STUDIO DELLA QUALITÀ DELL’ARIA NEL 2023

## COMPONENTI DEL GRUPPO:

## Vanessa Barbaro [MAT.725617]: [v.barbaro3@studenti.uniba.it](mailto:v.barbaro3@studenti.uniba.it)

## Stefano Todaro [MAT.720335]: [s.todaro4@studenti.uniba.it](mailto:s.todaro4@studenti.uniba.it)

## A.A 2023/2024

## Github: [https://github.com/VanessaBarb/ProgettoIcon/tree/main](%20https:/github.com/VanessaBarb/ProgettoIcon/tree/main)

## **Indice**

## **Introduzione**

L’obbiettivo della tesi è studiare e comprendere come i maggiori inquinanti presenti nell’aria influiscano e in che peso sulla qualità dell’aria, ed inoltre analizzare come ( o se) questa giochi un ruolo in alcuni fattori climatici

Come quantificatore per il calcolo della qualità dell’aria è stato selezionato l’**Air Quality Index** (**AQI**), ovvero un sistema per tradurre le misurazioni delle concentrazioni di inquinanti, a volte confuse o poco intuitive, in una scala di facile comprensione per rappresentare chiaramente il rischio per la salute posto dall'inquinamento dell'aria ambiente. L’AQI traduce le misurazioni delle concentrazioni di questi inquinanti in una scala che va da 0 a 500. I valori dell’AQI possono ricadere in 6 categorie distinte:

* Buono
* Moderato
* Malsano
* Molto malsano
* Pericoloso

Il calcolo dell’AQI si basa su 6 principali sostanze inquinanti presenti nell’aria:

* PM2.5
* PM10
* NO2
* SO2
* O3
* CO

Il dataset selezionato (e di conseguenza lo studio) analizza le misurazioni degli inquinanti e i dati climatici (temperatura, umidità, velocità del vento) in 20 grandi città nel mondo nell’arco di tutto il 2023.

### Requisiti funzionali:

Il progetto è stato realizzato in Python3.12 in quanto offre un vasto numero di librerie che permettono di trattare e manipolare i dati in modo semplice.

IDE utilizzato: PyScripter

### Librerie utilizzate:

* Pandas
* Sklearn
* Matplotlib
* Seaborn
* Imblearn
* Numpy
* Pgmpy

## **Creazione del Dataset**

Il dataset, già di base, conteneva già la maggior parte delle informazioni utili. I dati presenti erano informazioni geografiche e temporali accompagnate dai relativi valori ambientali e climatici;

* **City**: Riporta le città in cui sono state effettuate le misurazioni. Questa feature contiene le 20 più grandi e principali città nel mondo, come ad esempio Parigi, Tokyo, New York, ecc.
* **Country**: Contiene i paesi di appartenenza delle città. Ogni città appartiene ad un paese differente.
* **Date**: Indica le date in cui sono state effettuate le misurazioni. Queste sono state effettuate lungo tutto l’arco del 2023.
* **PM2.5**: Riporta le misurazioni medie nell’arco della giornata di **particolato fine**. Il particolato fine (Particulate Matter PM) è costituito da particelle solide e liquide aventi diametro aerodinamico variabile fra 0,1 e circa 100 μm che tendono a rimanere sospese in aria.  Il termine PM2.5 è relativo alle particelle con diametro aerodinamico inferiore o uguale ai 2.5 μm . Generalmente tali particelle sono costituite da una miscela di elementi quali carbonio, fibre, metalli. nitrati, solfati, composti organici , materiale inerte e particelle liquide.
* **PM10**: La definizione di questo elemento è molto simile alla precedente, con la differenza che questa feature misura la quantità media giornaliera di particolato fine delle particelle di diametro aerodinamico inferiore o uguale ai 10 μm (1 μm = 1 millesimo di millimetro)
* **NO2**: Mostra la misurazione giornaliera media del **biossido di azoto**, si forma in massima parte in atmosfera per ossidazione del monossido (NO), inquinante principale che si forma nei processi di combustione. Le emissioni da fonti antropiche derivano sia da processi di combustione che da processi produttivi senza combustione (produzione di acido nitrico, fertilizzanti azotati, ecc.)
* **SO2**: Riporta la quantità medià di **biossido di zolfo**. Questo si orma nel processo di combustione per ossidazione dello zolfo presente nei combustibili solidi e liquidi. Le fonti di emissione principali sono legate alla produzione di energia, agli impianti termici, ai processi industriali e al traffico. L'SO2 è il principale responsabile delle "piogge acide".
* **CO**: Misura la quantità media giornaliera di **monossido di carbonio**, un gas inodore e incolore che si forma dalla combustione incompleta degli idrocarburi presenti in carburanti e combustibili. Le concentrazioni in aria di questo inquinante possono essere ben correlate all'intensità del traffico.
* **O3**: Misura la quantità di **ozono** mediamente presente nell’aria. L'ozono è un gas incolore ed inodore, la sua presenza al livello del suolo dipende fortemente dalle condizioni meteoclimatiche. Le concentrazioni di Ozono più elevate si registrano normalmente nelle zone distanti dai centri abitati ove minore è la presenza di sostanze inquinanti con le quali, a causa del suo elevato potere ossidante, può reagire.
* **Temperature**: Riporta la temperatura media calcolata nella giornata in gradi Celsius (C°)
* **Humidity**: Indica il livello di umidità medio giornaliero misurato in percentuale
* **Wind Speed**: Mostra la velocità media del vento misurata in chilometri orari (km/h)

Dati gli obbiettivi dello studio è stato necessario aggiungere due colonne fondamentali, ovvero la colonna rappresentante l’AQI, aggiunta con il nome di **Air\_Quality**, e la colonna che contenente la categoria dell’indice precedente, nominata **Air\_Quality\_Label**.

Per il calcolo dell’AQI è stato selezionato il metodo standard:

per prima cosa si calcolare il valore di AQI per ciascun inquinante, tramite la formula standard

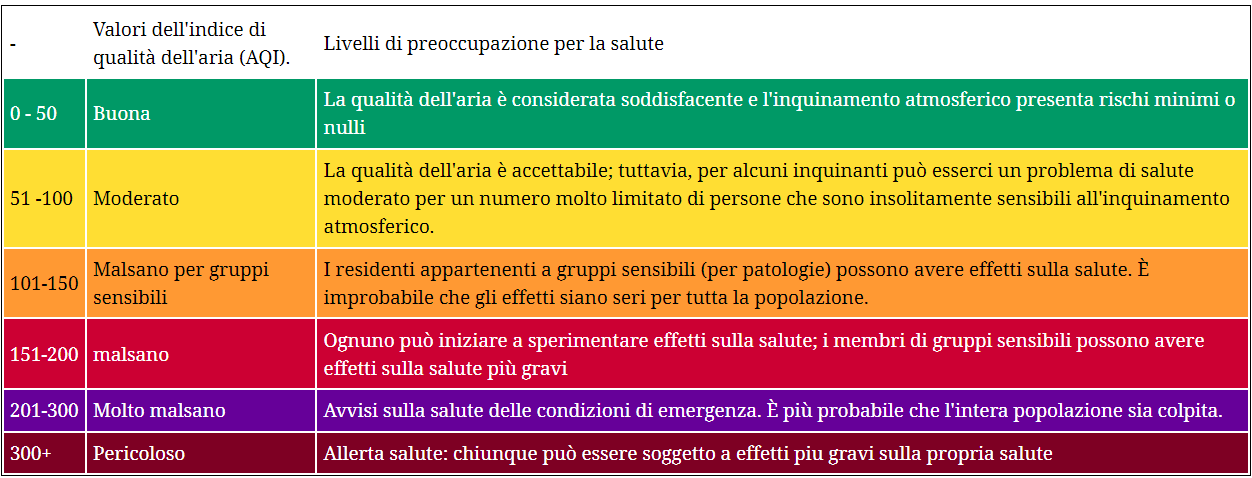
Dove:

* **C**: è la concentrazione misurata dell'inquinante.
* **Clo**: è il limite inferiore dell'intervallo di concentrazione in cui rientra la concentrazione **C**.
* **Chi**: è il limite superiore dell'intervallo di concentrazione in cui rientra la concentrazione **C**.
* **Ilo**: è il valore di AQI corrispondente a **Clo**
* **Ihi**: è il valore di AQI corrispondente a **Chi**

I limiti sono specifici per sostanza.

Una volta fatte queste misurazioni si prende il valore di AQI maggiore fra tutti gli inquinanti e questo determinerà la qualità dell’aria.

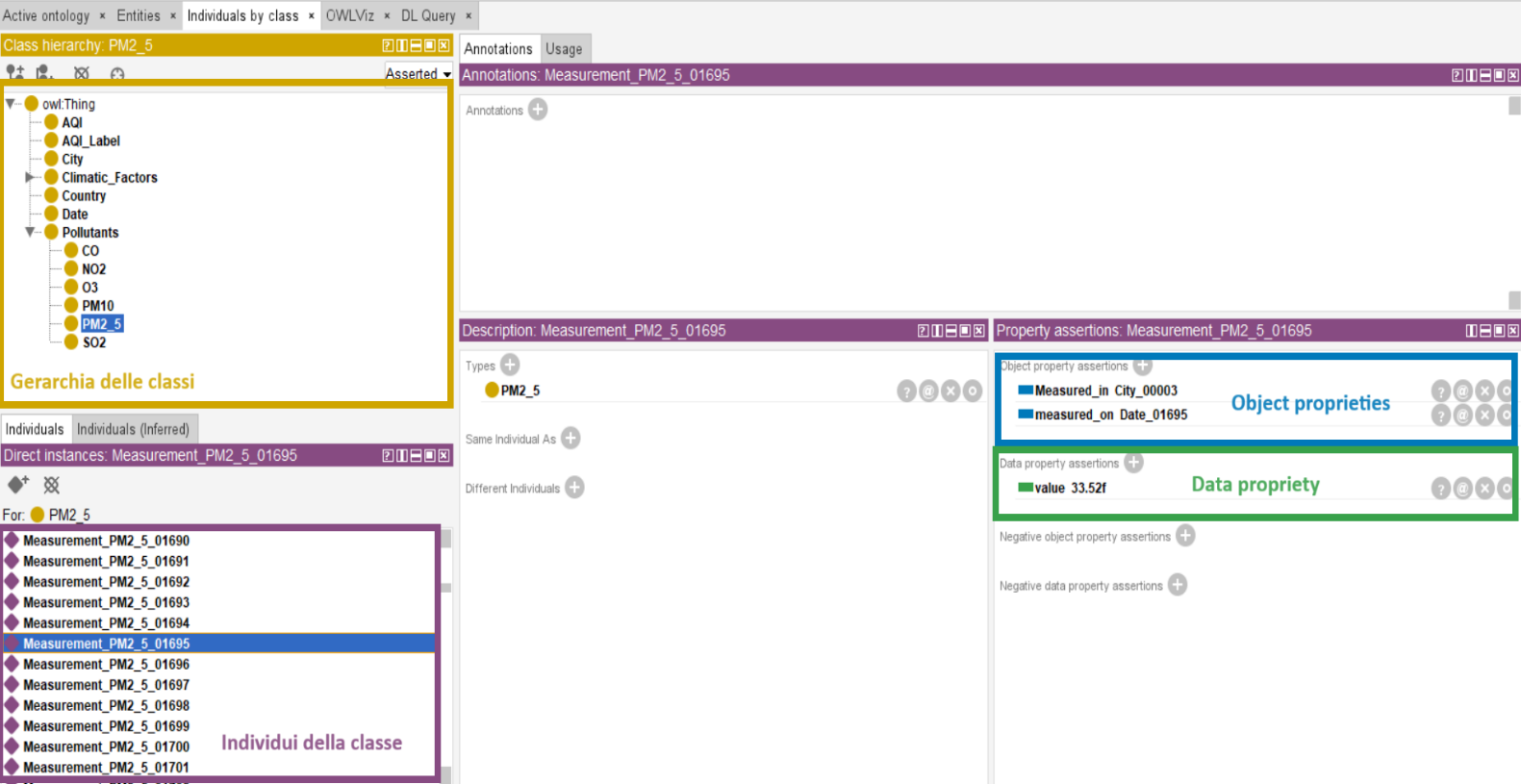
Una volta effettuati i calcoli per ogni riga presente nel dataset è stato necessario classificare ogni valore di AQI riportato, e per fare ciò si sfrutta la tabella di classificazione, che per ogni categoria descrivente la qualità dell’aria determina un range. Dipendente in quale range ricade il l’indice di qualità dell’aria è possibile determinare a quale categoria inscrivere la misurazione.



Tramite queste modifiche è stato creato il nuovo dataset **“globalAirNew.csv”** utilizzato per addestrare il modello ed effettuare le misurazioni.

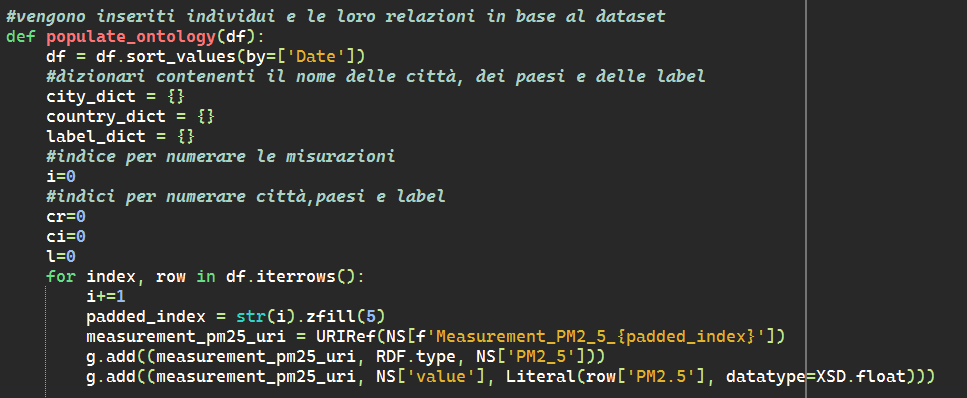
## **Ontologia**

In informatica, con il termine **ontologia** si intende una struttura formale che descrive in modo esplicito e organizzato un dominio di interesse, definendo le entità, i concetti e le relazioni che lo caratterizzano. Queste sono caratterizzate da una struttura fortemente gerarchica, con una relazione “is a” che va a collegare concetti più specifici con quelli generali. Queste sono rappresentati da linguaggi formali come ad esempio OWL. Queste forniscono un’ottima base di conoscenza.

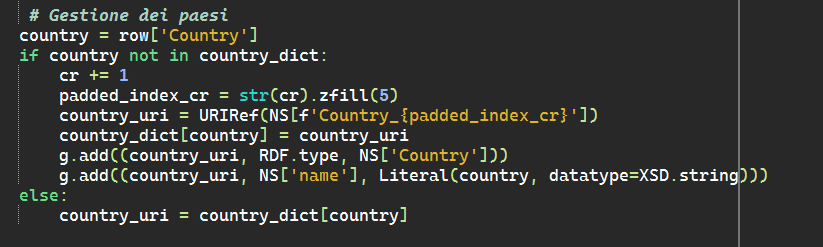
Per la creazione dell’ontologia è stato utilizzato il tool **Protegè-5.6.3**. Si tratta di un editor open-source specializzato nella creazione sistemi intelligenti.

L’ontologia è stata organizzata in classi, ognuna indicante le caratteristiche delle misurazioni (come luogo, data, inquinanti, caratteristiche climatiche ecc.). Ognuno di queste classi conterrà degli individui caratterizzati da delle **data proprieties**, che ne indicano le informazioni principali, e saranno in relazione tra loro, ove necessario, da delle **object proprieties**.

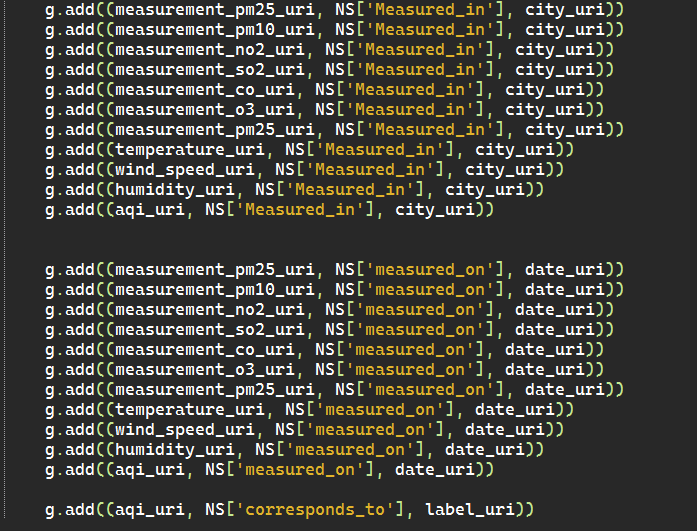
Per popolare l’ontologia con tutti i dati utili provenienti dal dataset è stata creata una funzione apposita: **populate\_ontology()**.



Questa funzione viene eseguita per ogni informazione all’interno del dataset. Per valori ripetuti spesso, come città, paese e categoria di AQI. Questi sono stati gestiti tramite dei dizionari appositi che mantengono le informazioni specifiche per classe, in modo da scongiurarne le ripetizioni.



Infine, terminato l’inserimento delle informazioni utili, si passa all’inserimento delle relazioni tra gli individui delle differenti classi



## **Apprendimento non supervisionato**

L’apprendimento non supervisionato (unsupervised learning) utilizza gli algoritmi di machine learning per analizzare e raggruppare in cluster i dataset senza etichette. I modelli di unsupervised learning vengono utilizzati per tre attività principali, il clustering, associazione e riduzione della dimensionalità.

### Clustering

È una tecnica che raggruppa i dati non etichettati in base alle similitudini e le differenze delle loro caratteristiche. Gli algoritmi di clustering possono essere raggruppati in varie tipologie, come algoritmi esclusivi, sovrapposti, gerarchici e probabilistici. Quelli più comunemente utilizzati sono gli algoritmi di tipo esclusivo e sovrapposto, noti anche come **hard** e **soft**.

Un algoritmo si definisce hard, se le classificazioni degli elementi del clustering questa avviene in maniera netta, inscrivendo ogni elemento ad un unico cluster. Mentre, un algoritmo è di tipo soft consenti ai punti di appartenere a più cluster con gradi di appartenenza differenti.

### Regole di associazione

Si tratta di metodi basati su regole utilizzati principalmente per trovare le relazioni tra le variabili di un dataset. Questi metodi vengono usati principalmente nelle analisi di mercato. I metodi principali sono algoritmi utilizzati per generare regole di associazione, come Apriori, Eclat e FP-Growth.

### Riduzione della dimensionalità

La riduzione della dimensionalità è una tecnica utilizzata quando il numero di caratteristiche, o dimensioni, in un determinato dataset è troppo elevato. Riduce il numero di input di dati a una dimensione gestibile preservando il più possibile l'integrità del dataset. È comunemente usata nella fase di preelaborazione dei dati e ci sono alcuni diversi metodi di riduzione della dimensionalità come l’analisi del componente principale (Principal component analysis o PCA), la decomposizione ai valori singolari (Singular value decomposition, SVD) e l’autoencoder.

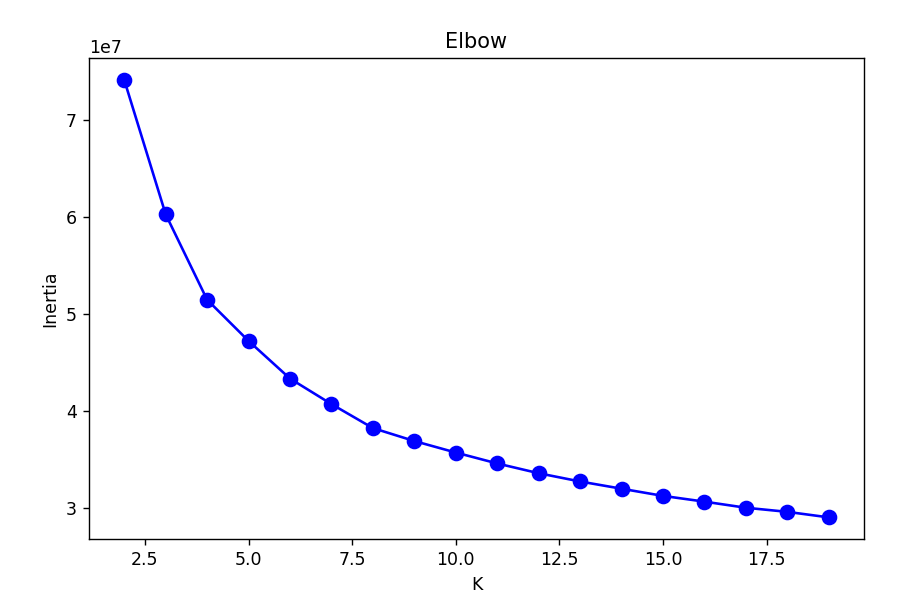
### Preprocessing dei dati

In un primo momento, dato l’obbiettivo dello studio, era stato deciso di effettuare una clusterizzazione delle misurazioni basate sulle temperature in relazione al valore dell’AQI. Ma questo è apparso molto poco pratico, in quanto tramite un’analisi della matrice di correlazione è apparso come, stando ai dati ritrovati, i fattori climatici siano influenzati in minima parte dalla presenza degli inquinanti nell’aria, riportando valori che rasentano lo zero.

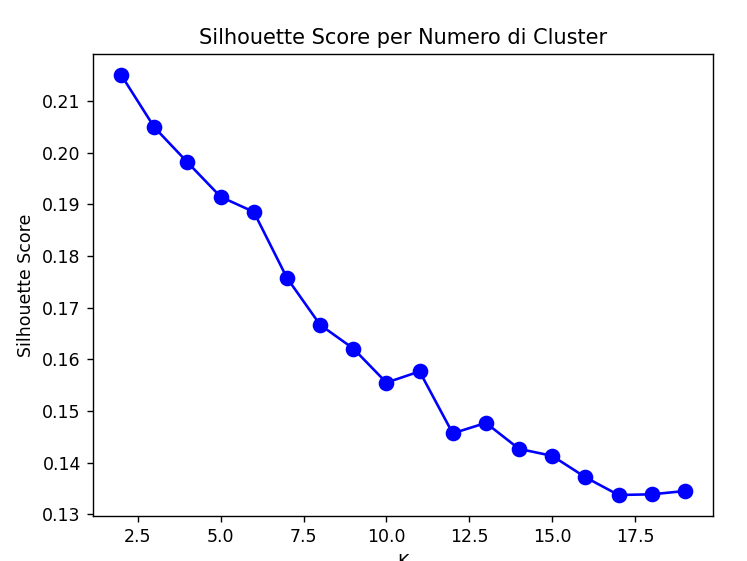
A questo punto si è deciso di clusterizzare i punti in base alle misurazioni degli inquinanti principali e la qualità dell’aria. Sempre tramite un’analisi della matrice di correlazione è stato osservato come l’elemento inquinante con maggiore impatto sulla qualità dell’aria sia il PM2.5, il quale è stato dunque scelto per la clusterizzazione dei dati.

È stato deciso di utilizzare un algoritmo di hard clustering: K-means. Il K-means è un algoritmo il cui obbiettivo è quello di clusterizzare i punti del dataset in un numero *k* specificato di cluster, calcolando per ogni punto la sua distanza dai centroidi di classe, rappresentanti i punti medi dei cluste, per poi assegnarlo al cluster con la distanza minore. Uno dei punti focali di questo algoritmo è la definizione del numero di cluster *k*, i quale è determinate per la qualità della distribuzione dei dati.

Per effettuare questi calcoli è stato necessario fare riferimento ai valori numerici del dataset.

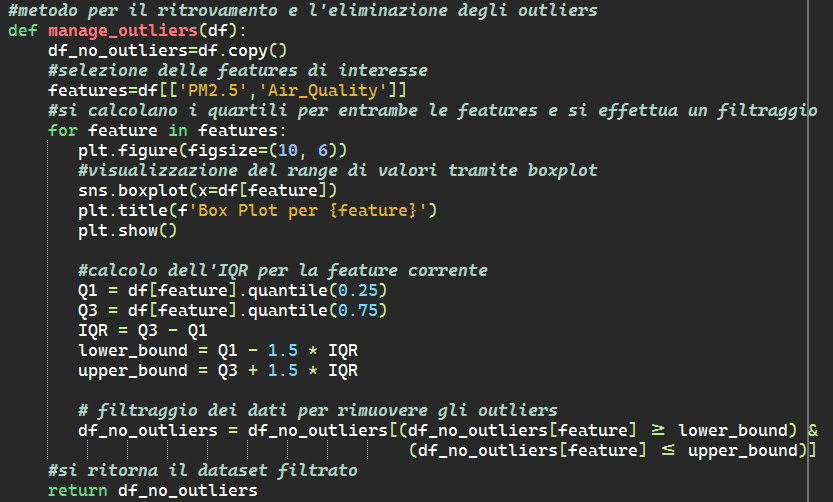
Per il ritrovamento del valore *k*, è stato utilizzato il metodo **find\_k()**, che sfrutta due algoritmi: il metodo più comunemente utilizzato, ovvero il **metodo del gomito**, in combinazione con il **silhouette score** calcolato in relazione al valore di k. Il metodo del gomito si basa sull’osservazione della diminuzione dell’inerzia all’aumentare del numero di cluster. Ciò accade in quanto, in generale, un numero maggiore di cluster rappresentano meglio la divisione dei dati. Con **inerzia** si definisce la somma delle distanze quadratiche tra ciascun punto e il centroide del cluster di appartenenza. Viene definito metodo del “gomito” in quanto nell’osservazione grafico riportante il rapporto tra inerzia e il valore di k, si seleziona il valore di k in cui il valore di inerzia rallenta la sua discesa andando a formare un certo angolo con i precedenti valori in discesa. 

Osservando il grafico si nota come questo sia rappresentato quasi da una curva, rendendo ardua l’identificazione del valore *k*. Per questo è stato ritenuto necessario utilizzare il *silhouette* *score*.

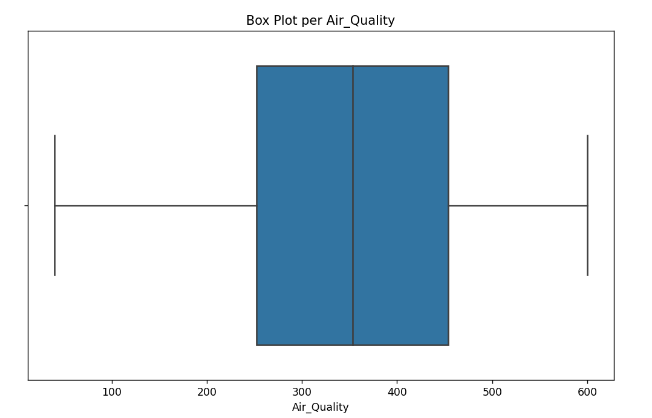
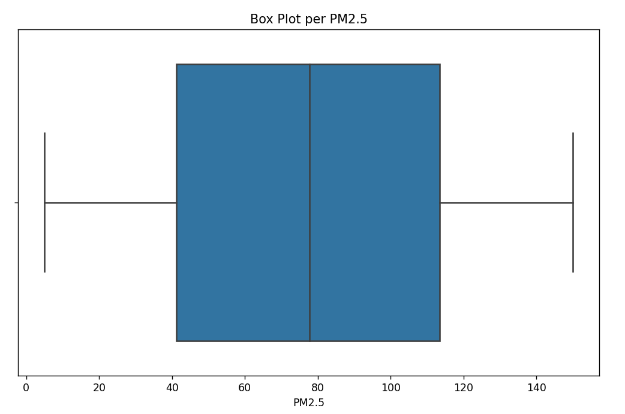
Il *silhouette score* definisce la distanza media tra i punti assegnati ai vari cluster, ovvero definisce quanto i cluster siano ben distinguibili gli uni dagli altri. Questo viene definito tramite un coefficiente di silhouette range [-1,1], con un valore 1 indica una separazione perfetta dei cluster, un valore 0 che indica come la separazione dei cluster non sia netta e anzi ci possono essere anche delle sovrapposizioni ed infine un valore negativo può indicare perfino delle assegnazioni errate. Duque, vengono effettuate delle misurazioni calcolando l’andamento del silhouette score in base al numero di *k* di cluster.

Si può notare come all’aumentare del numero di cluster si ha un repentino peggioramento del coefficiente di silhouette.

Mettendo i due grafici a confronto si è giunti alla conclusione che il valore ottimale per l’algoritmo è un numero di cluster *k*=3 (valore che verrà riconfermato anche da delle prove empiriche).

Prima di passare alla fase effettiva di clusterizzazione degli elementi del dataset, è stato necessario processare i dati in modo da ottenere dei risultati più performanti. Questo lavoro di ottimizzazione è stato seguito tramite il metodo **manage\_outliers()**.

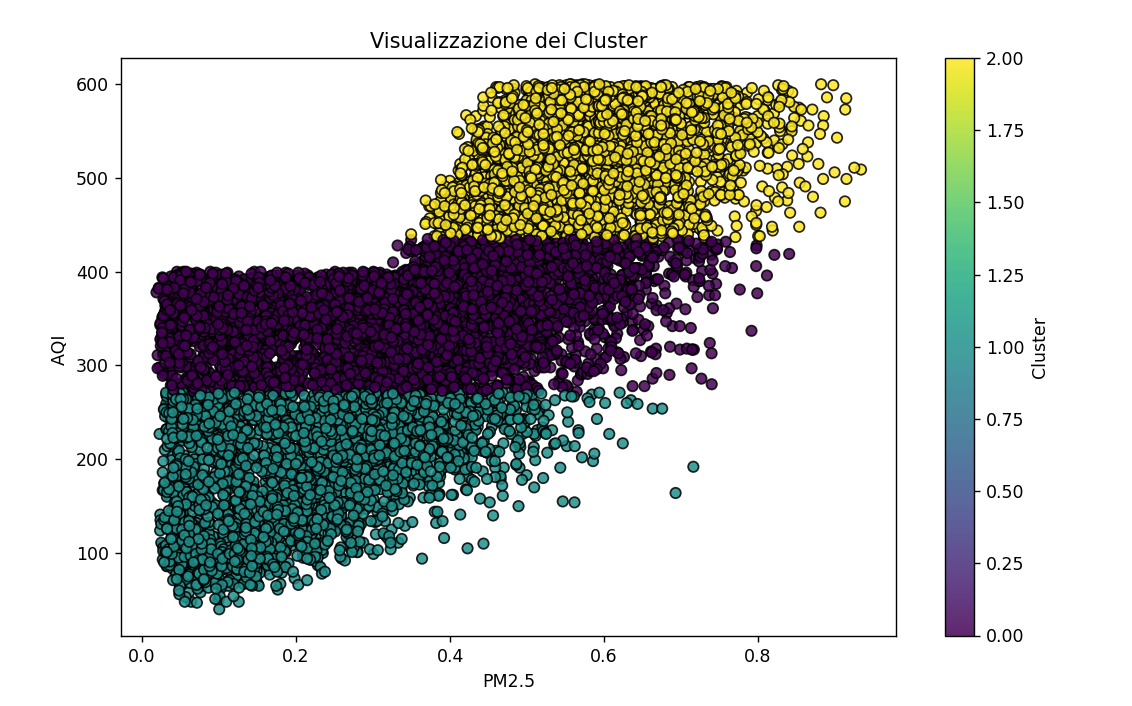
L’algoritmo di K-means è molto sensibile agli outliers, ovvero quei dati che si discostando dalla maggior parte dei valori presenti nel dataset. Per questo è stato utilizzato un metodo specifico per la gestione di questi valori. Per prima cosa sono stati creati dei grafici tramite la libreria Seaborn per visualizzazione della distribuzione dei valori principali per le classi di interesse.



In seguito, per ogni feature, sono stati calcolati i quartili, definendo il limite inferiore e superiore dei valori accettabili; i valori esterni a tale range vengono identificati come outliers e vengono rimossi dal DataFrame.

Il filtraggio dei valori viene seguito da un loro ulteriore raffinamento tramite lo scaler. La **scalarizzazione** (o **normalizzazione**) dei valori è una tecnica per trasformare le caratteristiche del dataset in un intervallo specifico e per l’ottenimento di una distribuzione più uniforme. Questo processo permette di assicurare la comparabilità dei valori e rende più accurata la misurazione delle distanze nel machine learning. Come algoritmo per la normalizzazione dei dati è stato selezionato l’algoritmo di **Unic Vector Transformation**, in quanto è il metodo è indipendente dalla magnitudine delle caratteristiche, fondando la distanza dei punti unicamente sulla direzione del vettore.

Effettuato questo ultimo processing dei dati, si passa all’esecuzione dell’algoritmo di *K-means*, ottenendo questa distribuzione:



Per avere un’idea più chiara delle performance del modello, è stato calcolato il coefficiente di silhouette score anche per cluster, ottenendo un valore di 0.59, che per quanto non sia ottimale si possa ritenere sufficientemente buono.