



Power Plant Generation Forecasting: A Comparative Study of Linear Regression, Random Forest and KNN

Buontempi Andrea, Bursi Nicole, Sepe Raffaele

Indice

1. INTRODUZIONE AL PROBLEMA
2. CREAZIONE DEL DATABASE
 - 2.1 Schema Logico
3. INPUT DEL PROBLEMA
4. STATISTICHE LEGATE ALLA DISTRIBUZIONE DI CENTRALI ELETTRICHE SUL TERRITORIO E PER TIPOLOGIA DI COMBUSTIBILE
5. PREVISIONE GENERAZIONE FUTURA DI ENERGIA
 - 5.1 Pre-Processing
 - 5.2 Previsione per Singola Centrale
 - 5.3 Previsione per Paese
 - 5.4 Previsione per Tipologia di Centrale
6. CONCLUSIONI

Introduzione al Problema

Il *World Resources Institute* ha sviluppato un database globale integrando dati governativi e rilevamenti satellitari sugli impianti energetici. In questo progetto, tali dati sono stati analizzati ed elaborati per sviluppare modelli predittivi della generazione elettrica futura, con livelli di dettaglio per singola centrale, per Paese e per tipologia di combustibile.

OBIETTIVI

- Creare un'infrastruttura dati strutturata per archiviare e gestire efficacemente i dati relativi agli impianti energetici di tutto il mondo.
- Sviluppare più modelli predittivi in grado di stimare la generazione futura per i prossimi 3 anni, valutandone la performance di ciascuno.

Creazione del Database

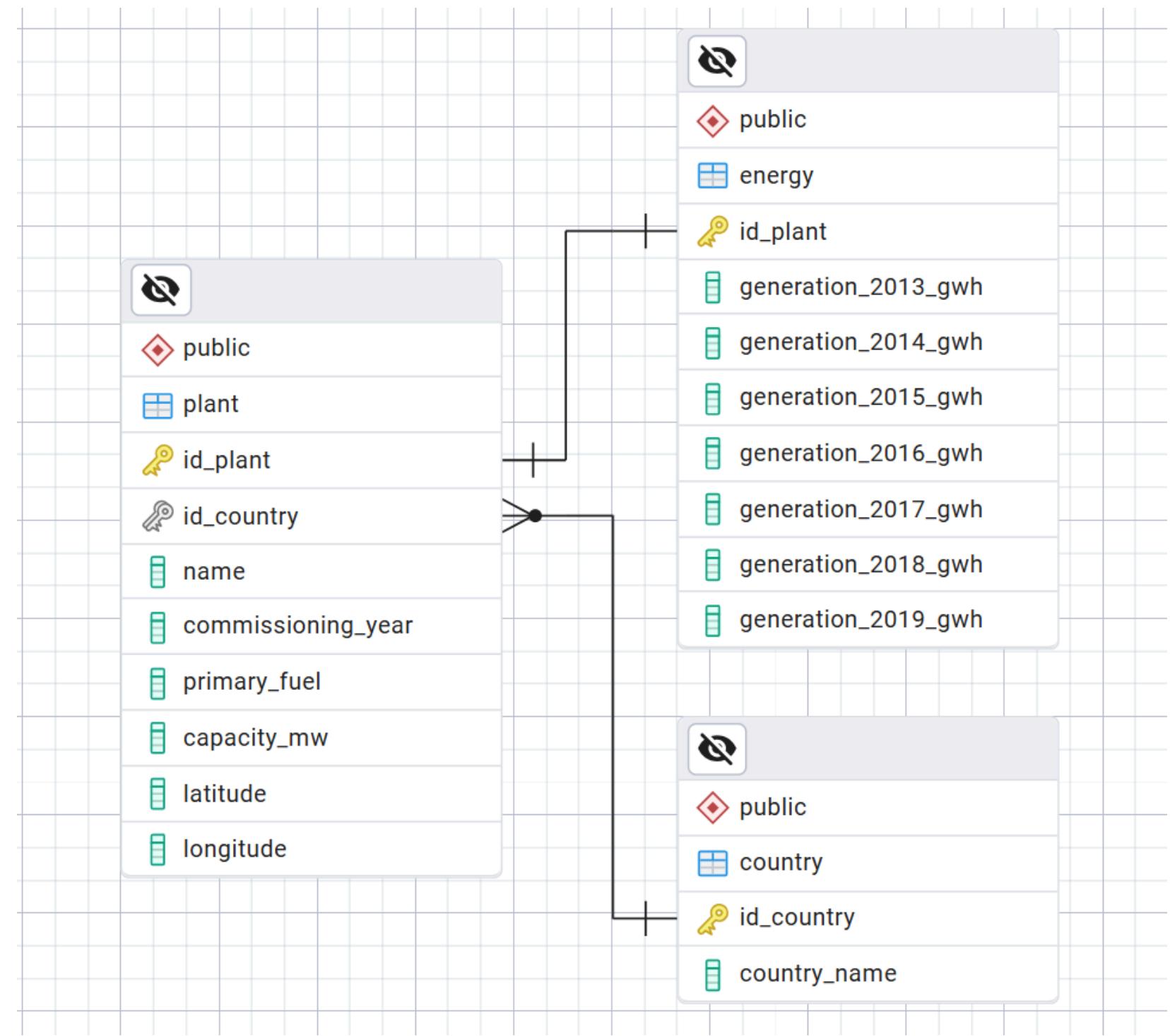
CREATE DATABASE PLANTS

Country (id_country, country_name);

Energy (id_plant, generation_2013_gwh, generation_2014_gwh, generation_2015_gwh, generation_2016_gwh, generation_2017_gwh, generation_2018_gwh, generation_2019_gwh);

Plant (id_plant, id_country, name, commissioning_year, primary_fuel, capacity_mw, latitude, longitude);

Creazione del Database - Schema Logico



Input del Problema

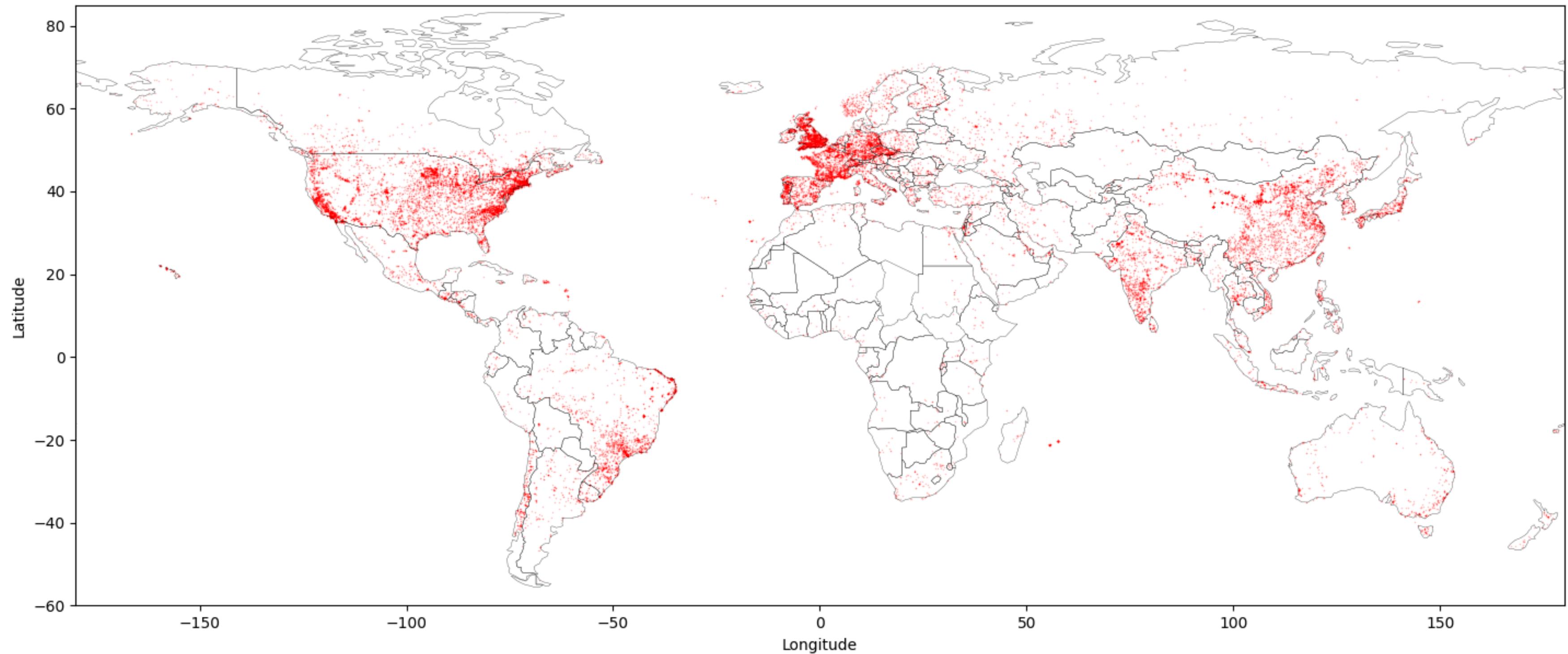
Il Database “PLANTS” è costituito da 3 tabelle: *country*, *plant* e *energy*.

Per popolare queste tabelle è stato usato il file *global_power_plant_database.csv*

L’interfaccia Python prende i dati attraverso una query che interroga il Database.

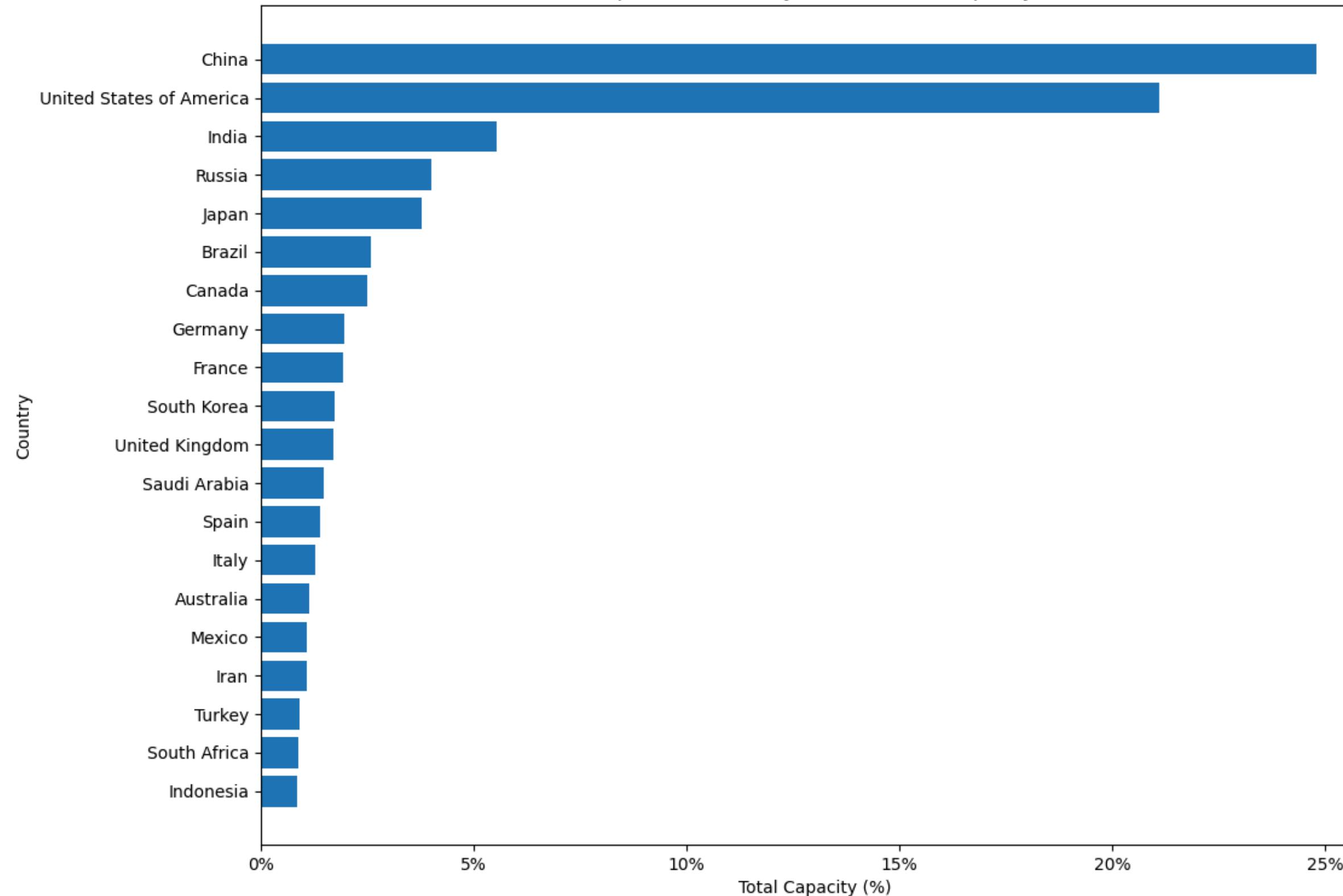
- *Country* (167 righe e 2 colonne) : Tabella dimensionale che contiene le referenze geografiche dei paesi censiti.
- *Plant* (34936 righe e 8 colonne) : Tabella anagrafica principale che raccoglie i metadati tecnici e spaziali di ogni singola centrale elettrica.
- *Energy* (34936 righe e 8 colonne) : Tabella contenente le serie storiche relative alla generazione di energia pregressa, collegate univocamente a ciascun impianto.

Distribuzione Geografica delle Centrali Elettriche



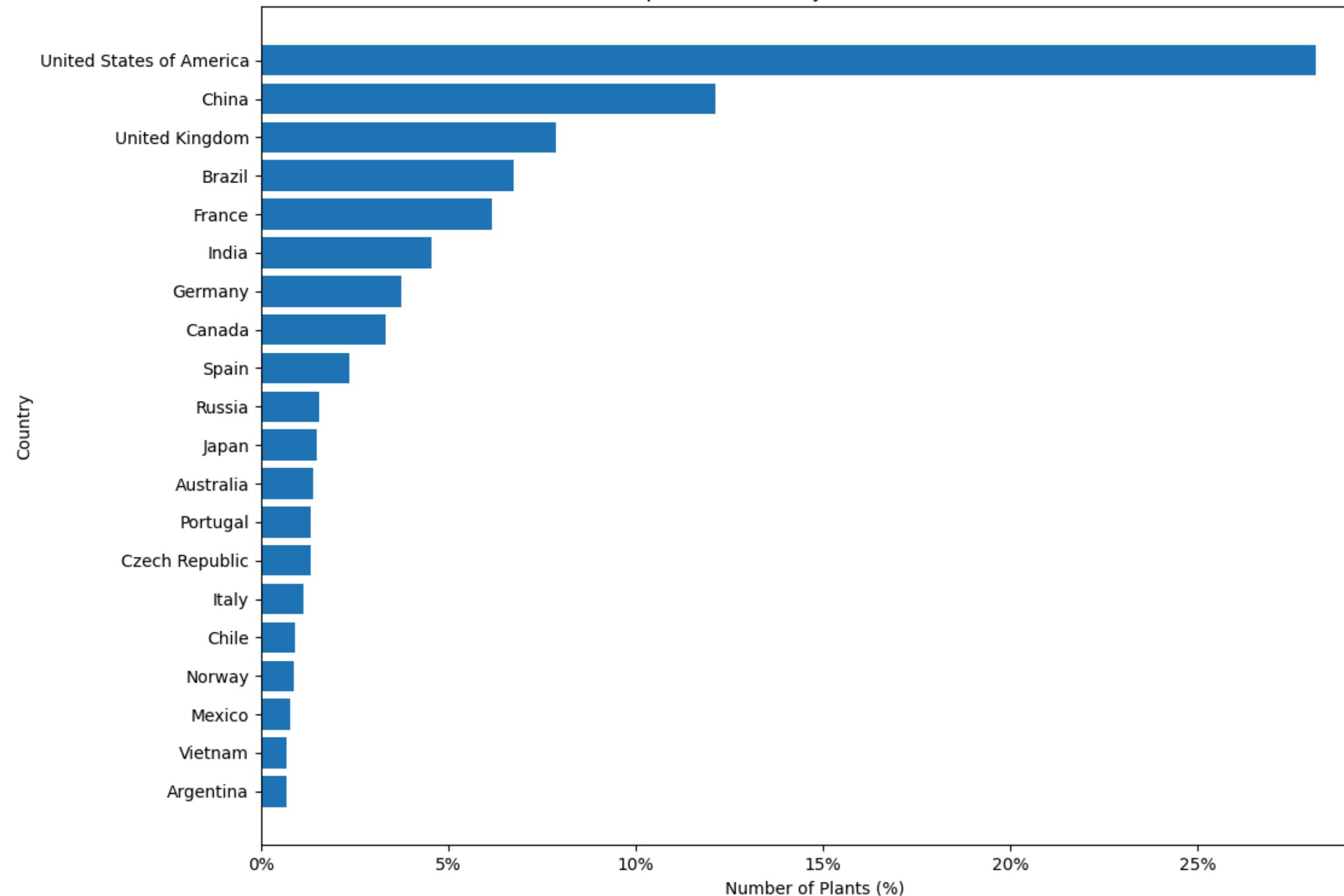
Statistiche per Paese

Top 20 Countries by Total Installed Capacity

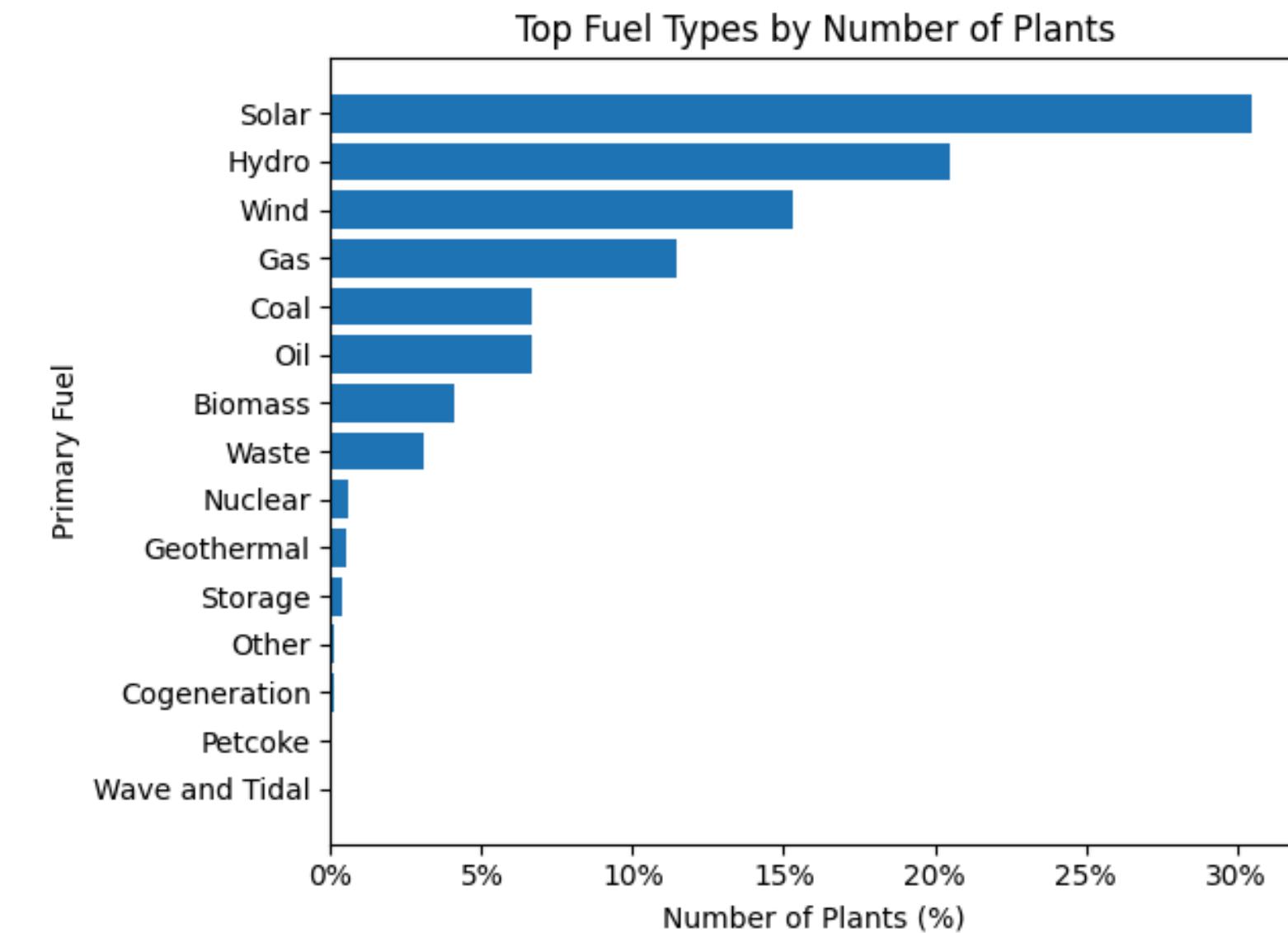
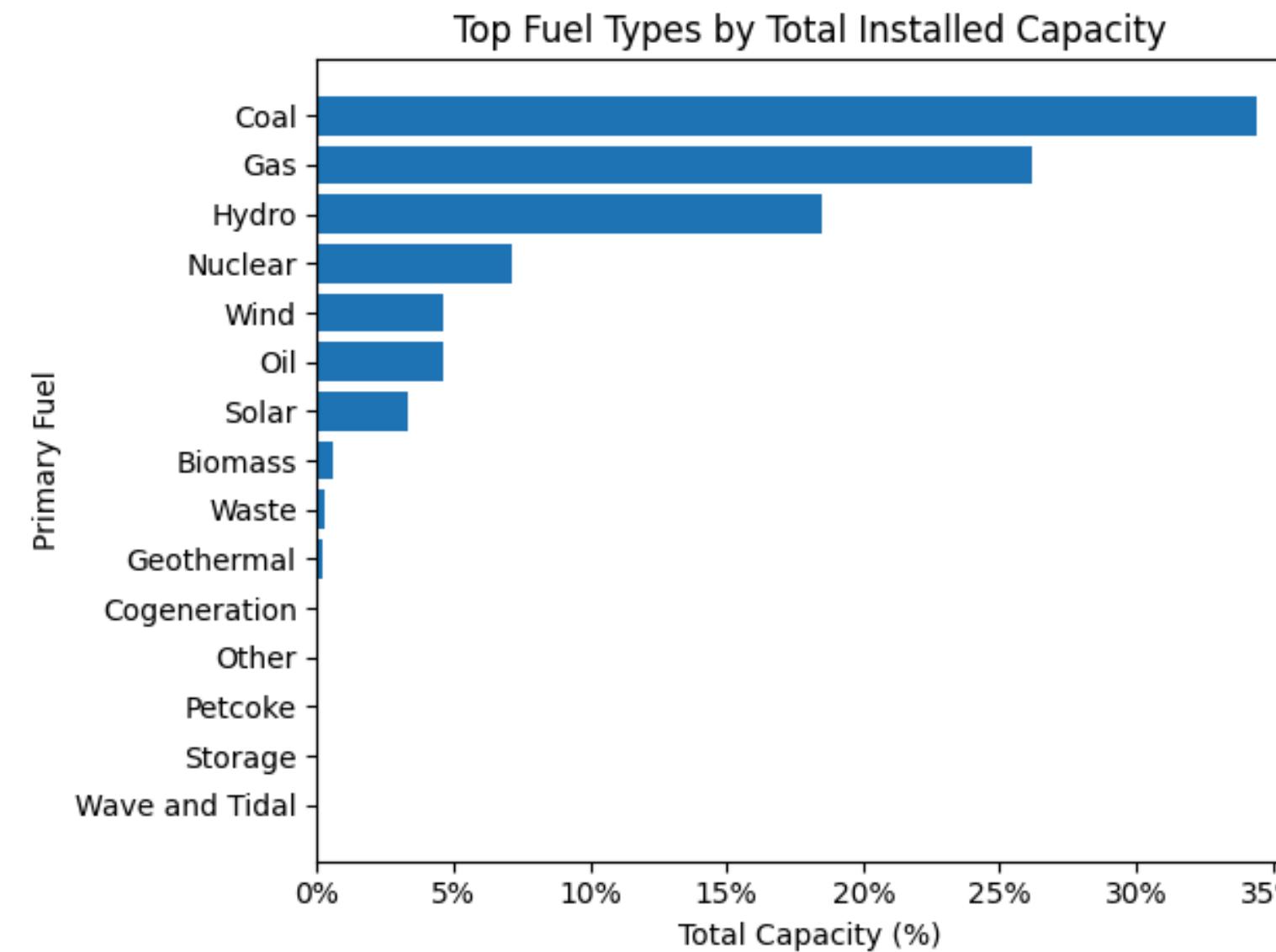


Statistiche per Paese

Top 20 Countries by Number of Plants



Statistiche per Tipo di Combustibile



PREVISIONE GENERAZIONE FUTURA DI ENERGIA

2017 - 2018 - 2019

Pre-Processing

La fase di Pre-Processing è stata progettata e differenziata a seconda della tipologia di forecasting affrontata.

In particolare, sono tre le modalità attraverso cui abbiamo effettuato una pulizia e riorganizzazione dei dati, a seconda che la previsione dovesse essere effettuata a livello di:

1. Singola Centrale
2. Centrali di un dato paese
3. Centrali di una data tipologia

Pre-Processing per Singola Centrale

Abbiamo apportato le seguenti modifiche:

- Trasformato le variabili *country_name* e *primary_fuel* in variabili categoriche identificate da un codice univoco;
- Abbiamo espresso la capacità della centrale in GW;
- Come dati di training abbiamo usato solo i record che hanno un valore in ogni features utilizzata;
- Abbiamo potuto farlo perchè, nonostante siano state eliminate molte righe, ne sono rimaste comunque un numero considerevole (più di 6000).

Pre-Processing Aggregato per Paese e Tipologia di Centrale

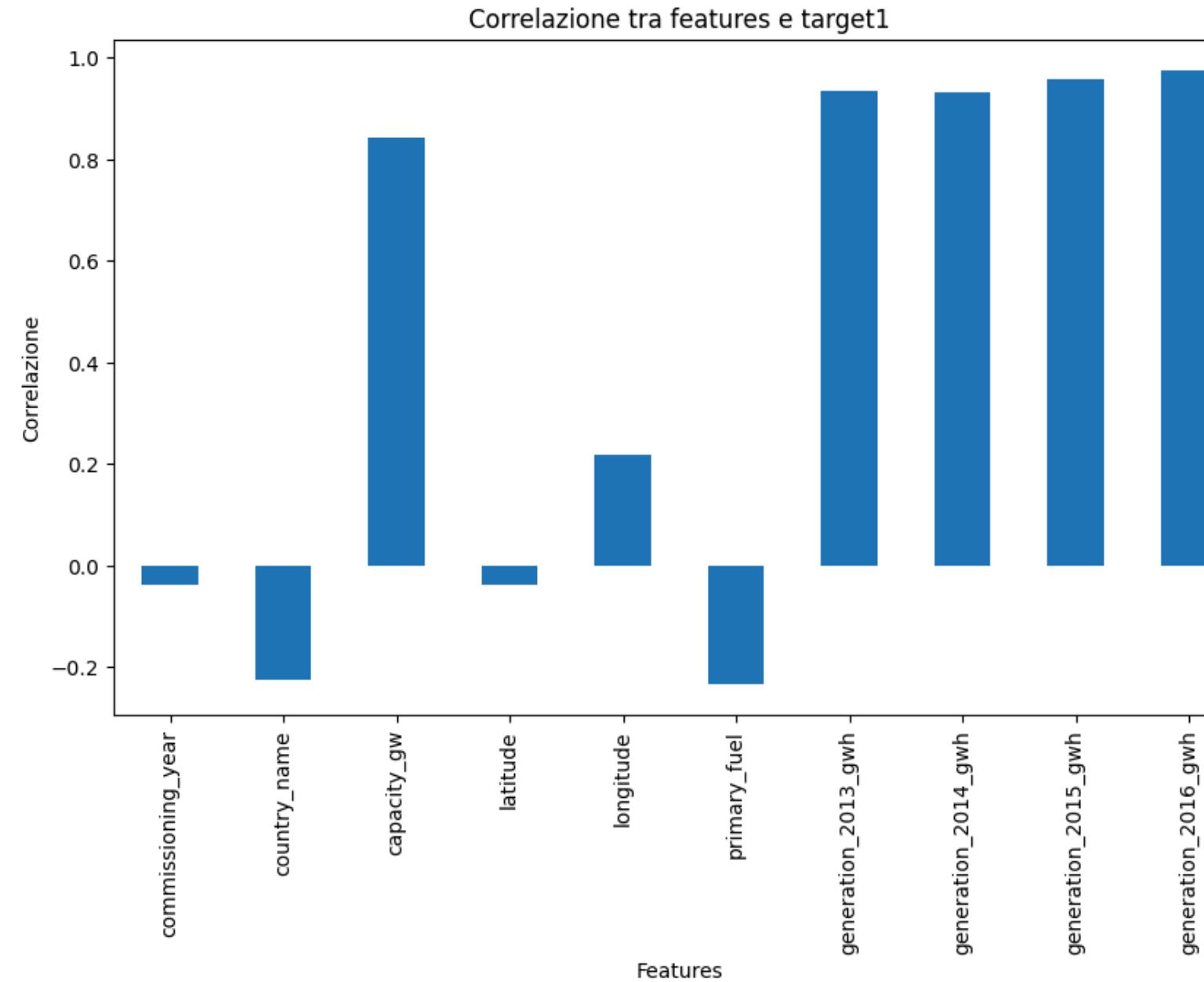
Abbiamo apportato le seguenti modifiche:

- Abbiamo espresso la capacità della centrale in GW;
- Nelle colonne delle generazioni che hanno somma zero abbiamo inserito delle stime di generazione che si basano sulla capacità della nazione moltiplicata per le ore medie di utilizzo annuale;
- Come dati di training abbiamo usato solo i record che hanno un valore in ogni feature utilizzata.



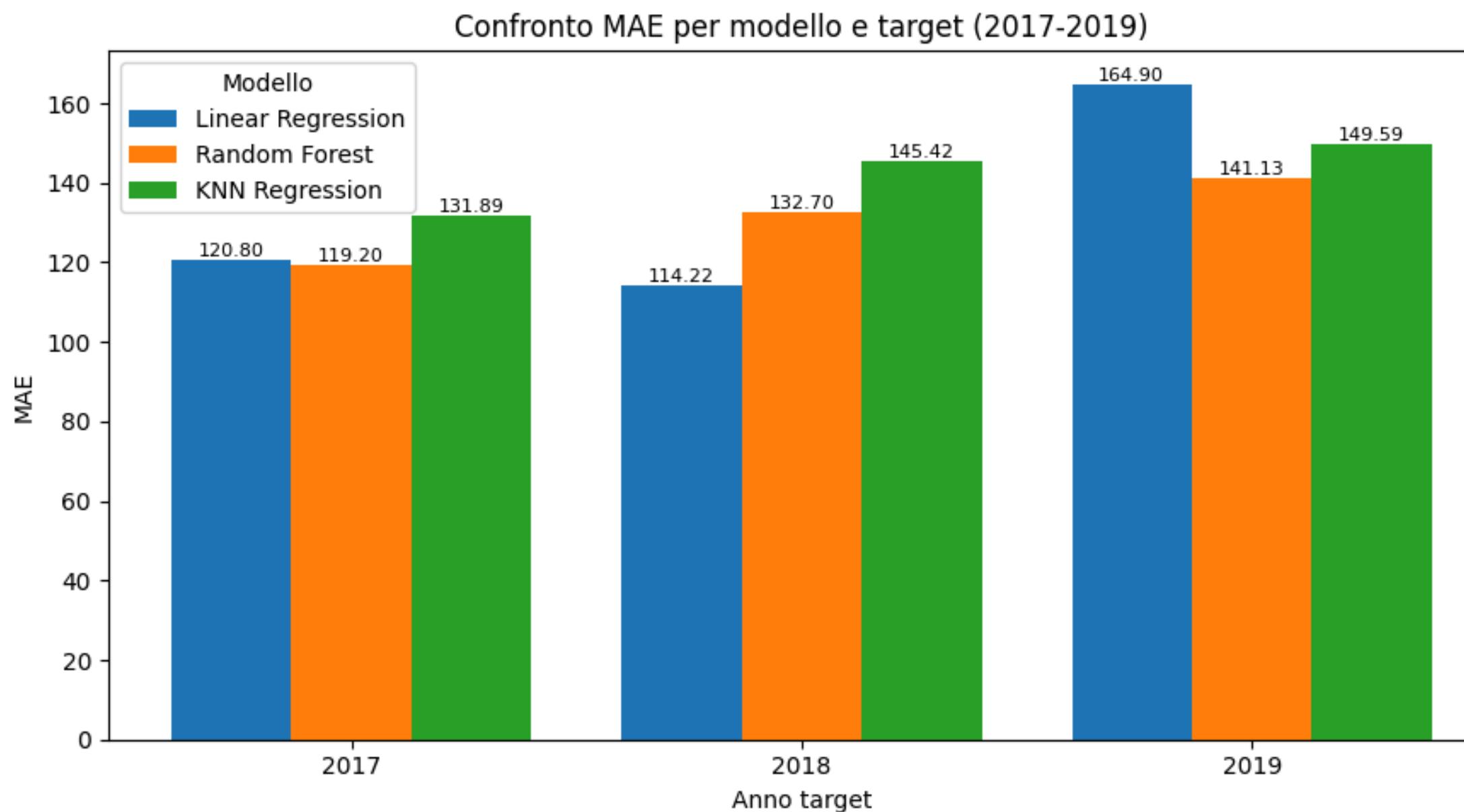
PREVISIONE PER SINGOLA CENTRALE

Analisi delle Correlazioni e Selezione delle Features



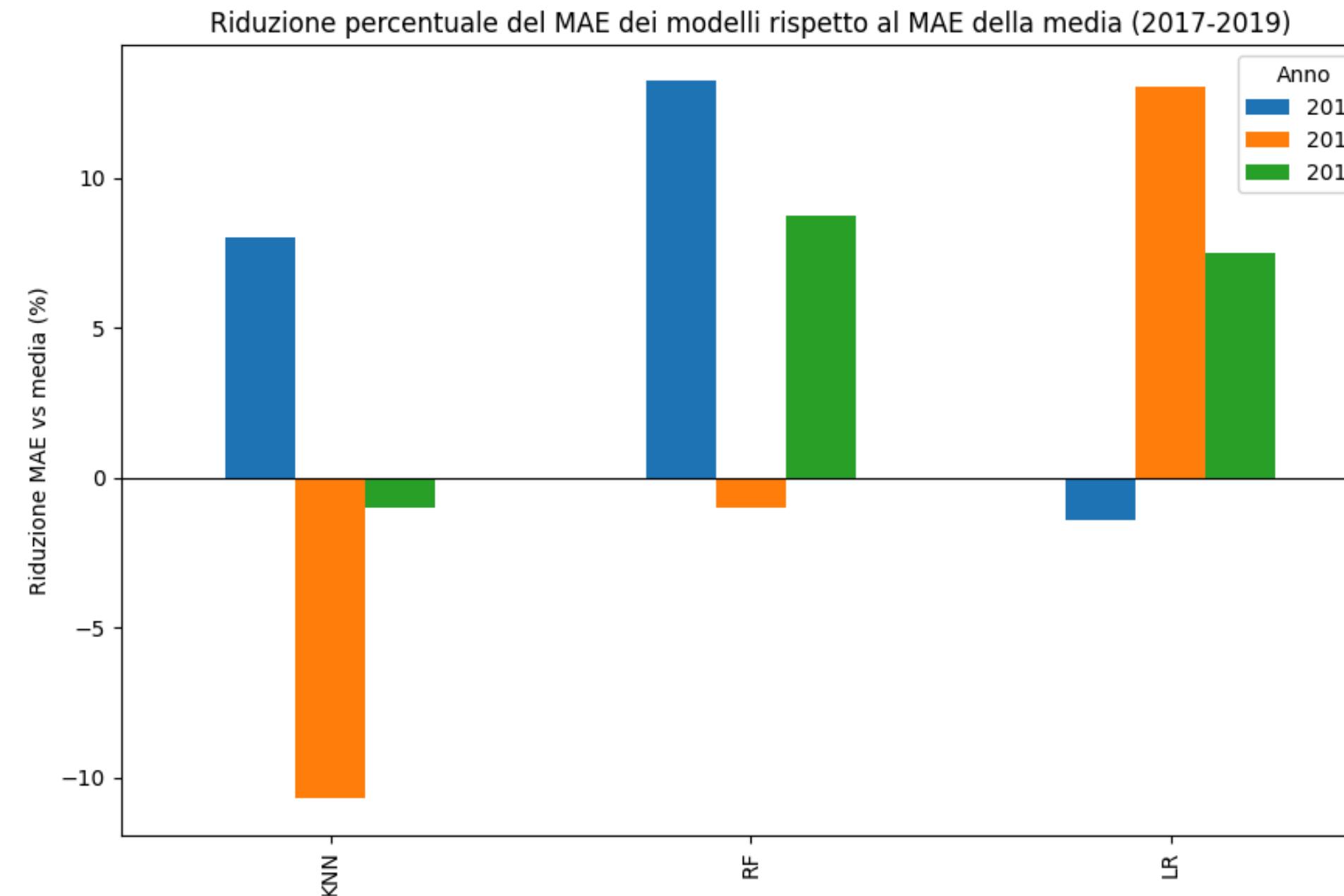
Obiettivo: Studiare la correlazione tra le features e il target

Training Modelli con Features Altamente Correlate



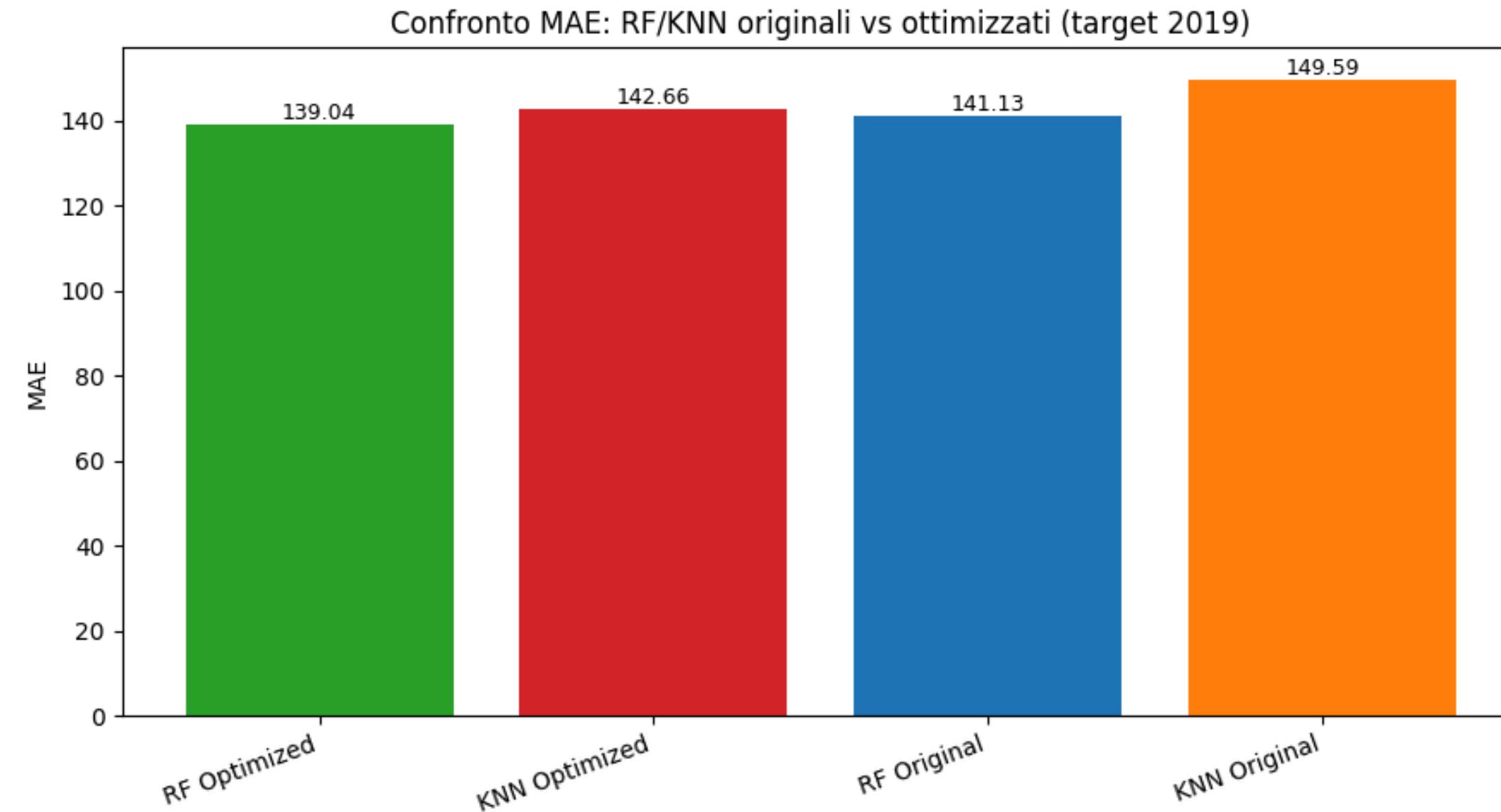
Le features utilizzate sono *capacity_mw* e *generation_gwh_2013-2016*

Confronto con la Baseline



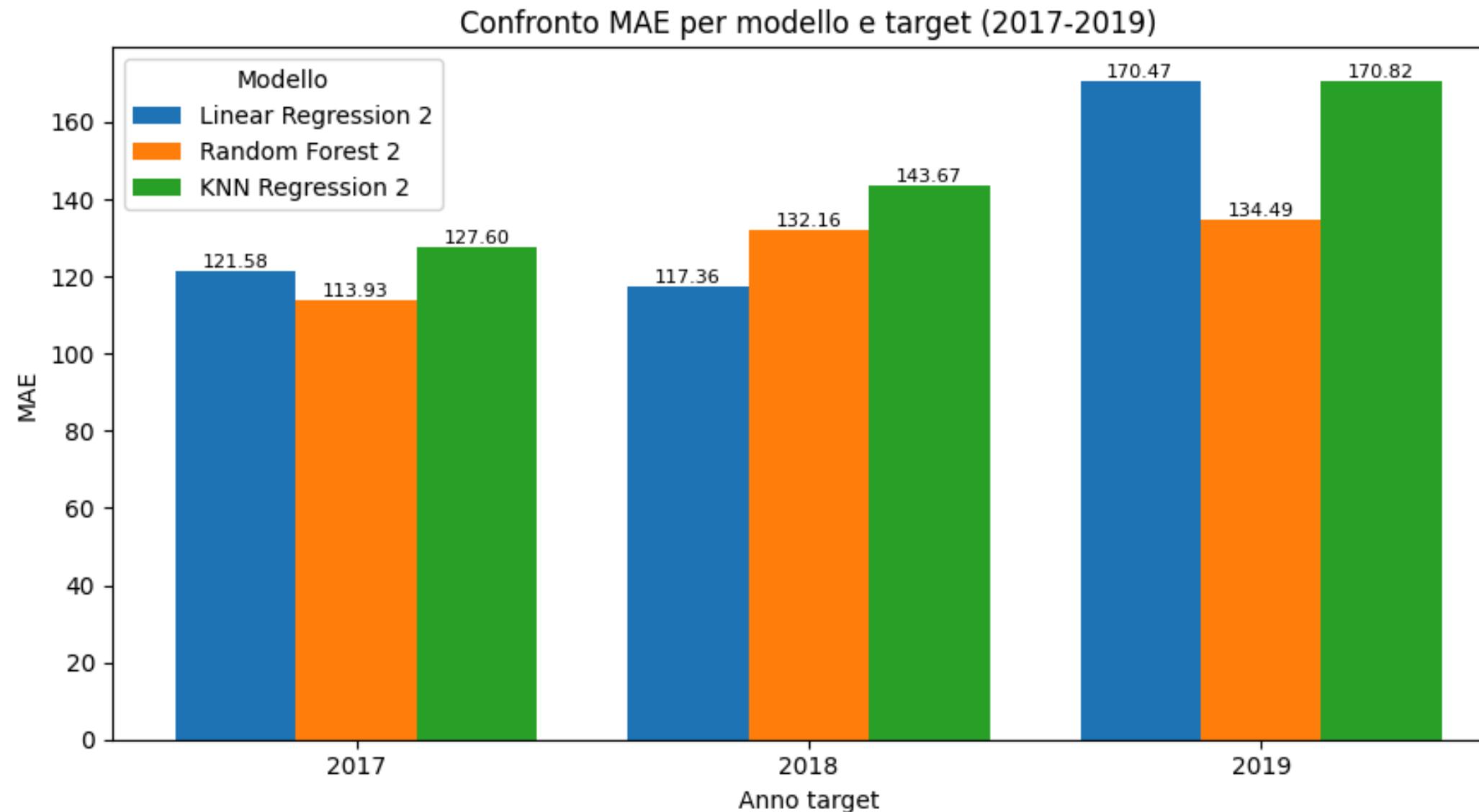
La baseline è stata costruita calcolando la media aritmetica dei valori osservati negli anni precedenti a quello oggetto della predizione.

Ottimizzazione degli iperparametri



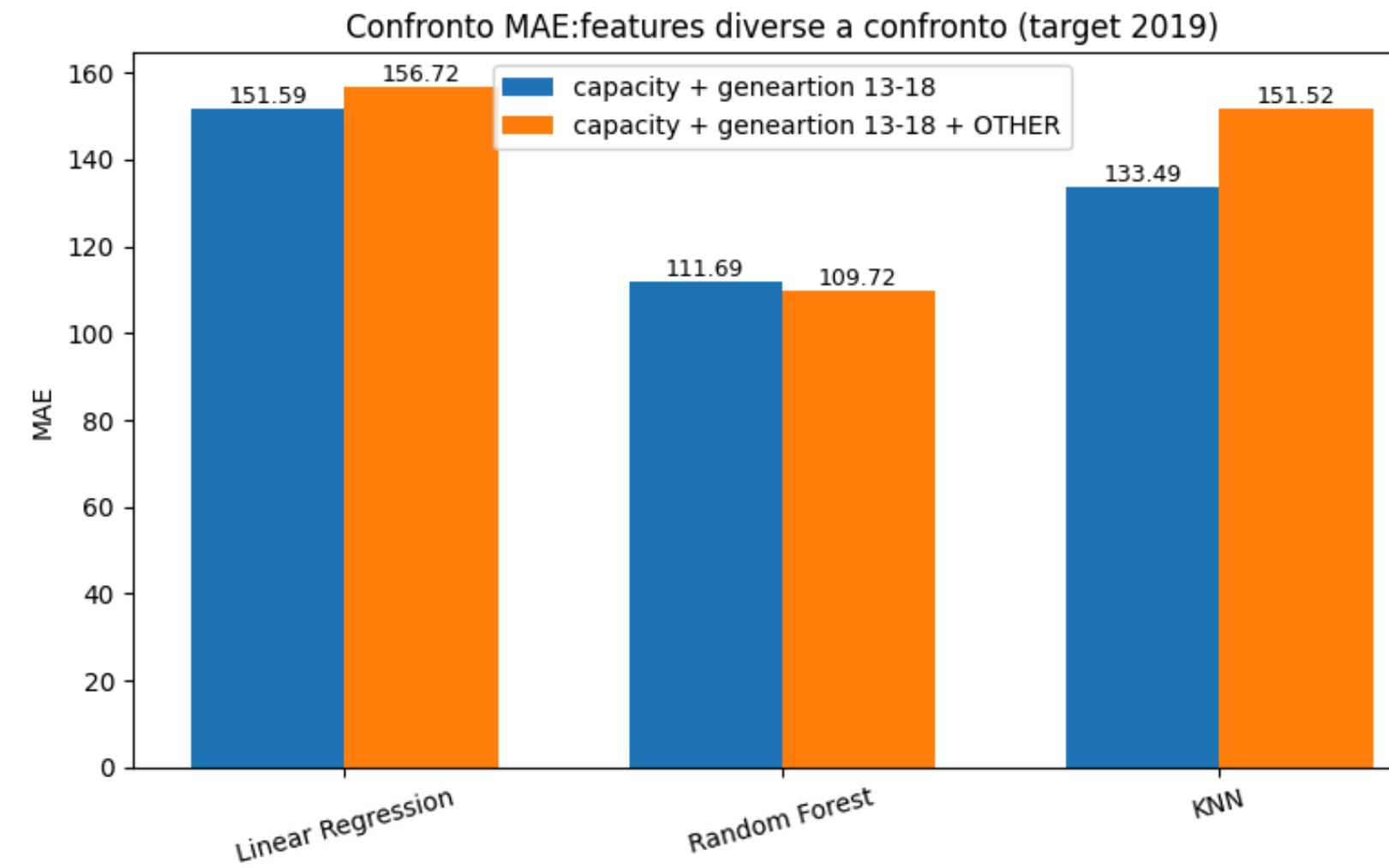
È stata implementata una procedura di ricerca automatica (su RF e KNN) per identificare la combinazione di iperparametri in grado di minimizzare la funzione.

Analisi dell'Impatto di Tutte le Features



Per il modello sono state utilizzate tutte le features a disposizione.
Si può notare un leggero peggioramento rispetto al caso iniziale

Analisi per Target 2019 con più Variabili Temporali



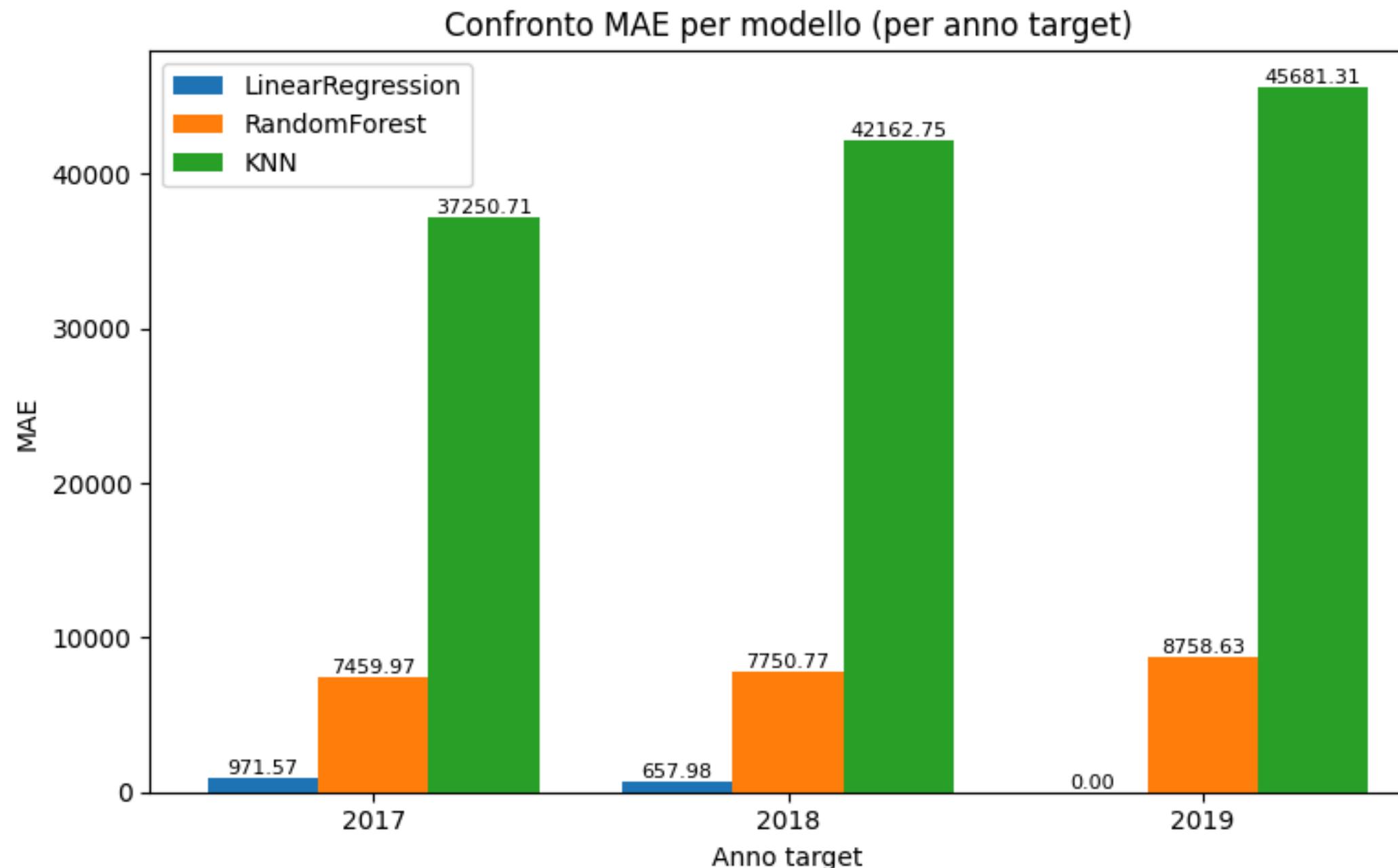
Sono stati messi a confronto due scenari di addestramento per la predizione del 2019:

1. Dataset Ridotto: Features ad alta correlazione + Target storici (2017, 2018) ;
2. Dataset Completo: Tutte le features + Target storici (2017, 2018).



PREVISIONE PER PAESE

Analisi MAE per Modello

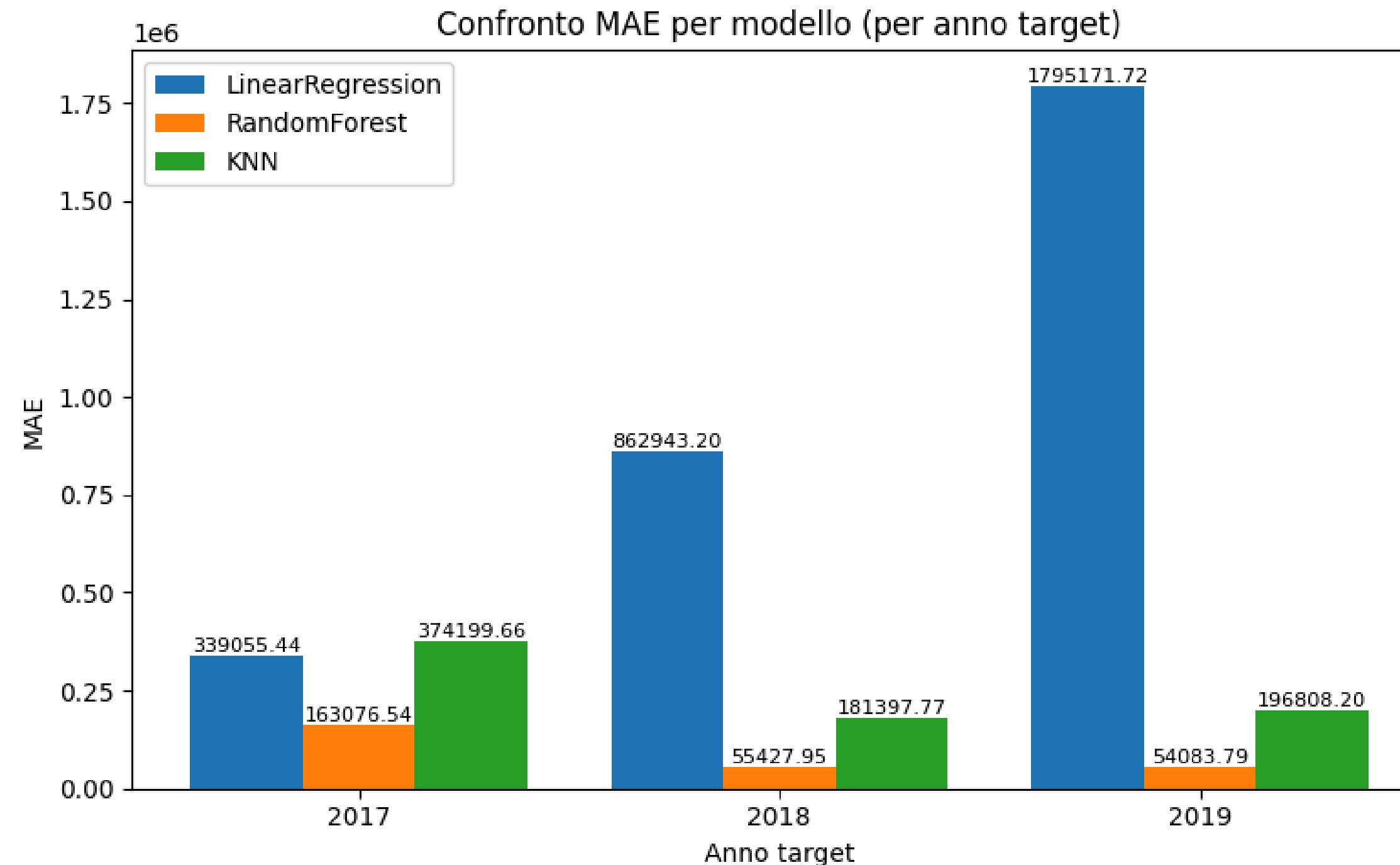


Le features utilizzate sono *capacity_mw* e *generation_gwh_2013-2016*.
Probabile overfitting per Linear Regression.

The background image shows a wide-angle aerial view of a massive industrial refinery or chemical plant. The facility is densely packed with a complex network of steel pipes, walkways, and various industrial structures. Large white storage tanks of different sizes are scattered throughout the site. In the center, several tall, thin vertical towers stand prominently. The refinery is situated in a rural area with green fields and some hills visible in the distance under a clear sky.

PREVISIONE PER TIPOLOGIA DI CENTRALE

Analisi MAE per Modello



Le features utilizzate sono *capacity_mw* e *generation_gwh_2013-2016*.

Conclusioni (1/2)

- Il KNN è il modello che funziona peggio rispetto alla baseline, mentre RF e LR performano discretamente bene;
- Il target 2017 è quello che viene predetto meglio, come risultato potevamo aspettarcelo in quanto più ci allontaniamo dalla finestra temporale del training set, più è probabile che i risultati siano meno precisi;
- Se come features utilizziamo anche variabili poco correllate i modelli funzionano leggermente peggio, l'unico che migliora è il RF;

Conclusioni (2/2)

- Se prediamo il 2019 utilizzando anche le variabili gen 17 e gen 18 i modelli hanno un errore minore, in quanto gli abbiamo fornito ulteriormente due variabili altamente correlate;
- Per il raggruppamento per country il modello che funziona meglio è la LR, mentre per il raggruppamento per fuel funziona meglio il RF. Questo risultato potrebbe essere imputabile alla differenza nella quantità di record in quanto nel raggruppamento per country abbiamo 167 record mentre in quello per fuel ne abbiamo solamente 15.



VI RINGRAZIAMO PER L'ATTENZIONE

Buontempi Andrea, Bursi Nicole, Sepe Raffaele