AQI - Indice Qualità dell'Aria Relazione Caso di Studio

Gruppo di lavoro

- Patrick Clark, 699646, p.clark@studenti.uniba.it
- Mario Giordano, 700373, m.giordano51@studenti.uniba.it

https://github.com/mariogiordano1/Progetto_ICON_22-23

AA 2022-2023

Introduzione

Il caso di studio verte sulla qualità dell'aria, partendo da dati recuperati sulla piattaforma Kaggle, e integrati con ulteriori dati provenienti dal web Semantico.

Il dataset originale contiene dati riguardanti latitudine e longitudine, città, nazioni, valore e categoria dell'AQI e delle sostanze inquinanti NO2, PM 2.5, O3 e CO.

Partendo dal dataset, sono stati integrati dati provenienti dal Web Semantico, creando nuovi esempi e integrando una nuova sostanza inquinante, il PM 10.

Infine, è stato svolto un task di classificazione in base alla qualità dell'aria, sfruttando diversi modelli e combinazioni di dati.

Elenco argomenti di interesse

- Web Semantico – SPARQL, interrogazione di Knowledge Base distribuite sul Web: Utilizzo del linguaggio di interrogazione SPARQL per interrogare le Knowledge Base distribuite sul Web, nello specifico le Knowledge Base utilizzate sono state DBPedia e Wikidata.
- Apprendimento Supervisionato: Utilizzo di KNN, Alberi di Decisione, Random Forest ed SVM per effettuare il task di Classificazione.

Analisi dei Dati ed Integrazione di nuovi Dati

Il dataset originale contiene dati riguardanti AQI, CO, O3, PM2.5 ed NO2 per diverse Città, oltre alla latitudine e alla longitudine del luogo di cui è stata effettuata la lettura.

Da una prima analisi, risultavano mancanti dei dati riguardanti le Nazioni, cui valori sono stati recuperati da Wikidata, mediante query SPARQL.

Per effettuare tutte le query SPARQL è stata utilizzata la libreria SPARQLWrapper, e gli endpoint utilizzati sono stati https://query.wikidata.org/sparql ed https://dbpedia.org/sparql

```
query = f"""SELECT ?label_en
    WHERE
    {{
      ?city wdt:P1566 "{cityname}".
      ?city wdt:P17 ?countryLabel .
      ?countryLabel rdfs:label ?label_en filter (lang(?label_en) = "en").
    }}""" + """ -H "Accept: text/csv" """
```

label_en
France

(Risultato per la query in alto, dove cityname è il GeoNamesID di Granville, Francia)

I risultati della query sono poi stati inseriti nel Dataset.

Successivamente, per espandere il dataset, si è deciso di trovare latitudine e longitudine delle varie università per capitali, ottenute da DBPedia tramite query SPARQL. Questo perché le varie capitali contengono un numero elevato di università, da cui è possibile estrarne latitudine e longitudine per ottenere dati sulla qualità dell'aria in punti diversi di una stessa città. Inoltre, molte università montano stazioni per monitorare la qualità dell'aria, di conseguenza si ottengono letture più precise e variegate.

Per iniziare, sono state isolate le nazioni presenti sul dataset originale. Successivamente, mediante query SPARQL su DBPedia, sono stati recuperati i nomi delle capitali per ogni nazione:

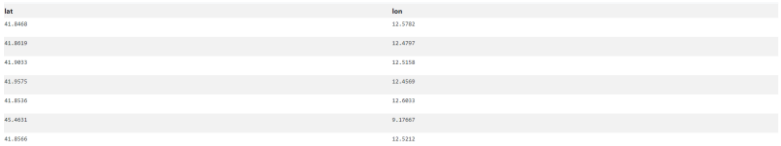
```
endpoint_url = "https://dbpedia.org/sparql"
file_output = open(f"""Database/CSVS/Capitals.csv""", 'a', newline="\n", encoding="utf-8")
print(file.readlines())
query = f"""SELECT ?citylabel
WHERE
{{
    ?country rdf:type dbo:Country.
    ?country rdfs:label "{countryname}" @en.
    ?country dbo:capital ?city.
    ?city rdfs:label ?citylabel.
    FILTER(LANG(?citylabel) = "en").
}}"""
```

citylabel
"Rome"@en

(Risultato per la query dove countryname = "Italy")

Trovate le capitali, è stata effettuata una nuova query, da dove sono state estratte latitudine e longitudine per ogni università per ogni capitale risultante dalla query precedente:

```
query = f"""SELECT DISTINCT ?lat ?lon
    WHERE
    {{
      ?univ rdf:type dbo:University.
      ?univ dbo:city dbr:{cityname}.
      ?univ geo:lat ?lat.
      ?univ geo:long ?lon.
      ?univ dbo:city ?city.
      ?city dbo:country ?country.
      ?country rdfs:label ?countrylabel.
      ?city rdfs:label ?citylabel.
      ?univ rdfs:label ?unilabel
      FILTER(LANG(?countrylabel) = "en").
      FILTER(LANG(?citylabel) = "en").
      FILTER(LANG(?unilabel) = "en").
    }}"""
```



lat	lon
41.8468	12.5782
41.8019	12.4797
41.9033	12.5158
41.9575	12.4569
41.8536	12.6033
45.4631	9.17667
41.8566	12.5212

(Risultato della query in alto dove cityname=Rome)

La query soprastante restituisce i valori presenti per le proprietà RDF geo:lat e geo:long per ogni università trovata. Questi valori sono poi stati utilizzati per recuperare le letture sulla qualità dell'aria in base alle coordinate dalla mappa online aqicn.org.

Per ottenere in dati, è stata effettuata una richiesta HTTPS ad aqicn.org:

```
r = requests.get(
    f"""https://api.waqi.info/feed/geo:{lat};{long}/?token=12f1f5758ff0027428dd7a64bbe1c50c7b10206b""")
```

che restituisce una lettura completa della qualità dell'aria per una specifica coppia latitudine e longitudine. I risultati della richiesta sono stati salvati all'interno di un nuovo dataset.

Dalla lettura ricevuta, si è notata la presenza di un nuovo agente inquinante, il PM 10, che sarebbe potuto tornare utile ai fini del task di classificazione. Per questa ragione, si è deciso di effettuare una nuova richiesta HTTPS, questa volta per ottenere una lettura che includa il PM10 per tutte le coppie latitudine e longitudine uniche del dataset originale.

È stata quindi introdotta una nuova feature nel dataset: il PM 10.

Infine, sono stati uniti i due dataset: quello originale con la nuova feature, e quello derivante dalla lettura della qualità dell'aria delle università.

Apprendimento Supervisionato

È stata identificata come feature per la classificazione la feature “AQI_Category”, che fornisce la categoria dell’AQI in base al valore corrispondente.

Sono stati eliminati gli esempi incompleti, in quanto mancavano ~300 valori per il PM10, ed una quantità simile per il CO.

Sono stati usati diversi modelli per la classificazione:

- Un modello KNN per le feature Latitudine e Longitudine, feature geografiche
- Un Albero di Decisione, per trovare le feature più utili per il task, ed un modello ensemble, RandomForest
- Una SVM

Per l’apprendimento è stato diviso il dataset in 2 parti: Test e Train, dato l’elevato numero di esempi, nello specifico è stato diviso in 75% train, 25% test.

Il dataset è però molto sbilanciato. AQI_Category contiene 6 valori diversi, quindi si tratta di una classificazione multiclass. Gli esempi per ogni categoria sono:

```
Good                                8004
Moderate                            7243
Unhealthy for Sensitive Groups       906
Unhealthy                            814
Very Unhealthy                       131
Hazardous                             62
```

Infatti, poco meno del 50% del dataset ha come valore in AQI_Category “Good”, mentre una esigua parte ha come valore “Hazardous” o “Very Unhealthy”.

Quando si lavora con dataset molto sbilanciati, la misura dell’accuratezza non è molto utile, in quanto è estremamente semplice raggiungere accuratezze elevate.

Per questa ragione, si è deciso di valutare le performance del sistema in base al F1-Score pesato. L’F1 score rappresenta la media armonica tra Precision e Recall:

$$F1\ score = 2 * \frac{Precision\ * Recall}{Precision + Recall}$$

Nell’F1 score pesato si pesa l’F1 score per ogni classe in base al numero di esempi per quella classe.

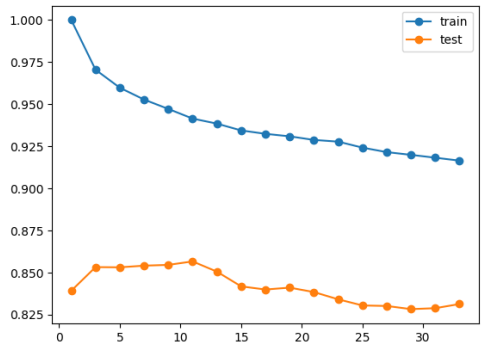
$$\text{Weighted F1 Score} = \sum_{i=1}^{N} w_i \times \text{F1 Score}_i$$

Come si vedrà in seguito, l’F1 score pesato e l’accuratezza avranno valori molto simili. Questo, molto probabilmente, a causa di un basso numero di false predizioni che rendono una componente dell’F1 score pesato relativamente piccola. Si nota che, per bassi numeri di esempi nel test set, la differenza tra Accuracy e F1 score pesato sarà a sua volta, probabilmente, molto bassa.

KNN

È stato utilizzato, per iniziare, un classificatore KNN, in particolare sulle Feature Geografiche, ovvero latitudine e longitudine. L'idea di base è osservare se la vicinanza di due posti incide sull'AQI.

Le feature di input utilizzate sono quindi soltanto latitudine e longitudine:

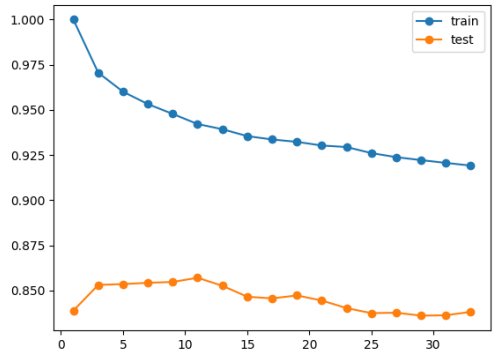



```
11---
0.8571095571095572 -- Accuracy
0.8566719368897278 -- F1
0.6605937692944726 -- Precision
0.673877067339582 -- Recall
```

Dal modello KNN, si nota come l'accuratezza e l'F1 pesato siano molto simili tra loro. Nonostante ci si aspettasse una accuratezza molto alta, dato lo sbilanciamento nei dati, si nota che anche l'F1 è abbastanza elevato. Nel grafico precedente è possibile vedere le prestazioni in termini di F1 score pesato al variare di K.

In termini di precision e recall, il sistema ha discrete prestazioni.

Nel grafico seguente invece si può osservare le prestazioni in termini di accuratezza. Ovviamente, essendo il dataset sbilanciato, ci si aspettava una elevata accuratezza. Si nota come i due grafici proposti siano molto simili, data la similarità nei valori delle due metriche.



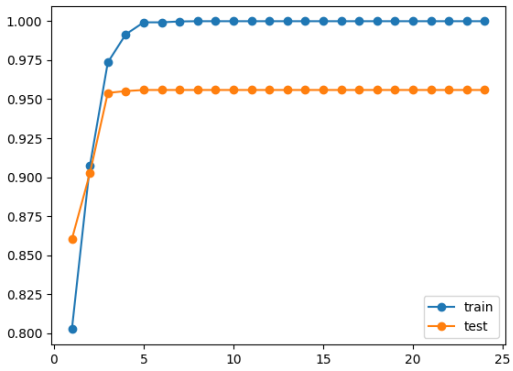
Di conseguenza, il modello KNN solo sulle feature geografiche si è rilevato essere abbastanza efficace, in base ai risultati ottenuti con F1 score pesato.

Alberi di Decisione

L'albero di decisione è stato usato principalmente per ritrovare le feature più importanti per il task di classificazione.

Per gli alberi si è deciso di misurarne le prestazioni al variare della profondità, con profondità che va da 1 a 25. Sono stati testati tutti i criteri disponibili, ma infine si è scelto il criterio 'entropy', in quanto riportava i risultati migliori.

Di seguito, sono riportati i valori per metriche di valutazione e i grafici risultanti dall'addestramento.



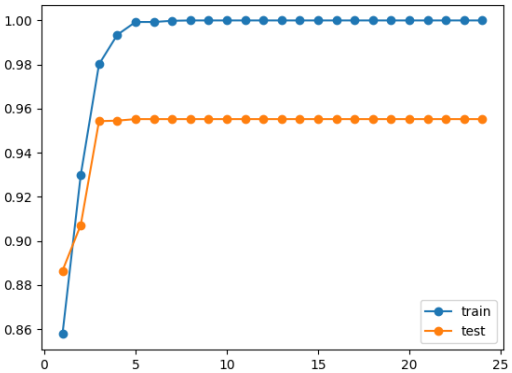

```
5---
0.9552447552447553 -- Accuracy
0.9559077699650049 -- F1
0.7970430495953317 -- Precision
0.8107294038796858 -- Recall
```

Dal grafico soprastante si possono osservare le performance dell'albero di decisione, al variare della profondità in base all'F1 score pesato.

Il sistema inoltre presenta ottime prestazioni in termini di precision e recall.

Si nota come già a profondità 5 si raggiunge uno score elevato, dopodiché non si riscontra alcun miglioramento nei dati di test, al variare dei dati di training. A profondità più elevate di 5 quindi il modello tende ad overfittare i dati.

Si riporta in seguito il grafico dell'accuratezza:



Anche nel grafico dell'accuratezza, il modello a profondità 5 raggiunge uno score elevato, dopodiché tende ad overfittare i dati.

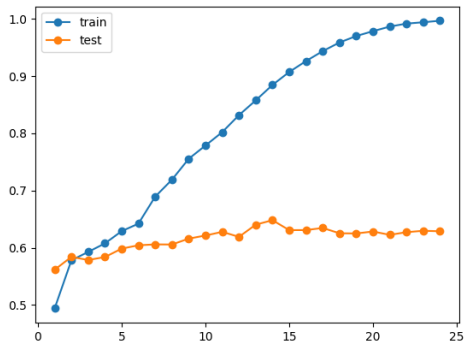
Per gli alberi si riportano le feature importance, ovvero l'importanza delle feature per gli alberi di decisione. Maggiore il valore di importanza, maggiore è l'importanza delle feature:

```
CO_AQI_Value:0.0009122586426664431
Ozone_AQI_Value:0.09949098478172637
NO2_AQI_Value:0.0
PM2_5_AQI_Value:0.8995967565756072
Lat:0.0
Lng:0.0
PM10:0.0
```

Da questi risultati, si nota come la feature PM2.5 abbia una rilevanza spropositata. Inoltre, latitudine e longitudine non hanno alcuna rilevanza, il che è in contrasto con i risultati ottenuti dal modello KNN.

Sulla base di questo dato, si è deciso di ispezionare nuovamente i dati, per capire come il PM2.5 potesse avere una rilevanza così elevata.

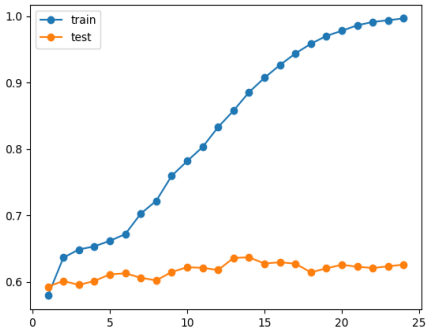
Come primo passo, si è provato a rimuovere la feature dal dataset, e di riaddestrare il modello senza di essa:




```
14---
0.6407925407925408 -- Accuracy
0.6481220624151978 -- F1
0.3935289557093879 -- Precision
0.5124536672740782 -- Recall
```

Dal grafico soprastante si possono osservare le performance dell'albero di decisione, al variare della profondità in base all'F1 score pesato, con la feature PM2.5 rimossa dal dataset.

Innanzitutto, si nota immediatamente come le performance siano di gran lunga inferiori. Inoltre, si nota come adesso l'albero raggiunga le massime prestazioni all'incirca ad una profondità pari a 14, dopodiché tende all'overfitting.



Dal grafico soprastante invece si osservano le performance in base all'accuratezza. Anche qui, l'accuratezza è calata di molto a seguito della rimozione della feature, e anche qui, intorno a profondità 13-14 l'albero raggiunge le massime prestazioni.

Di seguito si riportano le feature importances dopo la rimozione della feature PM2.5:

```
CO_AQI_Value:0.2171269391062218
Ozone_AQI_Value:0.18747482181590663
NO2_AQI_Value:0.072947604848822
Lat:0.235137766190862
Lng:0.20266453516351937
PM10:0.08464833287466816
```

Si nota come feature precedentemente quasi inutili, come per esempio latitudine e longitudine, ora hanno una importanza abbastanza elevata, per la precisione rispettivamente il 23.5% e il 20%.

Ne consegue quindi che il valore di PM2.5 "dominava". Si è allora deciso di ispezionare direttamente i dati contenuti nel dataset. Sono state quindi selezionate solo le colonne riguardanti il PM2.5 e l'AQI, e ne sono stati confrontati i valori. Si è notato che esiste quasi una corrispondenza uno a uno tra i valori di PM2.5 e AQI, che portava i modelli a utilizzare quasi esclusivamente la feature PM2.5 per effettuare le predizioni.

Nel modello KNN questo problema non si presenta, in quanto per l'addestramento vengono utilizzate solo le feature latitudine e longitudine.

In conclusione, l'albero di decisione, rimossa la feature PM2.5 risulta essere largamente inefficace.

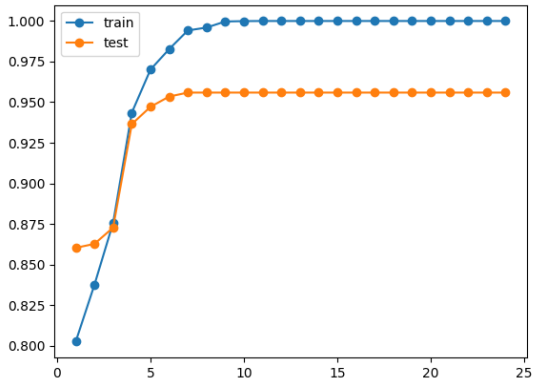
Anche per i successivi modelli si presenteranno sia i risultati con la feature PM2.5 e sia i risultati senza l'utilizzo di quest'ultimo.

Random Forest

Per il Random Forest si è utilizzato un approccio simile a quello per gli alberi di decisione.

Sono stati valutati Random Forest al variare della profondità degli alberi, usando il numero di alberi estimatori di default (100).

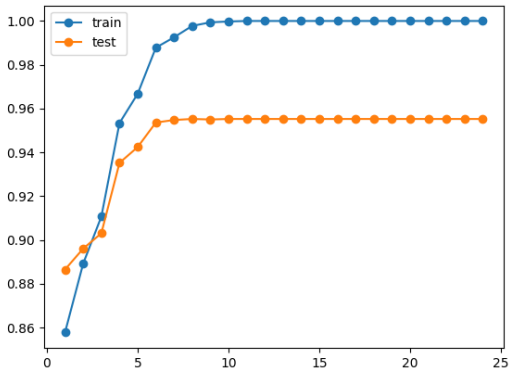
Si riportano di seguito i grafici per accuratezza ed F1 score pesato, sia con la feature PM2.5 sia senza.




```
7---
0.9547785547785548 -- Accuracy
0.9554521203514605 -- F1
0.7931968957491776 -- Precision
0.6996182927685748 -- Recall
```

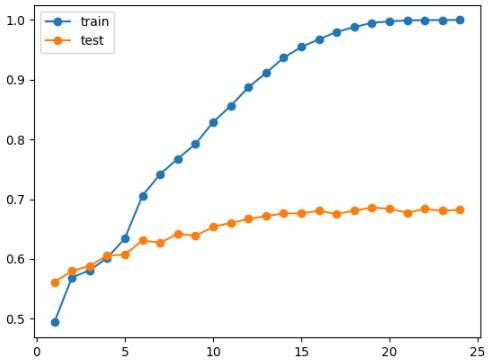
Nel grafico soprastante si osservano le performance in base all'F1 score pesato. Si nota che, come negli alberi, si raggiungono le massime prestazioni intorno a profondità 7, dopodiché tende ad overfittare i dati.

In termini di Precision e Recall invece osserviamo una Precision più elevata rispetto al Recall, ma comunque ottime prestazioni.



In questo grafico, invece, si osservano le prestazioni in termini di accuratezza. Notiamo come al solito una elevata accuratezza, derivata dallo sbilanciamento del dataset.

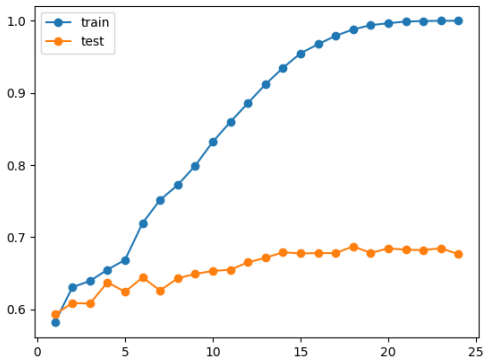
Di seguito si riportano i risultati del classificatore senza la feature PM2.5:




```
16---
0.6321678321678321 -- Accuracy
0.6221280395020368 -- F1
0.28299673917469964 -- Precision
0.35042478575856073 -- Recall
```

Il grafico rappresenta le performance in termini di F1 score pesato. Si può notare come le prestazioni siano indubbiamente inferiori, ma ora si ottengono prestazioni ottimali attorno a profondità 15.

La rimozione della feature PM2.5 comporta una elevatissima perdita di prestazioni in termini di Precision e Recall, come si può vedere dalle metriche.



La rimozione della feature PM2.5 ha lo stesso effetto sull'accuratezza del modello.

Si riportano anche i valori di importanza delle feature come per gli alberi, sia in presenza della feature PM2.5 sia in sua assenza.

```
CO_AQI_Value:0.1241985048648026
Ozone_AQI_Value:0.08703922637658999
NO2_AQI_Value:0.015032167119112663
PM2_5_AQI_Value:0.7022106571836809
Lat:0.022813460845753267
Lng:0.024321904966000551
PM10:0.024384078644055143
```

Anche qui si nota una rilevanza eccessiva del PM2.5, ma in maniera meno marcata. Le altre feature, a differenza degli alberi, hanno un peso seppur minimo.

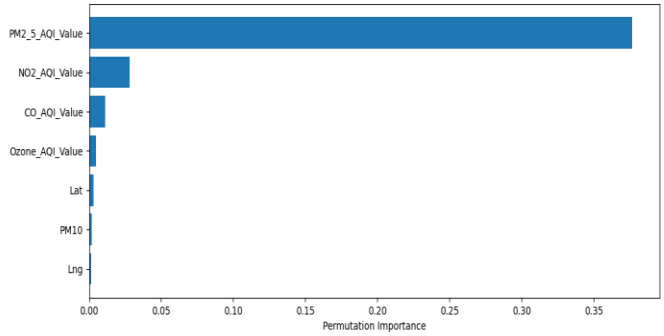
```
CO_AQI_Value:0.1791933675046857
Ozone_AQI_Value:0.18624876668236895
NO2_AQI_Value:0.0896373263531282
Lat:0.22891369554093688
Lng:0.20805136575718744
PM10:0.10795547816169294
```

Eliminando la feature PM2.5, come negli alberi di decisione, si nota che latitudine e longitudine passano da avere una bassa rilevanza ad essere le due feature più rilevanti.

In conclusione, il Random forest, una volta rimossa la feature PM2.5 risulta abbastanza inefficace. In presenza di essa, invece risulta essere molto efficace.

SVM

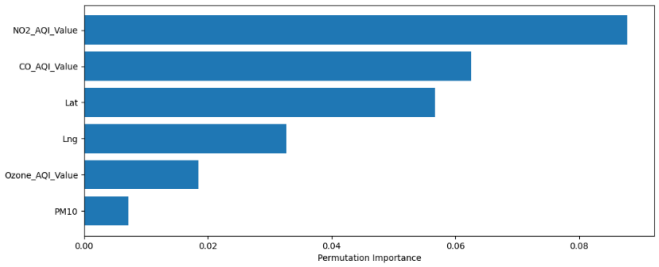
L'ultimo modello utilizzato per la classificazione è il classificatore SVM. Per quest'ultimo, sono stati provati tutti i kernel disponibili, infine è stato scelto quello con le performance migliori, ovvero il Kernel RBF (Radial Basis Function). Inoltre, è stato impostato il peso di ogni classe a "Balanced", maniera tale da pesare le classi in maniera inversamente proporzionale alla loro frequenza nei dati di input. Anche per l'SVM, si forniscono le importanze delle permutazioni effettuate dal classificatore in base alle feature di input.




```
SVM ----
0.8701631701631701 -- Accuracy
0.8713633265805227 -- F1
0.5138924533728909 -- Precision
0.5921676611734098 -- Recall
```

Addestrando il modello con la feature PM2.5 si osserva dal grafico in alto che, come negli altri classificatori, questa feature ha una elevata importanza, anche se in questo caso è molto più bassa rispetto agli altri classificatori.

In termini di prestazioni, il modello ha un ottimo F1 score, e discrete Precision e Recall.




```
---SVM
0.593006993006993 -- Accuracy
0.5912742014016991 -- F1
0.2899155673238128 -- Precision
0.298886682216628 -- Recall
```

Rimuovendo la feature PM2.5, in questo classificatore, la feature con la più elevata importanza nelle permutazioni diventa NO2, contrariamente ad alberi e random forest, dove quest'ultima rimane con una bassa rilevanza. La seconda invece è il CO, simile ad alberi e random forest.

In termini di prestazioni, la rimozione della feature PM2.5 comporta una perdita non irrilevante di prestazioni, rendendo il modello poco migliore di una classificazione casuale.

In conclusione, l'SVM in presenza della feature PM2.5 risulta essere un modello con ottime prestazioni. In assenza della feature PM2.5, invece, risulta essere largamente inefficace.

Conclusioni

In conclusione, il task di apprendimento supervisionato in presenza della feature PM2.5, risulta essere un task estremamente semplice data l'alta correlazione tra i valori dell'AQI e i valori del PM2.5, che tutti i modelli identificano. In presenza del PM2.5, per i modelli le altre feature risultano essere largamente irrilevanti.

In assenza della feature PM2.5, invece, il task di apprendimento supervisionato ha prodotto risultati mediocri o pessimi, in quanto le altre sostanze inquinanti hanno una bassa correlazione con i valori dell'AQI, poiché, data la loro pericolosità, raramente la loro concentrazione supera i limiti normativi vigenti in Europa. Infatti, in assenza della feature PM2.5 le feature più rilevanti risultano essere, quasi sempre, latitudine e longitudine.