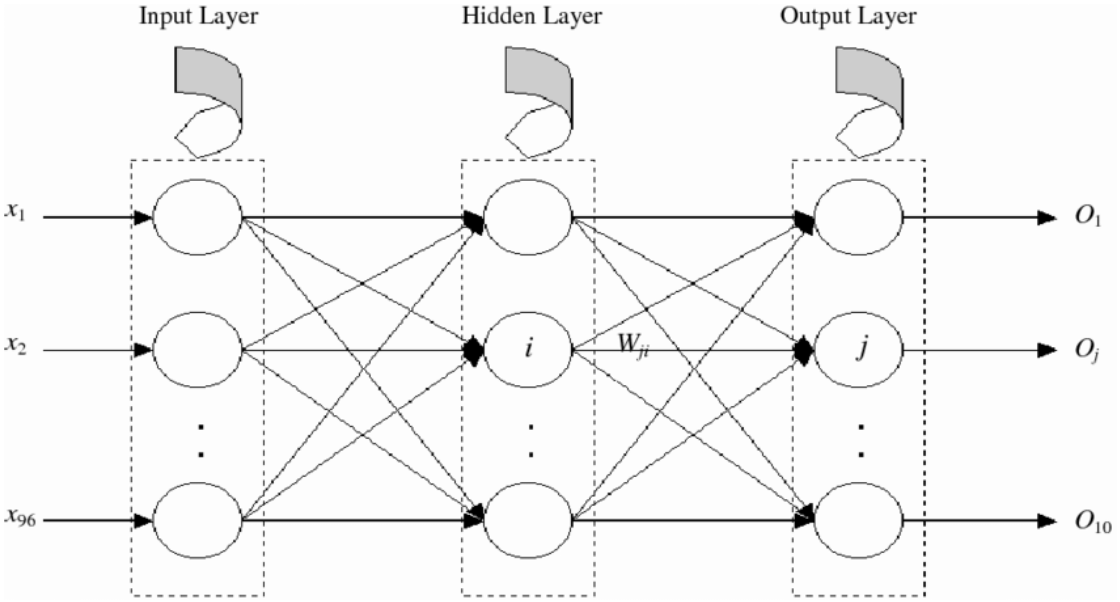


Figura 2.4: Esempio struttura MLP [32] .

2.7.2 Metodi di valutazione

I metodi di valutazione più utilizzati sono :

- RMSE(Root Mean Square Error) la radice della media dei quadrati degli scostamenti tra il valore vero e il valore predetto.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

- MSE (Mean Squared Error) indica la discrepanza quadratica media fra i valori dei dati osservati ed i valori dei dati stimati.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- MAE(Mean Absolute Error) è spesso conosciuto anche come L1 Loss, e matematicamente rappresenta la distanza tra il valore predetto e quello effettivo.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

- Il coefficiente di determinazione R^2 indica la proporzione tra variabilità dei dati e correttezza del modello.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2}$$

$\hat{y}_i$ —Valore predetto di y
y_i—Valore medio di y

2.7.3 Classificazione mediante algoritmi di machine learning

K-Nearest Neighbors

KNN è un algoritmo di machine learning non parametrico che utilizza attributi nominali e numerici dei dati attraverso la selezione dell'attributo più comune tra i vicini k più vicini.
In questo algoritmo la scelta di k è molto importante, nel caso in cui k fosse troppo piccolo, la classificazione sarebbe influenzata dal rumore dei dati, viceversa se k è troppo grande rende la classificazione non precisa. Non richiede la costruzione di un modello e rispetto ai sistemi basati su decision tree permette di costruire "contorni" delle classi non lineari e sono quindi più flessibili.

SVM

SVM è uno degli strumenti più utilizzati per la classificazione di pattern. Date due classi di pattern multidimensionali linearmente separabili, tra tutti i possibili iperpiani di separazione, SVM determina quello in grado di separare le classi con il maggior margine possibile. I pattern del training set che giacciono sul margine sono detti support vector, tali pattern, che costituiscono i casi più complessi, definiscono completamente la soluzione del problema. SVM prevede un'importante estensione della teoria inizialmente sviluppata per iperpiani. L'idea di base sta nell'aggiungere dimensioni per trovare la separazione. Le funzioni kernel più utilizzate sono : Polynomial kernel, Radial Basis Function(RBF), 2-layer Neural Network.

Random Forest.

Random Forest è un metodo versatile di machine learning, capace di affrontare sia compiti di classificazione che di regressione. Con il random forest è anche possibile applicare metodi per la riduzione della dimensionalità, gestire dati mancanti, valori

degli outlier ed altri passaggi essenziali di esplorazione dei dati, producendo buoni risultati. Appartiene alla famiglia di metodi di bagging(Boostrap aggregating). In random forest i singoli classificatori sono degli alberi di classificazione(albero binario in in cui ogni nodo divide i pattern sulla base di un criterio su una singola feature o dimensione). Random forest opera simultaneamente su due tipi di bagging, uno sui pattern del training set e uno sulle features. Un random forest quindi è un modello che si adatta a un certo numero di classificatori ad albero di decisione su vari sotto-campioni del set di dati e utilizza la media per migliorare la precisione predittiva e il controllo overfitting.

Naïve-Bayesian Algorithm

Naïve-Bayesian è un algoritmo di machine learning basato sulla classificazione di probabilità condizionata di un dataset. In questo algoritmo è applicato il teorema di Bayes' per valutare probabilità che si verifichi una variabile target data la probabilità del verificarsi dei suoi predittori.

2.8 Lavori correlati

L'argomento della qualità dell'aria all'interno del veicolo è una scelta molto specifica e nella letteratura non sono molti i lavori pubblicati, per questo è necessario visualizzare lavori correlati alla predizione della qualità dell'aria indoor e i pochi presenti. Pubblicazioni che possono essere molto utili nella progettazione sono progetti come "Real-Time In-Vehicle Air Quality Monitoring System Using Machine Learning Prediction Algorithm" [13], nel quale è stato sviluppato un sistema di monitoraggio real-time e cloud-based in grado di predire la qualità dell'aria all'interno del veicolo. Algoritmi di machine learning sono stati utilizzati per predire la sonnolenza e la fatica dei conducenti basandosi sulla qualità dell'aria presente nella cabina, misurata sfruttando parametri come la concentrazione di CO_2, particolato, velocità dell'auto, temperatura e umidità.
Per l'analisi dei dati e per predire la qualità dell'aria sono stati sviluppati modelli di multilayer perceptron, support vector regression e linear regression. Le prestazioni di questi modelli sono state valutate attraverso Root Mean Square Error, Mean Squared Error, Mean Absolute Error e il coefficiente di determinazione (R^2).

Risultati

Questi algoritmi sono stati testati e confrontati tra loro determinando il migliore.

Section				
	R^2	MSE	RMSE	MAE
SVM	0.9890	6.4513	2.5410	0.97194
LR	0.8137	109.9008	10.4833	5.1379
MLP	0.7151	212.4807	14.5767	11.5757

Tabella 2.4: Risultati dei modelli predittivi in sezioni

Month				
	R^2	MSE	RMSE	MAE
SVM	0.9981	3.6168	1.9018	0.4101
LR	0.9946	10.1875	3.1917	2.1348
MLP	0.9107	100.0034	9.0589	5.0422

Tabella 2.5: Risultati dei modelli predittivi in mesi

Il modello SVR con il kernel RBF aveva l' R^2 più alto e il più basso MSE, RMSE e MAE rispetto ad altri modelli. Il modello di previsione basato su SVR-RBF ha mostrato la massima accuratezza di previsione e migliori prestazioni di generalizzazione. Il coefficiente di determinazione R^2 ottenuto è stato di 0,9890 (sezione) e 0,9981 (mese).
Il sistema può essere utilizzato come misura potenziale per ridurre gli incidenti dovuti alla sonnolenza e alla fatica del conducente.
I risultati hanno dimostrato che l'SVR aveva il più alto tasso di rendimento in termini di R^2 e aveva meno tasso di errore. Ciò indica che il Il modello SVR ha registrato prestazioni predittive eccezionali.

Studi come "Development of Machine Learning-based Predictive Models for Air Quality Monitoring and Characterization" [3] propongono una metodologia di caratterizzazione della qualità dell'aria mediante la costruzione di modelli predittivi che mettono in relazione i valori forniti dai sensori ad un indice di qualità dell'aria. L'obbiettivo è trovare un modo alternativo per caratterizzare la qualità dell'aria attraverso l'uso di sensori di gas integrati e modelli predittivi utilizzando algoritmi di machine learning che possono essere utilizzati per ottenere dati per monitorare il rischio di inquinamento atmosferico. I modelli utilizzati sono k-nearest neighbors (KNN), support vector machine (SVM), Naïve-Bayesian classifier, random forest e neural network. La raccolta dei dati si basa sugli indici AQI precedentemente descritti con un totale di 750 misurazioni.
I risultati sono stati i seguenti:

Model	CV Performance	Confusion Accuracy	LogLoss Performance
k-nearest Neighbors	0.9872258	0.9867	0.0814
Polynomial SVM	0.9464590	0.9778	0.1530
Random Forest	0.9787313	0.9867	0.0662
Naïve-Bayesian Classifier	0.9443075	0.9422	0.2029
Neural Network	0.9924003	0.9956	0.0543

Tabella 2.6: Risultati classificazione

Sulla base dei dati e dei risultati, la metodologia proposta di per calcolare l'indice di qualità dell'aria con i modelli di machine learning è stata implementata con successo. Dei cinque i modelli di machine learning utilizzati, la rete neurale ha riportato i risultati migliori con una precisione del 99,56% e un 0,0543 di prestazioni logloss.
Oggigiorno le smart city sono sempre più comuni ed importanti. Una smart city è un comune urbano che utilizza le tecnologie dell'informazione e della comunicazione (ICT) per per fornire salute, trasporti e infrastrutture connesse all'energia in modo efficiente. In questo contesto esistono molti progetti sulla qualità dell'aria, "Comparative Analysis of Machine Learning Techniques for Predicting Air Quality in Smart Cities" [4] è un progetto che mira a presentare uno studio comparativo per determinare il modello migliore per prevedere accuratamente la qualità dell'aria con riferimento alla dimensione dei dati e al tempo di elaborazione. Dalla loro analisi sono stati scelti quattro modelli di regressione : Decision Tree regression, Random forest regression, Gradient boosting regression e Multi-Layer perceptron regression. Per la validazione sono stati utilizzati MAE e RMSE. L'applicazione di questo progetto è stato testato su diverse città : Beijin, Shanghai, Shemyamg, Guangzhou e Chengdu. Di seguito vengono confrontati i modelli migliori di ogni città:

	GBRT		MLP		DTR		RFTR	
City	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
Shanghai	17	0.07	13.84	0.03	21.74	0.09	17.68	0.05
Guangzhou	27.7	0.17	12.2	0.045	14.36	0.06	13.1	0.05
Chengdu	14.47	0.113	9.8	0.108	11.51	0.086	10.5	0.0828
Shenyang	17	0.102	13.65	0.062	21.3	0.0872	17.68	0.59
Beijing	29.30	0.0866	21.79	0.0806	19.03	0.07	16.92	0.075

Tabella 2.7: Risultati dei modelli predittivi di ogni città

Per confrontare i dataset hanno normalizziamo il valore RMSE :

$$NormalizeRMSE = \frac{RMSE}{Ymax--Ymin}$$

ymax = Valore massimo nel dataset
ymin = Valore minimo nel dataset

Dai risultati si osserva che:

- Beijing : città con i valori più alti di $PM_{2.5}$ all'interno del dataset, i modelli che hanno performato nel modo migliore sono DTR(decision tree regression) e RFR(random forest regression). Il valore RMSE ottenuto usando Random forest regression è di 0,0725 e dopo la normalizzazione un MAE del 16%.
- Shanghai: MLP ha raggiunto i valori migliori non solo nell'identificare i valori di picco ma anche nel raggiungimento del RMSE più basso (0,03) e il MAE più basso (13,84%). Random forest regression ha raggiunto dei risultati molto buoni come MLP, con RMSE di 0,05 e MAE del 17%.
- Shenyang: MLP ha preformato molto meglio degli altri modelli con un RMSE di 0.062 e MAE del 13.65 %.
- Guangzhou: MLP e random forest regression sono stati ancora ancora una volta i modelli migliori. MLP ha ottenuto MAE e RMSE rispettivamente del 12,2% e 0,045, mentre random forest regression ha prodotto un risultato leggermente peggiore con un MAE del 13,1% e un RMSE dello 0,05.
- Chengdu: Random forest regression è stato il metodo più accurato, ottenendo un RMSE dello 0,08 e un MAE del 10,5 %. MLP è il secondo miglior metodo con un MAE del 9,8 % e un RMSE dello 0,108.

La pubblicazione "Machine learning and statistical models for predicting indoor air quality" [43] espone una revisione della letteratura sull'uso di modelli per prevedere la qualità dell'aria interna. Sono stati esaminati i metodi più comunemente utilizzati, discussi i loro punti di forza e debolezza revisionando varie pubblicazioni. I modelli più comunemente utilizzati sono : regression models, partial least squares (PLS), decision trees (classification and regression trees), Bayesian hierarchical modeling, generalized boosting models, support vector machine, random forests, generalized linear models e artificial neural networks (ANN). Le variabili predette dai vari articoli pubblicati erano le concentrazioni interne dei vari inquinanti come PM, CO_2, CO, NO_x, radon e l'indice IAQ. Tra gli articoli presi in considerazione erano presenti anche alcuni lavori che facevano riferimento alla predizione della CO_2 e CO all'interno di veicoli pubblici come bus ma che si concentravano su variabili di input inerenti al veicolo, ventilazione e passeggeri, oltre ad informazioni

come temperatura e umidità. Prendendo in considerazione alcune analisi fatte all'interno di questa pubblicazione, si nota come il valore R^2 ottenuto dal modello MLR per la predizione del $PM_{2.5}$ è stato di 0.58, utilizzando come input il valore all'esterno del PM e il valore dell'umidità interna. Il valore è aumentato a 0.69 quando sono stati aggiunte come variabili di input la ventilazione, la velocità del vento e la temperatura, questo perché le informazioni relative al vento possono essere fortemente correlate al trasporto interno del particolato.
Lo studio "Integrating Statistical Machine Learning in a Semantic Sensor Web for Proactive Monitoring and Control" [1] presenta un approccio per ottenere un monitoraggio e controllo proattivo nell'ambito del semantic sensor web, integrando un modello predittivo basato sul machine learning, l'obbiettivo era di selezionare il modello più appropriato per una previsione a breve termine del $PM_{2,5}$ in un ambiente interno. Il modello MLP selezionato è stato integrato nel sistema utilizzando la libreria WEKA in Eclipse, un Itegrated Development Environment basato su Java. Questo studio suggerisce inoltre che un classificatore MLP e BN opportunamente addestrato può efficacemente prevedere l'andamento a breve termine di $PM_{2,5}$ con alta precisione e sensibilità. Negli esperimenti effettuati è stata raggiunta una precisione fino a 0,86 e sensibilità fino a 0,85 utilizzando un approccio "sliding window" per prevedere gli stati del $PM_{2.5}$ 30 minuti nel futuro.

L'Articolo "Indoor Air Quality Monitoring with IoT: Predicting PM10 for Enhanced Decision Support" [35] ha come obbiettivo quello di descrivere la funzionalità di un sistema di monitoraggio IAQ basato sull' internet of things per la misurazione di parametri come PM_{10}, $PM_{2.5}$, CO_2, VOC, temperatura e umidità. Il contributo principale di questo articolo è l'utilizzo dei dati raccolti per addestrare e progettare un sistema di previsione utilizzando il modello XGBoost. Questo sistema di monitoraggio è stato installato nella mensa del "National Institute of Technical Teachers Training and Research, Chandigarh" e la registrazione di questi inquinanti è avvenuta per mesi. Dopo aver effettuato una pulizia dei dati il dataset conteneva 17280 istanze, che sono state divise nel 33% in test set e 67% in training set. CO_2, VOC e umidità sono stati selezionati come input per prevedere la concentrazione del PM_{10}. I risultati mostrano un RMSE = 0.48, R^2 = 0.99, MSE = 0.234, MAE = 0.284 e MAPE = 3,24%.
"Internet of Things (IoT) Based Indoor Air Quality Sensing and Predictive Analytic" [23] propone una soluzione per un sistema di monitoraggio e predizione della qualità dell'aria interna basata su IoT e machine learning. Il progetto utilizza una serie di sensori Iot che misurano i seguenti agenti inquinanti: NH_3, CO, NO_2, CH_4, CO_2, $PM_{2.5}$, insieme con la temperatura ambiente e l'umidità dell'aria. Per la classificazione della qualità dell'aria interna sono stati utilizzati algoritmi di machine learning e deep learning tra cui Support vector Machine(SVM), K-Nerarest Neightbour(KNN), Naive Bayes(NB) e le Neural Network(NN). Per predire invece

la concentrazione della qualità dell'aria interna è stata utilizzata la rete Long and Short Term Memory (LSTM), che è una forma avanzata di Recurrent Neural network (RNN). Per misurare le performance dei classificatori sono state utilizzate F1 score, precision, recall e accuracy :

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Recall = \frac{TP}{TP + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$F1Score = \frac{2xPrecisionxRecall}{Precision + Recall}$$

dove TP denota veri positivi, TN denota veri negativi, FP denota falsi positivi e FN denota falsi negativi. La tabella mostra la precisione, il punteggio F1, l'accuratezza e la recall di ciascun algoritmo di classificazione applicato sul set di dati.

Model	Accuracy	Precision	Recall	F1 Score
SVM	95.35	95.1	94.9	94.99
NB	98.9	99	98	98.49
KNN	98.47	97.3	98.5	97.89
NN	99.1	99	99	99

Tabella 2.8: Valutazione classificatori

Si nota come in questo caso il classificatore SVM ha performato il modo molto peggiore rispetto agli altri classificatori, sapendo che tipicamente è utilizzato su set di dati di grandi dimensioni con variazione significativa e caratteristiche multiple, il modello naive Bayes ha sovraperformato sia SVM che KNN, grazie alla sua caratteristica di essere adatto peri problema multi-classe, mentre la rete neurale è stata la migliore. Successivamente è stato applicato il metodo LSTM per prevedere la concentrazione di ogni inquinante valutato dalle metriche MSE, MAE e RMSE.

Gass	MAE	MSE	RMSE
CO2	0.09324657	0.01546322	0.12435118
CO	0.02422077	0.00755213	0.08690299
NO2	0.13934513	0.03205927	0.17905103
PM 2.5	0.12350263	0.01681033	0.12965467
Temperature	0.07542943	0.01125118	0.10607159
Humidity	0.06494383	0.00875947	0.09359205

Tabella 2.9: Valutazione oraria delle prestazioni dell'LSTM.

In conscusione lo sviluppo di un sistema di monitoraggio indoor e i risultati basati sull' Iot e Machine Learning sono stati molto soddisfacenti dove NN ha fornito la precisione di classificazione del 99,1% e LSTM ha raggiunto dei risultati molto soddisfacenti per la predizione dei singoli inquinanti.

Questo capitolo è servito per esporre il background di questo progetto e analizzare i progetti correlati, le nozioni apprese dall'analisi della qualità dell'aria esterna e interna al veicolo serviranno per il progetto di questo studio, l'analisi dei lavori correlati evidenzia i migliori algoritmi per la classificazione e la regressione per individuare l'inquinamento dell'aria e prendendoli come esempio si andranno a confrontare i risultati.

Nel prossimo capitolo verrà esposto l'ambiente di lavoro, gli strumenti utilizzati e verranno proposti gli obbiettivi del progetto.

Capitolo 3

Progetto

3.1 Panoramica Progetto europeo NextPerception

Il progetto NextPerception è un progetto europeo avviano nel maggio 2020. Al progetto partecipano 43 partners di 7 paesi, tra cui l'università di Bologna.
NextPerception punta allo sviluppo di un modello di intelligenza distribuita per il monitoraggio umano sicuro ed affidabile da poter utilizzare nel settore sanitario e nel campo automobilistico.
Le moderne applicazioni utilizzate nel settore sanitario e automobilistico sono dei sistemi complessi che integrano componenti che interagiscono tra di loro con l'obbiettivo di essere sicuri, scalabili e integrabili. Alla base dei sistemi complessi ci sono concetti di cloud to cloud, Iot e intelligenza artificiale che sono nascoste anche all'interno di strumenti che quotidianamente utilizziamo, ma di cui forse non comprendiamo la portata. L'accuratezza e la tempestività delle decisioni dipendono dalla capacità dei sistemi di costruire una buona comprensione dell'ambiente, che si basa sulle osservazioni e sulla capacità di ragionare su di esse.
NextPerception porterà le tecnologie di rilevamento della percezione come le telecamere Radar, LiDAR e Time of Flight al livello successivo, migliorando le loro caratteristiche per consentire un rilevamento più accurato del comportamento umano e dei parametri fisiologici.
Oltre a soluzioni automobilistiche più accurate che garantiscono la vigilanza del conducente e la sicurezza di pedoni e ciclisti, questa innovazione aprirà nuove opportunità in termini di salute e benessere in grado di valutare lo stato della salute.
Per facilitare la costruzione dei sistemi complessi di rilevamento previsti e garantire il loro funzionamento sicuro e affidabile, il paradigma dell'intelligenza distribuita sarà migliorato e supportato da nuovi strumenti, sfruttando i vantaggi dell'Edge e del Cloud computing. Il progetto riunisce i principali attori industriali e partner

di ricerca per affrontare le principali sfide nel settore della salute, del benessere e automobilistico attraverso tre casi d'uso: monitoraggio integrale della vitalità e attività fisica, monitoraggio dei conducenti e fornitura di sicurezza e comfort agli utenti stradali [27] .
L'obbiettivo principale del progetto è sorpassare il moderno stato dell'arte dell'archiviazione e dei servizi che si possono ottenere da questi dati ottenuti dalle persone e l'ambiente.
I concetti principali dello sviluppo per questa tecnologia sono:

- Smart Perception Sensor.
  - Sensori Radar, Lidar e ToF necessari per osservare il comportamento umano e i parametri vitali, i dati raccolti sono combinati con altre informazioni complementari. Questa feature garantisce una facile integrazione con la parte d'intelligenza distribuita.
- Distributed Intelligence.
  - Questo paradigma consente la distribuzione delle analisi e il processo di decisione del sistema per ottimizzare l'efficienza, esecuzione e l'affidabilità.
- Proactive Behaviour and Physiological Monitoring.
  - Smart Perceptiopn Sensor e l'intelligenza distribuita saranno applicate per comprendere pienamente il comportamento umano per fare in modo che il sistema sia in grado di offrire funzionalità di supporto preservando la privacy dell'utente.

[25]

3.2 Università di Bologna in NextPerception

Il progetto NextPerception è un progetto molto ampio e con molti partner, l'università di Bologna come partner italiano si occupa di uno studio specifico, quello di analizzare diversi aspetti legati alle caratteristiche e elementi ambientali che possono incidere sulla guida.
L'elaborato di tesi pone le sue basi sugli argomenti di NextPerception focalizzandosi sull'analisi dei dati ambientali all'interno del veicolo come temperatura, pressione, umidità e sopratutto agenti inquinanti.

I dati in questione sono presenti all'interno di un dataset ancora in fase di sviluppo che utilizza un kit di sensori chiamato Canarin che è equipaggiato da varie tipologie di sensori tra i quali quello per il particolato (PM). In questo caso d'uso, NextPerception punta a sviluppare un Driver Monitoring System (DMS), in grado di classificare sia gli stati cognitivi del conducente (distrazione, affaticamento, carico di lavoro, sonnolenza) sia lo stato emotivo (ansia, panico, rabbia) del conducente e l'intenzione (svolta a sinistra o destra), nonché le attività e la posizione degli occupanti (compreso il conducente) all'interno dell'abitacolo. Queste informazioni verranno utilizzate per le funzioni di guida autonoma, tra cui la richiesta di acquisizione e il supporto al conducente.

Figura 3.1: Grafico del monitoraggio del guidatore [27].

3.3 Introduzione Progetto Canarin

In questo capitolo saranno descritti i principi di funzionamento, deployment e utilizzi all'interno dei vari progetti del kit di sensori Canarin, dal quale sono stati ricavati i dati per lo svolgimento del progetto.

3.4 Progetto Canarin

Il progetto Canarin è stato creato con lo scopo di mettere a disposizione un sensore portatile e robusto, convalidato per ricerca sulla salute pubblica, che richiede

continue informazioni sull'esposizione individuale.
L'obiettivo specifico del progetto Canarin è stato quello di mettere a disposizione uno strumento per valutare l'esposizione individuale al particolato (PM) in tempo reale e in relazione alla mobilità, descrivendo un approccio innovativo nella progettazione dell'architettura dei sensori specificamente finalizzato a migliorare l'interpretazione dell'esposizione personale all'inquinamento atmosferico.

A differenza dei sensori attualmente in commercio che spesso necessitano di un dispositivo esterno per la trasmissione dei dati, il progetto Canarin è stato realizzato con l'obbiettivo di risolvere i vari problemi di connettività, costi, dimensione, acquisizione dei dati e facilità d'uso per i partecipanti allo studio medico e sanitario.

3.4.1 Canarin

Il primo Canarin è stato creato nel 2016 dall'Asian Institute of Technology, Sorbona University e l'Università di Bologna, che hanno collaborato nell'ambito del Progetto SEA-HAZEMON.
Per lo sviluppo del progetto sono stati utilizzati una serie di Canarin che hanno agito come stazioni fisse di monitoraggio dell'inquinamento atmosferico dai quali hanno raccolto dati sul particolato (PM), vale a dire particelle di diametro 1, 2.5 e 10 μm (PM_1, $PM_{2.5}$, PM_{10}).
Nel 2018 è stato progettato il Canarin II aggiornando i sensori precedentemente utilizzati per includere multiple connessioni wireless e GPS(Global positioning system).
La capacità di visualizzare i dati geo-localizzati sull'inquinamento atmosferico in tempo reale è stato un importante miglioramento.
Altri dispositivi Canarin II sono stati implementati dal progetto POLLUSCOPE, che richiedeva soluzioni mobili e convenienti per valutare l'esposizione individuale all'aria inquinata nella regione dell'Île-de-France.
E' stato utile anche in altri progetti e test, come nella " Johannine Library of Coimbra" per monitorare l'impatto del particolato sull'ecosistema.

3.4.2 Canarin nano

Il Funzionamento di Canarin II si basa su linux così da poter permettere una buona potenza di elaborazione e sviluppo software. La scheda madre consuma approssimativamente sui 3 watt e richiede una batteria in grado di sostenere almeno una giornata intera, il peso è intorno a 1.2 kg.
I progressi tecnologici dei microcontrollori, sensori di gas e PM hanno reso possibile una diminuzione della dimensione senza compromettere la qualità dell'acquisizione

dei dati.

Figura 3.2: Canarin Nano [11].

Specifiche tecniche Canarin Nano

Il canarin Nano su basa su un processore ARM Cortex-M, sulla scheda sono presenti: modem 3G, flash memory, batteria da 400 mAh, modulo GPS, sensori per il particolato sottile(PM_1, $PM_{2.5}$, PM_{10}), temperatura e pressione.
Un sensore VOC, montato su un circuito stampato secondario (PCB) può essere facilmente collegato e scollegato dalla scheda principale. La memoria flash è stata utilizzata per prevenire la perdita di dati se il modem 3G di bordo perde la connessione.

Specifiche tecniche	
Components	References
Development board	Particle Electron 3G-U270
Microcontroller	STMicroelectronics STM32F205
GSM 1 module	U-BLOX SARA-U270
Cellular antenna	Molex Part Number 2072350100
IoT 6/M2M 7 SIM 8 card provider	ThingsMobile
GPS 2 module	U-BLOX NEO-M8N
GPS 2 antenna	Taoglass AGGBP.25B.07.0060A
Battery GEB battery	OEM 3 4000 mAh 4 Li-Po 5 battery
Charger CUI Inc.	SWI10-5-E-I38
Memory flash storage	Winbond Electronics W25Q64FWZPIG8MB
VOC 9 sensors	Sensirion SGP30, Bosch BME680

Tabella 3.1: Specifiche tecniche canarin nano

Il Canarin Nano è progettato per essere sempre in funzione, i vari tipi di utilizzatori non dovrebbero avere interazioni con il dispositivo se non per trasportarlo con loro. Avere un sensore connesso in modo permanente tramite 3G offre diversi vantaggi tra cui la possibilità di inviare i dati ricavati ogni minuto al servizio cloud, rendendoli immediatamente disponibili ai ricercatori per essere visualizzati o scaricati.
Gli utilizzatori possono essere monitorati più volte al giorno, limitando il rischio di perdita di dati. Le modifiche o gli aggiornamenti del firmware si possono effettuare a distanza.
La riduzione delle dimensioni e del peso del sensore ha contribuito a creare un dispositivo in grado di adattarsi meglio alle esigenze dei ricercatori medici rispetto ad altri sensori commerciali.

3.4.3 Accenni sull'utilizzo nel progetto Polluscope

Il progetto Polluscope è un progetto a lungo termine che mira all'utilizzo di sensori portatili a basso costo per misurare l'inquinamento atmosferico a livello individuale in funzione della mobilità.
La scelta di utilizzare Canarin II è dovuta dalla precisione dei sensori nella registrazione del particolato sottile(PM).
L'esperimento raggruppa tre differenti tipi di persone: Volontari dal VGP(Versailles Grand Parc), partecipanti RECORD e persone affette da problemi respiratori.
Le persone coinvolte avevano a disposizione un set composto da Canarin II, sensori come Electrochemical NO_2 e un tablet dedicato all'applicazione per la gestione e

per fungere da access point per il Canarin II.
L'analisi dei dati raccolti dal progetto POLLUSCOPE è in grado di fornirci informazioni estremamente rilevanti a livello scientifico. La raccolta dei dati è stata divisa in periodi con un totale di 384 ore. Il livello di inquinamento dell'aria è indicato in $\mu g/m^3$.
La grande mole di dati raccolti è in grado di mostrare:

- la distribuzione delle attività svolte dalle persone che indossavano i sensori durante il periodo.
- tracciamento nello spazio del PM_{10} ($\mu g/m^3$) dei vari periodi
  - si nota che l'esposizione al PM_{10} è più rilevante lungo la tangenziale di Parigi ("Boulevard Périphérique") e diminuisce lontano dalla strada principale.
- La concentrazione del PM_{10} durante l'attività di lavoro.
  - si nota quanto sia bassa la concentrazione di particolato sottile (PM_{10}) all'interno delle strutture confrontate rispetto allo spazio urbano circostante.
- Molti confronti rilevanti sulla linea dell'ultimo punto.

E' estremamente interessante, per il contesto di questo lavoro, l'esposizione media degli agenti inquinanti per ogni attività.

Activity	$PM_{1.0}$	$PM_{2.5}$	PM_{10}	NO_2	BC
Office	2.3	3.55	3.9	5.7	586.8
Home	10.5	15.1	16.4	6.2	610.0
Shopping	2.7	4.2	4.9	10.4	NA
Park	3.3	4.9	5.3	16.5	409.9
Restaurant	3.6	5.6	6.2	11.2	NA
Street	3.0	4.6	5.1	18.5	NA
Car	9.2	13.1	14.3	20.6	2937.8

Tabella 3.2: Esposizione media degli agenti inquinanti per ogni attività

Dalle registrazioni effettuate si nota che i livelli di NO_2 e BC più alti sono stati rilevati all'interno dell'auto [11] .

3.5 Caratterizzazione della qualità dell'aria e del comfort all'interno di un veicolo elettrico

Questo lavoro ha come obbiettivo quello di caratterizzare la qualità dell'aria all'interno di un veicolo elettrico. Il microclima della cabina del veicolo, come già detto in precedenza, dipende dalle condizioni esterne, riscaldamento, sistema di ventilazione e condizionamento dell'aria (HVAC).
La scelta del ricircolo dell'aria diventa consigliabile in alcune condizioni per risparmiare energia e per allungare il tempo di guida, bisogna comunque tener conto delle problematiche derivanti dall'accumulo di CO_2, composti organici (VOC), infiltrazione di contaminanti e problemi di odore.
La qualità dell'aria all'interno dell'abitatolo, soprattutto in fase di il ricircolo, dovrebbe essere monitorata e controllata, mantenendo i valori sotto specifici valori.
Il possibile utilizzo di sensori low-cost da parte delle catene di produzione potrebbe essere un grande salto di qualità, infatti sono sempre più utilizzati anche in svariati ambiti di indoor air quality.
E' estremamente necessario in questo contesto conoscere tutte le variabili dal quale dipende la qualità dell'aria, infatti la concentrazione del particolato e dei composti organici è stata misurata sia durante modalità del sistema HVAC sia nella modalità di ricircolo dell'aria.
I risultati fanno notare come la concentrazione del PM sia minore durante l'utilizzo di ricircolo dell'aria, ma la concentrazione del VOC è più alta rispetto all'ambiente esterno.

3.5.1 Specifiche di funzionamento hardware

Il sistema si basa su Arduino Mega 2560. L'interno della scheda è composto da un Real time clock(RTC), un data logger per salvare i dati sulla memoria flash e un display TFT .
Il sistema è capace di misurare :

- temperatura in 18 posizioni diverse all'interno della cabina (sensori Maxim integrated DS18B20, convertitore analogico digitale 12 bit con range da -55°C a 125°C)

- la concentrazione del particolato sottile (PM) (sensore Sensirion SPS30 ottimizzato per $PM_{2.5}$ e analisi di particelle ancora più piccole con un filtro HEPA per ridurre la contaminazione ottica con un range da 0 a 100 $\mu g/m^3$)

- concentrazione TVOC (Sensirion SGP30 TVOC, sensore di gas digitale "multipixel", i dati in uscita contengono misurazioni grezze di etanolo e H_2, calcola

i valori TVOC utilizzando un algoritmo interno con un range da 0 a 60000 ppb)

- concentrazione di CO_2 nell'aria (Winsen MH-Z19B non dispersivo nel sensore a infrarossi)
- concentrazione di formaldeide (sensore Winsen ZE08)
- umidità relativa e pressione (sensore Bosch BME280 ad alta precisione con range di pressione da 300 a 1100 hPa, temperatura da -40° a 85°C e da 0 a 100% di umidità)
- velocità del flusso d'aria
- posizione GPS, in un'unica posizione

Sono presenti due di questi sistemi di misurazione, uno esterno e uno interno, con l'unica differenza che a quello esterno mancano i 18 sensori di temperatura, GPS e il sensore di velocità dell'aria.
Entrambi i sistemi campionano i dati in modo indipendente ogni 10 secondi e tutti i sensori digitali utilizzati nel dispositivo di misurazione includono un microcontrollore che implementa l'ottimizzazione e algoritmi di auto-calibrazione.

3.5.2 Setup esperimento

Il veicolo su cui sono stati prodotti i test era una Nissan Leaf 40kWh, la cabina del veicolo è stata divisa in 3 parti: top, middle e bottom level, con 6 sensori in ogni parte.
La misurazione della temperatura nelle 3 parti del veicolo è stata utile anche per sviluppare un modello per la predizione del consumo energetico.
I sensori PM sono stati posizionati vicino al pomello del cambio, insieme al sistema di acquisizione.
Le condizioni esterne possono influenzare fortemente il microclima interno.
E' stato installato un sistema On - Board-Diagnostic(OBD) basato su linux per capire dettagliatamente il funzionamento delle capacità HVAC, in grado di raccogliere variabili differenti dall'unità di controllo tra cui:

- Consumo HVAC
- potenza utilizzata dalle apparecchiature ausiliarie
- Consumo dato dal riscaldamento

Analizzare questi dati è indispensabile per valutare come varia l'aria all'interno della cabina durante l'esperimento.
Sono state studiate due diverse condizioni di prova partendo da uno stato di equilibrio con l'ambiente esterno, ottenuto mantenendo tutti i sistemi spenti e tutte le porte aperte per 30 minuti. Una volta raggiunto l'equilibrio, è stato eseguito il test corretto mantenendo il riscaldamento attivo, la temperatura impostata al massimo di 30, la velocità della ventola al massimo, tutte le finestre e tutte le porte chiuse. Durante il primo test, il sistema di ricircolo era spento (il che significa che il sistema di ventilazione dell'aria era in configurazione open-air), mentre durante il secondo test il sistema di ricircolo era acceso.

3.5.3 Conclusioni esperimento

Sono stati effettuati confronti tra la qualità dell'aria ottenuta in cabina durante le modalità di configurazione del sistema di ventilazione.
Le misurazioni mentre il ricircolo dell'aria era disattivato fanno notare che la concentrazione di TVOC aumenta mentre il sistema HVAC funziona. Questo comportamento conferma la presenza di una fonte di VOC all'interno del veicolo e un fenomeno di accumulo durante il funzionamento HVAC. Le principali differenze delineate sono relative alle configurazioni di aria aperta e ricircolo. Con il ricircolo dell'aria acceso però TVOC ha raggiunto un valore simile al caso precedente, ma più lentamente. I valori del particolato evidenziano come le performace di filtraggio sono migliorate rispetto al ricircolo dell'aria. Riassumendo i risultati mostrano che, mentre i PM vengono filtrati, i VOC si accumulano durante il funzionamento in modalità di ricircolo [34] .
Molto interessante il metodo alternativo utilizzato per ottenere informazioni su IAQ del veicolo, proposto in [21] . L'indice di qualità (CAQI), è definito come segue:

$$CAQI = \frac{\int_{ti}^{tf} Cint(t)dt}{\int_{ti}^{tf} Cext(t)dt}$$

Cint è la concentrazione interna , Cext è la concentrazione esterna, ti è l'istante di inizio e tf è l'istante di fine.

3.6 Canarin II: Designing a Smart e-Bike Eco-System

La misurazione del particolato è sempre risultata complessa, costosa e con un limite importante sulla collezione di dati accessibili.
L'avvento dell'IoT e la possibilità di utilizzare sensori portatili a basso prezzo per la misurazione dell'inquinamento e la gestione dei dati in real time che possono essere utilizzati in diversi contesti ha contribuito alla nascita del concetto dello smart environment.
Lo smart environment ha come obbiettivo l'uso della tecnologia per incrementare la sostenibilità e la gestione delle risorse naturali. Oltre allo smart environment esistono diverse dimensioni simili come la smart mobility.
Un progetto interessante che si immerge completamente in questo scenario è quello di un prototipo di una smart e-bike equipaggiata con il Canarin II, un sistema per la rilevazione del particolato che colleziona dati e li condivide.
Disponendo di questi dati condivisi dagli utenti si è in grado di ampliare il progetto, in questo caso si è riusciti a "disegnare" tramite i dati ottenuti dagli utenti, una mappa dell'inquinamento urbano.
La possibilità di avere accesso a cosi tanti dati sull'inquinamento permette l'utilizzo di algoritmi di machine learning per progetti molto interessanti, come utilizzare reti neurali per identificare le diverse categorie di inquinamento per tipo di traffico.
Il prototipo della smart bike eco-system si basa su una bici elettrica, un canarin, uno smartphone per l'accesso ad internet e una web-app per la collezione dei dati.

In particolare l'architettura di Canarin II si basa su UDOO Neo Full e un Arduino-powered a scheda singola Android/Linux. Questa nuova versione mira a superare alcune limitazioni emerse in Canarin 1.0.

Miglioramenti Canarin I

- Aumento della memoria: salvataggio dei dati in una SD interna.
- Miglioramento della connessione : possibilità di stabilire connessione WPA2-Enterprise.
- Ottimizzazione della connessione: il sistema registra dati anche senza avere una connessione stabile, inviandoli quando la connessione viene stabilita.
- Miglioramento dei sensori del particolato sottile: colleziona dati di PM_1,$PM_{2.5}$ e PM_{10}, temperatura e umidità in modo più accurato ed è stato installato un sensore UVI(ultra violet index).

- Microprocessore ARM Cortex-A9 1GHz che utilizza linux.

La nuova architettura di sistema è quindi strutturata su due livelli: tutti i sensori girano su una piattaforma compatibile con Arduino UNO che ha un clock a 200 MHz, basato su un coprocessore I/O Cortex-M4, mentre un sistema operativo basato su Linux memorizza i dati in file e stabilisce una connessione al server per l'invio dei dati.
Il design del sensore si basa su un PCB interno che ospita la scheda ed i sensori saldati.
La comunicazione è basata su WiFi e la scheda può anche connettersi a reti EAP-SIM tramite un lettore SIM USB.
Il protocollo di comunicazione è di tipo UDP, così come il protocollo Canarin 1.0 e comunica con una versione avanzata del database MySQL.

L'esperimento è stato eseguito tra maggio e giugno 2017, è durato 4 ore percorrendo 40 kilometri nella zona urbana della città di Bologna.
La raccolta dei dati è stata eseguita ogni 40 secondi, per ogni registrazione si conoscono:

- data e ora.
- coordinate GPS.
- temperatura
- umidità
- pressione dell'aria.
- particolato - PM1,PM2.5,PM10

Questo lavoro, che si presenta come prototipo di una smart-bike, è stato presentato in occasione del G7 Ministerial Meeting on Environment.
E' un lavoro preliminare considerato come punto iniziale per progetti futuri per la qualità e la quantità dei dati raccolti, l'affidabilità e connettività [2] .

3.7 Valutazione dell'accuratezza dei modelli sull'inquinamento atmosferico che sfruttano sensori strategici

Il contenuto di questa pubblicazione è importante perché ha lo scopo di identificare i possibili problemi nella configurazione di test sull'inquinamento dell'aria, le

misurazioni e i modelli.
Esistono svariate teorie matematiche e strumenti numerici sotto la categoria della modellazione sull'inquinamento dell'aria con l'obiettivo comune di fare una valutazione dell'inquinamento su una determinata area utilizzando un insieme definito di dati.
La disponibilità dei sensori Canarin II è riuscita a offrire un set completo di dati per testare adeguatamente i modelli di diffusione dell'inquinamento in grado di poter definire modelli ancora più solidi e riuscendo a determinare la migliore configurazione necessaria in termini di numero di sensori e distanza tra un sensore all'altro.
Diversi fattori influenzano il risultato dato dall'applicazione di tecniche di modellazione dell'inquinamento atmosferico e devono essere considerati nella progettazione dei modelli per evitare risultati incerti.
Tali problemi includono:

- vegetazione urbana: la vegetazione può, direttamente e indirettamente, influenzare la qualità dell'aria.
- background concentration: questo fattore indica la concentrazione che verrebbe misurata se le fonti di inquinamento locali non fossero presenti.
- assetto urbano: gli edifici possono alterare i valori.
- terreno: la conformazione dell'area deve essere considerata per simulare il movimento degli inquinanti nell'atmosfera.
- fonti d'acqua: l'acqua, sotto forma di fiumi, laghi o oceani, può trasportare l'inquinamento per lunghe distanze anche in alte concentrazioni.
- dati meteorologici: le informazioni su velocità e direzione del vento, temperatura, umidità sono rilevanti per l'inquinamento atmosferico.

E' molto importante considerare questo progetto come strumento per informare le persone su come monitorare l'inquinamento.
Un esperimento è stato messo in atto al campus di Cesena dell'Univesità di Bologna. Sulla misurazione incidono la vicinanza ad un parco, ad un'area residenziale, ad una superstrada e ad una zona industriale non operativa.

La stazioni dei sensori(Canarin II) hanno registrato i diversi valori della qualità dell'aria e condizioni ambientali, come PM_1 (≤ 1 micron), $PM_{2.5}$ (≤ 2.5 micron) , PM_{10} ($\leq$10 micron), temperatura, umidità relativa e pressione dell'aria.
Le stazioni di sensori sono state disposte in tre locazioni al di fuori dell'università,

ogni stazione comprende dei Canarin II posizionati in linea ogni 20 metri. In un mese di test sono stati raccolti 714,240 dati complessi all'interno del database, ogni dato racchiude diversi valori (come $PM_{1.0}$, PM $_{2.5}$, PM $_{10}$, umidità, temperatura, direzione e velocità del vento, pressione atmosferica) contestualizzati nello spazio (coordinate GPS) e tempo.

Questi esperimenti preliminari, riescono ad evidenziare quanto la qualità dell'aria sia influenzata da diversi fattori, tra cui vegetazione, diverse fonti mobili di inquinamento, disposizione urbana e la vicinanza ad una fiume naturale. Questi dati puntano ad indirizzare l'attenzione su tre possibili problemi:

- raccolta accurata dei dati.
- distribuzione strategica delle configurazioni dei sensori di qualità dell'aria e ambientali.
- sviluppare modelli per misurare l'accuratezza e testare l'efficienza di tecniche di modellazione dell'inquinamento atmosferico.

[41]

3.8 Obbiettivi progetto

Questo progetto presenta una proposta di sviluppo di un monitoraggio della qualità dell'aria che ne rende possibile la predizione all'interno del veicolo tramite l'utilizzo di algoritmi di machine learning. Nello specifico si vuole, tramite le informazioni sugli agenti inquinanti e sostanze ricavate dall'utilizzo del kit di sensori canarin descritto in precedenza, predire la concentrazione degli inquinanti interni più incisivi, in modo da riuscire a prevenire gli effetti dannosi che possono mettere in pericolo il conducente durante la guida. Lo stato dell'arte indica quali siano gli algoritmi di regressione più performanti nella predizione delle concentrazioni di agenti inquinanti, fornendo inoltre importanti indicazioni sulla fase di analisi. Lo stato dell'arte ha inoltre evidenziato la novità che questo progetto punta a raggiungere, quello di riuscire ad utilizzare gli algoritmi di regressione sulla base dei dati esterni, poiché i lavori presenti mirano alla predizione della qualità dell'aria interna al veicolo considerando solo i valori interni senza prendere in considerazione l'importante correlazione con l'ambiente esterno al veicolo. L'idea iniziale del progetto era quella di riuscire a ricavare un set di dati conforme a quelli in possesso dell'università di Bologna per svolgere il progetto e poi utilizzare i nostri dati come test, purtroppo essendo un caso d'uso molto specifico non si è riusciti a trovarlo, per cui la scelta è stata quella di utilizzare unicamente i nostri dati.

3.8.1 Collezione dei dati

Le informazioni messe a disposizione fino a questo punto sono sufficienti per capire come strutturare il progetto. Essendo un progetto in evoluzione le registrazioni effettuate non sono in grande quantità ma mette le basi per progetti futuri molto interessanti. Le registrazioni sono state effettuate i giorni 23/11/2021, 24/11/2021 e il 14/12/2021, in fasce orarie pressoché differenti (10:00 - 13:00,17:00-20:00,18:00 - 21:00). Alla macchina a disposizione dell'università di Bologna, una Nissan Leaf 40 kWh, sono stati montati due canarin, uno interno e uno esterno al veicolo, i dati a disposizione sono:

Parametri	Unità di misura	Descrizione
Datetime(UTC+1)	-	Data e ora in formato UTC
Coordinate GPS	-	Latitudine, Longitudine, altezza.
Umidità	%	Umidità.
Temperatura	C°	Temperatura.
Particolato	$\mu g/m^3$	PM1,PM2.5,PM4,PM10
Pressione dell'aria	hPa	Pressione dell'aria.
TVOC	ppb	Total volatile organic compounds.
Co2	ppm	Anidride carbonica.
H2	-	Idrogeno.
Ethanol	-	Etanolo.
vento	(m/s)	Velocità del vento.
HCHO	ppm	Formaldeide.

Tabella 3.3: Parametri, unità di misura e descrizione delle registrazioni interne ed esterne

I dataset a nostra disposizione quindi sono due, uno interno e uno esterno, le registrazioni effettuate dal canarin interno eseguivano un campionamento ogni 10 secondi, dai quali sono stati ricavati all'incirca 2500 dati, il canarin esterno invece eseguiva un campionamento ogni 5 secondi. La macchina in questione ha effettuato un percorso nella zona urbana di Bologna, seguendo la seguente tratta:

Figura 3.3: Percorso automobile durante le registrazioni.

3.8.2 Requisiti essenziali per il benessere del guidatore

Analizzando lo stato dell'arte abbiamo capito l'importanza del monitoraggio degli agenti inquinanti per il benessere della persona e soprattutto i rischi che possono comportare. Anche se non è il nostro obbiettivo primario quello di prevedere l'AQI lo utilizzeremo per controllare che l'indice di ogni agente inquinante predetto segua la distribuzione di quello reale in modo da poterli analizzare singolarmente in un futuro. Il calcolo di ogni inquinante si basa sulla formula dell'EPA AQI descritta nel primo capitolo utilizzando i valori della tabella seguente.

TVOC(ppb)	CO_2 (ppm)	PM2.5($\mu g/m^3$)	PM10($\mu g/m^3$)	IV-AQI	classe
0 - 65	340–600	0.0–12.0	0–54	0-50	Buona
66 - 220	601–1000	12.1–35.4	55–154	51-100	Moderata
221-660	1001–1500	35.5–55.4	155–254	101-150	Dannosa per gruppi sensibili
661 - 2200	1501–2500	55.5–150.4	255–354	151-200	Dannosa
2201 - 5501	2501–5000	150.5–250.4	355–424	201-300	Molto dannosa

Le informazioni sui livelli di particolato sono prese dal [30],i livelli di Co2 dal lavoro correlato [13], mentre i livelli di TVOC sono stati presi da lavori che si riferivano alla German Federal Environmental Agency [6] .

Tabella 3.4: Breakpoints per IAQ.

3.8.3 Prossimo capitolo

Il progetto successivamente descritto partirà con un pulizia dei dati in modo da strutturare il dataset adeguatamente per la realizzazione delle previsioni. Prima di applicare gli algoritmi di machine learning verrà eseguita un'analisi dei dati per capire la distribuzione degli agenti inquinanti all'interno del set di dati e le loro correlazioni. L'applicazione dei modelli di regressione verrà effettuata gradualmente, partendo da semplici modelli di regressione lineare per poi passare a modelli più complessi come il multilayer perceptron. Nella parte finale dell'elaborato si propone un'analisi dei risultati ottenuti.

Capitolo 4

Implementazione

4.1 Preparazione dei dati

Non possiamo adattare e valutare algoritmi di apprendimento automatico su dati grezzi, dobbiamo trasformare i dati per soddisfare i requisiti dei singoli algoritmi. Inoltre, bisogna rappresentare i dati nel miglior modo possibile comprendendone il significato dato che gli algoritmi non conoscono la struttura dei dati.
Per il progetto è stato utilizzato Google Colab un prodotto di Google Research che consente di scrivere ed eseguire codice Python arbitrario tramite il browser ed è particolarmente adatto per machine learning e data analysis. Più tecnicamente, Colab è un servizio di notebook Jupyter ospitato che non richiede alcuna configurazione per l'uso, fornendo al contempo l'accesso gratuito alle risorse di elaborazione, comprese le GPU. In questo progetto viene utilizzato pandas [28], una libreria di python di uso comune che offre svariate strutture dati e funzionalità per l'analisi di dati strutturati. I dati vengono serviti su un formato csv, il formato più comune per la rappresentazione dei dati tabulari, ogni riga viene separata da un carattere specifico, spesso la virgola. I dati a nostra disposizione sono stati caricati su Bitbucket e successivamente caricati su Google Colab.
I dati grezzi importati contengono i seguenti valori: Unixtime , data, ora, latitudine, longitudine, temperatura(C), umidità(%), Pressione(hPa), Co2(ppm), RawH2 , RawEthanol, TVOC(ppb), HCHO(ppm), PM1(μg/m^3), PM2.5(μg/m^3), PM4(μg/m^3), PM10(μg/m^3), vento1(m/s) e vento2(m/s). La parte di preprocissing si occupa di modificare il dataset in modo da strutturarlo adeguatamente alla realizzazione delle previsioni. In particolare si eliminano features non note nel momento in cui la previsione viene effettuata e alcuni record per i quali non è possibile effettuare la previsione perché presentano valori nulli nelle variabili target. All'interno dei dataframe sono presenti valori errati come ad esempio giorni e ore segnati come 1665/165/165 165:65:65, per questo queste righe verranno scartate, l'importazio-

ne del csv con il divisore dei valori con le virgole ha impostato valori float come PM_1, $PM_{2.5}$, PM_4, PM_{10}, umidità etc come Object, per questo si andranno a gestire tramite delle funzioni in modo da riuscire a conventirli nel formato corretto. Le registrazioni con dati nulli, che all'interno del dataframe sono segnati come Nan(Not a number) sono state scartate in modo da non compromettere i modelli di regressione che si andranno ad applicare. Infine, durante la fase di preprocessing è stata eseguita un'operazione molto importante, cioè quella di associare ad ogni registrazione interna la corrispettiva registrazione esterna, la funzione che è stata applicata si basa su una variabile aggiunta "datetime" ottenuta dai dati giorno e orario, le associazioni ottenute avevano uno scarto temporale al massimo di 2 secondi.
A seguito di una visualizzazione dei dati forniti dai sensori canarin disponibili dall'università di Bologna per dei possibili lavori futuri che saranno spiegati nelle conclusioni, si è notato che non sono disponibili le informazioni riguardanti al vento, per cui in questo progetto non sono state utilizzate.

4.2 Analisi dei dati

Prima di fornire i dati ad un algoritmo machine learning, è necessario ispezionarli per identificare i problemi e ottenere informazioni su di essi. La capacità predittiva del modello è elevata solo se anche la qualità dei dati forniti è elevata. Lo scopo dell'analisi dei dati per gli algoritmi che si andranno ad utilizzare si focalizza sulla modellazione e scoperta di conoscenza per scopi predittivi piuttosto che descrittivi. L'analisi dei dati è un'attività scientifica per comprendere e predire fenomeni di diverso interesse, la preparazione dei dati si occupa di raccogliere i dati dalle sorgenti, comprenderne la struttura ed il significato, di effettuare la trasformazione e pulizia dei dati, successivamente dobbiamo occuparci dell'estrazione del modello di conoscenza dai dati(modelli predittivi), della validazione e interpretazione della conoscenza estratta. Per l'esplorazione delle singole feature viene utilizzata la libreria matplotlib, una libreia python per la creazione di grafici di dati, matplotlib è spesso usato tramite la API pyplot con cui si creano i grafici aggiungendo progressivamente elementi, imitando le istruzioni per creare grafici in Matlab.

4.2.1 Esplorazioni singole feature

Si analizzano ora le singole variabili, in particolare si porrà attenzione sulla distribuzione delle variabili target. Partendo da un'analisi statistica che riassumono la tendenza centrale, la dispersione e la forma della distribuzione del set di dati.

4.2.2 Particolato

Una delle nostre variabili target di maggiore importanza è il particolato, le tabelle successive indicano l'analisi del particolato all'interno dell'auto(4.1) e all'esterno (4.2):

	PM1	PM2.5	PM4	PM10
Mean	8.110400	8.611582	8.640022	8.645709
sdt	4.729427	5.013274	5.024194	5.026499
Min	1.009114	1.067101	1.067102	1.067102
Max	27.006462	28.558334	28.558353	28.558353
Percentile 75%	10.742303	11.398311	11.452472	11.456490

Tabella 4.1: Analisi del particolato all'interno dell'auto

	PM1	PM2.5	PM4	PM10
Mean	16.814183	17.887896	17.974948	17.992354
sdt	9.456473	9.975060	10.024981	10.042415
Min	4.292527	4.539185	4.539190	4.539190
Max	48.873272	51.683353	51.684795	51.685051
Percentile 75%	22.756501	24.423219	24.553162	24.617472

Tabella 4.2: Analisi del particolato all'esterno dell'auto

Una prima analisi del particolato ne descrive i valori indicando la media, deviazione standard, valore minimo, valore massimo e percentile. Questa analisi fa notare che i valori esterni sono in media il doppio più alti di quelli interni, ma sopratutto che la distribuzione è molto simile. Alcuni valori massimi esterni sono relativamente alti con una massima per il PM_{10} di 51.685051, ma comunque il percentile indica che il 75% delle registrazioni è inferiore a 24.617472.
Sempre prendendo come esempio il PM_{10} visualizziamo i valori esterni e interni utilizzando un grafico a barre che mostra una sequenza di valori tramite barre di altezza proporzionale usato spesso nelle distribuzioni dei valori.

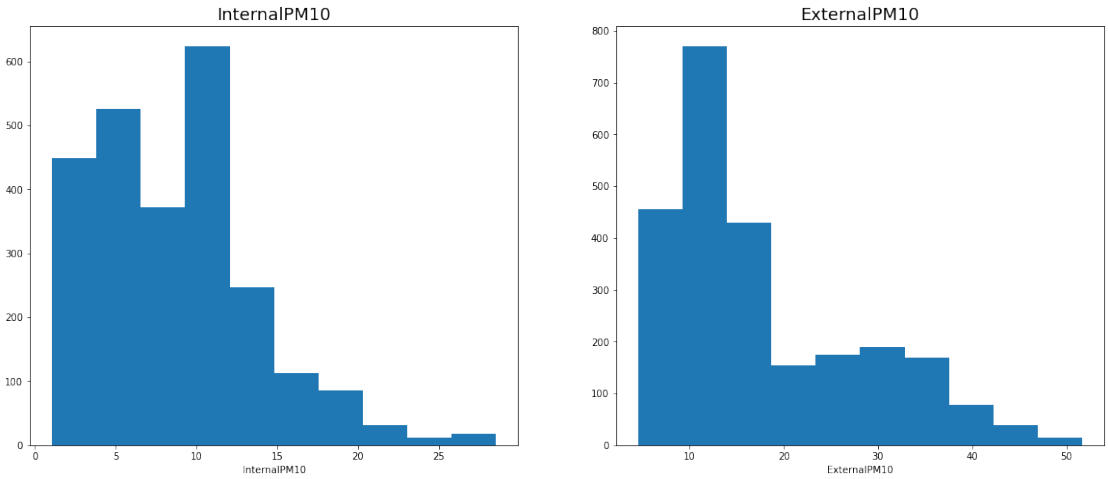


Figura 4.1: Istogramma dei valori interni ed esterni del PM_{10}

La distribuzione dei valori molto simili ci fa già capire che anche le correlazioni con le altre feature saranno molto simili e quindi anche gli score riportati dagli algoritmi di regressione.

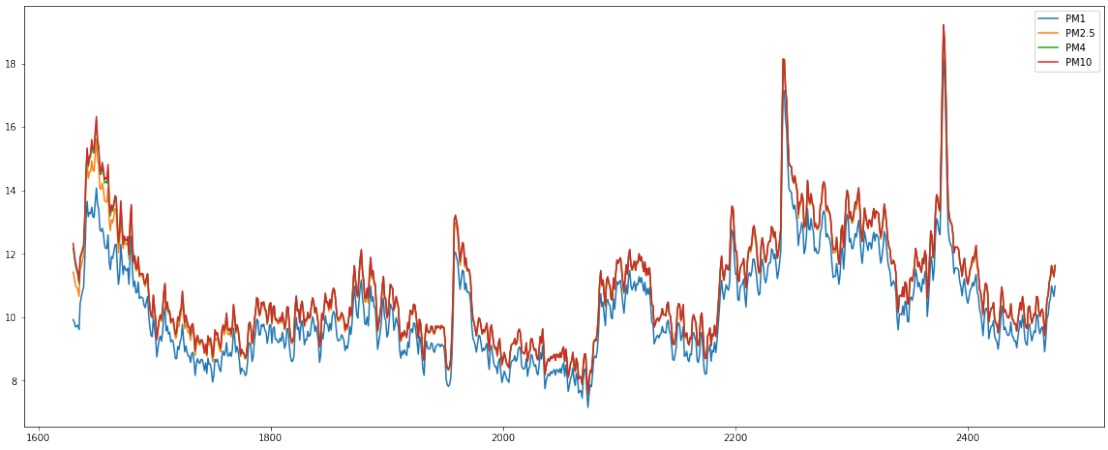


Figura 4.2: Distribuzione particolato nell'ambiente interno.

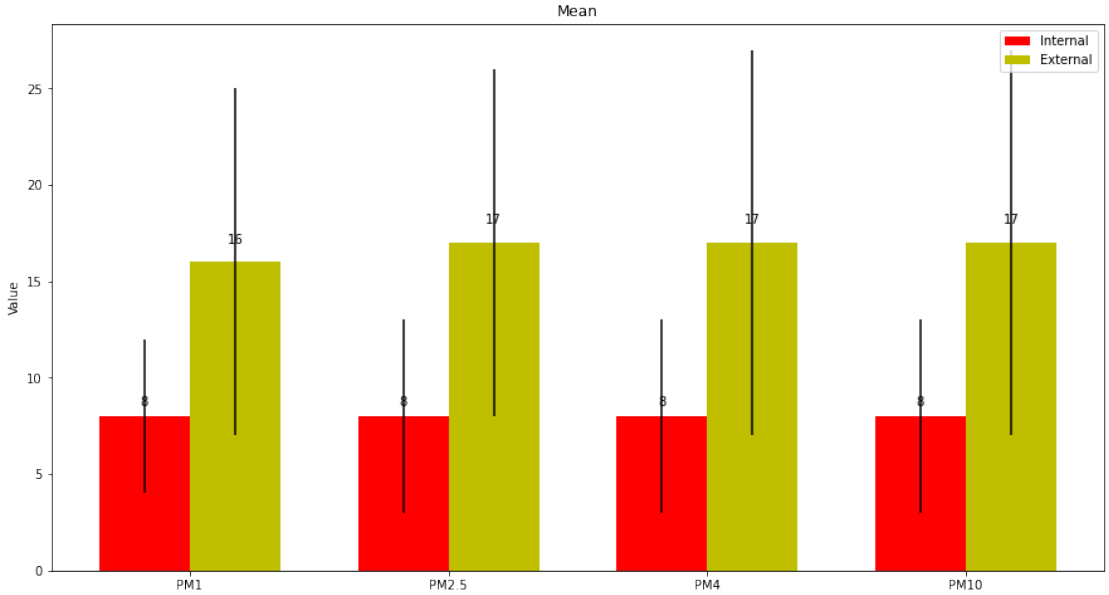


Figura 4.3: Confronto media dei valori interni ed esterni del particolato con deviazione standard.

4.2.3 Tvoc

Le registrazioni di TVOC descritte nella prossima tabella contengono anche valori TVOC a 0, la scelta di tenere queste registrazioni è data dalla volontà di non diminuire ancora di più il dataset e perché queste registrazioni possono essere veritiere nel momento in cui la macchina è stata appena messa in moto.

	Mean	std	Max	Percentile 75%
TVOC Interni	111.147	89.448	439	166
TVOC Esterni	236.193	275.590	2076	327

Tabella 4.3: Analisi valori TVOC esterni e interni

Anche i valori di TVOC interni ed esterni hanno una distribuzione simile a quella del particolato, con la differenza che i massimali raggiunti all'esterno sono molto più alti di quelli interni, che non superano il massimo di 439 ppb.

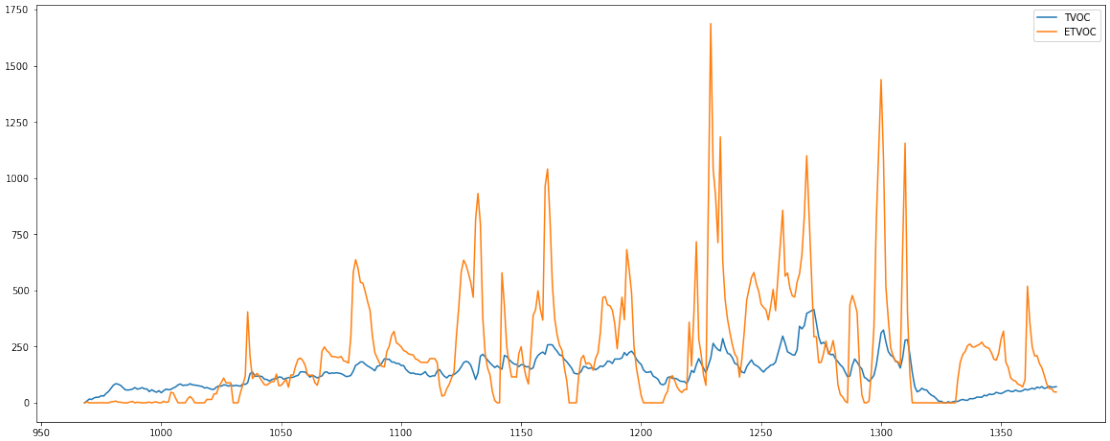


Figura 4.4: Visualizzazione grafica della distribuzione interna ed esterna del TVOC nel giorno 2021/11/23

Notiamo che le registrazioni interne di TVOC effettuate nella mattinata del 2021/11/23 seguono l'andamento di quelle esterne ma comunque senza raggiungere i picchi di quelli esterni.

4.2.4 CO2, H2, Ethanol e HCHO

A differenza del particolato e TVOC, le registrazioni di etanolo e H2 non hanno questa distribuzione, infatti i livelli interni ed esterni sono molto simili tra loro. Nel caso invece del HCHO e della CO_2 la relazione è contraria, infatti i valori interni sono più alti, nel caso dell'HCHO quasi il doppio, mentre i valori di Co2 interni sono quasi il triplo più alti.

	Mean	std	Max	Min	Percentile 75%
Etanolo Interno	19048.399	290.553	19749	18570	19246
Etanolo Esterno	19628.021	361.894	20776	18453	19907
H2 Interno	14286.471	298.965	13581	14962	14962
H2 Esterno	14223.348	248.874	14731	13370	14431
HCHO Interno	0.019723	0.010428	0.050000	0.00	0.031000
HCHO Esterno	0.010029	0.001309	0.014000	0.008000	0.011000
Co2 Interna	1549.419	837.137	3672	692	1977.5
Co2 Esterna	547.981	65.036	701	401	598.5

Tabella 4.4: Analisi degli agenti inquinanti e sostanze

4.2.5 Correlazione tra le variabili

Per misurare la correlazioni tra le variabili viene utilizzato il coefficiente di correlazione di Pearson, una misura specifica usata nell'analisi della correlazione per quantificare la forza della relazione lineare tra due variabili. Date due variabili casuali X e Y, il coefficiente è il rapporto tra la loro covarianza(un valore numerico che fornisce una misura di quanto le due varino assieme, ovvero della loro dipendenza) σ_{XY} e il prodotto delle deviazioni standard(una stima della variabilità di una variabile) σ_X e σ_Y

$$\rho(X, Y) = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

Data una serie di campioni $(x_1, y_1), \ldots, (x_n, y_n)$ delle due variabili con medie μ_X e μ_Y, la correlazione si può stimare come

$$\frac{\sum_{i=1}^{n} (x_i - \mu_X) \cdot (y_i - \mu_Y)}{n \cdot \sigma_X \cdot \sigma_Y}$$

Il coefficiente di correlazione r è un valore compreso tra -1 e 1. Più r si avvicina a zero, più la correlazione lineare è debole. Un valore r positivo è indice di una correlazione positiva, in cui i valori delle due variabili tendono ad aumentare in parallelo. Un valore r negativo è indice di una correlazione negativa, in cui il valore di una variabile tende ad aumentare quando l'altra diminuisce. I valori 1 e -1 rappresentano le correlazioni "perfette", una positiva e l'altra negativa. Due variabili perfettamente correlate mutano insieme a velocità fissa. In questo caso, si dice che hanno una relazione lineare perché, se inseriti in un grafico a dispersione, tutti i punti di dati possono essere collegati tra loro tramite una linea retta.

Servendosi di una heatmap viene ora mostrato il grado di correlazione esistente tra le variabili.

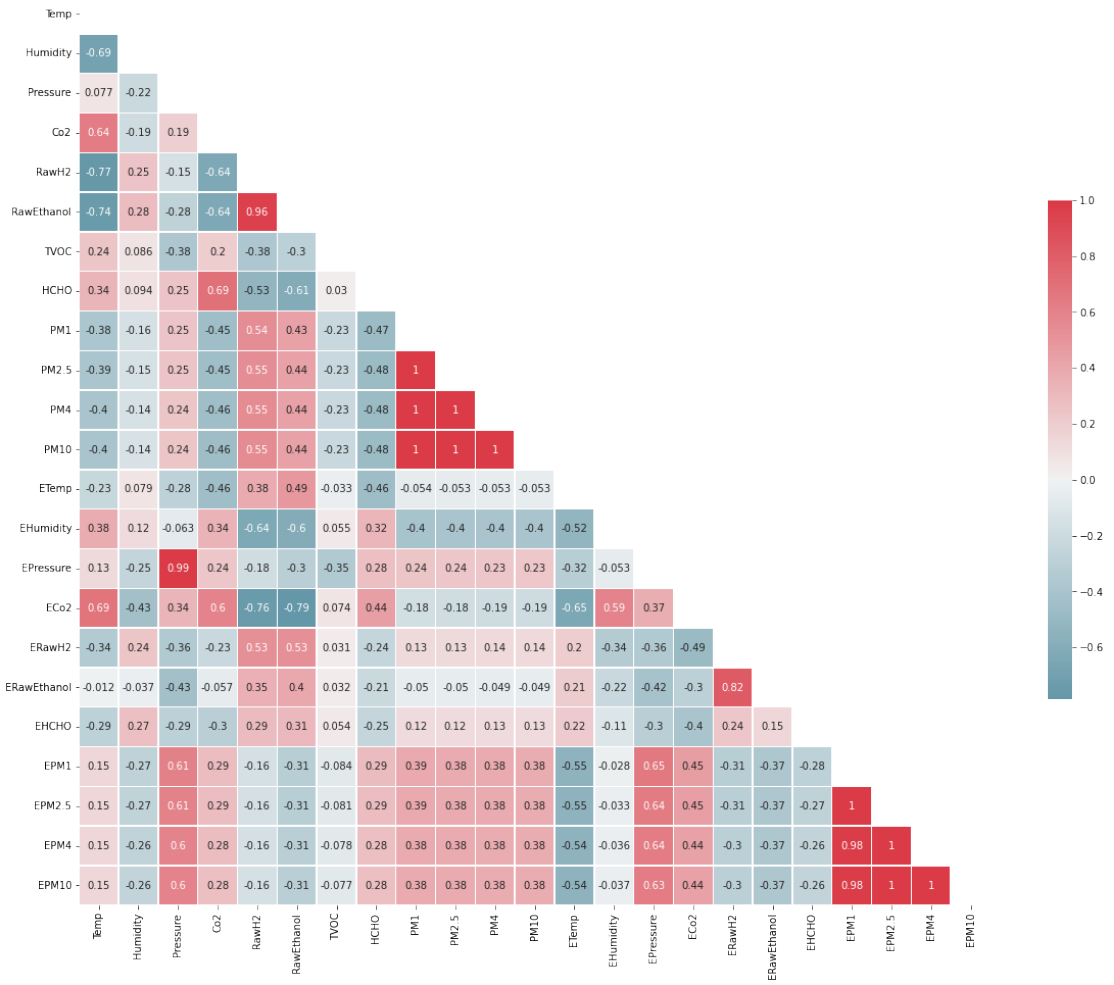


Figura 4.5: Heatmap con il grado di correlazione esistente tra le variabili.

4.3 Preparazione dataset

Esistono librerie che implementano algoritmi di machine learning per i vari tipi di regressione, tra queste è nota scikit-learn, una libreria general purpose per il machine learning, i passi generali per l'uso di scikit learn sono: creare un modello definendo i parametri dell'algoritmo di addestramento(detti anche iperparametri), addestrare il modello fornendo un insieme di dati da cui apprendere i migliori parametri del modello, effettuare predizioni fornendo nuovi dati al modello appreso e verificare le risposte che fornisce.
Rappresentate le feature e le loro relazioni con le variabili target, si passa ora alla preparazione del dataset per le fasi successive. Il training set è utilizzato per adde-

strare il modello di regressione, minimizzando l'errore su di esso. Il test set è usato dopo l'addestramento per verificare l'errore del modello su dati ignoti(capacità di generalizzazione).

```
X_train, X_Test, Y_train, Y_Test = train_test_split(X, Y
    , test_size=0.33, random_state=99)
```

In questo caso abbiano diviso il set di dati applicando la funzione train_test_split di scikit-learn per partizionare casualmente un set di dati in due set disgiunti, in questo caso è stata specificata come matrice di input X contenente i dati del nostro dataset senza le variabili target mentre Y sono i vettori degli output attesi (le nostre variabili target), il training set è costituito dal 67% dei dati, mentre il test set dal 33%, è stato inoltre specificato il seed random_state = 99(specifica la sequenza dei numeri random) che indica la mescolanza applicata ai dati prima di applicare la divisione, per ottenere gli stessi risultati indipendentemente dell'esecuzione.
Dal training set verrà estratto il modello mentre dal test set verranno misurate le prestazioni su dati ignoti in fase di addestramento. Inoltre verrà utilizzato lo StandardScaler che standardizza le variabili numeriche per far si che tutte le variabili abbiano la stessa distribuzione e la stessa importanza. Per ogni applicazione dei modelli che si andranno ad utilizzare bisogna considerare che all'aumentare della complessità del modello di learning si riduce l'errore, ma dopo una certa soglia di complessità l'errore sul validation set torna a crescere, bisogna quindi cercare di evitare problemi di underfitting dove il modello è troppo semplice e quindi inadeguato a rappresentare i dati (è insoddisfacente sia l'errore sul training, sia sul test) o overfitting dove il modello è troppo complesso, l'errore sul training è significativamente inferiore a quello sul validation quindi il modello non si generalizza dal training e non rappresenta adeguatamente dati ignoti.

4.4 Machine Learning Algorithms

In questa sezione si procede alla realizzazione di modelli di regressione in grado di predire la concentrazione interna del singolo inquinante. Per ogni addestramento di un modello viene utilizzata la K-Fold Cross Validation, dove i dati sono sono divisi in k sottoinsiemi disgiunti, un sottoinsieme è usato come validation set e gli altri k-1 come training set. Dalla K-Fold Cross Validation si ottengono k modelli di learning dai quali si sceglie la combinazione degli iperparametri che ottiene l'accuratezza migliore e si estrae il modello eseguendo l'addestramento dell'intero training set per poi valutarlo sul test set. La scelta di utilizzare la K-cross fold validation è stata presa per avere una scelta più robusta degli iperparametri viste le dimesioni ridotte del dataset. Nel nostro caso si è scelto di utilizzare una divisione in 5 fold implementando la seguente fuzione:

```
def grid_search_with_cross_validation(model, grid, X_train,
    y_train,X_val,y_val,n_folds=5):
    kf = KFold(n_folds,shuffle=True,random_state=99)
    grid_search = GridSearchCV(model, grid, cv=kf, n_jobs=-1)
    grid_search.fit(X_train, y_train)
    score = grid_search.score(X_val, y_val)
    print("Best cross validation score:{}\n".format(grid_search
        .best_score_))
    print("Test set score: {}\n".format(score))
    print("Best params: {}\n".format(grid_search.best_params_))
    return grid_search
```

Per misurare l'accuratezza dei modelli sono state utilizzate cinque metriche diverse:

- MAE(Mean Absolute Error) è spesso conosciuto anche come L1 Loss, e matematicamente rappresenta la distanza tra il valore predetto e quello effettivo.
- MSE (Mean Squared error) ovvero la media dei quadrati delle differenze tra ciascun valore reale e la corrispondente predizione, l'esponente determina una particolarità dell'errore quadratico medio dove i grandi errori sono fortemente penalizzati.
- RMSE (Root Mean Square Error) la radice della media dei quadrati degli scostamenti tra il valore vero e il valore predetto.
- Il coefficiente di determinazione R^2 indica la proporzione tra variabilità dei dati e correttezza del modello.
- L'errore relativo è stato inserito poiché è una metrica d'errore facilmente interpretabile che misura intuitivamente di quanto il modello si sbaglia in percentuale rispetto al valore reale.

4.4.1 Modelli di regressione lineare

Il primo algoritmo preso in considerazione è una semplice regressione lineare. Si è scelto di partire da un algoritmo semplice per avere una stima iniziale dell'accuratezza che verrà poi migliorata adottando man mano modelli più complessi. In un modello di regressione lineare, il valore della variabile dipendente è previsto come combinazione lineare delle variabili indipendenti, ciascuna variabile indipendente x_i è moltiplicata per un coefficiente θ_i, viene aggiunto un termine noto (intercetta) θ_0.

$$\hat{y} = \theta_0 + \theta_1 \cdot x_1 + \ldots + \theta_n \cdot x_n$$

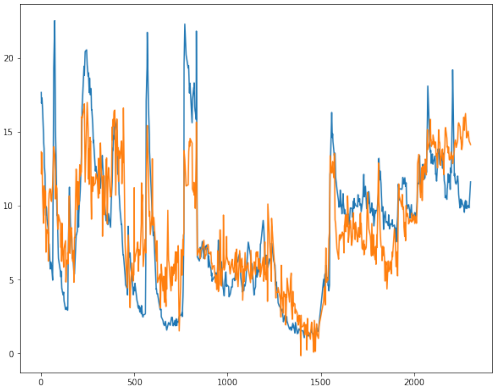
Per ogni singolo inquinante e sostanza è stato raggiunto questo risultato:

	MAE	MSE	Relative Error	R^2	RMSE
PM1	2.0307	6.9908	36.50407%	0.63158	2.644
PM2.5	2.156	7.8656	36.60028%	0.63108	2.8046
PM4	2.1633	7.9093	36.68115%	0.63062	2.8123
PM10	2.1649	7.9185	36.69821%	0.63053	2.814
Co2	345.92	-	28.06574%	0.72972	418.67
TVOC	50.933	-	-	0.479158	65.755
HCHO	0.0052555	-	-	0.6084	0.0064085
RawH2	67.026	7840.0	0.47013%	0.91358	88.544
RawEthanol	61.283	6867.3	0.32218%	0.92001	82.869

Tabella 4.5: Valutazione dei modelli di regressione lineare

Si nota come un semplice modello lineare non sia adatto alla previsione precisa del singolo agente inquinante nei casi in cui i dati non hanno un andamento interamente lineare, infatti molti agenti inquinanti e sostanze come il particolato, CO_2, TVOC e HCHO hanno dei valori relativamente bassi rispetto ai valori che si andranno a descrivere con modelli più complessi. Al contrario si nota come H2 e etanolo producano dei buoni risultati anche tramite una semplice regressione lineare, questo risultato può derivare dalla semplicità nel riuscire a predire queste sostanze. Per avere un'idea in forma grafica delle regressioni effettuate vengono mostrate nella figura (a) la regressione lineare della CO_2, mentre nella figura (b) la regressione lineare del PM_{10}.

(a) Visualizzazione grafica della regressione Lineare per la CO_2



(b) Visualizzazione grafica della regressione Lineare del PM_{10}

4.4.2 Modelli polinomiali

Proviamo ora a considerare dei modelli non lineari, con lo scopo di descrivere più accuratamente i dati. La regressione polinomiale è una generalizzazione di quella lineare in cui il modello include termini di grado superiore. La regressione polinomiale corrisponde in pratica a quella lineare con l'aggiunta di variabili derivate da quelle esistenti.
Il primo modello preso in considerazione è un modello polinomiale dove sono stati testati i polinomi di grado 2,4 e 6. In questa fase, non utilizzando funzioni Kernel, si tiene il grado basso per evitare la proliferazione delle feature. Successivamente si prova ad aggiungere regolarizzazione per ridurre la complessità e mantenere basso il valore dei coefficienti. Proviamo aggiungendo regolarizzazione L1, L2 e una combinazione delle 2 con un modello ElasticNet. La regressione ElasticNet combina insieme le regolarizzazioni L2 e L1 usate in ridge e lasso. Si applica in scikit-learn tramite la classe 'ElasticNet', per cui l'errore è calcolato come:

$$E = \underbrace{\frac{1}{2m}||X\theta - y||_2^2}_{\text{errore sui dati}} + \underbrace{\alpha\rho||\theta||_1}_{\text{L1}} + \underbrace{\frac{\alpha(1-\rho)}{2}||\theta||_2^2}_{\text{L2}}$$

I parametri impostabili sono:

- alpha(α) che determina il peso generale della regolarizzazione
- l1ratio (ρ, compreso tra 0 e 1) che determina il peso di L1 relativo al totale (con $\rho = 1$ si ha la regressione lasso, con $\rho = 0$ la ridge)

```
grid = [
        {
        "poly__degree" : [2,4,6],
        "reg" : [LinearRegression()]
    },
    {
        "reg" : [ElasticNet()],
        "reg__l1_ratio": [0.2,0.5,0.8],
        "reg__alpha": [0.01, 0.1, 1, 10]
    },
    {
        "poly__degree" : [2,4,6],
        "reg" : [Lasso(),Ridge()],
        "reg__alpha": [ 0.01, 0.1, 1, 10]
    }
]
```

```
polynomial_model = Pipeline([
        ("scale", StandardScaler()),
        ("poly", PolynomialFeatures(include_bias=False)),
        ("reg",LinearRegression())
])
```

Questa applicazione è stata utilizzata su tutti gli agenti senza ottenere miglioramenti significativi. L'utilizzo della regolarizzazione è stata comunque molto utile poiché si era notato per la previsione delle concentrazioni di alcuni agenti inquinanti un comportamento irregolare del modello in casi estremi ed aumentando la regolarizzazione è stato raggiunto un comportamento regolare.

Come ultimo modello non lineare viene preso in considerazione l'utilizzo delle funzioni kernel che ci permettono di aumentare il grado senza introdurre maggiore complessità senza l'aggiunta di variabili, gli iperparametri presi in considerazione sono stati : il grado(fino ad un massimo di 10), il tipo di kernel(poly o rbf) e l'alfa che indica la forza della regolarizzazione, migliora il condizionamento del problema e riduce la varianza delle stime.

Nella tabella seguente (Tab. 4.6) si mostrano i risultati del modello migliore per ogni variabile:

	MAE	MSE	Relative Error	R^2	RMSE
PM1	0.80249	1.7625	13.20299%	0.90711	1.3276
PM2.5	0.84554	1.9731	13.06806%	0.90746	1.4047
PM4	0.84911	1.9886	13.08205%	0.90713	1.4102
PM10	0.84974	1.9908	13.08367%	0.90711	1.411
Co2	87.02	-	6.05762%	0.96538	149.84
TVOC	149.84	-	-	0.87405	32.335
HCHO	0.0025696	-	-	0.87578	0.0036093
RawH2	39.833	3593.7	0.27947%	0.96039	59.947
RawEthanol	37.317	2765.5	0.19590%	0.96779	52.588

Tabella 4.6: Valutazione dei modelli di regressione polinomiale per ogni variabile

Si nota come l'aumento di grado restituisca modelli più accurati, per cercare di migliorare ulteriormente i modelli si testa ora l'utilizzo degli alberi di regressione.

4.4.3 Alberi di regressione

Il primo modello preso in considerazione è un semplice DecisionTree, l'obiettivo del DecisionTree è creare un modello che preveda il valore di una variabile target

apprendendo semplici regole decisionali dedotte dalle caratteristiche dei dati.
Gli iperparametri che abbiamo scelto di valutare sono:

- max_depth: indica la massima profondità del Decision Tree, definita in livelli.
- min_samples_split: numero minimo di istanze che un nodo deve avere prima di essere diviso;

La complessità del modello aumenta all'aumentare della profondità dell'albero di regressione, bisogna fare comunque attenzione perché aumentando la complessità si può generare overfitting. Dai primi risultati ottenuti si nota che un modello ad albero potrebbe essere più indicato per la previsione dei singoli inquinanti, di seguito invece di visualizzare tutti i valori di valutazione viene indicato solo il coefficiente di valutazione R^2 sapendo che è possibile ancora migliorare il risultato con più alberi di regressione.

	PM1	PM2.5	PM4	PM10	Co2	TVOC	HCHO	RawH2	RawEthanol
R^2	0.93	0.93	0.93	0.93	0.96	0.89	0.92	0.97	0.98

Tabella 4.7: Coefficiente di determinazione R^2 ottenuti dal DecisionTree

I prossimi due modelli proposti invece che utilizzare un singolo albero di regressione utilizzano foreste. Un gruppo di Decision Trees, ognuno con un diverso sottoinsieme casuale del training set, che forniscono una previsione finale è chiamato Random Forest ed è uno dei più potenti algoritmi di Machine Learning, nonostante la sua semplicità. Random Forest combina un certo numero di alberi decisionali, addestra ognuno di essi su un insieme leggermente differente di osservazioni ed effettua le partizioni considerando un numero limitato di caratteristiche. Le predizioni finali del Random Forest derivano dalla media delle predizioni di ogni albero decisionale, il che significa che tutti gli alberi hanno la stessa influenza nel determinare la decisione finale. Un iperparametro estremamente importante da utilizzare nelle foreste di regressione è n_estimator che sono il numero di alberi creati dal random forest, l'implementazione su sckikit-learn offre di default un numero abbastanza limitato (100) che è stato aumentato per aumentare l'accuratezza. L'utilizzo di random forest ha migliorato l'accuratezza del modello, perciò si è provato ad ottenere un risultato ancora migliore utilizzando algoritmi di gradient boosting, in particolare XGBoost. XGBoost, come il random forest, si basa sul concetto di ensemble, ovvero sfrutta una aggregazione di modelli di conoscenza per ottenere una predizione più accurata, ogni nuovo albero viene addestrato sugli errori del precedente in un processo detto boosting. Si tratta di un processo altamente efficiente, in quanto aiuta a implementare procedure di valutazione più efficaci, inoltre XGBoost utilizza un processo di gradient boosting in cui riduce al minimo gli errori attraverso l'algoritmo di discesa del gradiente.

Di seguito viene raffigurata un esempio degli iperparametri scelti:

```
grid = [
        {
            'tree__colsample_bytree': [0.5,0.6],
            'tree__learning_rate': [0.01],
            'tree__max_depth': [10,15,20,25],
            'tree__min_child_weight': [10,12],
            'tree__n_estimators': [400,600,800,1000],
            'tree__reg_alpha': [0,1,2],
            'tree__reg_lambda': [0,1,3]
        }
]
xgb_model = Pipeline([
      ("tree", XGBRegressor(random_state=99,objective='
         reg:squarederror'))
])
```

Gli iperparametri passati all'XGBoost sono:

- colsample_bytree: che indica il rapporto di sottocampioni delle colonne durante la costruzione di ogni albero.
- leatnin_rate: inidica la velocità di apprendimento, è un parametro di ottimizzazione che determina la dimensione del passo ad ogni iterazione mentre si sposta verso un minimo di una funzione di perdita.
- max_depth : indica la massima profondità.
- min_child_weight: indica il peso di ogni nodo figlio.
- n_estimator : numero di alberi da testare.
- reg_alpha : Regolarizzazione L1.
- reg_alpha : Regolarizzazione L2.

Di seguito vengono riportati i risultati migliori per ogni variabile prodotti dal random forest o XGBoost:

	MAE	MSE	Relative Error	R^2	RMSE	Model
PM1	0.44305	0.461	6.51901%	0.9757	0.67897	RF
PM2.5	0.47994	0.54658	6.67303%	0.97436	0.73931	RF
PM4	0.48015	0.55448	6.57497%	0.9741	0.74463	RF
PM10	0.47253	0.53039	6.51195%	0.97525	0.72828	RF
Co2	34.477	3613.7	2.17131%	0.99443	60.114	XGB
TVOC	10.362	301.2	-	0.96372	17.355	RF
HCHO	0.0012301	-	-	0.96171	0.0020039	XGB
RawH2	19.361	1129.4	0.13548%	0.98755	33.607	RF
RawEthanol	18.228	864.89	0.09562%	0.98993	29.409	XGB

Tabella 4.8: Valutazione delle regressioni migliori ottenute da XGBoost e Random forest

Utilizzando RandomForestRegression, si nota che nei casi della CO_2 e TVOC, il criterio "absolute_error", utilizzato al posto dello "squared_error" di default, ottiene risultati migliori. Nella figura successiva si nota come la regressione del PM_4 viene fatta con dei risultati molto buoni ed evidenzia la fatica nel riuscire ad identificare i valori di picco.

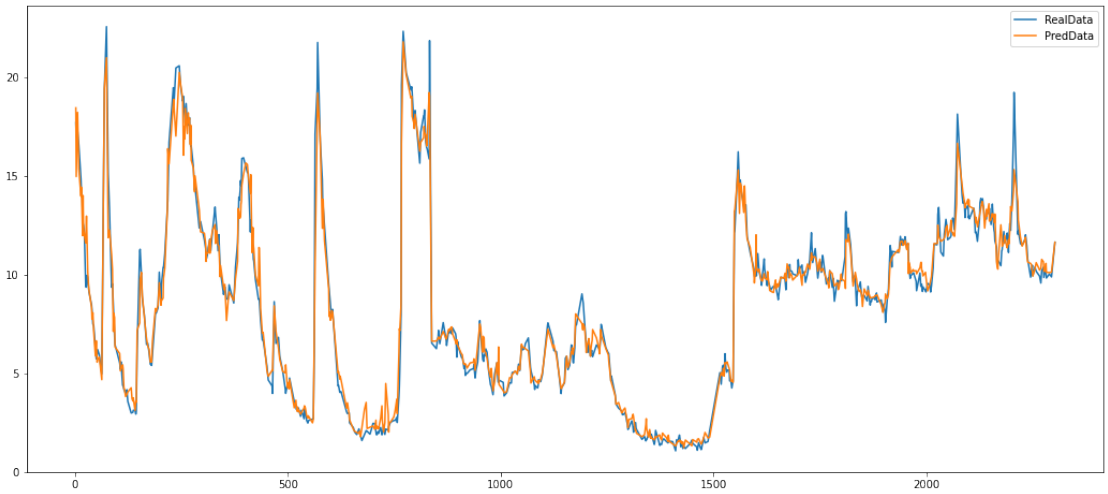


Figura 4.7: Differenze valori real e predetti del PM_4.

Una delle regressioni migliori è stata sicuramente la CO_2 con un coefficiente R^2 altissimo e un MAE di 34.477 che su valori medi molto alti della CO_2 significa che è un risultato notevole.

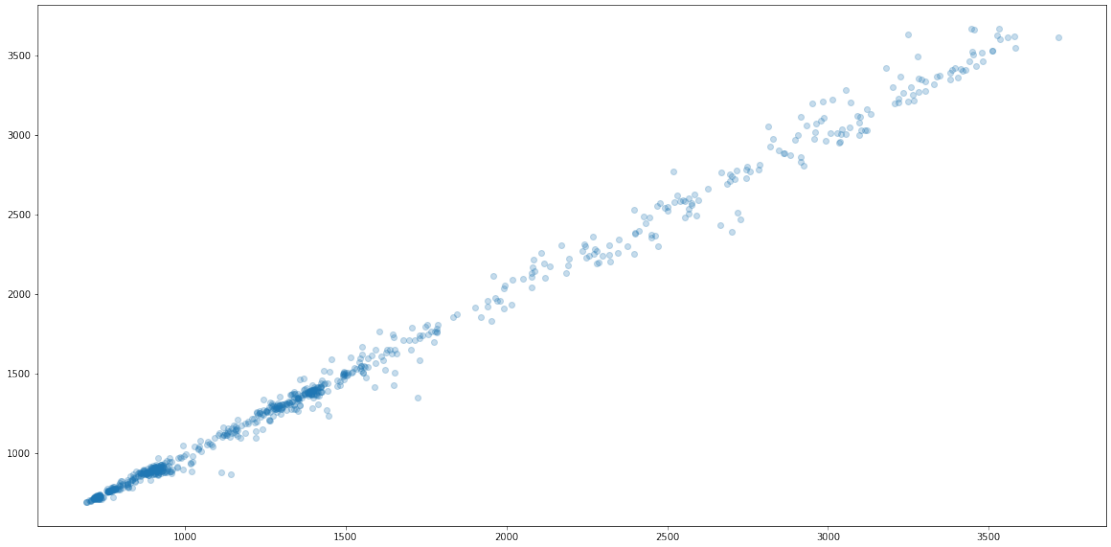


Figura 4.8: Grafico scatter della CO_2 con valori reali e predetti.

4.4.4 Rete neurale MLP

Quest'ultima applicazione è ancora in fase di sviluppo in quanto necessita di conoscenze profonde per ottenere risultati ottimali, comunque viene proposta per dare un'indicazione sui risultati che può ottenere. Una rete neurale è un modello di apprendimento costituito da molteplici livelli di neuroni elementari, ciascun neurone è in pratica un modello di regressione, i cui input sono forniti dal livello precedente e il cui output al quale è stata applicata una funzione di attivazione viene passato ai neuroni del livello successivo. MLP(Multilayer Perceptron) è una rete feedforward dove le connessioni collegano i neuroni di un livello con i neuroni di un livello successivo, MLP ha almeno 3 livelli e utilizza funzioni di attivazione non lineari. Ci sono molti aspetti personalizzabili (iperparametri) nella configurazione e nell'addestramento di una rete, in particolare nella definizione della struttura: numero di livelli, numero di nodi in ciascun livello e la funzione di attivazione. Un'importante iperparametro è l'algoritmo utilizzato per l'apprendimento dei pesi delle connessioni, il quale necessita di ulteriori iperparametri ad esempio utilizzando un algoritmo sgd sono : il learning rate, il peso della regolarizzazione e le dimensioni del minibach. Alcuni degli iperparametri utilizzati per ottenere la migliore configurazione sono il numero di neuroni negli hidden layer, le funzioni di attivazione testate sono state la relu(Unità lineare rettificata), la funzione della tangente iperbolica "tanh" e la funzione di attivazione logistica(sigmoide). Per l'ottimizzatore del peso chiamato "solver" sono stati utilizzati sgd(discesa stocastica del gradiente) e adam. E' stato impostato "early stopping" che indica

l'interruzione anticipata quando il punteggio sul validation set non migliora.
Di seguito viene proposto un esempio di implementazione :

```
grid = {
                "MLP__hidden_layer_sizes":
                   [50,75,100,120],
                "MLP__activation": ["tanh","relu"],
                "MLP__solver": ["sgd","adam"],
                "MLP__early_stopping": [True],
                "MLP__alpha": [0.0001,0.0005],
                "MLP__learning_rate": ['adaptive',"
                   constant"],
                "MLP__learning_rate_init": [0.01,0.001],
                "MLP__batch_size": [3,5],
                "MLP__max_iter": [1000],

}
MLPmodel = Pipeline([
      ("scale",StandardScaler()),
      ("MLP", MLPRegressor(random_state=99))
])
```

Di seguito sono riportati i risultati ottenuti:

	MAE	MSE	Relative Error	R^2	RMSE
PM1	0.72016	1.141	11.42073%	0.94648	1.0682
PM2.5	0.67092	0.9765	10.82378%	0.9542	0.98818
PM4	0.69814	1.1553	10.82302%	0.94605	1.0748
PM10	0.7046	1.1194	11.40084%	0.94777	1.058
Co2	85.746	-	6.01466%	0.97157	135.79
TVOC	17.176	768.86	-	0.90738	27.728
HCHO	0.0031379	-	-	0.82423	0.0042935
RawH2	34.615	2670.9	0.24273%	0.97056	51.681
RawEthanol	40.159	3078.8	0.21094%	0.96414	55.487

Tabella 4.9: Valutazioni del modello MLP

Un'analisi dei risultati ottenuti ci indica che i risultati sono buoni con R^2 molto alti ed errori relativamente bassi, il fatto che non abbia raggiunto risultati ottimali come gli alberi di regressione possiamo imputarlo alla dimensione ristretta del

dataset, perché essendo MLP un algoritmo di machine learning ad alta complessità probabilmente per ottenere risultati ottimali necessità di più dati.

4.4.5 Visualizzazione grafica delle regressioni ottenute

Di seguito sono riportati i grafici delle regressioni degli agenti inquinanti e sostanze che nella fase di analisi del background di questo progetto sono definiti come i più incisivi per la salute del guidatore cioè : $PM_{2.5}$, PM_{10}, TVOC e CO_2. Nei prossimi grafici si rappresenta la retta che approssima i dati reali(linea rossa) confrontata a quella dei valori ottenuti dalla regressione(linea verde), insieme ad uno scatterPlot dei valori con un alpha a 0.25.

Regressione migliore PM_{10}

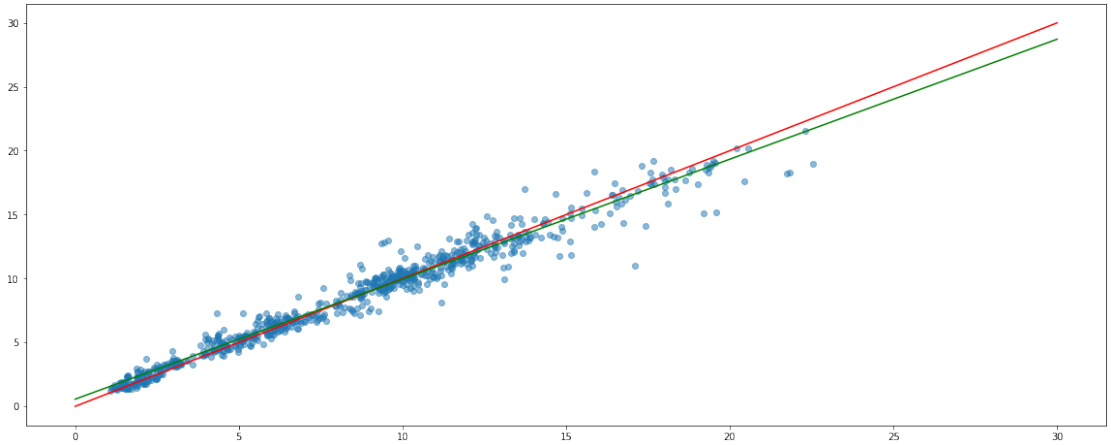


Figura 4.9: Visualizzazione grafica della regressione migliore per il PM_{10}.

Regressione migliore $PM_{2.5}$

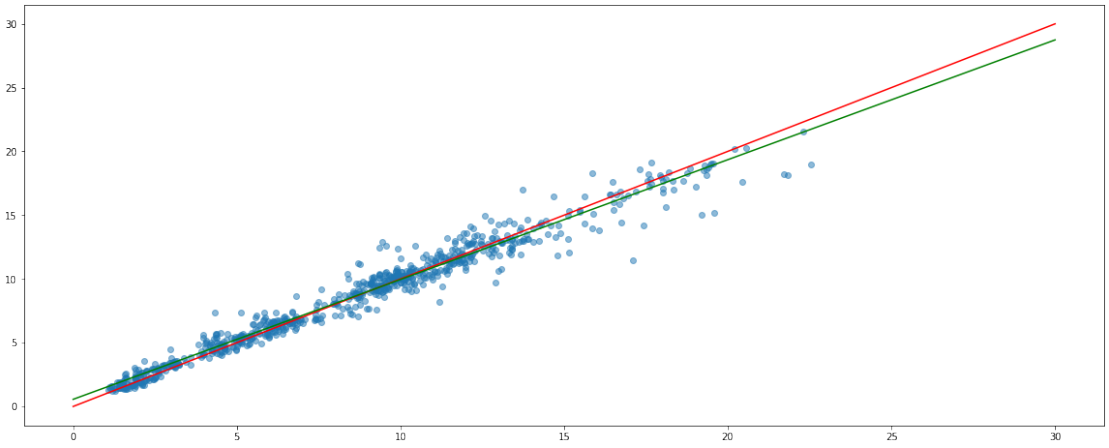


Figura 4.10: Visualizzazione grafica della regressione migliore per il $PM_{2.5}$.

Regressione migliore CO_2

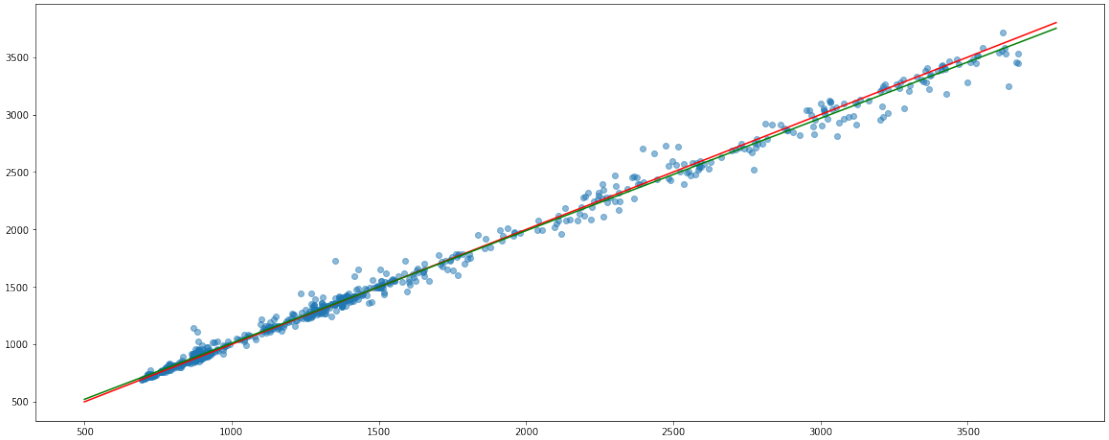


Figura 4.11: Visualizzazione grafica della regressione migliore per la CO_2.

Regressione migliore TVOC

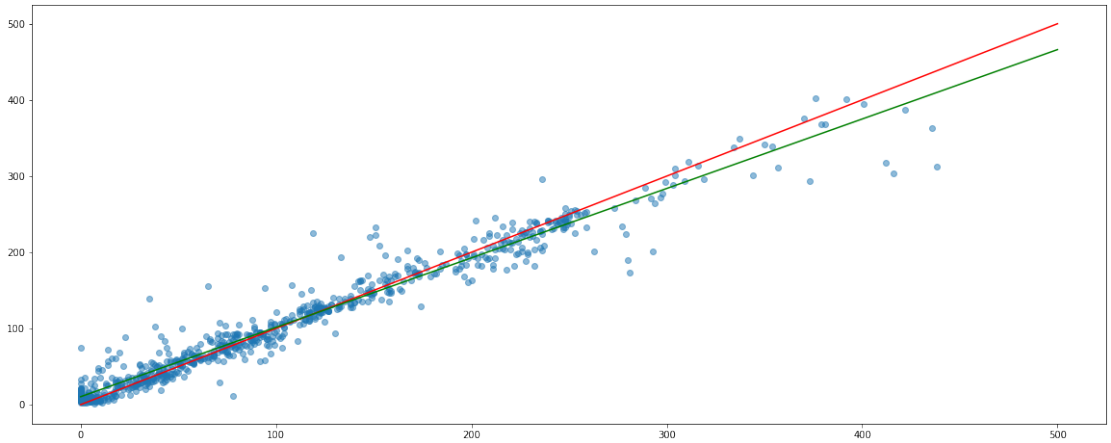


Figura 4.12: Visualizzazione grafica della regressione migliore per il TVOC.

4.4.6 Confronto regressioni con solo valori esterni

In questa sezione facciamo una valutazione delle regressioni ottenute confrontate con le regressioni fatte con i soli valori esterni.
Le regressioni fatte fino ad ora venivano prodotte inserendo in input le registrazioni esterne contenenti particolato, CO_2, TVOC, HCHO, etanolo, H2, temperatura, pressione e umidità, aggiungendo la temperatura, pressione e umidità interna. Questa scelta di aggiungere questa parte di registrazioni interne è stata obbligatoria in quanto il tema centrale di questo progetto sono degli agenti inquinanti che possono nuocere alla salute e necessitano di più precisione possibile. Inoltre pensando a possibili applicazioni future di questo progetto, dati come temperatura, pressione e umidità possono essere facilmente reperibili attraverso dispositivi come smartphone oppure dalle automobili. Di seguito vengono riportate le regressioni migliori ottenute dagli alberi di regressione con i soli valori esterni.

	MAE	MSE	Relative Error	R^2	RMSE
PM10	1.0243	2.6321	19.86346%	0.87719	1.6224
PM2.5	0.93637	2.486	19.40436%	0.86899	1.5767
PM4	1.028	2.6669	19.81078%	0.87545	1.6331
PM1	0.93555	2.5062	19.31690%	0.86792	1.5831
Co2	147.12	-	9.51253%	0.88715	270.53
TVOC	21.743	1157.3	-	0.86059	34.019
HCHO	0.0022614	-	-	0.86644	0.0037426
RawH2	40.509	4286.7	0.28468%	0.95275	65.473
RawEthanol	34.828	3141.1	0.18319%	0.96341	56.046

Tabella 4.10: Valutazioni dei modelli con soli valori esterni

Si nota come l'accuratezza delle regressioni ottenute con i soli valori esterni sia molto più imprecisa, infatti per il particolato gli errori relativi vengono più che triplicati ed il MAE quasi raddoppiato, prendendo in considerazione l'etanolo e H2 le regressioni risultano più che buone ma comunque peggiorate se confrontate con le regressioni ottenute con la temperatura, umidità e pressione interna. Importante notare soprattutto il valore MSE che si alza di molto, il che significa che ci sono differenza tra valori previsti e osservati che differiscono sostanzialmente.

4.4.7 IAQ Singolo Inquinante

Come ultima parte viene considerato il confronto tra l'AQI(Air Quality index)calcolato con i dati reali e l'AQI calcolato con i dati predetti. Non sono stati utilizzati gli algoritmi di regressione per predire l'AQI generale della qualità dell'aria interna perché i dati raccolti non sono abbastanza significativi. Infatti i livelli di particolato e TVOC registrati non rappresentano una minaccia per il guidatore, mentre i livelli di CO_2 sono significativamente alti per cui l'AQI interno sarebbe strettamente correlato ai livelli di CO_2 e perderebbe il suo significato. Viene comunque calcolato l'AQI del singolo inquinate per controllare che la distribuzione dei valori predetti segua in maniera soddisfacente l'AQI reale e soprattutto per aprire studi interessanti futuri che possano utilizzare questi risultati sull'indice di qualità dell'aria interna.

Per calcolare questo indice è stato utilizzata la formula fornita dall'US EPA((U. S. Environmental Protection Agency) spigata nel primo capitolo.

$$I_p = \frac{I_{Ji} - I_{Lo}}{BP_{Hi} - BP_{Lo}}(C_p - BP_{Lo}) + I_{Lo}$$

I_p = indice per inquinante p
C_p = la concentrazione arrotondata dell'inquinante p
BP_{Hi} = il breakpoint maggiore o uguale a Cp
BP_{Lo} = il breakpoint inferiore o uguale a Cp
I_{Hi} = il valore AQI corrispondente a BPHi
I_{Lo} = il valore AQI corrispondente a BPLo

Ed i seguenti valori:

TVOC(ppb)	CO_2 (ppm)	PM2.5($\mu g/m^3$)	PM10($\mu g/m^3$)	IV-AQI	classe
0 - 220	340–600	0.0–12.0	0–54	0-50	Buona
221-660	601–1000	12.1–35.4	55–154	51-100	Moderata
661-2200	1001–1500	35.5–55.4	155–254	101-150	Dannosa per gruppi sensibili
2201-5500	1501–2500	55.5–150.4	255–354	151-200	Dannosa
$\geq$ 5501	2501–5000	150.5–250.4	355–424	201-300	Molto dannosa

Le informazioni sui livelli di particolato sono prese dal [30],i livelli di Co2 dal lavoro correlato [13], mentre i livelli di TVOC sono stati presi da lavori che si riferivano alla German Federal Environmental Agency [6] .

Tabella 4.11: Breakpoints per l'IAQI.

Dai risultati si nota come gli indici siano molto simili, contando i problemi nel predire i valori di picco degli agenti inquinanti si nota come non sempre riesca a seguire l'indice reale ma comunque è un buon risultato e un punto di partenza per lavori futuri. Di seguito vengono visualizzati alcuni esempi in forma grafica:

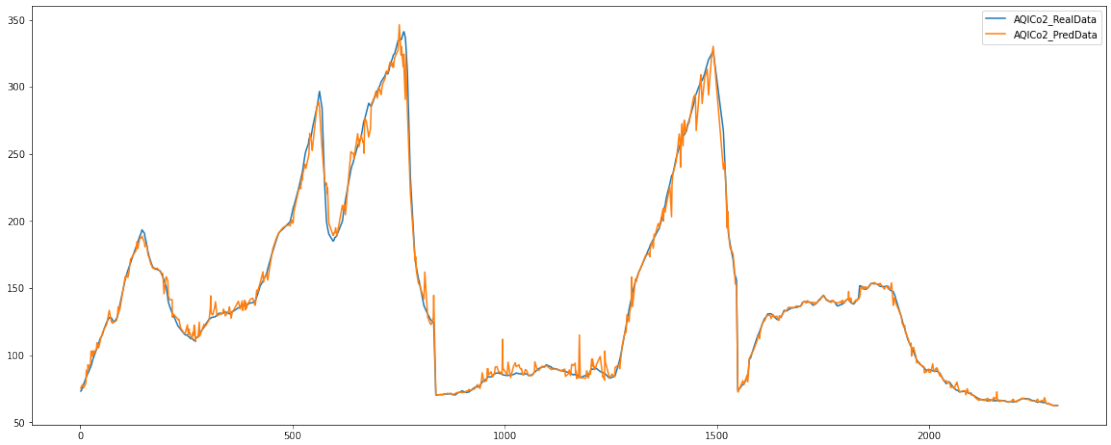


Figura 4.13: Confronto AQI della CO_2 calcolato con i valori reali e predetti.

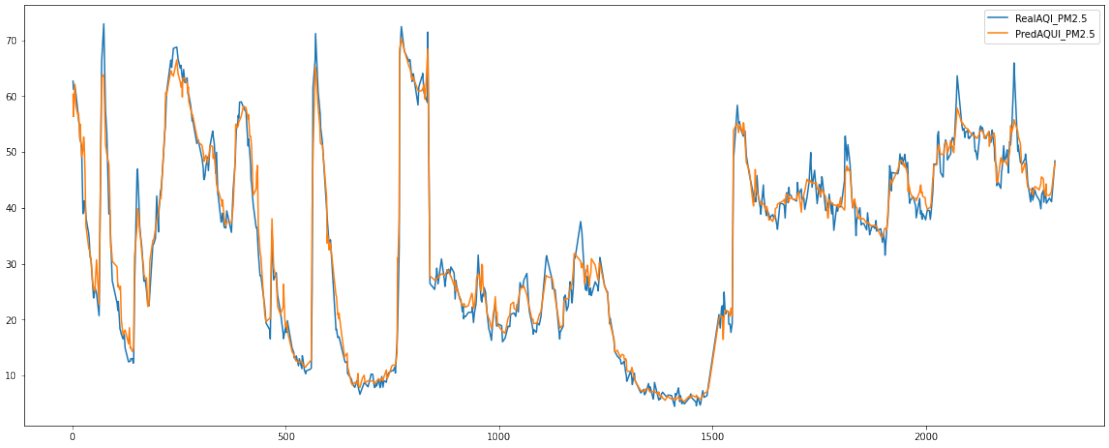


Figura 4.14: Confronto AQI del PM_{10} calcolato con i valori reali e predetti.

4.5 Considerazioni sui risultati ottenuti

Dall'analisi dei dati si nota come la distribuzione dei diversi tipi di particolato(PM_1, $PM_{2.5}$, PM_4, PM_{10}) sia molto simile, una distribuzione cosi simile può essere il risultato di una bassa sensibilità dei sensori nel momento delle registrazioni, comunque non si aspettavano grandi differenze da particelle di diametro $\leq$ di 10 micron. Dall' heatmap delle correlazioni si nota come i valori del particolato interno abbiano una discreta correlazione con i valori del particolato esterno, la temperatura e l'umidità esterna. Il primo algoritmo di regressione utilizzato evidenzia come una semplice regressione lineare non sia in grado di predire in modo preciso i valori con un errore relativo del 36%, infatti il coefficiente di determinazione R^2 che indica la proporzione tra variabilità dei dati e correttezza del modello ci dà un valore medio di 0.63, il che indica che ci sono molti margini di miglioramento. L'utilizzo di modelli non lineari incrementa notevolmente le prestazioni riuscendo a ottenere dei coefficienti R^2 sopra il 0.90, abbassando l'errore relativo intorno al 13% e con MAE e MSE rispettivamente del 0.84 e 1.9 circa. I risultati migliori si sono ottenuti con l'utilizzo degli alberi di regressione tra i quali sono stati testati i modelli di DecisionTree, RandomForest e per ultimo si è provato ad ottenere un risultato migliore utilizzando algoritmi di gradient boosting. Il risultato migliore è stato prodotto utilizzando il modello RandomForest (MAE 0.47 di media, errore relativo del 6.5% e R^2 sopra al 0.97). Anche la rete neurale MLP ha performato un modo ottimale ottenendo risultati leggermente peggiori (MAE del 0.7, errore relativo di circa 11% e R^2 del 0.95). All'interno del dataset abbiamo H2 e etanolo, la cui distribuzione e le correlazioni sono molto simili, anche la differenza delle registrazioni interne ed esterne sono minime, infatti l'etanolo ha una media

interna di 19048 ed esterna di 19628, mentre H2 ha una media interna di 14286 e una media esterna di 14223. La rivelazione più importante è che dopo l'utilizzo del primo algoritmo di regressione (regressione lineare) si hanno già dei risultati ottimali con un coefficiente di determinazione R^2 molto alto, per l'etanolo 0.92 e per l'H2 0.91, i valori MAE sembrano alti (61,283 per l'etanolo e 67.026 per l'H2), ma in realtà sono bassi in relazione all'ordine di grandezza in cui si presentano. MLP ha raggiunto risultati molto alti con un R^2 sopra a 0.96 per entrambi, ma il risultato migliore è stato ottenuto utilizzando gli alberi di regressione. Nello specifico XGBoost ha ottenuto il risultato migliore per l'etanolo con un MAE di 18.228, errore relativo del 0.09562% e un R^2 di 0.989 mentre per H2 ha performato meglio il random forest con un MAE del 19.361 , errore relativo pari allo 0.135% e R^2 di 0.987. TVOC si presenta all'interno del dataset con le correlazioni più basse tra tutti gli agenti inquinanti, infatti la correlazione diretta più alta si ha con la temperatura (0.24) mentre quella inversa più alta si ha con la pressione (-0.38). Questa informazione ci fa capire che la predizione risulterà più difficile, infatti partendo dalla regressione lineare è l'inquinante con il coefficiente di determinazione più basso tra tutti gli inquinanti, con un valore pari a 0.47, lo stesso anche per i modelli polinomiali. Un miglioramento della predizione è avvenuto con gli alberi di regressione, nello specifico il random forest, utilizzando come criterio (la funzione che misura la qualità della divisione dei dati) il Mean Absolute Error al posto del Mean Squared Error di default. Il risultato migliore è stato prodotto proprio dal random forest ottenendo un R^2 pari a 0.96, un MAE pari a 10.362 e un RMSE di 17.355. Non è stato calcolato l'errore relativo in quanto si è deciso in fase di progettazione di contare anche le registrazioni pari a 0 per due motivi principali: 1) Non si voleva diminuire ulteriormente il dataset, 2) potrebbero essere veritiere se fatte nel momento in cui si iniziavano le registrazioni. Anche per il HCHO non viene calcolato l'errore relativo in quanto si parla di ordini di grandezza estremamente piccoli, infatti la registrazione massima è di 0.05 ppm. La particolarità della formaldeide è che all'interno di questo dataset le misurazioni interne sono quasi il doppio più alte di quelle esterne. La predizione di HCHO tramite l'utilizzo di MLP non ha performato in modo ottimale raggiungendo un valore di R^2 pari a 0.82. La migliore regressione è avvenuta con XGBoost con un R^2 pari a 0.96 , MAE pari a 0.0012 e RMSE pari a 0.002. L'ultima sostanza presa in considerazione è la CO_2, la CO_2 insieme alla formaldeide ha dei valori interni più alti rispetto all'esterno, addirittura quasi il triplo più alti, con una media di 1549.419 rispetto a 547 dell'esterno. Nell'heatmap delle correlazioni infatti notiamo che ha delle forti correlazioni con quasi tutte le feature, è la sostanza che ha raggiunto il coefficiente di determinazione R^2 più alto in assoluto pari a 0.99443 e un errore relativo del 2.17131% utilizzando XGBoost. Ha comunque performato in maniera efficiente con MLP ottenendo R^2 pari a 0.97, ma comunque avendo

un errore relativo più alto. Inoltre, la CO_2 è la sostanza con i valori più alti in assoluto rispetto alle indicazioni per la qualità dell'aria interna, infatti se venisse calcolato l'AQI all'interno del veicolo la maggioranza dei valori sarebbero correlati strettamente alla CO_2.

Capitolo 5

Conclusioni

Il progetto che propone lo sviluppo di un sistema di monitoraggio della qualità dell'aria che ne rende possibile la predizione all'interno del veicolo tramite l'utilizzo di algoritmi di machine learning ha dato dei risultati positivi. Dai risultati ottenuti si può confermare che è possibile predire la concentrazione dei singoli agenti inquinanti all'interno del veicolo basandoci principalmente sulle registrazioni esterne. Questi risultati possono essere un punto di svolta per garantire la sicurezza stradale e sulle condizioni di salute del guidatore. La possibilità di essere utili a progetti europei come NextPerception amplifica ancora di più le direzioni che questo progetto può prendere, infatti si potranno prendere in considerazione le regressioni delle concentrazioni dei singoli inquinanti in modo da analizzarli uno ad uno ed eseguire dei controlli più specifici su di essi. Il progetto in questo momento non è del tutto completo e non può essere direttamente applicato ad un sistema reale in quanto ci possono essere dei miglioramenti e si potrebbero testare altri algoritmi presenti nello stato dell'arte. Una delle necessità più importanti dal quale si deve partire è quello di ampliare il dataset in quanto questo progetto è stato sviluppato su un dataset di piccole dimensioni. Un dataset più grande offrirà dei miglioramenti importanti per gli algoritmi di regressione utilizzati, soprattutto per i modelli più complessi come le reti neurali. Nel momento della scrittura di questa tesi l'università di Bologna ha già mosso i primi passi mettendo a disposizione dei kit di sensori Canarin in grado di eseguire registrazioni all'interno e all'esterno di un veicolo. La possibilità di eseguire nuove registrazioni può portare a sviluppare progetti correlati, mantenendo gli stessi obbiettivi di questo progetto, ma offrendo la possibilità di concentrarsi sulle caratteristiche del veicolo e ad esempio a varie correlazioni con il numero di occupanti.

Bibliografia

[1] Jude Adekunle Adeleke, Deshendran Moodley, Gavin Rens, and Aderemi Oluyinka Adewumi. Integrating statistical machine learning in a semantic sensor web for proactive monitoring and control. *Sensors*, 17(4), 2017.

[2] Davide Aguiari, Giovanni Delnevo, Lorenzo Monti, Vittorio Ghini, Silvia Mirri, Paola Salomoni, Giovanni Pau, Marcus Im, Rita Tse, Mongkol Ekpanyapong, and Roberto Battistini. Canarin ii: Designing a smart e-bike ecosystem. In *2018 15th IEEE Annual Consumer Communications Networking Conference (CCNC)*, pages 1–6, 2018.

[3] Timothy M. Amado and Jennifer C. Dela Cruz. Development of machine learning-based predictive models for air quality monitoring and characterization. In *TENCON 2018 - 2018 IEEE Region 10 Conference*, pages 0668–0672, 2018.

[4] Saba Ameer, Munam Ali Shah, Abid Khan, Houbing Song, Carsten Maple, Saif Ul Islam, and Muhammad Nabeel Asghar. Comparative analysis of machine learning techniques for predicting air quality in smart cities. *IEEE Access*, 7:128325–128338, 2019.

[5] Marjan Asgari, Mahdi Farnaghi, and Zeinab Ghaemi. Predictive mapping of urban air pollution using apache spark on a hadoop cluster. ICCBDC 2017, page 89–93, New York, NY, USA, 2017. Association for Computing Machinery.

[6] Atmotech. Standards for indoor air quality (iaq).

[7] Natasha Barnes, Tsz-Wai Ng, Kwok Ma, and Ka Lai. In-cabin air quality during driving and engine idling in air-conditioned private vehicles in hong kong. *International Journal of Environmental Research and Public Health*, 15:611, 03 2018.

[8] Jae-joon Chung and Hyun-Jung Kim. An automobile environment detection system based on deep neural network and its implementation using iot-enabled in-vehicle air quality sensors. *Sustainability*, 12(6), 2020.

[9] Aaron J. Cohen, Michael Brauer, Richard T. Burnett, Hugh Ross Anderson, Joseph Jon Frostad, Kara Estep, Kalpana Balakrishnan, Bert Brunekreef, Lalit Dandona, Rakhi Dandona, Valery L. Feigin, Greg Freedman, Bryan Hubbell, Amelia Jobling, Hai dong Kan, Luke D. Knibbs, Yang Liu, Randall V. Martin, Lidia Morawska, C. Arden Pope, Hwashin H. Shin, Kurt Straif, Gavin Shaddick, Matthew L. Thomas, Rita Van Dingenen, Aaron van Donkelaar, Theo Vos, Christopher J. L. Murray, and Mohammad Hossein Forouzanfar. Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: an analysis of data from the global burden of diseases study 2015. *Lancet (London, England)*, 389:1907 – 1918, 2017.

[10] Ministero della Salute. Composti organici volatili (cov), 2015.

[11] Boris Dessimond, Isabella Annesi-Maesano, Jean-Louis Pepin, Salim Srairi, and Giovanni Pau. Academically produced air pollution sensors for personal exposure assessment: The canarin project. *Sensors*, 21(5), 2021.

[12] Daghan Ekmekcioglu and S. Sinan Keskin. Characterization of indoor air particulate matter in selected elementary schools in istanbul, turkey. *Indoor and Built Environment*, 16(2):169–176, 2007.

[13] Chew Cheik Goh, Latifah Munirah Kamarudin, Ammar Zakaria, Hiromitsu Nishizaki, Nuraminah Ramli, Xiaoyang Mao, Syed Muhammad Mamduh Syed Zakaria, Ericson Kanagaraj, Abdul Syafiq Abdull Sukor, and Md. Fauzan Elham. Real-time in-vehicle air quality monitoring system using machine learning prediction algorithm. *Sensors*, 21(15), 2021.

[14] Anna Goshua, Cezmi Akdis, and Kari C Nadeau. World health organization global air quality guideline recommendations: Executive summary, 2021.

[15] Ade Silvia Handayani, Nyayu Latifah Husni, Rosmalinda Permatasari, and Carlos R Sitompul. Implementation of multi sensor network as air monitoring using iot applications. In *2019 34th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC)*, pages 1–4, 2019.

[16] IQAir. In-car air pollution, 2021.

[17] Marilena Kampa and Elias Castanas. Human health effects of air pollution. *Environmental Pollution*, 151(2):362–367, 2008. Proceedings of the 4th

International Workshop on Biomonitoring of Atmospheric Pollution (With Emphasis on Trace Elements).

[18] Balz Maag, Zimu Zhou, and Lothar Thiele. A survey on sensor calibration in air pollution monitoring deployments. *IEEE Internet of Things Journal*, 5(6):4857–4870, 2018.

[19] Ioannis Manisalidis, Elisavet Stavropoulou, Agathangelos Stavropoulos, and Eugenia Bezirtzoglou. Environmental and health impacts of air pollution: A review. *Frontiers in Public Health*, 8, 2020.

[20] Christian Meyer. Overview of tvoc and indoor air quality. *Retrieved from San Jose, CA*, 2018.

[21] Nick Molden, Sam Boyle, Kent Johnson, and Heejung Jung. Development of a standard testing method for vehicle cabin air quality index. *SAE International Journal of Commercial Vehicles*, 12(2):151–161, may 2019.

[22] L Mølhave, Geo Clausen, B Berglund, J De Ceaurriz, A Kettrup, T Lindvall, M Maroni, AC Pickering, U Risse, H Rothweiler, et al. Total volatile organic compounds (tvoc) in indoor air quality investigations. *Indoor Air*, 7(4):225–240, 1997.

[23] Rafia Mumtaz, Syed Mohammad Hassan Zaidi, Muhammad Zeeshan Shakir, Uferah Shafi, Muhammad Moeez Malik, Ayesha Haque, Sadaf Mumtaz, and Syed Ali Raza Zaidi. Internet of things (iot) based indoor air quality sensing and predictive analytic—a covid-19 perspective. *Electronics*, 10(2), 2021.

[24] Zulauf N. ndoor air pollution in cars: An update on novel insights. *Int J Environ Res Public Health*, 2019.

[25] NextPerceptionSite. Nextperceptionsite.

[26] World Health Organization. Occupational and Environmental Health Team. Who air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulfur dioxide : global update 2005 : summary of risk assessment, 2006.

[27] TEKNOLOGIAN TUTKIMU SKESKUS VTT OY. Nextperception - next generation smart perception sensor and distributed intelligente for proactive human monitoring in health,wellbeing and automotive systems.

[28] The pandas development team. pandas-dev/pandas: Pandas, February 2020.

[29] EPA United States Environmental Protection. Particulate matter (pm) basics, 1999.

[30] EPA United States Environmental Protection. aqi-technical-assistance-document-, 2018.

[31] EPA United States Environmental Protection. What are the trends in indoor air quality and their effects on human health?, 2018.

[32] Kaushik Roy, Chitrita Chaudhuri, Mahantapas Kundu, Mita Nasipuri, and Dipak Basu. Comparison of the multi layer perceptron and the nearest neighbor classifier for handwritten numeral recognition. *Journal of Information Science and Engineering*, 21:1247–1259, 11 2005.

[33] Mariantonietta Ruggieri and Antonella Plaia. An aggregate aqi: Comparing different standardizations and introducing a variability index. *Science of The Total Environment*, 420:263–272, 2012.

[34] Luigi Russi, Paolo Guidorzi, Beatrice Pulvirenti, Giovanni Semprini, Davide Aguiari, and Giovanni Pau. Air quality and comfort characterisation within an electric vehicle cabin. pages 169–174, 2021.

[35] Jagriti Saini, Maitreyee Dutta, and Gonçalo Marques. Indoor air quality monitoring with iot: Predicting pm10 for enhanced decision support. In *2020 International Conference on Decision Aid Sciences and Application (DASA)*, pages 504–508, 2020.

[36] Iqbal Sarker. Data science and analytics: An overview from data-driven smart computing, decision-making and applications perspective. *SN Computer Science*, 2, 09 2021.

[37] Iqbal Sarker. Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science*, 2, 05 2021.

[38] MA Shehab and FD Pope. Effects of short-term exposure to particulate matter air pollution on cognitive performance. *Scientific reports*, 9(1):1–10, 2019.

[39] Amol S Shinde. Air pollution: Challenges and control measures.

[40] Breeze Technologies. Calculating an actionable indoor air quality index, 2020.

[41] Rita Tse, Lorenzo Monti, Catia Prandi, Davide Aguiari, Giovanni Pau, and Paola Salomoni. On assessing the accuracy of air pollution models exploiting a strategic sensors deployment. page 55–58, 2018.

[42] Peng Huiping Lima Aranildo R Teakles Andrew Jin Jian Cannon Alex J Hsieh William W. Evaluating hourly air quality forecasting in canada with nonlinear updatable machine learning methods. *Air Quality, Atmosphere and Health*, 2, 03 2017.

[43] Wenjuan Wei, Olivier Ramalho, Laeticia Malingre, Sutharsini Sivanantham, John Little, and Corinne Mandin. Machine learning and statistical models for predicting indoor air quality. *Indoor Air*, 29, 06 2019.

[44] Bin Xu, Xiaokai Chen, and Jianyin Xiong. Air quality inside motor vehicles' cabins: A review. *Indoor and Built Environment*, 27, 11 2016.

[45] Wei Ying Yi, Kin Ming Lo, Terrence Mak, Kwong Sak Leung, Yee Leung, and Mei Ling Meng. A survey of wireless sensor network based air pollution monitoring systems. *Sensors*, 15(12):31392–31427, 2015.