Alma Mater Studiorum · Università di Bologna

SCUOLA DI SCIENZE
Corso di Laurea in Informatica per il Management

INQUINAMENTO AMBIENTALE: ANALISI ESPLORATIVA E PREDITTIVA DELLE EMISSIONI GLOBALI DAL 2000 AD OGGI

Relatore:
Prof. Elena Loli Piccolomini

Candidato:
Filippo Lupatelli

Anno Accademico 2023/2024

Sessione Unica

Indice

Introduzione

Al Gore, ex Vice Presidente degli Stati Uniti e attivista ambientale, afferma *"Non abbiamo un pianeta B"*, questa dichiarazione invita a riflettere sulla vulnerabilità del nostro pianeta e ricorda che non esiste un piano di riserva se danneggiamo irreparabilmente il nostro habitat e il clima che lo regola.

Il cambiamento climatico è un problema che non può più essere trascurato. Richiede politiche e azioni concrete, sottolineando la necessità di invertire il pericoloso trend attuale e ripristinare un clima favorevole per lo sviluppo naturale del nostro pianeta. Viviamo in un'era di cambiamenti climatici dettati dall'imprevedibilità e da eventi atmosferici estremi che mettono in pericolo comunità e distruggono interi ecosistemi.

I gas serra, che svolgono un ruolo cruciale nel regolare il clima del pianeta, sono presenti in modo naturale nell'atmosfera in quantità adeguate per sostenere la vita sulla Terra. Tuttavia, l'attività umana ha alterato questo equilibrio, portando all'accumulo e alla produzione di gas in eccesso nell'ambiente e alla generazione di composti ancora più dannosi.

L'analisi dei dati sulle emissioni, supportata dalle tecnologie avanzate di Machine Learning (ML), può svolgere un ruolo fondamentale nel comprendere le principali fonti di inquinamento. Questo approccio permette di implementare correzioni mirate per affrontare le problematiche più gravi, monitorando costantemente le emissioni e analizzando i dati in modo autonomo. In questo modo, si può attuare un controllo efficace per ridurre le emissioni e minimizzare i danni già recati al nostro pianeta.

L'obiettivo di questo elaborato è analizzare come la scienza dei dati possa contribuire alla lotta contro il cambiamento climatico e alla riduzione delle emissioni. Nel primo capitolo, si esamina il concetto di emissione, descrivendo le metodologie di misurazione, le cause e gli impatti sull'ambiente e sulla salute umana. Inoltre, si elencano gli accordi e le normative più importanti per affrontare il cambiamento climatico e le relative soluzioni adottate. Il secondo capitolo delinea un'analisi esplorativa di un dataset sulle emissioni globali, coprendo un arco temporale di due decenni, dal 2000 al 2020. Tale studio traccia un quadro dettagliato dell'evoluzione di tali emissioni nel corso del tempo, individuando le variazioni significative e mettendo in luce le situazioni più preoccupanti. Nel terzo capitolo, ci si concentra sull'analisi di un dataset contemporaneo che include indicatori chiave relativi a vari aspetti sociali, ambientali, economici e sanitari. Dopo una fase iniziale di esplorazione dati, si adottano algoritmi di Machine Learning per formulare previsioni delle emissioni di anidride carbonica (CO2) basate su parametri rilevanti relativi all'anno di riferimento. Questa regressione non solo coinvolge l'utilizzo di diversi algoritmi per determinare il più idoneo ai nostri obiettivi, ma costituisce anche il fondamento per la creazione di un modello predittivo che può essere utilizzato per stimare le emissioni future.

1 Emissioni e Cambiamento Climatico

1.1 Definizione e Misurazione delle Emissioni

Con il termine **emissione** si intende qualsiasi sostanza solida, liquida o gassosa introdotta nell'atmosfera che possa causare inquinamento atmosferico.

L'**approccio ideale** per creare un inventario completo delle emissioni comporterebbe la misurazione diretta di tutte le emissioni e di tutte le diverse tipologie di sorgenti nell'area di interesse durante il periodo oggetto di studio. Tuttavia, questo metodo è spesso poco pratico per diverse ragioni. In primo luogo, i dati riguardanti le emissioni coprono spesso vaste aree territoriali, come regioni o interi paesi, e pertanto la raccolta di dati diretti su così ampia scala può essere problematica. Inoltre, alcune tipologie di emissioni, a causa della loro natura, sono difficili da misurare in maniera completa attraverso le comuni tecniche di misurazione.

Per questo motivo, spesso si utilizza un **approccio analitico**. Questo approccio è particolarmente utile per gli impianti industriali di grandi dimensioni che dispongono di sistemi di monitoraggio continuo. Questi sistemi forniscono dati affidabili e rappresentativi dell'emissione totale della sorgente. Tuttavia, per impianti industriali di medie e piccole dimensioni, il monitoraggio diretto può essere più problematico a causa dei costi e delle difficoltà operative, soprattutto per rispettare le normative ambientali.

Quando non è possibile effettuare misurazioni dirette, è necessario stimare le emissioni. Questa stima si basa sull'uso di un indicatore che rappresenta l'attività della sorgente e un fattore di emissione specifico per il tipo di sorgente, il processo industriale e la tecnologia di depurazione utilizzata. La formula generale utilizzata è la seguente:

$$E_i = A * FE_i$$

dove E_i rappresenta l'emissione dell'inquinante in tonnellate per anno, A è l'indicatore dell'attività, che può essere legato alla quantità prodotta, al consumo di combustibile e altro e FE_i è il fattore di emissione specifico per l'inquinante considerato. La precisione di questa stima dipende dalla qualità e dalla precisione dei fattori di emissione, che variano in base al tipo di impianto. Per i processi di combustione viene generalmente scelto come indicatore di attività il consumo di combustibile, mentre per i processi industriali gli indicatori privilegiati sono la quantità di prodotto processata nell'unità di tempo o il numero di addetti nel settore di cui si vuole stimare l'emissione.

La prima fase nell'analisi delle emissioni coinvolge il **campionamento**. Questo processo è critico e complesso, poiché i metodi di campionamento sono spesso unici per ciascun inquinante. Questo evidenzia anche la delicatezza della procedura di campionamento, che talvolta richiede di operare a quote elevate, in spazi ristretti o in condizioni atmosferiche avverse. Tutte queste considerazioni rendono il campionamento delle emissioni un'attività altamente specializzata. Il personale coinvolto deve essere adeguatamente formato, il che richiede un notevole impegno in termini di tempo e costi, oltre a richiedere un buon stato di salute.

Durante il processo di campionamento, vengono utilizzati strumenti specializzati, come trappole, per estrarre campioni di aria contenenti gli inquinanti. Questi campioni vengono quindi inviati in laboratorio per l'analisi. In alternativa, quando si utilizzano analizzatori portatili, non è necessario effettuare l'estrazione del campione poiché gli strumenti misurano direttamente il contenuto di inquinanti, calcolando sempre un volume di aria noto da campionare.

1.2 Analisi dei Principali Inquinanti Atmosferici

Gli inquinanti atmosferici possono essere suddivisi in due principali categorie:

1. I **macroinquinanti**: questi sono inquinanti presenti nell'atmosfera a concentrazioni rilevanti, espressi solitamente in milligrammi per metro cubo (mg/Nm^3). La categoria dei macroinquinanti comprende inquinanti tradizionali derivanti dai processi di combustione, come il monossido di carbonio (CO), gli ossidi di azoto (NOx), il biossido di zolfo (SO2), gas acidi come l'acido cloridrico (HCl) e l'acido fluoridrico (HF), il materiale particolato, le sostanze organiche volatili e altri composti.
2. I **microinquinanti**: questa categoria include inquinanti presenti in concentrazioni molto più basse, misurate in microgrammi per metro cubo ($\mu g/Nm^3$) o nanogrammi per metro cubo (ng/Nm^3). Nonostante le basse concentrazioni, i microinquinanti possono rappresentare una minaccia ambientale significativa a causa della loro tossicità e persistenza. I microinquinanti si suddividono in inorganici, come i metalli pesanti (come piombo, cadmio e mercurio), e organici, che includono diossine, idrocarburi policiclici aromatici (IPA) e policlorobifenili (PCB).

I gas serra sono componenti atmosferici che contribuiscono all'effetto serra, causando l'innalzamento della temperatura globale. Tra i gas serra, i più comuni presenti in natura sono:

- Anidride Carbonica (CO2): Questo gas è prodotto sia da fonti naturali, come la respirazione degli animali e la decomposizione della biomassa, che da attività umane, in particolare dalla combustione di combustibili fossili. Le piante, durante la fotosintesi, assorbono CO2 dall'atmosfera, contribuendo al suo bilancio.
- Metano (CH4): Il metano è il principale componente del gas naturale ed è emesso da attività legate alla produzione e al trasporto di combustibili fossili come carbone, gas naturale e petrolio. Altre fonti di metano includono il bestiame, le pratiche agricole, l'uso del suolo e la decomposizione dei rifiuti organici nelle discariche.
- Protossido di Azoto (N2O): Questo gas serra è prodotto da processi microbici nel suolo, dall'uso di fertilizzanti azotati, dalla combustione del legno e da processi chimici industriali. Le emissioni di N2O provengono da attività agricole, industriali, uso del suolo, combustione di combustibili fossili e trattamento delle acque reflue.

Inoltre, esistono i cosiddetti gas fluorurati a effetto serra, che sono estremamente potenti e persistenti rispetto ai gas serra tradizionali. Questi includono:

- drofluorocarburi (HFC): Utilizzati in refrigerazione, condizionamento dell'aria e pompe di calore, così come negli estintori. L'UE sta lavorando per eliminarli gradualmente entro il 2050.
- Perfluorocarburi: Sono composti artificiali ampiamente utilizzati nei processi di produzione industriale.
- Esafluoruro di Zolfo (SF6): Utilizzato nell'isolamento delle linee elettriche.
- Trifluoruro di Azoto (NF3): Impiegato come gas di pulizia della camera nei processi di produzione, per rimuovere accumuli indesiderati da parti e circuiti del microprocessore.

Questi gas serra, in particolare i gas fluorurati, sono noti per il loro potere di riscaldamento globale estremamente elevato, superiore a migliaia di volte rispetto alla CO2. La comprensione di queste diverse categorie di inquinanti e gas serra è fondamentale per affrontare le sfide legate al cambiamento climatico e all'inquinamento atmosferico.

1.3 Cause e Fonti delle Emissioni

I **gas serra** possono essere il risultato di processi naturali, come nel caso del vapore acqueo, o di processi naturali e artificiali, come l'anidride carbonica (CO2) e il metano (CH4), oppure di processi soltanto artificiali, come i gas fluorati. La maggiore responsabilità, quando si parla di riscaldamento globale, è l'anidride carbonica, che rappresenta oltre il 75% delle emissioni causate dall'uomo. L'aumento delle concentrazioni di questi gas serra nell'atmosfera sta conducendo a un riscaldamento globale senza precedenti. Quest'ultimo è provocato anche dall' abbattimento delle foreste, **deforestazione**, poichè gli alberi aiutano a regolare il clima assorbendo CO2 dall'atmosfera. Abbattendoli, quest'azione viene a mancare e la CO2 immagazzinata negli alberi viene rilasciata nell'atmosfera, alimentando in tal modo l'effetto serra.

Secondo il Comitato sul Cambiamento Climatico delle Nazioni Unite (IPCC), entro il 2030, il riscaldamento potrebbe superare i 1,5 gradi Celsius, considerati la soglia massima di sicurezza per evitare impatti climatici gravi e difficilmente gestibili.

Le emissioni di gas serra sono il risultato di varie attività umane, e la loro suddivisione nei diversi settori può variare in base a diversi fattori. Alcune stime indicano quanto ciascun settore contribuisca alle emissioni globali di gas serra. Le stime possono variare a seconda dei parametri considerati. Ecco alcune stime sui settori e le relative percentuali di contributo alle emissioni globali di gas serra:

1. Produzione di elettricità e calore: Questo settore rappresenta circa il 25% delle emissioni globali ed è legato alla combustione di carbone, gas naturale o petrolio per generare energia.
2. Agricoltura, allevamento e deforestazione: Questo settore contribuisce al 24% delle emissioni globali. Gli allevamenti, in particolare quelli di bovini per carne e latte, sono responsabili del 14% delle emissioni dovute all'attività umana. La produzione di mangimi e la digestione degli animali sono le principali fonti di emissioni.
3. Industria: Il settore industriale contribuisce al 21% delle emissioni globali di gas serra.
4. Trasporti: Questo settore rappresenta circa il 14% delle emissioni globali. Le automobili, i furgoni e i piccoli camion sono responsabili del 59% delle emissioni legate ai trasporti.
5. Consumo di combustibili fossili per uso residenziale e commerciale: Questo settore contribuisce al 6% delle emissioni globali.
6. Altre attività: Questo comprende l'estrazione di combustibili fossili, la raffinazione del petrolio, la lavorazione e il trasporto del petrolio e rappresenta il 10% delle emissioni.

Le emissioni per paese mostrano che la Cina è il principale responsabile globale con circa il 28% delle emissioni totali, seguita dagli Stati Uniti con il 15%.

Le principali tipologie di **fonti emissive** si dividono in:

- emissione puntuale: la fonte emissiva è localizzata, tipico per esempio dei camini industriali.
- emissione lineare: la fonte è un tratto di strada cui sono associate le emissioni degli autoveicoli che la percorrono.
- emissione aerale: la fonte è un serbatoio dove evapora un certo agente inquinante.
- emissioni diffuse: ovvero distribuite sul territorio, stimate attraverso l'uso di opportuni indicatori e fattori di emissione, tramite la formula mostrata prima.

Le sorgenti emissive possono essere classificate anche come continue o discontinue in base alle modalità di funzionamento nel tempo e in fisse o mobili a seconda della loro dislocazione nello spazio.

1.4 Effetti dell'Ambiente e sulla Salute Umana

Gli effetti evidenti sull'ambiente di tali emissioni sono numerosi:

Il fenomeno dell'**effetto serra**, essenziale per la regolazione della temperatura, consente di trattenere parte delle radiazioni infrarosse emesse dalla Terra, contribuendo così a mantenere condizioni adatte alla vita. Tuttavia, il riscaldamento globale che stiamo vivendo è in gran parte attribuito all'aumento innaturale degli "inquinanti" atmosferici noti come gas serra, causato dalle attività umane. Questi gas serra influenzano il bilancio energetico del pianeta e contribuiscono all'aumento delle temperature.

Le **temperature in aumento** sono responsabili di ondate di calore, che possono portare a morti premature, riduzione della produttività e danni alle infrastrutture. Le fasce più vulnerabili della popolazione, come gli anziani e i neonati, sono particolarmente a rischio.

La **siccità**, causata dalla mancanza di precipitazioni e dall'evaporazione aumentata dovuta alle temperature elevate, provoca carenze d'acqua e può avere impatti devastanti su infrastrutture, agricoltura, approvvigionamento idrico e biodiversità. Le siccità più frequenti e gravi aumenteranno il rischio di incendi boschivi, specialmente nelle regioni mediterranee.

I cambiamenti climatici prevedono un aumento delle precipitazioni in diverse regioni, che si traduce in due tipi di **inondazioni**: fluviali causate da lunghi periodi di piogge e pluviali scatenate da precipitazioni intense in brevi periodi. I temporali violenti diventeranno più frequenti e intensi a causa delle temperature più elevate, aumentando il rischio di inondazioni improvvise.

L'**innalzamento del livello del mare**, causato principalmente dallo scioglimento dei ghiacci e dall'espansione termica degli oceani, minaccia le comunità costiere e la biodiversità. Le aree costiere sono spesso densamente popolate e contribuiscono in modo significativo all'economia, ma sono vulnerabili all'innalzamento del livello del mare, alle inondazioni e all'erosione.

I cambiamenti climatici influenzano anche la **biodiversità**, causando cambiamenti nella distribuzione e nell'abbondanza delle specie, nonché nella struttura degli habitat e nei processi degli ecosistemi. Gli impatti negativi sulla biodiversità includono la perdita di habitat, l'inquinamento e la diffusione di specie invasive.

Gli effetti delle emissioni ricadono non solo sull'ambiente ma anche sulla **salute umana**, con conseguenze significative per il nostro benessere. Durante l'estate, con le nuove ondate di caldo elevato, si assiste a un aumento della mortalità e delle malattie legate al calore, mentre in inverno si riducono le morti legate al freddo. I cambiamenti climatici, come già detto, ci espongono a un aumento degli eventi meteorologici estremi, con un conseguente aumento del rischio di incidenti e impatti sulla salute e sul rischio di mortalità. Inoltre, l'espansione delle malattie trasmesse da vettori come zanzare, malattie legate a roditori e i cambiamenti nella qualità dell'acqua e dell'approvvigionamento alimentare costituiranno una minaccia per la salute umana, con malattie che variano o riemergono, influenzando la salute delle persone.

Non meno gravi sono gli impatti dell'inquinamento dell'aria sulla **salute dei bambini**, come riconosciuto dall'OMS (Organizzazione Mondiale Sanità) e dall'ERS (European Respiratory Society). L'esposizione all'inquinamento dell'aria è stata collegata a un aumento della mortalità infantile e al rischio di SIDS (Sindrome della Morte Improvvisa del Lattante). I bambini esposti all'inquinamento dell'aria possono sperimentare una riduzione della funzionalità polmonare, il che può avere effetti a lungo termine sulla loro salute respiratoria e un maggior rischio di malattie quali asma e le infezioni delle vie respiratorie. Infine, l'inquinamento atmosferico è un cancerogeno certo, come riporta l'Agenzia Internazionale per la Ricerca sul Cancro (IARC) che ha comprovato una relazione diretta tra esposizione al traffico e tumori dei bambini, tra cui la leucemia.

1.5 Normative e Regolamentazioni negli anni

Di seguito un quadro generale sui negoziati più significativi nella riduzione delle emissioni e nel cambiamento climatico:

1972 - Conferenza delle Nazioni Unite sull'ambiente umano a Stoccolma: prima conferenza su questioni climatiche internazzionali e politiche ambientali.

1979 - Conferenza Mondiale sul Clima a Ginevra: evento essenzialmente scientifico che conduce ad un programma sugli effetti del clima.

1987 - Protocollo di Montreal: limita l'uso di sostanze che danneggiano lo strato di ozono.

1990 - Gruppo IPCC: prima valutazione scientifica sul cambiameno climatico.

1992 - Convenzione quadro delle Nazioni Unite sui cambiamenti climatici (UNFCCC) a New York: stabiliti obiettivi vincolanti di riduzione delle emissioni di gas.

1997 - Protocollo di Kyoto: pone la riduzione delle emissioni di una media del 5% dal 2008 al 2012.

2010 - Accordi di Cancun: assistenza alle nazioni in via di sviluppo che si trovano ad affrontare il cambiamento climatico.

2015 - Accordo di Parigi sul clima: viene stabilito l'obiettivo di limitare il riscaldamento globale "ben al di sotto" dei 2°C rispetto ai livelli preindustriali, con uno sforzo per limitare il riscaldamento a 1,5°C. Gli Stati membri hanno presentato obiettivi nazionali di riduzione delle emissioni, noti come Contributi Nazionali Determinati (NDC), e si sono impegnati a rivederli periodicamente al rialzo. Tale accordo è entrato in vigore nel 2016.

2018 - Relazione speciale 1.5° dell'IPCC: L'IPCC dell'ONU dichiara un aumento del riscaldamento di 1.5° rispetto ai livelli preindustriali, sottolineanado l'esigenza di misure che riducano le emissioni.

2019 - Parlamento europeo dichiara l'emergenza climatica nel mondo.

2019 - UE sulla strada per diventare climaticamente neutrale entro il 2050: L'UE basandosi sulla legge Green Deal ha come obiettivo la neutralità climatica entro il 2050

2021 - Legge Europea sul Clima: nuovo più ambizioso obiettivo per il raggiungimento della neutralità climatica del 2050 con la riduzione delle emissioni di almeno il 55% entro il 2030.

2022 - COP27 a Sharm-el-Sheikh: fondo per perdite e danni di paesi colpiti da disastri climatici.

1.6 Soluzioni per la Riduzione delle Emissioni

L'Unione europea ha adottato una serie di misure e iniziative per affrontare il cambiamento climatico e ridurre le emissioni di gas serra. Di seguido una panoramica delle principali azioni.

In primo luogo, l'Unione europea ha implementato il **Mercato delle Emissioni** (ETS), sistema di scambio delle emissioni che riguarda centrali elettriche ed industrie. Questo sistema richiede alle aziende di acquistare permessi per emettere CO2, creando così un incentivo economico a ridurre le emissioni. L'ETS copre il 40% delle emissioni totali di gas serra nell'UE, dunque questa misura dimostrando l'impegno per ridurre le emissioni industriali.

Per adeguare l'ETS agli obiettivi del Green Deal, si è rivisto il sistema per ridurre le emissioni industriali del 62% entro il 2030. Inoltre, si è ampliato l'ETS per includere settori precedentemente esclusi. Tali settori, come l'edilizia, l'agricoltura e la gestione dei rifiuti, contribuiranno entro il 2030 ad un aumento delle riduzioni dal 29% al 40%. In particolare, il settore dei trasporti è al centro delle politiche con l'obiettivo di raggiungere zero emissioni di CO2 per i nuovi mezzi, e nel trasporto marittimo di ridurre le emissioni.

L'attenzione riguarda anche norme sull'**assorbimento di carbonio** delle foreste, con un aumento del 15% dei pozzi di carbonio entro il 2030, contribuendo così alla lotta contro il cambiamento climatico.

Si sono adottate norme per un meccanismo di adeguamento del carbonio alle frontiere. Questo meccanismo stabilirà un **prezzo del carbonio** per le importazioni da industrie ad alta intensità di carbonio provenienti da paesi meno ambiziosi nella lotta al cambiamento climatico.

Si incentiva la promozione di **energia rinnovabile** ed efficienza energetica tramite leggi riguardanti l'efficienza energetica degli edifici ed elettrodomestici, con obiettivi di riduzione del consumo energetico. La quota di energie rinnovabili finale dovrebbe aumentare al 42,5% entro il 2030, con obiettivi diversi per i singoli paesi.

L'UE promuove il consumo sostenibile e l'**economia circolare** per ridurre il consumo di risorse e le emissioni di gas serra. Le misure riguardano vari settori, tra cui imballaggi, tessuti, elettronica, costruzioni, batterie, alimenti, materie prime e il riutilizzo di beni. L'economia circolare è un modello economico che si contrappone all'approccio tradizionale, noto come economia lineare, mentre in quest'ultima i prodotti vengono fabbricati, utilizzati e poi smaltiti come rifiuti, nell'economia circolare l'obiettivo è ridurre al minimo lo spreco e massimizzare il riutilizzo, il riciclaggio e il ripristino dei materiali e delle risorse.

Infine, sono state adottate norme **contro la deforestazione** per consentire la conservazione della biodiversità e ripristino degli habitat naturali.

2 Analisi Esplorativa sulle Emissioni (2000 - 2020)

Questo capitolo ha come argomento l'analisi esplorativa dei dati (EDA) su un dataset che contiene i valori di emissione dei principali tipi di elementi inquinanti responsabili del riscaldamento globale e del cambiamento climatico, suddivisi per paese e nel range compreso dal 2000 al 2020. Il dataset utilizzato per questa analisi è stato fornito dall'Organizzazione delle Nazioni Unite per l'Alimentazione e l'Agricoltura (FAO), un'importante fonte di informazioni sulle questioni ambientali e agricole. Questo dataset comprende una vasta gamma di dati, tra cui le emissioni dei tre principali gas serra: anidride carbonica (CO2), protossido di azoto (N2O) e metano (CH4). Ognuno di questi gas serra svolge un ruolo significativo nel contribuire all'effetto serra e, di conseguenza, al riscaldamento globale.

2.1 Descrizione del dataset

L'analisi esplorativa si basa sul dataset "Total Emissions Per Country (2000-2020)" (*https://www.kaggle.com/datasets/justin2028/total-emissions-per-country-2000-2020*). L'obiettivo EDA è quello di analizzare, all'interno del range di anni compreso fra il 2000 e il 2020, come è variato lo scenario delle emissioni anno per anno, cosa è cambiato dal primo periodo di rilevamento fino al 2020, quali sono i principali paesi responsabili di tali emissioni e quali elementi influiscono maggiormente. I dati sono riportati in una tabella contenete 25 colonne, le prime 4 comprendono il paese, la causa delle emissioni, tipi di elementi e unità di emissione, le altre 21 sono relative agli anni:

- Area (Country)
- Item (Source of Emission)
- Element (Type of Emission)
- Unit (Emissions = Kilotonnes)
- Year (Total Emissions for Each Year, 2000 - 2020)

Il dataset, inoltre, presenta nella colonna "Area" differenti tipologie. Nella prima parte delle righe fa riferimento ai paesi, nella seconda parte fa riferimento a dati mondiali (*'Area' == World*) e nella terza parte fa riferimento a Continenti e Regioni mondiali.

2.2 Analisi del dataset

Inizialmente, dopo aver importato le librerie che consentono la visualizzazione e la manipolazione del dataset, oltre che la creazione di grafici, si importa il dataframe salvato precedentemente in formato *csv*.

[2]:
```
#Importo dataset e visualizzo le prime 5 righe
df = pd.read_csv('/Users/filippolupatelli/Desktop/'
                 'Total Emissions Per Country.csv')
```

[2]:
```
          Area           Item                         Element  \
0  Afghanistan  Crop Residues          Direct emissions (N2O)
1  Afghanistan  Crop Residues        Indirect emissions (N2O)
2  Afghanistan  Crop Residues                 Emissions (N2O)
3  Afghanistan  Crop Residues  Emissions (CO2eq) from N2O (AR5)
4  Afghanistan  Crop Residues          Emissions (CO2eq) (AR5)

      2000      2001      2002  ...      2018      2019      2020
0    0.520    0.5267    0.8200  ...    0.8988    1.2176    1.3170
1    0.117    0.1185    0.1845  ...    0.2022    0.2740    0.2963
2    0.637    0.6452    1.0045  ...    1.1011    1.4916    1.6133
3  168.807  170.9884  266.1975  ...  291.7838  395.2689  427.5284
4  168.807  170.9884  266.1975  ...  291.7838  395.2689  427.5284

[5 rows x 25 columns]
```

2.3 Manipolazione dei dati

Il comando *.info()* è utile per visualizzare un riassunto generale della struttura del dataset. Oltre al numero di colonne e righe presenti, per ogni colonna si riporta il numero di valori non nulli e il tipo di dato che contiene. Si nota che il numero di righe non nulle delle colonne è differente, questo indica che sono presenti valori non specificati che potrebbero influenzare negativamente la validità delle stime statistiche. Si provvede quindi all'eliminazione delle righe che presentano tali valori nulli per garantire un'analisi basata su dati completi e rappresentativi in modo da ottenere risultati affidabili. In seguito si eliminano colonne superflue o non rilevanti ai fini della propria indagine e si rinominano.

[3]:
```
#Informazioni relative ai dati del dataset
df.info()
```

```
RangeIndex: 58765 entries, 0 to 58764
Data columns (total 25 columns):
 #   Column   Non-Null Count  Dtype
---  ------   --------------  -----
 0   Area     58765 non-null  object
 1   Item     58765 non-null  object
 2   Element  58765 non-null  object
 3   Unit     58765 non-null  object
 4   2000     55577 non-null  float64
 5   2001     54038 non-null  float64
 6   2002     54137 non-null  float64
 7   2003     54158 non-null  float64
 ...
 22  2018     54024 non-null  float64
 23  2019     53988 non-null  float64
 24  2020     53671 non-null  float64
dtypes: float64(21), object(4)
memory usage: 11.2+ MB
```

[4]:
```
#Elimino righe contenenti valori mancanti
df1 = df.dropna(how="any")

#Elimino colonne non pertinenti
del df1["Unit"]
```

[5]:
```
#Rinomino colonne
df1.columns = ["Paese","Causa","Elemento","kt2000","kt2001",
               "kt2002", ... ,"kt2019","kt2020"]
```

Controllando nel sito del FAO, si deduce che eventuali valori negativi riportati nel dataset possono essere un errore considerando che il dataset riporta valori in kilotonnellate annue e non il risultato in rapporto agli anni precedenti. Si procede dunque all'eliminazione di tali valori in quanto incoerenti col tipo di dato rappresentato e non conforme, poichè non possono esserci emissioni negative nella realtà.

[6]:
```
#Considero solo righe con valori non negativi
df2 = df1[(df1.kt2000 >= 0) & (df1.kt2001 >= 0) &
          (df1.kt2002 >= 0) & (df1.kt2003 >= 0) &
          ...
          (df1.kt2020 >= 0)]
```

Per non avere statistiche o grafici con dati non omogenei, si procede a dividere le righe del dataset in tre blocchi andando a creare tre dataframe differenti, uno relativo agli stati, uno relativo alle analisi globali e l'ultimo relativo a continenti e regioni del mondo. I dati sono stati divisi esplicitando il range di valori *[indice inferiore : indice superiore]* e tramite la funzione *.shape* si visualizzano le dimensioni della matrice per ogni dataset.

[7]:
```
#Dati relativi a Stati
df_stati = df2[0:43759]
df_stati.shape
```

[7]:
```
(43759, 24)
```

[8]:
```
#Dati Globali
df_mondo = df2[43760:44036]
df_mondo.shape
```

[8]:
```
(276, 24)
```

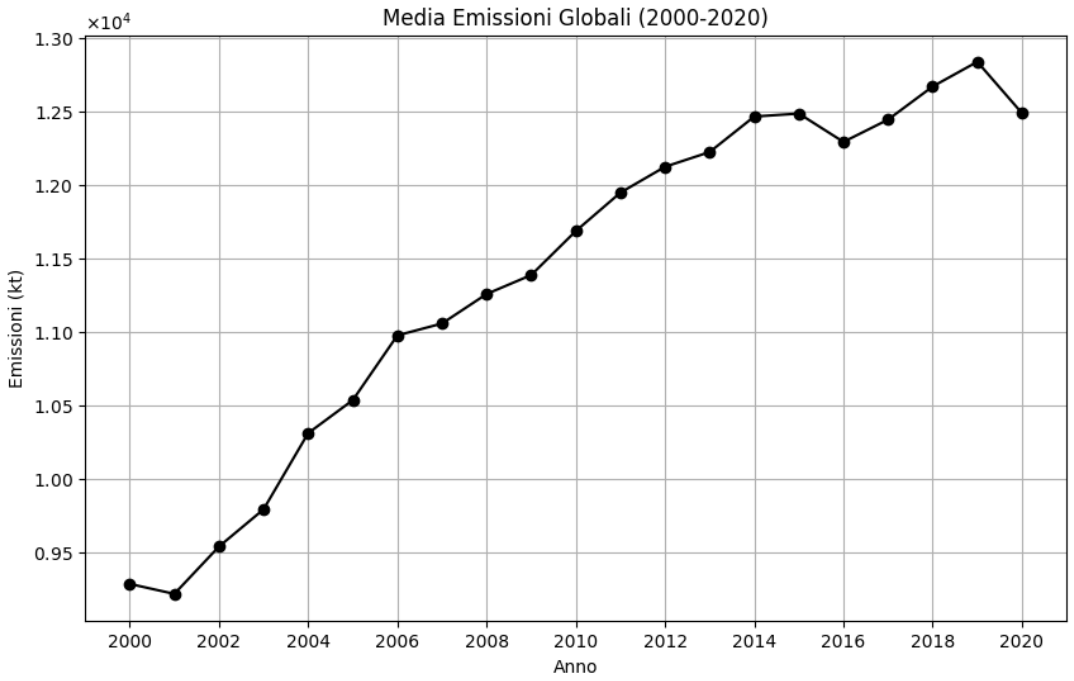
[9]:
```
#Dati relativi a Regioni del Mondo
df_reg_m = df2[44037:51685]
df_reg_m.shape
```

[9]:
```
(7648, 24)
```

2.4 Analisi delle Emissioni Medie Globali

Per avere un'idea generale di come è variato il trend medio di valori dall'anno 2000 al 2020 si procede con un'analisi anno per anno fra i valori medi calcolati per ogni colonna.

[10]:
```
#Media Emissioni Globali dal 2000 al 2020
data_stati = {'Media':[df_stati['kt2000'].mean(),
             df_stati['kt2001'].mean(),df_stati['kt2002'].mean(),
             ...
             df_stati['kt2019'].mean(),df_stati['kt2020'].mean()]}
df_mean = pd.DataFrame(data_stati)
```



Il grafico a linee riporta sull'asse delle ascisse gli anni e in quello delle ordinate le emissioni, rappresentate in kilotonnellate (kt). La linea collega i punti dei dati corrispondenti agli anni dal 2000 al 2020 e rappresenta l'andamento delle emissioni nel corso di questi anni. Si può osservare che la media delle emissioni è aumentata in maniera molto veloce fino al 2010 per poi crescere in modo più lento e diminuire dopo il 2019. Queste informazioni possono essere utilizzate per prendere decisioni in merito a politiche ambientali, agli obiettivi di riduzione delle emissioni o per scopi di monitoraggio.

2.5 Analisi dei Paesi con Emissioni Maggiori

Per capire quali sono i paesi maggiormente interessati nella produzione di emissioni, si procede isolando il dataset con solo le colonne numeriche relative agli anni e si mostra in output il valore massimo di emissione per ogni anno con il relativo paese.

[12]:
```
#Isolamento colonne numeriche
colonne = ['kt2000','kt2001', ... 'kt2018','kt2019','kt2020']

#Stampo valori medi massimi per anno col paese corrispondente
for colonna in df_stati_n.columns:
    massimo_valore = df_stati_n[colonna].max()
```

```
    print(df_stati[['Paese', colonna]].loc[df_stati[colonna] ==
          massimo_valore])
```

```
                        Paese       kt2000
48147                 America  7089877.342
                        Paese       kt2001
48147                 America  6949210.715
                        Paese       kt2002
48147                 America  6996378.210
                        Paese       kt2003
48147                 America  7041271.290
                        Paese       kt2004
9250                    China  7577858.028
                        Paese       kt2005
9250                    China  8375211.122

...
                        Paese       kt2017
9250                    China  13236412.68
                        Paese       kt2018
9250                    China  13526644.50
                        Paese       kt2019
9250                    China  13894626.13
                        Paese       kt2020
9250                    China  14091446.31
```

Dall'output si evince che fino al 2003 i picchi di massima emissione sono dovuti agli Stati Uniti, mentre negli anni a seguire questo spiacevole primato appartiene alla Cina.
Si procede confrontando i valori di emissioni medie globali con le emissioni medie dei primi paesi, classificati per produzione di emissioni decrescenti.

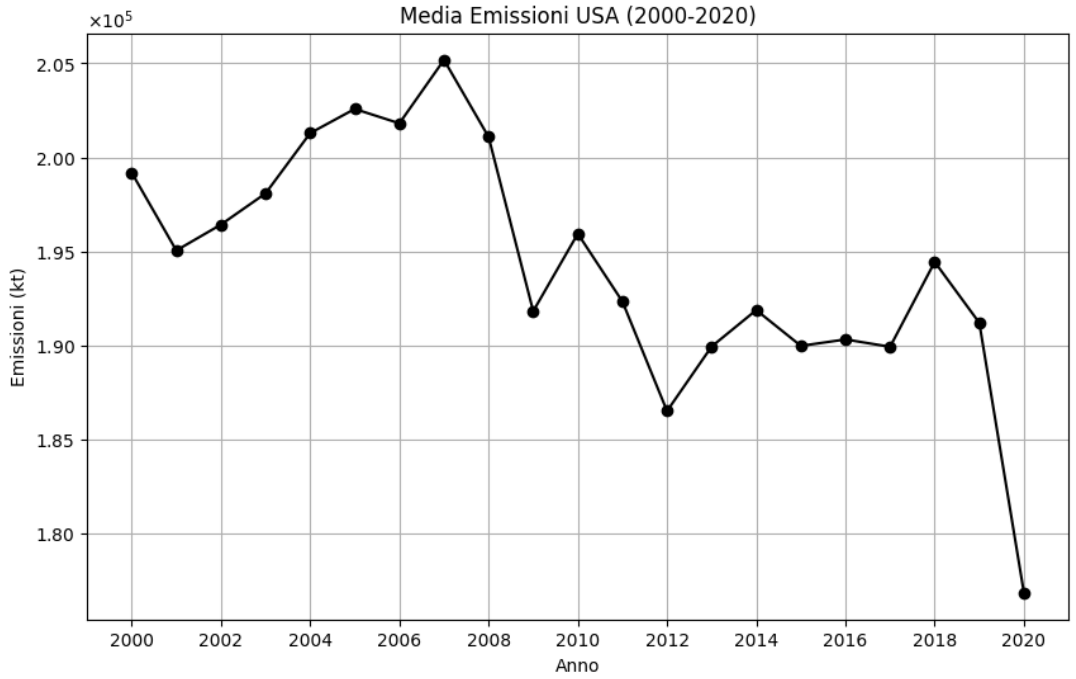
[13]:
```
#Classifica 10 paesi con Emissioni maggiori nel 2020
df_stati_sub=df_stati.groupby(['Paese'])[['kt2000',
                     'kt2020']].sum().reset_index()
df1_stati_sub = df_stati_sub.sort_values(by='kt2020',
                     ascending=False).reset_index()
df1_stati_sub.head(10)
```

[13]:

	index	Paese	kt2000	kt2020
0	43	China	4.278635e+07	9.770429e+07
1	47	China, mainland	4.074325e+07	9.544711e+07
2	220	United States of America	5.258552e+07	4.669152e+07
3	98	India	2.070323e+07	3.418404e+07
4	27	Brazil	2.801345e+07	2.211231e+07
5	175	Russian Federation	1.639577e+07	1.851246e+07
6	99	Indonesia	1.601555e+07	1.612555e+07
7	58	Democratic Republic of the Congo	6.013150e+06	9.687607e+06
8	107	Japan	9.986418e+06	9.032342e+06
9	100	Iran (Islamic Republic of)	4.211822e+06	7.609467e+06

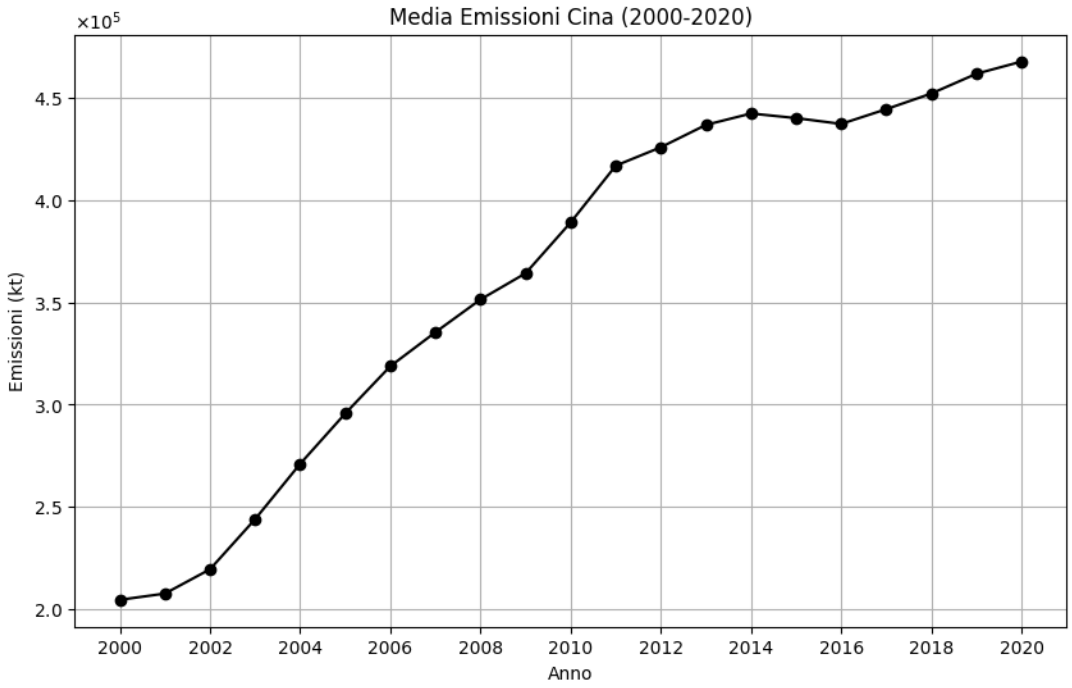
In questa classifica i primi due paesi per emissione sono sempre la Cina e gli Stati Uniti, si procede quindi a rappresentare i grafici delle medie dei due paesi singolarmente, per poi confrontarli col grafico delle emissioni medie globali.

[14]:
```
#Isolamento righe relative agli USA
df_USA = df_stati[df_stati.Paese == 'United States of America']

#Valori medi annui in USA (2000 - 2020)
data_USA = {'Media':[df_USA['kt2000'].mean(),df_USA['kt2001'].mean(),
                     ... df_USA['kt2020'].mean()]}
df_mean_USA = pd.DataFrame(data_USA)
```

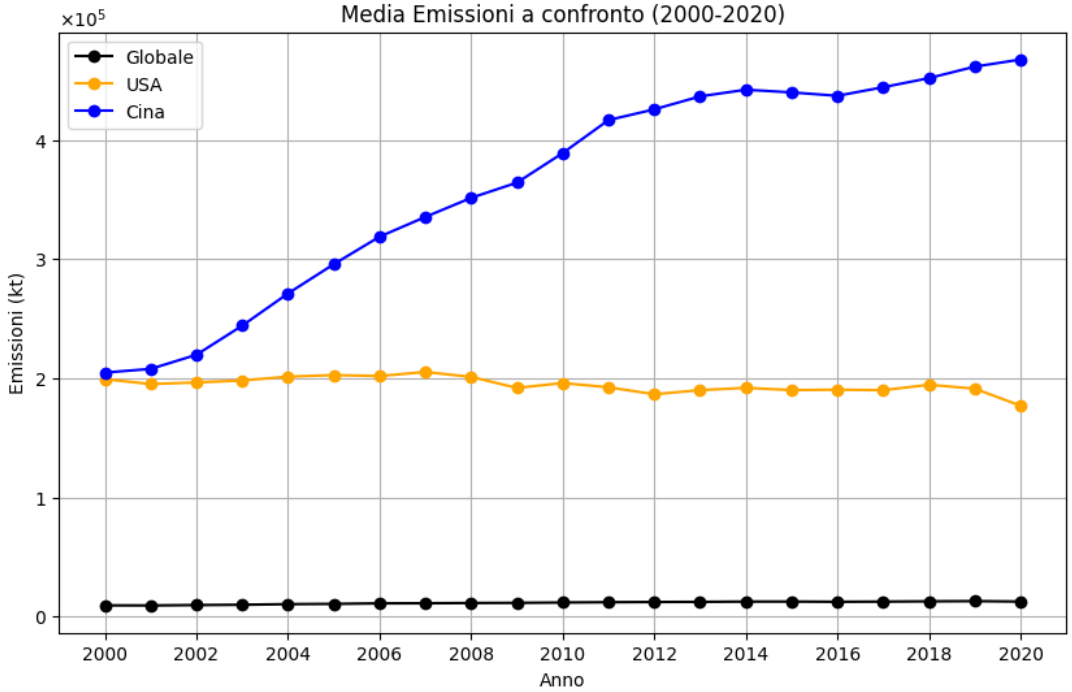


Il Grafico delle emissioni medie negli Stati Uniti evidenzia, nel periodo di tempo compreso fra il 2000 e il 2020 un andamento variabile, nel 2007 si assiste al punto di svolta: le emissioni erano costantemente cresciute fino a quell'anno, per poi iniziare a decrescere in maniera quasi costante, arrivando al 2020 a una sostanziale riduzione delle emissioni rispetto al primo periodo di valutazione dell'anno 2000.
Lo stesso procedimento si attua per la Cina.

[16]:
```
#Isolamento righe relative alla Cina
df_China = df_stati[df_stati.Paese == 'China']

#Valori medi annui in Cina (2000 - 2020)
data_China = {'Media':[df_China['kt2000'].mean(), ...
                       df_China['kt2020'].mean()]}
df_mean_China = pd.DataFrame(data_China)
```



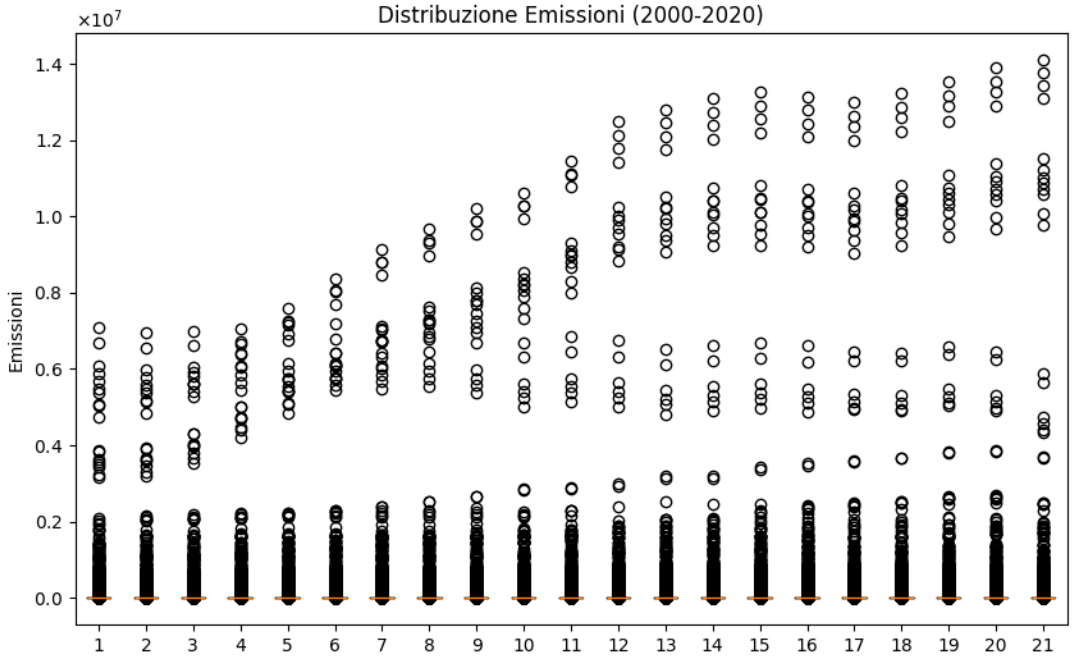
In questo caso si può notare come le emissioni medie della Cina, nel range di anni considerato, tende ad aumentare molto rapidamente fino al 2011 per poi rallentare, ma evidenziando un trend complessivo in crescita.
Per confrontare i grafici generati fino ad ora si visualizzano in un unico piano.



Si può notare come, rispetto alla media globale (nero), i due grafici di Cina (blu) e Stati Uniti (arancione) siano molto più in alto. Questo fa capire come ci sia uno squilibrio di emissioni fra alcuni paesi che producono enormi quantità di emissioni rispetto alla maggior parte dei paesi che produce emissioni per lo più intorno ai valori medi globali.

2.6 Analisi della Distribuzione dei Dati

Per come è strutturato il grafico delle medie globali a confronto, risulta interessante visualizzare come si distribuiscono i valori dal 2000 al 2020. Per questo si usa uno Scatterplot che prensenta sempre gli anni in ascissa e le emissioni in ordinata e marca con un pallino ogni valore contenuto nel dataset, evidenziando i punti di maggiore concentrazione di dati dai punti con valori di emissioni più rade.



Ogni punto nel grafico rappresenta un'osservazione o un'istanza dei dati ed è posizionato sul grafico in base al valore delle due variabili: l'ascissa contenente l'anno e l'ordinata contenente il valore di emissione. Un grafico di dispersione è utilizzato per valutare la distribuzione dei valori e osservare eventuali punti di cluster, insieme di dati concentrati o se sono presenti outlier, punti fuori dall'andamento principale. In questo grafico si nota che la maggior parte dei punti è concentrato nella parte inferiore del grafico contenente la media dei valori. Gli outliers in questo caso sono tutti quei paesi che producono enormi quantità di emissioni rispetto agli altri e vanno a posizionare i loro valori nella parte alta del grafico. Inoltre si possono notare delle linee di tendenza, curve che evidenziano l'andamento dei dati e i trend di valori nel tempo.
Il metodo *.describe()* è utile a visualizzare le statistiche generiche di un determinato dataset.

[20]:
```
#Statistiche generali stati
pd.options.display.float_format = '{:.2f}'.format
df_stati.describe()
```

[20]:

	kt2000	kt2001	...	kt2018	kt2019	kt2020
count	43759.00	43759.00	...	43759.00	43759.00	43759.00
mean	9284.50	9216.45	...	12672.62	12840.77	12491.35
std	114729.13	114783.64	...	210959.19	215230.88	214638.90
min	0.00	0.00	...	0.00	0.00	0.00
25%	0.04	0.03	...	0.05	0.06	0.05

```
50%         8.44       7.63     ...       10.59       11.16       10.86
75%       440.09     422.01     ...      582.08      589.10      584.15
max   7089877.34 6949210.71     ... 13526644.50 13894626.13 14091446.31
[8 rows x 21 columns]
```

Queste statistiche comprendono:

- count: il numero totale di osservazioni non nulle presenti in ciascuna colonna, che aiuta a capire se ci sono dati mancanti, nel nostro caso il numero di valori è uguale quindi non sono presenti dati nulli.
- mean: il valore medio delle osservazioni in ogni colonna numerica, che offre un'idea della tendenza centrale dei dati.
- std: la deviazione standard per misurare la dispersione dei dati rispetto alla media, se la deviazione standard è un valore molto alto, questo indica una forte variabilità nei dati, nel nostro grafico si può notare come nelle prime colonne la deviazione standard è più bassa e ci sono meno valori dispersi in alto, lontani dalla media, mentre nelle ultime colonne aumenta con l'aumentare della dispersione dei dati.
- min: valore minimo per ogni colonna.
- 25%: il primo quartile è il valore che separa il primo 25% dei dati più bassi dagli altri dati, se il valore delle emissioni per un determinato anno si trova al di sotto di questo valore significa che è tra le emissioni più basse dell'intervallo temporale considerato.
- 50%: il secondo quartile è il valore centrale dei dati, ovvero il punto in cui il 50% dei dati è al di sopra e il 50% è al di sotto.
- 75%: il terzo quartile è il valore che separa il primo 75% dei dati più bassi dal restante 25%, le emissioni superiori a questo valore comprendono le emissioni più alte dell'intervallo temporale considerato.
- max: il valore massimo per ogni colonna.

2.7 Analisi Percentuale di Emissioni per Continente

Un'altra interessante analisi serve ad evidenziare il contributo percentuale di cui ogni continente è responsabile nello scenario globale delle emissioni. Si isolano quindi i soli dati relativi ai continenti e si stampano i valori di emissioni totali per ogni continente riguardanti gli anni 2000 e 2020.

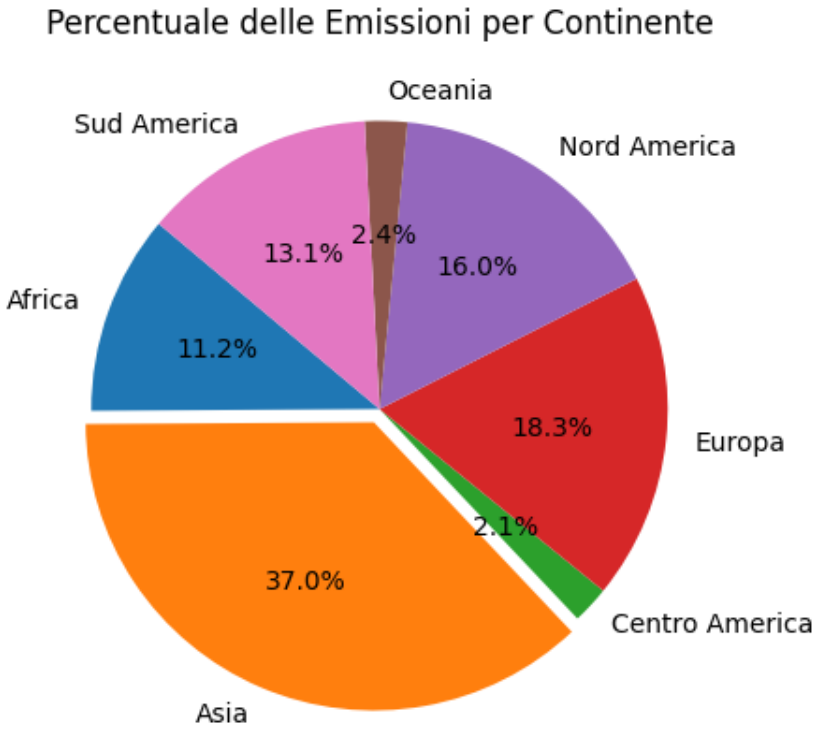
[21]:
```python
#DataFrame relativo ai continenti
df_continenti = df_reg_m[(df_reg_m.Paese == 'Africa') |
                         (df_reg_m.Paese == 'Northern America') |
                         (df_reg_m.Paese == 'Central America') |
                         (df_reg_m.Paese == 'South America') |
                         (df_reg_m.Paese == 'Asia') |
                         (df_reg_m.Paese == 'Europe') |
                         (df_reg_m.Paese == 'Oceania')]
```

[22]:
```
#DataFrame con somma totale emissioni 2000 - 2020 per Continente
df_continenti_sub = df_continenti.groupby(['Paese'])[['kt2000',
                    'kt2020']].sum().reset_index()
df_continenti_sub = df_continenti_sub.reset_index()
df_continenti_sub.head(7)
```

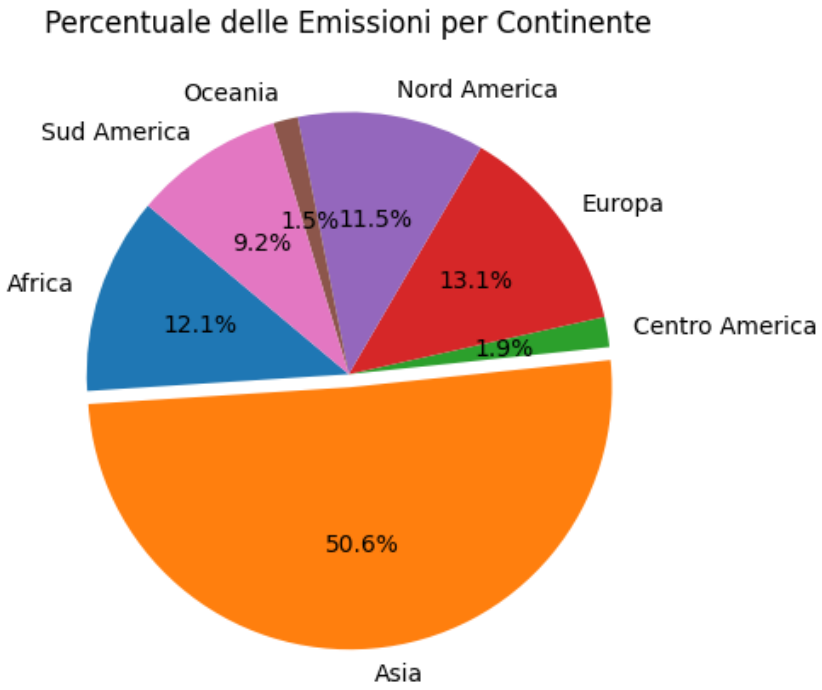
[22]:

	index	Paese	kt2000	kt2020
0	0	Africa	39866263.18	53563862.53
1	1	Asia	131456628.99	223558619.73
2	2	Central America	7461836.72	8201327.63
3	3	Europe	65200615.11	57744007.03
4	4	Northern America	56864549.55	50932575.33
5	5	Oceania	8425663.67	6721309.43
6	6	South America	46475289.92	40779874.32

Per una migliore visualizzazione dei dati ci si aiuta con un grafico a torta che esplicita le percentuali per continente.



Il grafico sopra è relativo all'anno 2000, si nota come i principali responsabili in questo anno per le emissioni sono l'Asia al primo posto col 37%, l'Europa col 18.3% e il Nord America col 16%. Insieme questi tre continenti sono responsabili di quasi 3/4 delle emissioni globali.



Questo grafico invece comprende i dati relativi all'anno 2020, si nota come i primati siano invariati e detenuti da Asia, Europa e Nord America, il valore più evidente risulta quello dell'Asia con un contributo in termini di emissioni pari al 50% delle emissioni globali.

2.8 Analisi delle Tipologie di Emissioni 2020

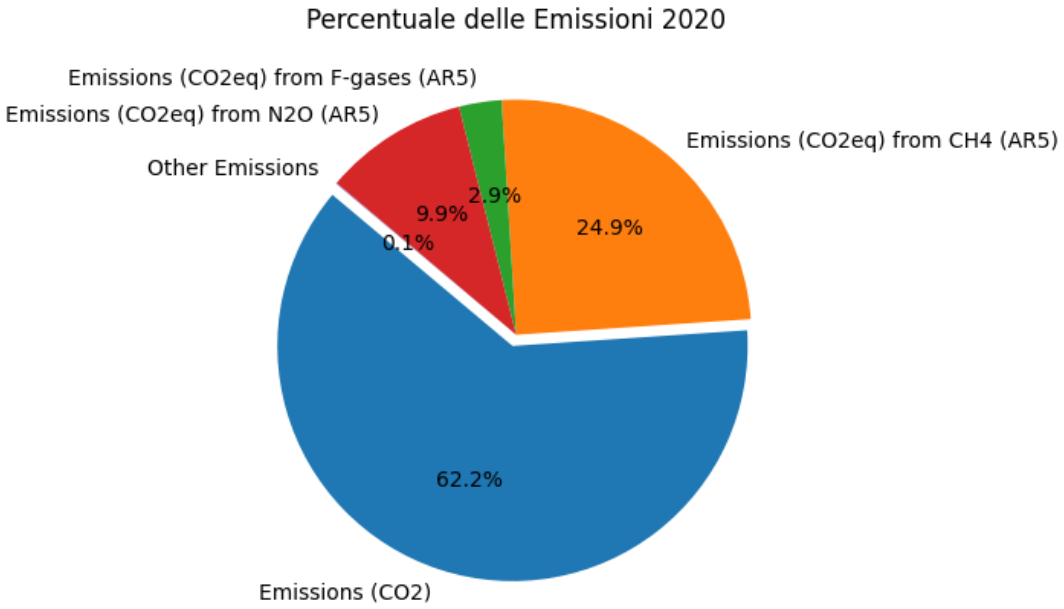
Concentrandosi sull'ultimo anno di analisi contenuto nel dataset, il 2020, è interessante studiare quali sono le tipologie di gas maggiormente responsabili del cambiamento climatico e del riscaldamento globale. Di seguito vengono riportati i valori totali di emissione nel 2020 raggruppati per tipo di elemento.

[25]:
```
#DataFrame tipologie di Emissioni nell'anno 2020
df_continenti_sub = df_continenti.groupby(['Elemento'])[[
    'kt2020']].sum().reset_index()
df_continenti_sub.head(9)
```

[25]:

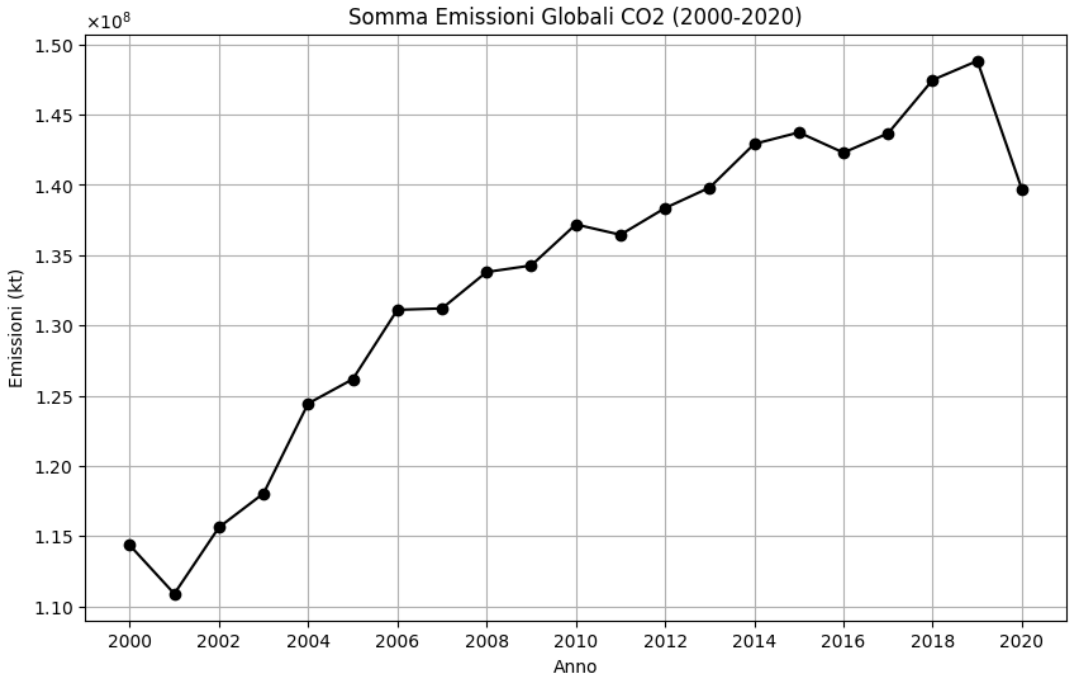
	Elemento	kt2020
0	Direct emissions (N2O)	20365.76
1	Emissions (CH4)	1956540.38
2	Emissions (CO2)	136644224.55
3	Emissions (CO2eq) (AR5)	219776135.73
4	Emissions (CO2eq) from CH4 (AR5)	54783130.72
5	Emissions (CO2eq) from F-gases (AR5)	6419512.19
6	Emissions (CO2eq) from N2O (AR5)	21813338.78
7	Emissions (N2O)	82314.49
8	Indirect emissions (N2O)	6013.41

In questa tabella sono riportati sia i valori effettivi sia quelli in CO2eq. Il CO2eq è l'abbreviazione di "Carbon Dioxide Equivalent", si tratta di una misura utilizzata per esprimere l'effetto di riscaldamento globale causato da diverse sostanze chimiche oltre all'anidride carbonica (CO2). I gas serra hanno diversi potenziali di riscaldamento globale, il CO2eq viene utilizzato per standardizzare il valore di ogni elemento e confrontare l'effetto di riscaldamento globale di queste sostanze in modo che possano essere valutate in una scala comune. L'acronimo "AR5", che si nota fra parentesi in alcune etichette, fa riferimento al "Quinto Rapporto di Valutazione" ("Fifth Assessment Report"), rapporto pubblicato nel 2014. Esso fornisce una revisione completa delle conoscenze scientifiche sul cambiamento climatico, inclusi dati sulle emissioni di gas serra, impatti ambientali, modelli climatici e scenari futuri. Questi rapporti sono utilizzati per informare le decisioni politiche a livello mondiale e nazionale riguardo alle politiche climatiche e alle azioni per affrontare il cambiamento climatico.



Si nota che la CO2 è il principale elemento inquinante con una percentuale molto maggiore rispetto alle altre emissioni, oltre 62%. A seguire si ha il metano CH4 con una percentuale di circa il 25% e il protossido di azoto N2O con una percentuale del 10% circa.
Possiamo quindi visualizzare il trend globale di anidride carbonica dal 2000 al 2020 per vedere come è variato.

[27]:
```
#Dataset con Emissioni CO2 Globali dal 2000 al 2020
df_CO2 = df_mondo[(df_mondo.Elemento == 'Emissions (CO2)')]
data_CO2 = {'Somma':[df_CO2['kt2000'].sum(),df_CO2['kt2001'].sum(),
                     df_CO2['kt2002'].sum(),df_CO2['kt2003'].sum(),
                     df_CO2['kt2004'].sum(),df_CO2['kt2005'].sum(),
                     df_CO2['kt2006'].sum(),df_CO2['kt2007'].sum(),
                     df_CO2['kt2008'].sum(),df_CO2['kt2009'].sum(),
                     ...
                     df_CO2['kt2018'].sum(),df_CO2['kt2019'].sum(),
                     df_CO2['kt2020'].sum()]}
df_sum_CO2 = pd.DataFrame(data_CO2)
```



Dal grafico si vede come la variazione delle emissioni di CO2 è aumentata molto rapidamente anno per anno dal 2001 al 2019 ed è drasticamente diminuita nel 2020.

2.9 Analisi delle Principali Cause di Emissione

La seconda colonna del dataset riporta tutte le cause principali di inquinamento dal 2000 al 2020. E' utile visualizzare una classifica delle prime fonti di Emissioni ordinate partendo da quella di maggiore impatto sulle emissioni.

[29]:
```
#DataFrame principali cause Emissioni
df_mondo_cause = df_mondo.groupby(['Causa'])[['kt2000',
              'kt2020']].sum().reset_index()
df1_mondo_cause = df_mondo_cause.sort_values(by='kt2020',
                    ascending=False).reset_index()
df1_mondo_cause.head(15)
```

[29]:

	index	Causa	kt2000	kt2020
0	3	All sectors with LULUCF	77909300.62	104398752.73
1	11	Energy	53126197.75	71334820.24
2	4	All sectors without LULUCF	37307030.79	50989635.53
3	1	Agri-food systems	29910451.11	32477811.71
4	10	Emissions on agricultural land	22141843.56	21237934.71
5	23	IPCC Agriculture	22704799.84	15299657.26
6	0	AFOLU	14231114.24	15004237.21
7	13	Farm-gate emissions	13330697.46	14944400.30
8	35	Pre- and post- production	7768607.55	11239877.00
9	24	IPPU	4652336.05	9967715.07

Dall'output si può intuire come le principali cause inquinanti a livello di emissioni siano legate ai settori che sfruttano il suolo e le foreste (LULUCF), all'Energia e relativi alla produzione agro-alimentare.

2.10 Considerazioni

L'analisi esplorativa dei dati (EDA) sulle emissioni svolge un ruolo fondamentale nell'interpretare il contenuto del dataset e nell'acquisire una comprensione più approfondita delle dinamiche delle emissioni di gas serra nel periodo considerato. Questo processo può essere paragonato a una sorta di esplorazione in cui i dati grezzi vengono esaminati e visualizzati in componenti più significative. In particolare, ciò che emerge da questa analisi può essere ulteriormente affinato e utilizzato come base per sviluppare strategie di mitigazione e politiche per affrontare il cambiamento climatico.

Uno degli aspetti principali dell'EDA in questo dataset è l'analisi delle tendenze temporali delle emissioni. Ciò implica la creazione di grafici che mostrano come le emissioni di CO2, N2O e CH4 siano cambiate nel corso degli anni, dal 2000 al 2020. Questi grafici consentono di individuare se ci sono aumenti costanti, diminuzioni o fluttuazioni significative nel tempo. Ad esempio, da quello che è emerso, le emissioni sono notevolmente cresciute dai primi anni di rilevamento, il che rappresenta una sfida significativa per la lotta al cambiamento climatico.

Tale analisi può inoltre suggerire anomalie o eventi straordinari che hanno causato picchi o cali significativi nelle emissioni, evidenziando i punti dove attuare un analisi più dettagliata. Ad esempio, una crescita improvvisa potrebbe essere associata a un aumento delle attività industriali, queste informazioni possono aiutare a comprendere meglio il panorama delle infrastrutture nei paesi studiati.

L'EDA consente anche di esaminare le emissioni su base geografica, suddividendo i dati per paese o continenti. Ciò rivela le differenze nelle emissioni tra nazioni, evidenziando i principali emettitori. Questo può essere utile per identificare i paesi che contribuiscono in modo significativo alle emissioni globali e determinare le aree in cui è necessario concentrare politiche di mitigazione.

L'analisi esplorativa aiuta anche a comprendere quali sono i principali gas serra (CO2, N2O e CH4), come interagiscano tra di loro e quali settori causano maggiormente inquinamento atmosferico.

Infine, l'EDA può essere un efficace strumento alla sensibilizzazione e alla comunicazione visuale delle tendenze delle emissioni, in modo da sviluppare piani mirati a invertire i trend più preoccupanti. Queste azioni possono includere politiche di riduzione delle emissioni, promozione delle energie rinnovabili, aumento dell'efficienza energetica, sostegno all'agricoltura sostenibile come suggerito nel primo capitolo. In sintesi, l'analisi esplorativa dei dati sulle emissioni è un primo passo fondamentale verso la comprensione del problema del cambiamento climatico e verso la progettazione di soluzioni efficaci per diminuire l'impatto delle attività umane sul clima globale.

3 Analisi Predittiva sulle Emissioni Globali 2023

Questo capitolo è incentrato sull'analisi predittiva dei dati relativi al 2023, raccolti da numerosi paesi in tutto il mondo, e comprendenti un vasto insieme di indicatori e attributi, tra cui statistiche demografiche, indicatori economici, fattori ambientali, parametri sanitari e dati sull'istruzione.
Inizialmente, si procede con una panoramica sul Machine Learning, esaminando definizioni, tipologie di apprendimento e il processo di formazione dei modelli. Di seguito, si attua una generica analisi descrittiva del dataset preso in esame, ampiamente discussa nel capitolo precedente. L'obiettivo principale di questo capitolo è applicare il Machine Learning per stimare le emissioni di CO2 per ciascun paese, utilizzando diversi algoritmi al fine di determinare quali di essi sono in grado di generare stime che si avvicinano maggiormente ai valori effettivi delle emissioni di CO2 rilevati nel 2023.

3.1 Introduzione al Machine Learning

"Il Machine Learning è il campo di studio che offre ai computer la capacità di apprendere senza essere esplicitamente programmati" (Arthur Samuel, 1959), più semplicemente è un'area dell'informatica che si concentra sullo sviluppo di algoritmi e modelli che possono analizzare dati, riconoscere modelli e fare previsioni o prendere decisioni basate su tali modelli. Un **algoritmo** rappresenta una procedura computazionale che viene applicata ai dati per sviluppare un modello di machine learning. Il **modello**, quindi, costituisce l'output derivante dall'applicazione di un algoritmo ai dati. Questo modello rappresenta ciò che il sistema ha imparato o estratto dai dati stessi ed è in grado di effettuare previsioni, classificazioni o altre operazioni specifiche in base all'apprendimento dai dati di addestramento. In sostanza, l'algoritmo è il motore di apprendimento, mentre il modello è il risultato dell'elaborazione dei dati e dell'apprendimento che l'algoritmo ha compiuto.

I principali tipi di Machine Learning per l'apprendimento sono:

- **Apprendimento supervisionato**: il modello è addestrato su un insieme di dati di addestramento etichettati, in cui le risposte corrette sono già conosciute. Questo tipo di addestramento è estremamente utile per compiere compiti di classificazione, con l'obiettivo di assegnare ciascun punto dei dati a una categoria o classe specifica, oppure di regressione, utilizzato per prevedere un valore numerico o un target in base a una serie di caratteristiche chiamate predittori.
- **Apprendimento non-supervisionato**: i dati di addestramento non sono etichettati e il sistema cerca di imparare in maniera autonoma. Tipologia utile per fare clustering, ovvero provare a rilevare gruppi di dati simili e trovare connessioni in modo indipendente.
- **Apprendimento per rinforzo**: il modello apprende in un ambiente dinamico, può osservare l'ambiente, selezionare ed eseguire azioni per ottenere ricompense positive o penalità, deve quindi imparare da solo qual è la migliore politica per ottenere la massima ricompensa.

- **Apprendimento online o batch**: se online il sistema si addestra in modo incrementale con flussi di dati in sequenza, ogni fase di apprendimento è veloce ed economica, quindi il sistema può apprendere i nuovi dati al volo. Nella tipologia batch il sistema non è in grado di apprendere in modo incrementale, deve essere addestrato utilizzando tutti i dati disponibili, richiede tempo e risorse di calcolo, quindi in genere l'addestramento viene eseguito offline.
- **Apprendimento basato su istanze o modelli**: nel caso di istanze il modello impara esempi a memoria, generalizza a nuovi casi utilizzando una misura di somiglianza per confrontarli con gli esempi appresi. Un altro modo per generalizzare da una serie di esempi è costruire un modello di questi esempi e utilizzarlo per fare previsioni.

Con il set di dati che si va ad analizzare, si applica un'approccio supervisionato. Di seguito i passi per attuare il processo di apprendimento automatico che poi si studieranno in dettaglio andando avanti nel capitolo:

1. Raccolta dei dati: acquisizione dei dati rilevanti per lo studio da effettuare.
2. Elaborazione dei dati: pulizia e trasformazione dei dati da passare all'algoritmo.
3. Creazione di un modello: scegliere un algoritmo di machine learning appropriato e addestrare il modello sui dati di addestramento.
4. Valutazione e ottimizzazione del modello: misurare e migliorare le prestazioni del modello tramite metriche appropriate.
5. Implementazione: utilizzare il modello addestrato per fare previsioni o prendere decisioni.

3.2 Raccolta dei dati

La seconda analisi esplorativa, più generica, si basa sul dataset "Global Country Information Dataset 2023" (*https://www.kaggle.com/datasets/nelgiriyewithana/countries-of-the-world-2023/data*). L'obiettivo è quello di analizzare, durante l'anno 2023, quanto hanno influito i principali attributi riportati nel dataset riguardo alle emissioni di CO2, per poi andare a predire tali valori. I dati sono riportati in una tabella di 35 colonne, contenenti dati di aspetto sociale, economico, ambientale e di salute riguardo a ogni paese del mondo. Inizialmente si importano le librerie utili a visualizzare e manipolare i dati e si visualizza il dataset preso in esame.

[2]:
```
#Importo dataset e visualizzo le prime 5 righe
df = pd.read_csv('/Users/filippolupatelli/Desktop/'
                 'world-data-2023.csv')
```

[2]:

	Country	Density\n(P/Km2)	...	Latitude	Longitude
0	Afghanistan	60	...	33.94	67.71
1	Albania	105	...	41.15	20.17
2	Algeria	18	...	28.03	1.66
3	Andorra	164	...	42.51	1.52
4	Angola	26	...	-11.20	17.87

[5 rows x 35 columns]

3.3 Elaborazione dei dati e Matrice di Correlazione

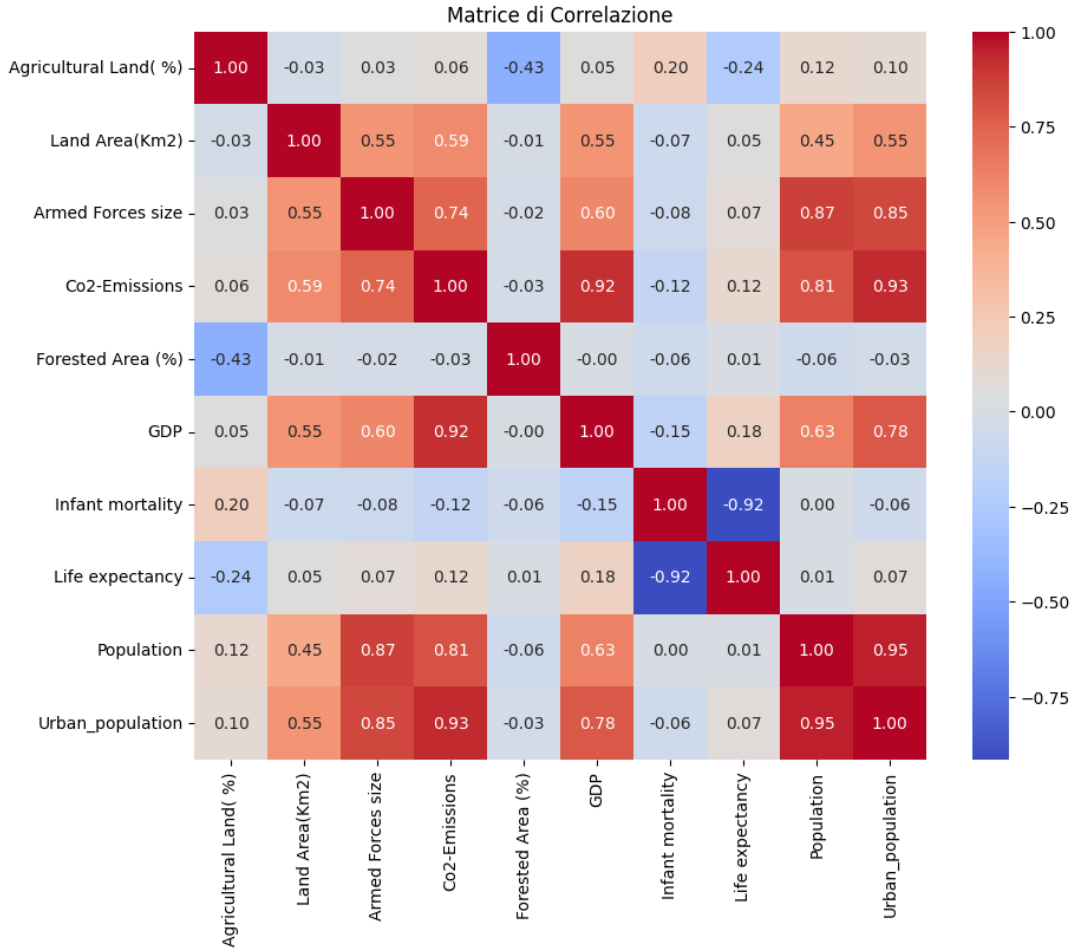
La creazione di una matrice di correlazione in un dataset è una tecnica comune nell'analisi dei dati che offre una panoramica delle relazioni tra le variabili presenti. Consiste in una tabella quadrata in cui le righe e le colonne rappresentano le variabili del dataset e gli elementi della tabella contengono i coefficienti di correlazione tra le coppie di variabili. Il valore dei coefficienti di correlazione sono utili per valutare la forza delle relazioni. I valori nella matrice di correlazione possono variare tra -1 e 1, un valore di 1 indica una correlazione positiva perfetta, -1 indica una correlazione negativa perfetta, mentre un valore di 0 indica assenza di correlazione. La diagonale principale contiene tutti coefficienti uguali ad 1 poichè si compara una variabile con se stessa.
In primo luogo, la matrice di correlazione permette di visualizzare il collegamento fra le diverse variabili nel dataset, approfondisce date connessione andando ad identificare se tale rapporto è positivo o negativo e può aiutare a ridurre le dimensioni del dataset, andando ad escludere relazioni deboli o quasi inesistenti.
Si procede quindi andando a creare la matrice, si considerano solo le colonne numeriche, trasformando anche tutte quelle colonne che contengono % o $ in variabili di tipo float. Inoltre si va a riempire le caselle dei valori nulli con la media della colonna, non andando coì ad influire sul risultato.

[3]:
```
#Tolgo simboli non numerici
for variable in NumericalVariables:
    if (DataSet[variable].dtypes == 'object'):
        DataSet[variable] = DataSet[variable].str.replace(',', '')
        DataSet[variable] = DataSet[variable].str.replace('%', '')
        DataSet[variable] = DataSet[variable].str.replace('$', '')
        DataSet[variable] = DataSet[variable].astype(float)

#Dati nulli sostituiti con media della colonna
for variable in NumericalVariables:
    DataSet[variable].fillna(DataSet[variable].mean(), inplace=True)
```

Si considerano solo le colonne numeriche di interesse che sono correlate alle emissioni:

- Agricultural Land(%)
- Land Area(Km2)
- Armed Forces size
- Co2-Emissions
- Forested Area (%)
- GDP
- Infant mortality
- Life expectancy
- Population
- Urban_population

Si realizza la matrice di correlazione e si visualizza.



Come si può evincere dalla tabella, guardando la colonna o la riga relativa alle emissioni di CO2, la variabile meglio correlata è se stessa, con un valore di 1.0. Altre variabili positivamente correlate risultano le colonne relative alla popolazione e alla popolazione urbana (0.81 e 0.93), ciò significa che all'aumentare della popolazione, le emissioni di CO2 tendono ad aumentare. Altro dato interessante è il valore rispetto al GDP (0.92), il livello di sviluppo economico di un paese, misurato dal GDP (Prodotto Interno Lordo), può influenzare significativamente le emissioni di CO2, i paesi con economie forti tendono a avere un maggior accesso all'energia e a una produzione industriale più elevata, il che contribuisce ad emissioni di CO2 più elevate. Altri dati da non trascurare sono le dimensioni delle forze armate (0.74) probabilmente in correlazione alle spese dovute a questo settore che influenzano positivamente le emissioni e le dimensioni del territorio (0.59) poichè in genere, le nazioni più estese, possono avere emissioni più elevate semplicemente perché hanno più spazio per ospitare industrie e attività.

3.4 Introduzione alla Regressione

La regressione è una tecnica statistica che viene utilizzata per studiare la relazione tra una variabile dipendente, chiamata anche risposta, e una o più variabili indipendenti, dette predittori. L'obiettivo della regressione è modellare e comprendere la relazione tra le variabili, nonché prevedere il valore della variabile dipendente sulla base delle variabili indipendenti.

Si importano le librerie utili per l'analisi predittiva:

[6]:
```
#Importo librerie per regressioni
from sklearn import datasets
from sklearn.neighbors import KNeighborsRegressor
from sklearn.ensemble import RandomForestRegressor
from sklearn.tree import DecisionTreeRegressor

from sklearn.model_selection import train_test_split
from sklearn.model_selection import GridSearchCV
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import mean_squared_error
from sklearn.metrics import r2_score
from sklearn.metrics import mean_absolute_error
```

Il termine sklearn che si nota nel codice fa riferimento a "Scikit-Learn", una libreria di machine learning in Python essenziale per sviluppare modelli di machine learning in modo semplice ed efficiente. In questo elaborato si utilizzano i seguenti algoritmi di regressione: k-Nearest Neighbors, Decision Tree e Random Forest. I passi che vengono effettuati per ogni algoritmo sono:

1. Divisione del set di train e di test: Il dataset viene diviso in un set di addestramento (80%) e un set di test (20%) utilizzando la funzione *train_test_split*. Il set di addestramento viene utilizzato per insegnare all'algoritmo, mentre il set di test per valutare le prestazioni del modello con target non noto.
2. Standardizzazione delle features: questo processo garantisce che tutte le variabili abbiano la stessa scala, evitando che variabili con scale diverse influenzino in modo distorto il modello.
3. Ricerca degli iperparametri: si esplorano diverse combinazioni di iperparametri attraverso una griglia definita da *GridSearchCV* per identificare la combinazione ottimale di parametri che massimizza le prestazioni del modello.
4. Addestramento del modello tramite il metodo *.fit* e predizione dei valori tramite il metodo *.predict*.
5. Valutazione dell'algoritmo: si valuta utilizzando diverse metriche di valutazione, tra cui il coefficiente di determinazione (R-Squared) che fornisce una misura della bontà di adattamento, l'Errore Quadratico Medio (MSE), la sua radice (RMSE) e l'Errore Assoluto Medio (MAE) che forniscono la qualità della precisione del modello.
6. Calcolo e visualizzazione dei residui.

Si riassumono di seguito le metriche di valutazione degli algoritmi per l'interpretazione dei risultati ottenuti:

Il coefficiente di determinazione **R-Squared** è una misura statistica che fornisce quanto bene le variabili indipendenti spiegano la variabilità della variabile dipendente. Il valore solitamente è compreso tra 0 e 1, uguale a 1 spiega completamente la variabile dipendente, 0 non la riesce a spiegare. Se il valore è positivo, indica che il modello è migliore di un modello che predice la media della variabile dipendente, in caso negativo, il modello è peggio di un modello di stima media.

$$R^2 = 1 - (Varianza\ residua / Varianza\ totale\ variabile\ dipendente)$$

La varianza residua è la somma dei quadrati delle differenze tra i valori predetti dal modello e i valori effettivi della variabile dipendente, mentre la varianza totale della variabile dipendente è la somma dei quadrati delle differenze tra ogni valore della variabile dipendente e la sua media.

L'errore quadratico medio **MSE** è una misura della differenza media quadratica tra i valori predetti da un modello e i valori effettivi della variabile dipendente. Tale valore fornisce una misura della precisione di un modello di previsione rispetto ai dati di addestramento.

$$MSE = 1/n \sum_{i=1}^{n} (y_i - y_i^*)^2$$

L'**RMSE** è la radice quadrata del MSE, entrambe metriche utili per valutare la precisione di un modello di regressione.

L'Errore Medio Assoluto **MAE** è anch'esso una metrica utilizzata per valutare la precisione del modello di previsione. A differenza dell'MSE, il MAE non eleva gli errori al quadrato, ma considera le differenze assolute tra i valori predetti e quelli effettivi.

$$MAE = 1/n \sum_{i=1}^{n} |y_i - y_i^*|$$

Rappresenta dunque la media delle differenze assolute tra valori predetti e quelli effettivi.

3.5 Regressione k-Nearest Neighbors

L'algoritmo k-Nearest Neighbors (kNN) è un algoritmo di machine learning utilizzato principalmente per problemi di regressione, l'idea di base è che i "vicini" tendono ad avere comportamenti simili, quindi, se un punto nello spazio di input è vicino a molti altri punti, è probabile che condivida le stesse caratteristiche. Nella regressione ogni punto ha un valore numerico associato e, quando si deve predire il valore per un nuovo punto, kNN calcola la media dei valori nei k punti più vicini e utilizza quel valore come previsione per il nuovo punto. Risulta quindi importante la scelta dei parametri della distanza, ad esempio quella euclidea, e dei k, numero di vicini da considerare. Un k piccolo rende il modello più sensibile al rumore nei dati, mentre un k grande può rendere il modello troppo generale.

Divisione del set di train e di test: il valore che l'algoritmo deve predire è l'emissione di CO2, si attribuiscono quindi le features alla variabile X, escludendo la colonna dell'emissione di CO2 che si attribuisce alla variabile y.

[7]:
```
#Definizione features e colonna target
X = df2[['Agricultural Land( %)', ... 'Urban_population']]
y = df2['Co2-Emissions']
```

Il dataset si divide per l'80% dei dati nel set di train che addestra l'algoritmo, passando le variabili features come input e avendo il valore target noto. In questo modo l'algoritmo riesce a memorizzare i risultati basati sulle variabili in ingresso che restituiscono il valore di emissione dato. Il restante 20% dei dati di test si passa all'algoritmo senza conoscere il valore target, questo viene predetto secondo le correlazioni individuate nei dati del set di addestramento.

[8]:
```
#Divisione dataset in set di addestramento e test
X_train, X_test, y_train, y_test = train_test_split(X, y,
                                test_size=0.2, random_state=42)
```

Standardizzazione delle features: la standardizzazione è importante quando le features hanno scale e unità di misura diverse. Gli algoritmo di regressione tendono ad assegnare pesi maggiori alle features con scale più grandi. La standardizzazione aiuta a garantire che tutte le features contribuiscano in modo eguale alla predizione del modello.

[9]:
```
#Standardizzazione delle features
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)
```

Ricerca degli iperparametri: questi valori sono importanti poichè determinano la buona riuscita del modello di predizione o una stima con significativi errori. Per la regressione KNN si cerca una combinazione di iperparametri ottimale che comprende il numero dei vicini da considerare durante la predizione, il peso che viene assegnato ai vicini, in modo uniforme o in base alla loro distanza, ciò significa che i punti più vicini contribuiranno maggiormente alla predizione rispetto ai vicini più lontani. Infine si considera la distanza tra i punti, quella più comune è di tipo euclideo, ma altre opzioni includono la distanza di Manhattan, somma del valore assoluto delle differenze di coordinate.

[10]:
```
#Definizione modello KNN
knn = KNeighborsRegressor()

#Definizione griglia degli iperparametri da esplorare
param_grid = {
    'n_neighbors': [3, 5, 7],
    'weights': ['uniform', 'distance'],
    'p': [1, 2]  #1 distanza di Manhattan, 2 euclidea
}
```

```
Migliori parametri: {'n_neighbors': 5, 'p': 1, 'weights': 'distance'}
```

Addestramento e predizione: Il modello di regressione viene addestrato utilizzando il set di addestramento, durante questa fase, il modello analizza le relazioni tra le features (variabili indipendenti) e la variabile target (emissione di CO2) in base agli iperparametri ottimali calcolati precedentemente. Una volta addestrato, il modello viene utilizzato per effettuare predizioni sul set di test.

[11]:
```
#Esecuzione addestramento
grid_search.fit(X_train_scaled, y_train)

#Valori predetti
y_pred = best_knn.predict(X_test_scaled)
```

[11]:
```
array([151413.24581696,  ...   61204.76643932, 170755.88168625])
```

Valutazione dell'algoritmo: si procede a calcolare i valori per valutare la bontà e la precisione dell'algoritmo KNN sul set di dati in esame.

[12]:
```
#Valutazione MSE sul set di test
mse1 = mean_squared_error(y_test, y_pred)
```

```
Mean Squared Error sul set di test: 1628400016.7878804
```

[13]:
```
#Valutazione RMSE sul set di test
rmse1 = np.sqrt(mean_squared_error(y_test, y_pred))
```

```
RMSE: 40353.4387232102
```

[14]:
```
#Valutazione R-Square sul set di test
r21 = r2_score(y_test, y_pred)
```

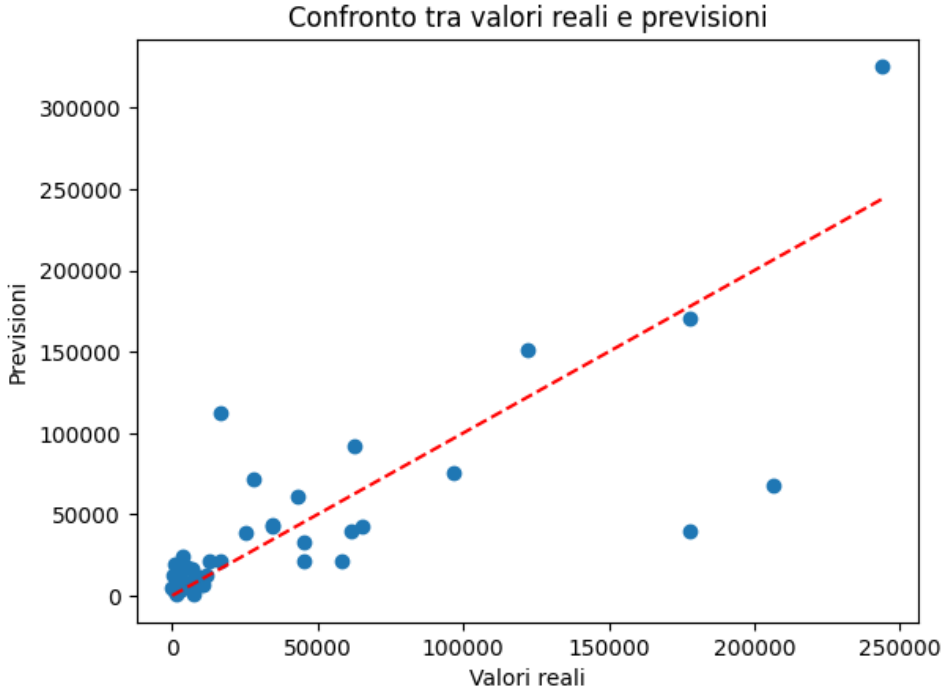
```
R-squared: 0.563007376067648
```

[15]:
```
#Valutazione MAE sul set di test
mae1 = mean_absolute_error(y_test, y_pred)
```

```
MAE: 22909.552085461535
```

Il valore calcolato di RMSE di circa 40.4 k tonnellate indica che, in media, le previsioni del modello possono deviare di circa 40.000 tonnellate rispetto ai valori effettivi delle emissioni. È importante considerare che questo valore può essere significativo quando si trattano dati di emissioni nell'ordine delle migliaia di tonnellate, mentre può sembrare meno rilevante per valori di emissioni nell'ordine dei milioni di tonnellate. Complessivamente, per l'interpretazione dei risultati si deve tener conto della scala delle emissioni e delle specifiche caratteristiche del contesto. In questo elaborato ci si concentra maggiormente sui Paesi con emissioni estremamente elevate in modo da valutare l'adeguatezza del modello per le situazioni più critiche. Lo stesso ragionamento può essere esteso al MAE, che è di circa 22.9 k tonnellate.
Per quanto riguarda il valore di R-squared di circa 0.56, questo indica che il 56% della variazione nelle emissioni è spiegato dalle variabili indipendenti nel modello. Un R-squared di 0.56 suggerisce una capacità moderata del modello di spiegare la variazione nelle emissioni.

Per visualizzare la relazione tra i dati reali e quelli predetti, si crea un grafico di confronto.

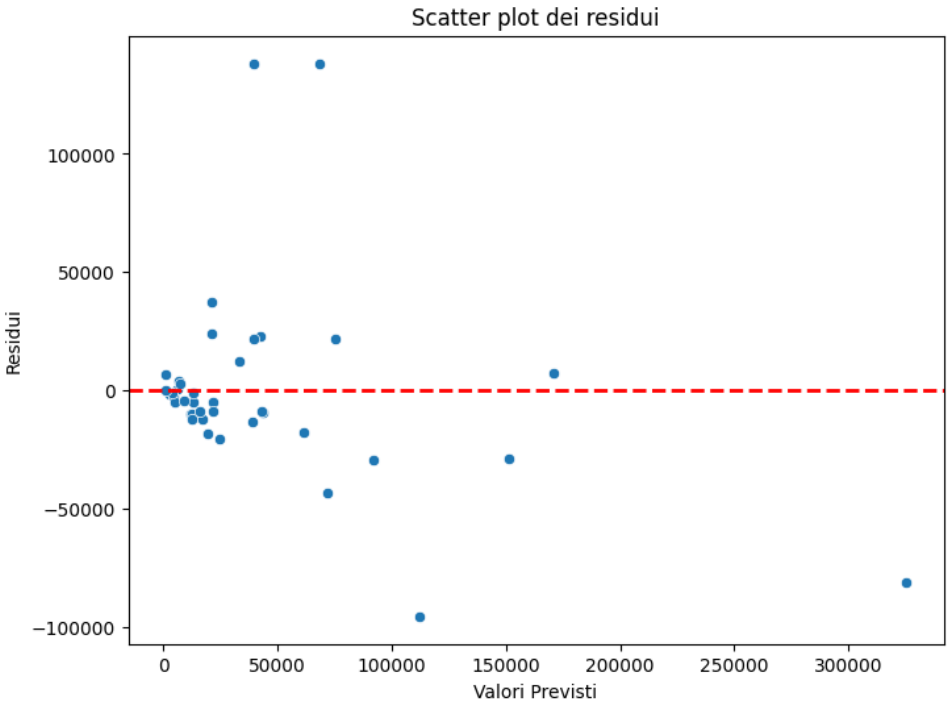


In un scenario perfetto in cui i dati predetti rispecchiano perfettamente quelli reali, tutti i punti si trovano lungo l'asse tratteggiato, se i punti nel grafico non si discostano significativamente dall'asse tratteggiato, ciò suggerisce che il modello è in grado di approssimare bene i dati. Nella figura sopra la maggior parte dei dati si trova in prossimità o di poco spostata rispetto alla linea, nel complesso i valori predetti quindi approssimano bene i valori reali delle emissioni.

Di seguito, i risultati sono presentati anche sotto forma di tabella per una visione più dettagliata, come si visualizza nel grafioco sopra, prendendo alcuni dati casuali fra quelli predetti e confrontandoli con valori reali si nota come per alcuni dati le previsioni sono molto simili, mentre in alcuni casi si discostano maggiormente.

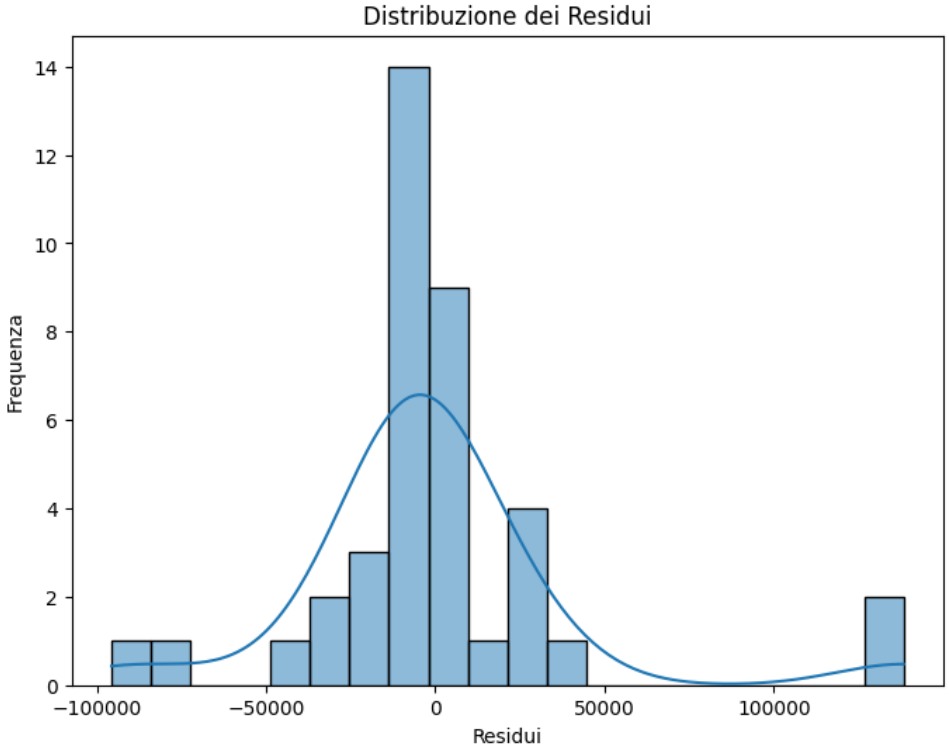
	Valore Effettivo	Previsioni	...	Valore Effettivo	Previsioni
138	122287.00	151413.25	172	11973.00	12802.84
16	96889.00	75195.12	127	28284.00	71821.04
155	1093.00	2719.31	169	34477.00	43124.53
96	1386.00	4151.56	19	1261.00	1162.29
68	16777.00	112390.98	168	43252.00	61204.77
153	45221.00	21311.79	73	177799.24	170755.88
...					

Per identificare potenziali outliers nel modello, è comune esaminare i residui, che rappresentano le differenze tra i valori reali e quelli previsti dal modello. Un modo efficace per visualizzare i residui è utilizzare un grafico residui, dove i punti sono rappresentati rispetto all'asse $y = 0$, linea orizzontale tratteggiata in rosso. Se i punti sono concentrati attorno a questa linea, indica che il modello ha gestito accuratamente i dati. Al contrario, i punti che si discostano notevolmente dalla linea possono essere considerati come residui outliers.



La maggioranza dei punti risiede vicino alla linea tratteggiata quindi il modello è buono per rappresentare i dati del set in esame.

Per visualizzare in modo più chiaro la presenza di outliers, si utilizza un grafico che rappresenta la distribuzione dei residui. La distribuzione dei residui dovrebbe essere approssimativamente normale e centrata intorno a zero per suggerire che il modello sta facendo previsioni accurate, se la distribuzione è distorta o ha code pesanti, può indicare la presenza di problemi, inclusi potenziali outliers.



Nel caso in esame, la distribuzione dei residui è approssimativamente normale e centrata intorno a zero. Questa osservazione suggerisce che il modello sta approssimando i dati in modo corretto, senza la presenza di outliers evidenti. Una distribuzione normale dei residui è un buon indicatore della validità del modello e della sua capacità di catturare le variazioni nei dati senza deviazioni significative.

3.6 Regressione Decision Tree

Nella regressione basata su alberi decisionali, l'obiettivo è prevedere un valore numerico, durante la fase di addestramento, l'albero cerca di dividere il set di dati in modo che la media dei valori target nei nodi foglia sia il più vicino possibile ai valori effettivi. Dopo l'addestramento, l'albero può essere utilizzato per fare previsioni. Un'istanza viene attraversata dall'albero, seguendo le regole di divisione, fino a raggiungere un nodo foglia. Il valore restituito dal nodo foglia è la previsione del modello per quella specifica istanza. Si procede con i medesimi passi attuati per l'algoritmo precedente.

Divisione del set di train e di test: anche in questa regressione il valore che l'algoritmo deve predire è l'emissione di CO2, si attribuiscono quindi le features alla variabile X, escludendo la colonna dell'emissione di CO2 che si attribuisce alla variabile y.

[20]:
```
#Definizione features e colonna target
X1 = df2[['Agricultural Land( %)', ... 'Urban_population']]
y1 = df2['Co2-Emissions']
```

Come per l'algoritmo precedente il dataset si divide per l'80% dei dati nel set di train che addestra l'algoritmo e il restante 20% dei dati di test si passa all'algoritmo senza conoscere il valore target, questo viene predetto secondo le correlazioni individuate nei dati del set di training. Si nota che il parametro *random state* è sempre 42 come per l'algoritmo KNN, questo parameto permette di scegliere gli stessi valori casuali per avere lo stesso set di addestramento in ogni algoritmo.

[21]:
```
#Divisione dataset in set di addestramento e test
X1_train, X1_test, y1_train, y1_test = train_test_split(X1, y1,
                                test_size=0.2, random_state=42)
```

Standardizzazione delle features: di nuovo si procede alla standardizzazione per garantire che tutte le features contribuiscano in modo eguale alla predizione del modello.

[22]:
```
#Standardizzazione delle features
scaler = StandardScaler()
X1_train_scaled = scaler.fit_transform(X1_train)
X1_test_scaled = scaler.transform(X1_test)
```

Ricerca degli iperparametri: per la regressione di alberi decisionali si cerca una combinazione di iperparametri ottimale che comprende la profondità massima dell'albero, il numero minimo di campioni richiesti in una foglia, per controllare la complessità dell'albero evitando foglie con un numero ridotto di campioni. Infine, il numero minimo di campioni richiesti per suddividere un nodo interno, questo parametro controlla la complessità

dell'albero impedendo suddivisioni che avrebbero portato a un numero insufficiente di campioni in uno dei nodi figli.

[23]:
```
#Definizione modello Regression Tree
tree = DecisionTreeRegressor()

#Definizione griglia degli iperparametri da esplorare
param_grid = {
    'max_depth': [None, 5, 10, 15],
    'min_samples_split': [2, 5, 10],
    'min_samples_leaf': [1, 2, 4]
}
```

```
Migliori parametri: {'max_depth': 15, 'min_samples_leaf': 1,
                     'min_samples_split': 2}
```

Addestramento e predizione: Il modello di regressione viene addestrato utilizzando il set di addestramento che viene utilizzato per ogni algoritmo di questo elaborato in base agli iperparametri ottimali calcolati precedentemente. Una volta addestrato, il modello viene utilizzato per effettuare predizioni sul set di test.

[24]:
```
#Esecuzione ricerca degli iperparametri
grid_search.fit(X1_train_scaled, y1_train)
#Valori predetti
y1_pred = best_tree.predict(X1_test_scaled)
```

[24]:
```
array([2.38560000e+05,   ...   1.70780000e+05, 1.77799239e+05])
```

Valutazione dell'algoritmo: si procede a calcolare i valori per valutare la bontà e la precisione dell'algoritmo Decision Tree sul set di dati in esame.

[25]:
```
#Valutazione MSE sul set di test
mse2 = mean_squared_error(y1_test, y1_pred)
```

```
Mean Squared Error sul set di test: 83292824435.72992
```

[26]:
```
#Valutazione RMSE sul set di test
rmse2 = np.sqrt(mean_squared_error(y1_test, y1_pred))
```

```
RMSE: 288604.96259719774
```

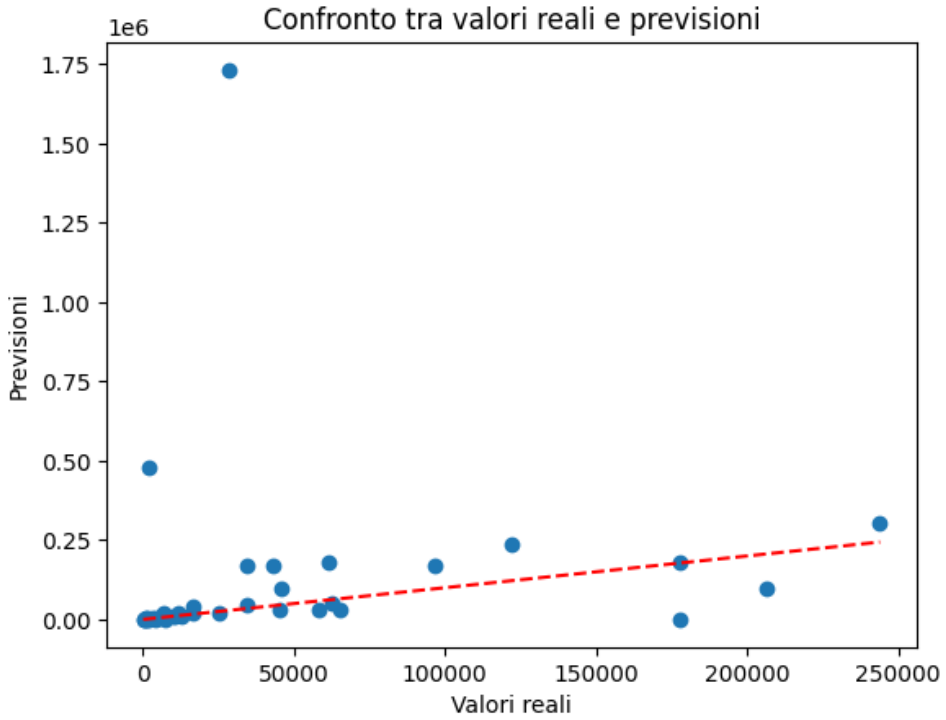
[27]:
```
#Valutazione R-Square sul set di test
r22 = r2_score(y1_test, y1_pred)
```

```
R-squared: -21.3522166111889
```

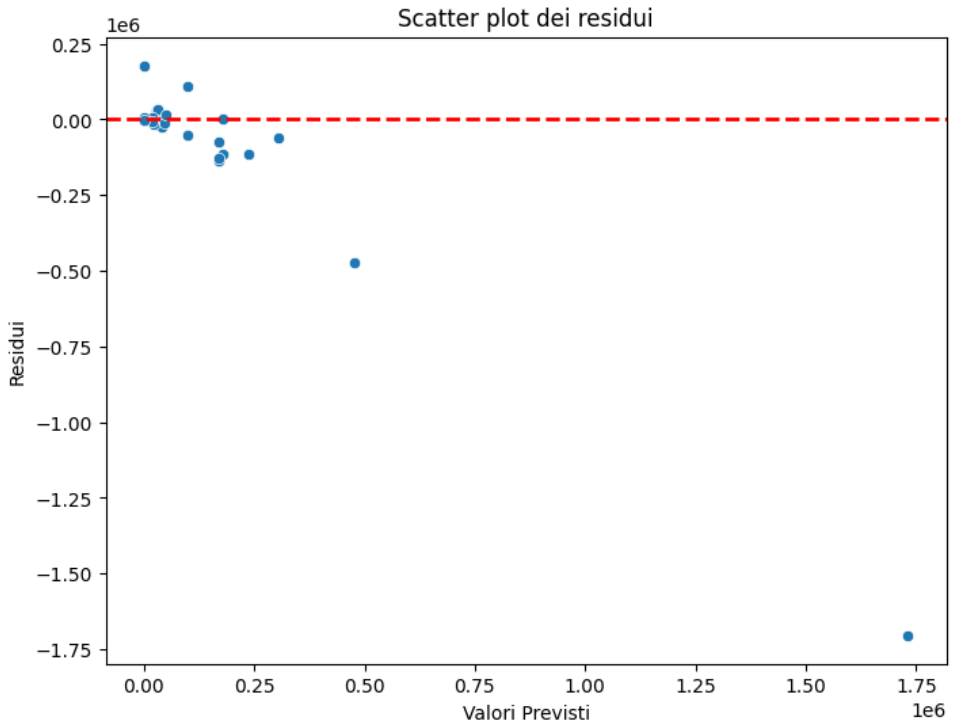
[28]:
```
#Valutazione MAE sul set di test
mae2 = mean_absolute_error(y1_test, y1_pred)
```

```
MAE: 85457.89261683942
```

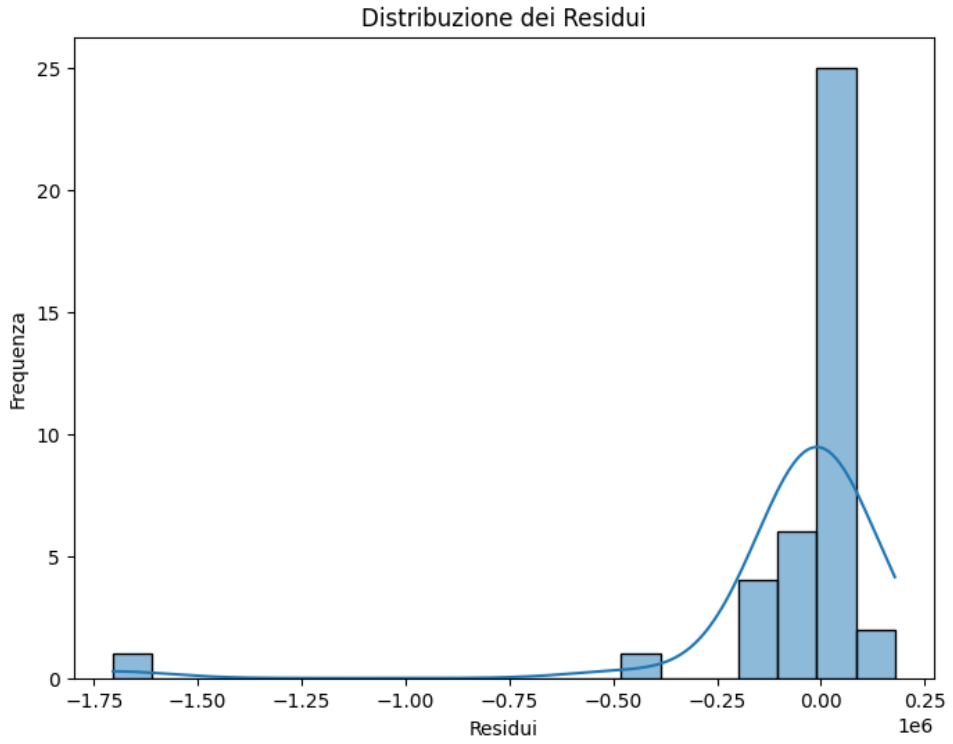
Dai valori calcolati, un RMSE di circa 288.6 k tonnellate è un indicatore elevato che suggerisce una scarsa accuratezza del modello. Lo stesso discorso si applica al MAE calcolato di circa 85.5 k tonnellate, il quale conferma la mancanza di precisione nelle previsioni del modello.
Il valore R-Squared di -21.4 indica che il modello non è adatto per i dati o che potrebbero esserci possibili outliers che influenzano negativamente le previsioni del modello. In generale, un R-Squared negativo suggerisce che il modello non riesce a spiegare la variazione nei dati in modo significativo.
Per visualizzare meglio la relazione tra i dati reali e quelli predetti, si realizza un grafico di confronto dove è evidente la presenza di possibili outliers che si discostano notevolmente dalla linea di perfetta corrispondenza.



Il calcolo e la visualizzazione dei residui confermano ulteriormente la presenza di problemi nel modello. L'individuazione di punti molto distanti dalla linea rossa tratteggiata indica la presenza di residui fortemente influenti. Di seguito, il grafico dei residui mostra chiaramente la presenza di punti che si discostano notevolmente dalla linea orizzontale, indicando residui outlier.



Anche la distribuzione dei residui è influenzata da questi punti e, come risultato, si ha una distribuzione distorta con valori outliers concentrati nella parte negativa del grafico.



Per individuare ed eliminare gli outliers in un modello di regressione basato su un albero decisionale si può utilizzare la tecnica dell'Intervallo Interquartile (IQR), l'IQR è la differenza tra il terzo quartile (Q3) e il primo quartile (Q1). Dopo aver calcolato questo valore si crea una serie booleana indicante se ciascun valore nella colonna 'Co2-Emissions' è considerato un outlier. Gli outlier sono definiti come quei valori che sono al di sotto di

(Q1 - 1.5 * IQR) o al di sopra di (Q3 + 1.5 * IQR).

[32]:
```
#Individuazione degli outlier utilizzando l'IQR
Q1 = df2['Co2-Emissions'].quantile(0.25)
Q3 = df2['Co2-Emissions'].quantile(0.75)
IQR = Q3 - Q1

outliers1 = ((df2['Co2-Emissions'] < (Q1 - 1.5 * IQR)) |
             (df2['Co2-Emissions'] > (Q3 + 1.5 * IQR)))
```

Dopo aver eliminato gli outliers si calcolano di nuovo le predizioni di emissioni utilizzando gli stessi passi seguiti per la stima prima dello studio degli outliers e si valuta nuovamente il modello.

[38]:
```
#Valutazione MSE sul set di test
mse3 = mean_squared_error(y1_test_no, y1_pred_no)
```

```
Mean Squared Error sul set di test: 2208272768.318957
```

[39]:
```
#Valutazione RMSE sul set di test
rmse3 = np.sqrt(mean_squared_error(y1_test_no, y1_pred_no))
```

```
RMSE: 46992.262855910194
```

[40]:
```
#Valutazione R-Square sul set di test
r23 = r2_score(y1_test_no, y1_pred_no)
```

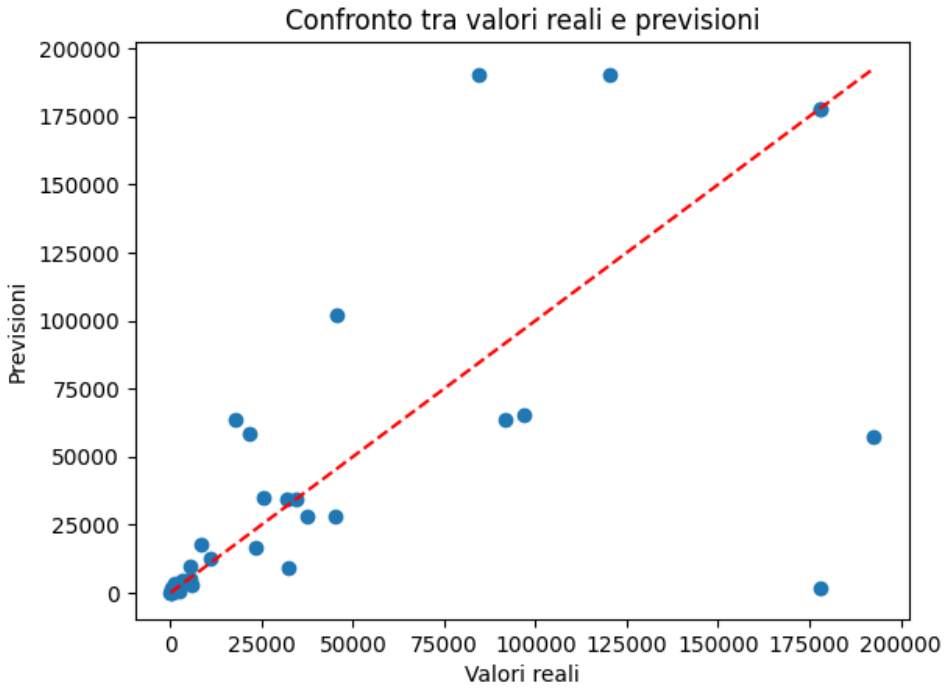
```
R-squared: 0.36374159263133754
```

[41]:
```
#Valutazione MAE sul set di test
mae3 = mean_absolute_error(y1_test_no, y1_pred_no)
```

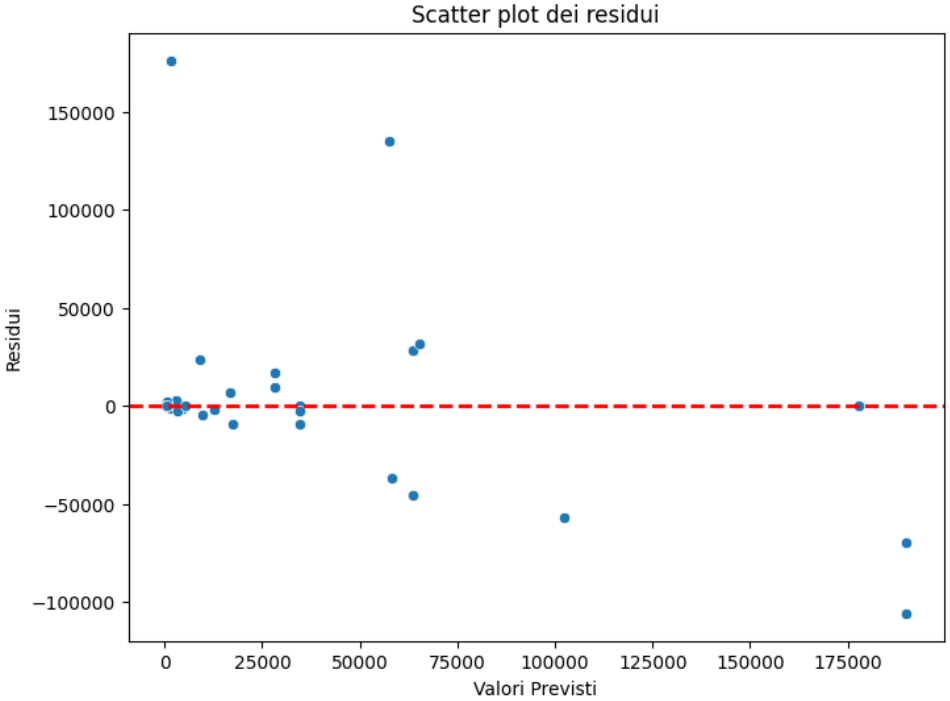
```
MAE: 22965.65409887359
```

Il calcolo del nuovo RMSE di circa 46.9 k tonnellate evidenzia un miglioramento della stima una volta rimossi gli outliers. Lo stesso miglioramento si riflette nel MAE calcolato, che è ora di circa 22.9 k tonnellate.
Il valore R-Squared di circa 0.36 indica che il 36% della variazione nelle emissioni è spiegato dalle variabili indipendenti. Tuttavia, questa percentuale suggerisce che il modello ha una capacità limitata di spiegare la variazione nelle emissioni.

Per visualizzare meglio la relazione tra i dati reali e quelli predetti dopo la rimozione degli outliers, è possibile creare nuovamente un grafico di confronto.

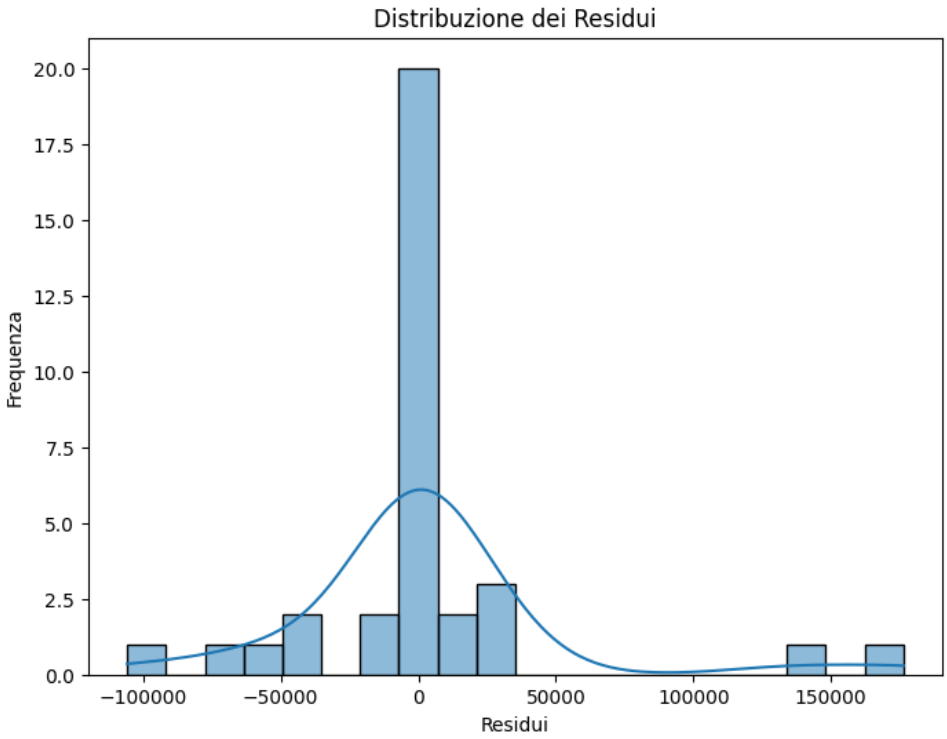


È notevole come, dopo le modifiche, i punti non si discostano significativamente dall'asse, indicando una migliore aderenza del modello ai dati reali rispetto alla stima precedente. Si procede a visualizzare i residui che, dopo l'eliminazione dei valori outliers sono meglio concentrati attorno alla linea e si mostrano i risultati reali e predetti in una tabella.



	Valore Effettivo	Previsioni	...	Valore Effettivo	Previsioni
33	297.00	495.00	46	31786.00	34477.00
134	10715.00	12633.00	62	532.00	495.00
165	23362.00	16777.00	171	5310.00	9787.00
56	177799.24	177799.24	126	120369.00	190061.00

Infine, si visualizza la distribuzione che risulta centrata attorno allo zero dopo l'eliminazione degli outliers. Si notano ancora punti che si discostano dallo 0 ma che mantengono la distribuzione normale.



3.7 Regressione Random Forest

Un Random Forest è un tipo di modello di machine learning basato su alberi decisionali ma, a differenza di un singolo albero decisionale, un Random Forest combina l'output di più alberi per migliorare la generalizzazione. Ciascun albero decisionale viene addestrato su un sottoinsieme casuale dei dati di addestramento e di feature, aiutando a creare alberi diversi e non correlati. In una regressione con Random Forest, ogni albero decisionale restituisce una previsione numerica. La previsione finale del Random Forest è la media delle previsioni di tutti gli alberi. Inoltre, è possibile ottimizzare le prestazioni del modello regolando gli iperparametri come il numero di alberi o la profondità massima degli alberi. Si procede con i medesimi passi attuati per gli algoritmi precedenti.

Divisione del set di train e di test: anche in questa regressione il valore che l'algoritmo deve predire è l'emissione di CO2, si attribuiscono quindi le features alla variabile X, escludendo la colonna dell'emissione di CO2 che si attribuisce alla variabile y.

[46]:
```
#Definizione features e colonna target
X2 = df2[['Agricultural Land( %)', ... 'Urban_population']]
y2 = df2['Co2-Emissions']
```

Come per l'algoritmo precedente il dataset si divide per l'80% dei dati nel set di train che addestra l'algoritmo e il restante 20% dei dati di test si passa all'algoritmo senza conoscere il valore target, questo viene predetto secondo le correlazioni individuate nei dati del set di training. Il *random state* è sempre 42 come per gli algoritmi precedenti per avere lo stesso set di addestramento.

[47]:
```
#Divisione dataset in set di addestramento e test
X2_train, X2_test, y2_train, y2_test = train_test_split(X2, y2,
                                    test_size=0.2, random_state=42)
```

Standardizzazione delle features: di nuovo si procede alla standardizzazione per garantire che tutte le features contribuiscano in modo eguale alla predizione del modello.

[48]:
```
#Standardizzazione delle features
scaler = StandardScaler()
X2_train_scaled = scaler.fit_transform(X2_train)
X2_test_scaled = scaler.transform(X2_test)
```

Ricerca degli iperparametri: per la regressione random forest si cerca una combinazione di iperparametri ottimale che comprende il numero totale di alberi nella foresta, aumentare il numero di alberi può migliorare la performance del modello, ma può comportare rendimenti decrescenti, la profondità massima dell'albero, il numero minimo di campioni richiesti in una foglia, per controllare la complessità dell'albero evitando foglie con un numero ridotto di campioni. Infine, il numero minimo di campioni richiesti per suddividere un nodo interno, questo parametro controlla la complessità dell'albero impedendo suddivisioni che avrebbero portato a un numero insufficiente di campioni in uno dei nodi figli.

[49]:
```
#Definizione modello Random Forest
rf = RandomForestRegressor()

#Definizione griglia degli iperparametri da esplorare
param_grid = {
    'n_estimators': [50, 100, 200],
    'max_depth': [None, 10, 20],
    'min_samples_split': [2, 5, 10],
    'min_samples_leaf': [1, 2, 4]
}
```

```
Migliori parametri: {'max_depth': 20, 'min_samples_leaf': 1,
                     'min_samples_split': 5, 'n_estimators': 50}
```

Addestramento e predizione: Il modello di regressione viene addestrato utilizzando il set di addestramento che viene utilizzato per ogni algoritmo di questo elaborato in base agli iperparametri ottimali calcolati precedentemente. Una volta addestrato, il modello viene utilizzato per effettuare predizioni sul set di test.

[50]:
```
#Esecuzione ricerca degli iperparametri
grid_search.fit(X2_train_scaled, y2_train)
#Valori predetti
y2_pred = best_rf.predict(X2_test_scaled)
```

[50]:
```
array([1.94255040e+05,    ...    3.52845024e+05, 2.03349156e+05])
```

Valutazione dell'algoritmo: si procede a calcolare i valori per valutare la bontà e la precisione dell'algoritmo Random Forest sul set di dati in esame.

[51]:
```
#Valutazione MSE sul set di test
mse4 = mean_squared_error(y2_test, y2_pred)
```

```
Mean Squared Error sul set di test: 23882295047.796272
```

[52]:
```
#Valutazione RMSE sul set di test
rmse4 = np.sqrt(mean_squared_error(y2_test, y2_pred))
```

```
RMSE: 154538.97582097622
```

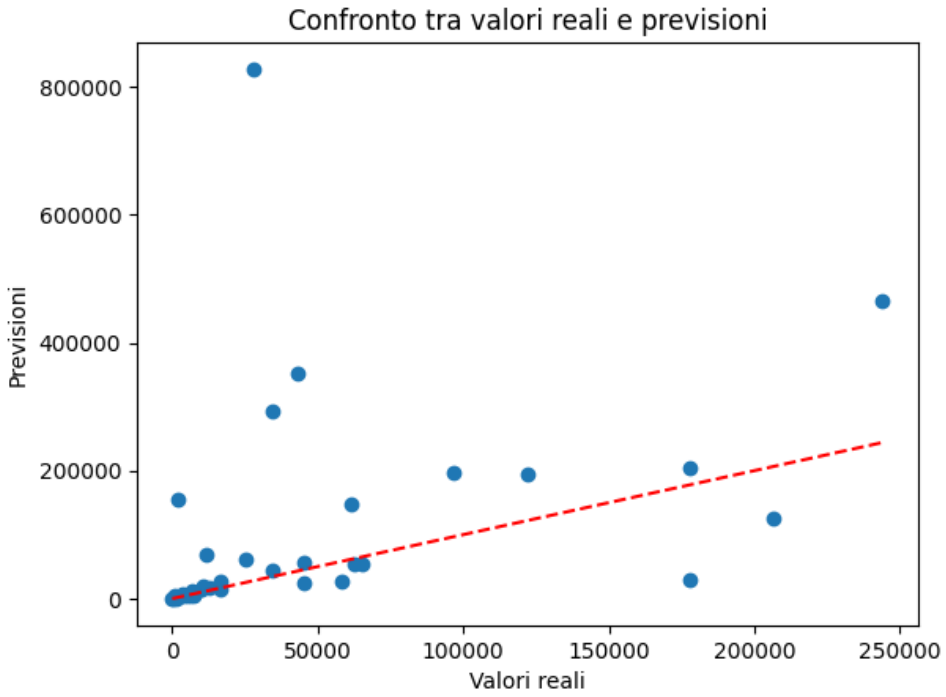
[53]:
```
#Valutazione R-Square sul set di test
r24 = r2_score(y2_test, y2_pred)
```

```
R-squared: -5.408982234629012
```

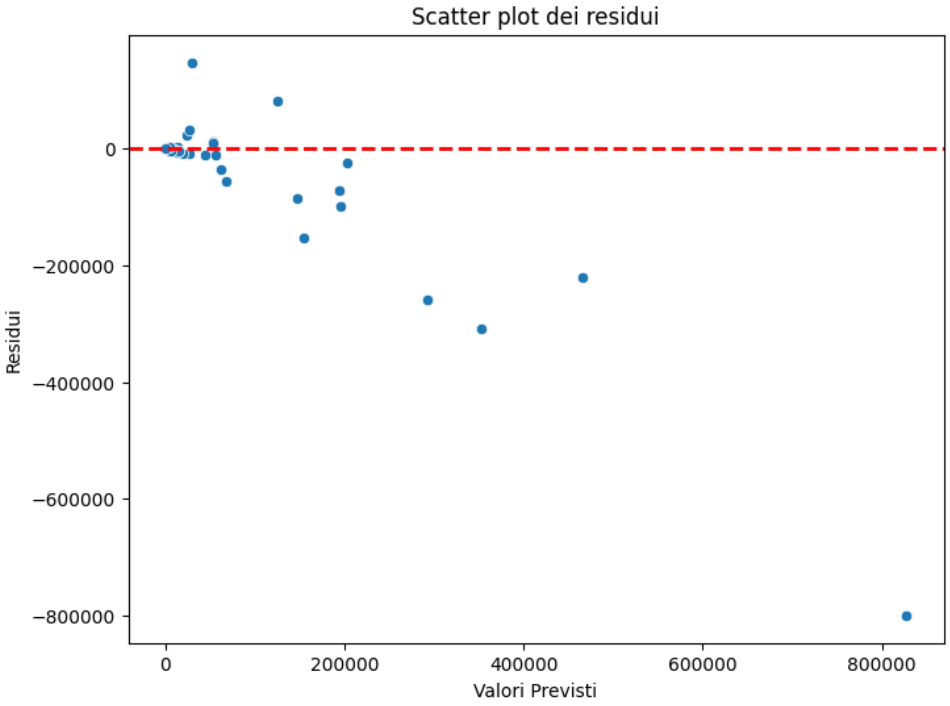
[54]:
```
#Valutazione MAE sul set di test
mae4 = mean_absolute_error(y2_test, y2_pred)
```

```
MAE: 63972.89596780524
```

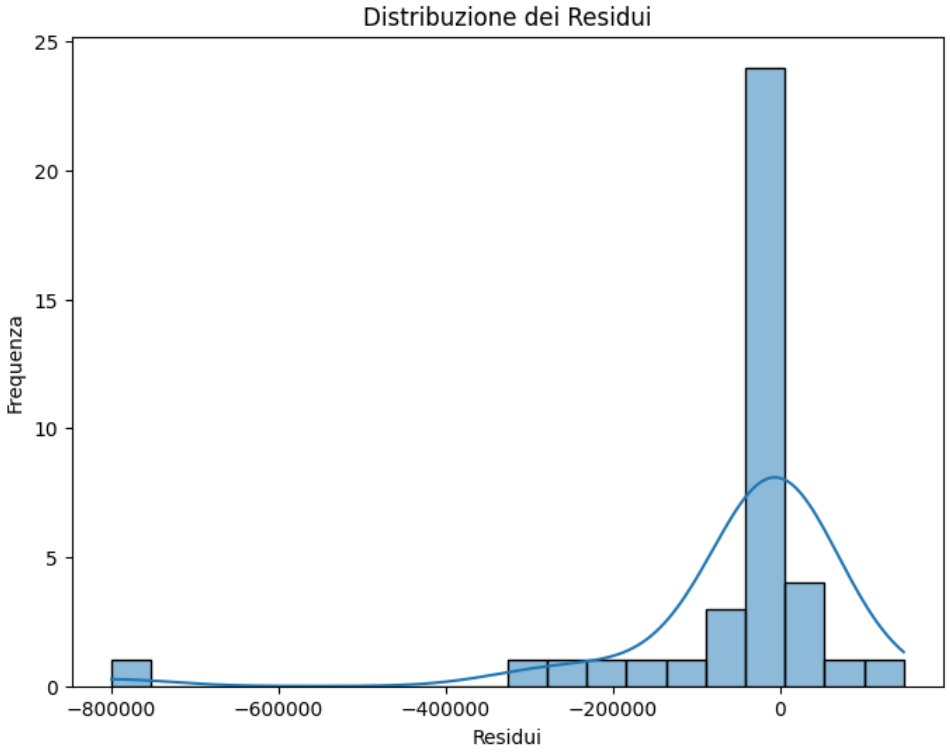
Dai valori calcolati, un RMSE di circa 154.5 k tonnellate è un indicatore elevato che suggerisce una scarsa accuratezza del modello. Lo stesso discorso si applica al MAE calcolato di circa 63.9 k tonnellate, il quale conferma la mancanza di precisione nelle previsioni del modello.
Il valore R-Squared -5.4 indica che il modello non è adatto per i dati o che potrebbero esserci possibili outliers che influenzano negativamente le previsioni del modello. In generale, un R-Squared negativo suggerisce che il modello non riesce a spiegare la variazione nei dati in modo significativo.
Per visualizzare meglio la relazione tra i dati reali e quelli predetti, si realizza un grafico di confronto dove è evidente la presenza di possibili outliers che si discostano notevolmente dalla linea di perfetta corrispondenza.



Il calcolo e la visualizzazione dei residui confermano ulteriormente la presenza di problemi nel modello. L'individuazione di punti molto distanti dalla linea rossa tratteggiata indica la presenza di residui fortemente influenti. Di seguito, il grafico dei residui mostra chiaramente la presenza di punti che si discostano notevolmente dalla linea orizzontale, indicando residui outlier.



Anche la distribuzione dei residui è influenzata da questi punti e, come risultato, si ha una distribuzione distorta che porta a visualizzare outliers nella parte negativa del grafico.



Per individuare ed eliminare gli outliers in un modello di regressione si può utilizzare la tecnica dell'Intervallo Interquartile (IQR) come nell'algoritmo precedente. Dopo aver eliminato gli outliers si ricalcolano le predizioni e si valuta nuovamente il modello.

[64]:
```
#Valutazione MSE sul set di test
mse5 = mean_squared_error(y2_test_no, y2_pred_no)
```

```
Mean Squared Error sul set di test: 1367654457.7114987
```

[65]:
```
#Valutazione RMSE sul set di test
rmse5 = np.sqrt(mean_squared_error(y2_test_no, y2_pred_no))
```

```
RMSE: 36981.81252604446
```

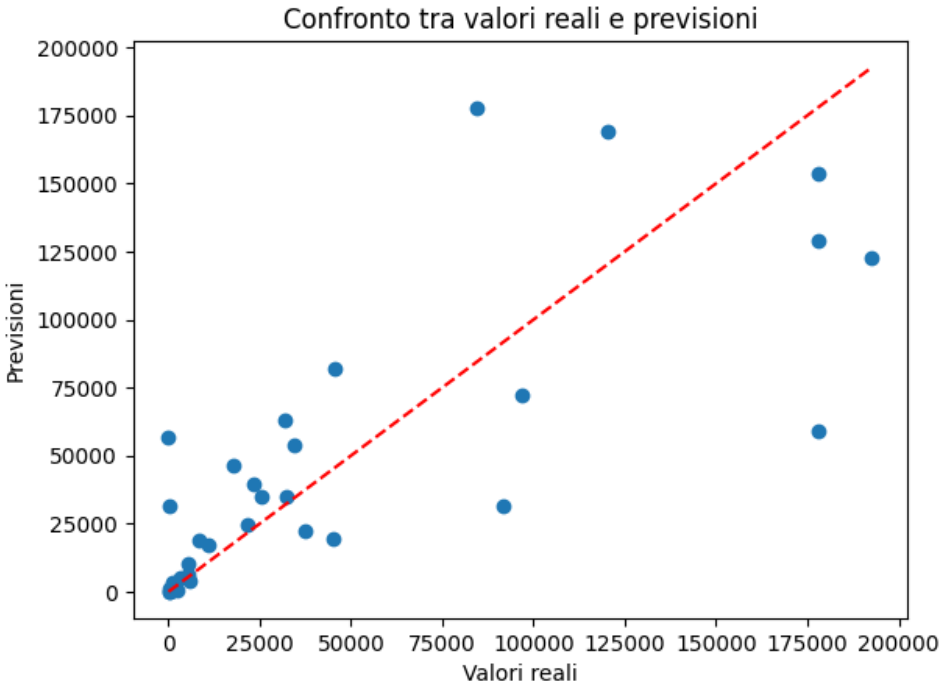
[66]:
```
#Valutazione R-Square sul set di test
r25 = r2_score(y2_test_no, y2_pred_no)
```

```
R-squared: 0.6059446733310061
```

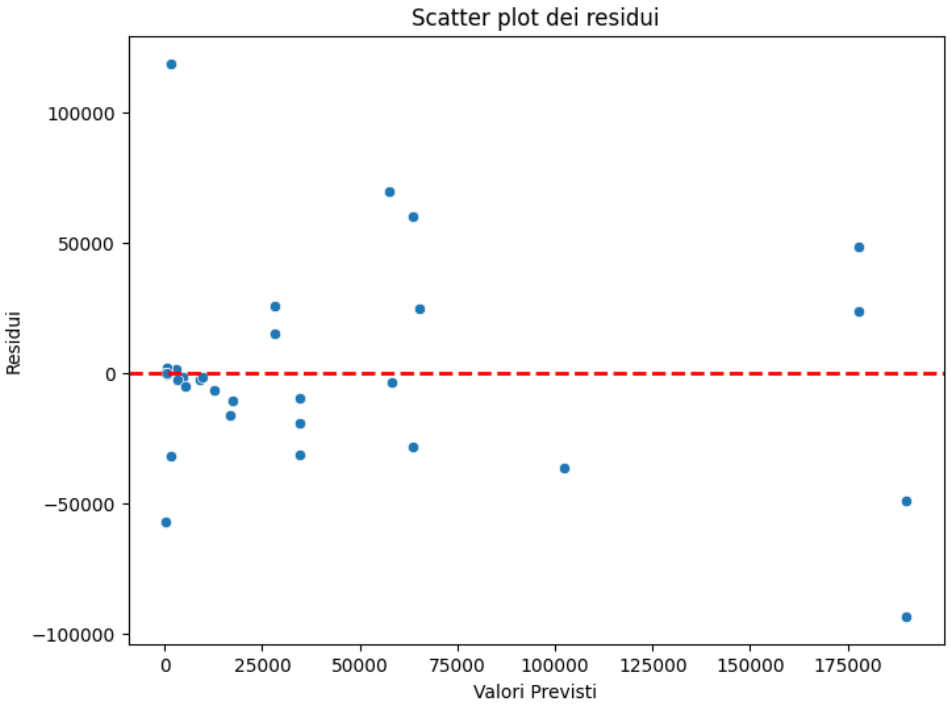
[67]:
```
#Valutazione MAE sul set di test
mae5 = mean_absolute_error(y2_test_no, y2_pred_no)
```

```
MAE: 23450.62386576972
```

Il calcolo del nuovo RMSE di circa 36.9 k tonnellate evidenzia un miglioramento della stima una volta rimossi gli outliers. Lo stesso miglioramento si riflette nel MAE calcolato, che è ora di circa 23.5 k tonnellate.
Il valore R-Squared di circa 0.61 indica che il 61% della variazione nelle emissioni è spiegato dalle variabili indipendenti. Questa percentuale suggerisce che il modello ha una notevole capacità di spiegare la variazione nelle emissioni.
Per visualizzare meglio la relazione tra i dati reali e quelli predetti dopo la rimozione degli outliers, è possibile creare nuovamente un grafico di confronto.

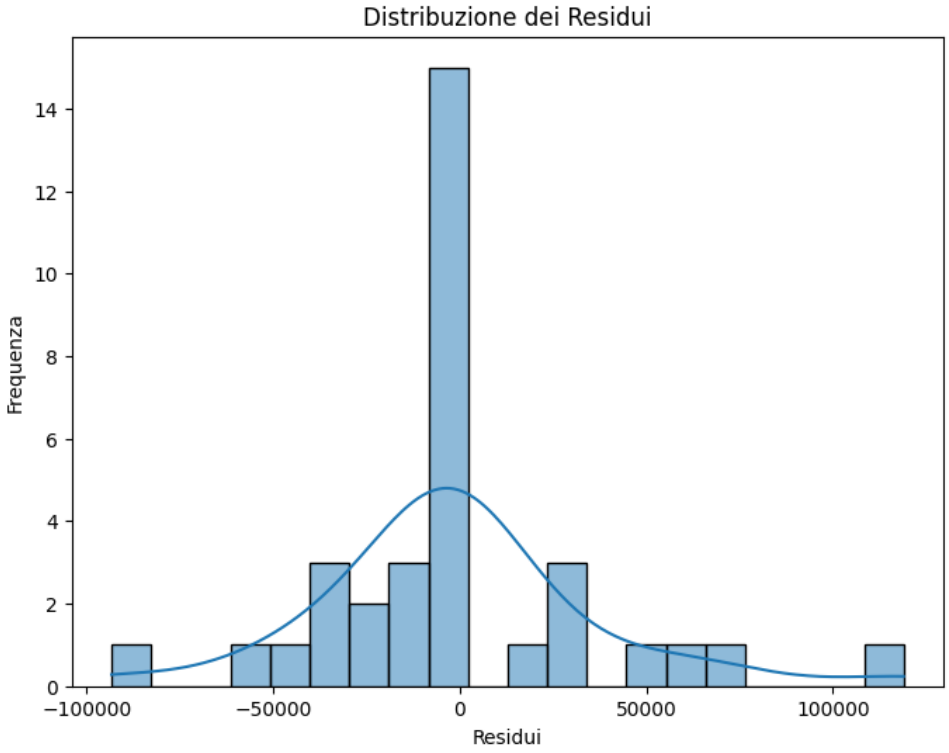


È notevole come, dopo le modifiche, i punti non si discostano significativamente dall'asse, indicando una migliore aderenza del modello ai dati reali rispetto alla stima precedente. Si procede a visualizzare i residui che, dopo l'eliminazione dei valori outliers sono meglio concentrati attorno alla linea e si mostrano i risultati reali e predetti in una tabella.



	Valore Effettivo	Previsioni	...	Valore Effettivo	Previsioni
157	32424.00	34770.65	88	17910.00	46330.48
120	177799.24	153847.13	123	34382.00	53708.72
67	268.00	217.10	20	21606.00	24874.07
16	96889.00	72322.72	17	568.00	561.97
75	45537.00	81844.46	34	1016.00	3663.70
98	51.00	31626.62	10	37620.00	22303.77
26	3418.00	4967.72	95	2512.00	688.42
124	5592.00	4190.87	89	66.00	152.34
61	5321.00	10285.08	46	31786.00	62924.86
62	532.00	582.06	171	5310.00	6551.96
126	120369.00	169282.53			

Infine, si visualizza la distribuzione dei residui che risulta centrata attorno allo zero dopo l'eliminazione degli outliers.



3.8 Algoritmi a Confronto

Si crea una tabella riassuntiva dei valori di R-Squared, RMSE e MAE per ciascun algoritmo (k-Nearest Neighbors, Decision Tree e Random Forest) in modo da trarre qualche considerazione finale e analizzare quello che meglio approssima il set di dati preso in esame con i valori predetti piu attendibili.

[72]:
```
#Valutazioni valori Algoritmi a confronto
data = {
    'Algoritmo': ['KNN', 'Decision Tree','Random Forest'],
    'R-Squared': [r21, r23, r25],
    'RMSE': [rmse1, rmse3, rmse5],
    'MAE': [mae1, mae3, mae5]
}
```

Algoritmo	R-Squared	RMSE	MAE
KNN	0.56	40353.44	22909.55
Decision Tree	0.36	46992.26	22965.65
Random Forest	0.61	36981.81	23450.62

L'algoritmo Random Forest ha il valore più alto di R-Squared (0.61), mostrando che, tra gli algoritmi esaminati, è quello che meglio spiega la variazione nei dati rispetto a KNN (0.56) e Decision Tree (0.36), quindi sembra avere una maggiore capacità di adattamento ai dati in esame.

Inoltre, l'algoritmo di Random Forest ha il valore più basso di RMSE (37.9 k), indicando una maggiore precisione nelle previsioni, in relazione alla grandezza dei dati, rispetto a KNN (40.4 k) e Decision Tree (46.9 k). Un RMSE più basso suggerisce una migliore accuratezza nel rappresentare i dati rispetto agli altri algoritmi.

Il KNN ha il valore più basso di MAE (22.90 k), indicando una migliore precisione delle previsioni, a livello assoluto, rispetto a Random Forest (23.45 k) e Decision Tree (22.97 k). Mentre L'RMSE è più sensibile agli errori più grandi, poiché calcola la radice quadrata della media dei quadrati degli errori, il MAE tratta tutti gli errori nello stesso modo, calcolando la media degli errori assoluti.

Di conseguenza, l'algoritmo di regressione Random Forest, sembra avere le migliori prestazioni per la previsione delle emissioni globali sul dataset di studio.

3.9 Considerazioni

Il Machine Learning è un potente strumento quando si hanno dati complessi, in quanto riesce ad apprendere da essi e individuare modelli che sarebbero difficili da rappresentare tramite formule matematiche tradizionali.

- E' utile quando si hanno grandi quantità di dati, sia in termini di dimensioni che di caratteristiche, soprattutto quando è necessario analizzare dati complessi provenienti da diverse fonti.
- E' utile per la capacità di adattamento automatico ai cambiamenti nei dati, in maniera significativa con informazioni dinamiche o soggette a variazioni nel tempo.
- E' utile nel riconoscimento di relazioni fra dati, offrendo la possibilità di scoprire connessioni che potrebbero sfuggire a un'analisi tradizionale.

Quindi, il Machine Learning è essenziale per le stime predittive, tra i molti algoritmi che offre bisogna poi cercare quello che meglio approssima il modello di studio in base alle caratteristiche e al contesto che si deve esaminare.

Conclusione

Nel primo capitolo di questo elaborato, si conduce una breve introduzione sul tema delle emissioni. Si inizia esponendo i principali agenti inquinanti, analizzando le cause principali che ne determinano la presenza nell'ambiente e le conseguenze derivanti da queste emissioni. Inoltre, si esaminano le soluzioni adottate fino ad oggi per affrontare e mitigare l'impatto negativo di tali agenti sull'ambiente e sulla salute umana. Questa panoramica iniziale serve da fondamento per la comprensione delle dinamiche legate alle emissioni e alle relative strategie di gestione e controllo.

Nel secondo capitolo, l'analisi esplorativa del dataset, che copre il periodo dal 2000 al 2020 e include i valori delle emissioni globali, ha confermato e approfondito le informazioni presentate nel capitolo precedente. I risultati dell'analisi hanno messo in evidenza la preoccupante situazione in alcuni paesi, permettendo di quantificare la percentuale di contributo di ciascun continente all'inquinamento atmosferico. In particolare, l'analisi fornisce una chiara identificazione delle tipologie di emissioni coinvolte e individua i settori che sono maggiormente responsabili di questo inquinamento. Questa fase di esplorazione del dataset è risultata essenziale dando una panoramica più dettagliata sulle situazioni più a rischio.

L'ultimo capitolo evidenzia un approccio più avanzato nell'analisi dei dati, andando oltre l'interpretazione descrittiva. Si introduce l'analisi predittiva, un'ulteriore fase che supera la semplice descrizione e visualizzazione dei dati per estrarre informazioni più approfondite e predire possibili sviluppi. In questa fase dell'elaborato, attravenso un dataset sulle emissioni di CO2 nel 2023, correlato a vari aspetti sociali, ambientali, economici e sanitari, si intraprende il calcolo di stime predittive sui valori di anidride carbonica. L'impiego di diversi algoritmi di Machine Learning consente di individuare il modello che meglio approssima i dati reali, fornendo così un modello predittivo in grado di stimare le emissioni per gli anni futuri. Questa fase avanzata di analisi non solo arricchisce la comprensione del fenomeno delle emissioni di CO2, ma fornisce anche strumenti pratici per formulare previsioni utili nella pianificazione e nella gestione ambientale. La combinazione di analisi descrittiva e predittiva costituisce un approccio completo per esplorare e comprendere le dinamiche delle emissioni atmosferiche e per sviluppare strategie efficaci per intervenire nelle situazioni più disastrose.

L'analisi dei dati, dunque, abbinata alle tecnologie avanzate di Machine Learning fornisce gli strumenti necessari per affrontare le sfide ambientali in modo intelligente, automatico e mirato, aprendo prospettive concrete per un futuro più sostenibile.

Bibliografia

[1] Aurélien Géron. *Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow*. September 2019.

[2] Jake VanderPlas. *Python Data Science Handbook*. December 2016.

[3] ARPAV. *Emissioni di Inquinanti*. URL: "https://www.arpa.veneto.it/temi-ambientali/aria/emissioni-di-inquinanti/emissioni-di-inquinanti". Settembre 2022.

[4] INEMAR. *Metodologia*. URL: "https://www.inemar.eu/xwiki/bin/view/Inemar DatiWeb/Metodologia".

[5] ARPAT. *Tecniche di campionamento emissioni in atmosfera e valutazione dei risultati*. URL: "https://www.isprambiente.gov.it /files2020/controlli-ambientali /presentazione-sarrini_arpat.pdf".

[6] Europarl. *Cambiamento climatico: gas a effetto serra che causano il riscaldamento globale*. URL: "https://www.europarl.europa.eu/news/it/headlines/society/20230316STO77629/cambiamento-climatico-gas-a-effetto-serra-che-causano-il-riscaldamento-globale". Marzo 2023.

[7] ilPOST. *Da dove arrivano le emissioni inquinanti*. URL: "https://www.ilpost.it/2019/09/14/cause-emissioni-gas-serra-settori/". Settembre 2019.

[8] Commissione europea. *Conseguenze dei cambiamenti climatici*. URL: "https://climate.ec.europa.eu/ climate-change/consequences-climate-change_it".

[9] Apotecanatura. *Inquinamento atmosferico: gli effetti sulla nostra salute*. URL: "https://www.apotecanatura.it /news/inquinamento-atmosferico-gli-effetti-sulla-nostra-salute/".

[10] Europarl. *I negoziati sul cambiamento climatico*. URL: "https://www.europarl.europa.eu/infographic/climate-negotiations-timeline/index_it.html".

[11] Consilium. *Cambiamenti climatici: il contributo dell'UE*. URL: "https://www.consilium.europa.eu/it/policies/climate-change/".

[12] Europarl. *Le soluzioni dell'UE per contrastare i cambiamenti climatici*. URL: "https://www.europarl.europa.eu/news/it/headlines/society/20180703STO07129/le-soluzioni-dell-ue-per-contrastare-i-cambiamenti-climatici".