LMU LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

INSTITUT FÜR INFORMATIK
LEHRSTUHL FÜR DATENBANKSYSTEME
UND DATA MINING

Bachelor Thesis
in Computer Science

Active Learning for Air Quality in Europe

Paola Këllezi

Aufgabensteller:	Prof. Dr. Thomas Seidl
Betreuer:	Philipp Jahn
Abgabedatum:	18.03.2025

Declaration of Authorship

I hereby declare that the thesis submitted is my own unaided work. All direct or indirect sources used are acknowledged as references.
This paper was not previously presented to another examination board and has not been published.

Munich, 18.03.2025

Paola Këllezi

Abstract

Air pollution poses a major health risk worldwide.[11][58] To improve air quality reporting and thereby reduce the associated health risks, the World Health Organization (WHO) and the European Union (EU) together with its affiliated European Environment Agency (EEA), have implemented air quality monitoring measures and policies for reporting units.[59][14] Despite these efforts, significant gaps in data collection and reporting persists, particularly in low- and middle-income countries.[31][25] This is concerning, given that 80% of the population exposed to high $PM_{2.5}$ levels live in low- and middle-income countries.[41]

In this thesis we explore the role of Active Learning (AL) as a resource-efficient machine learning approach for air quality monitoring.[56][52] Our goal is to investigate whether AL, compared to random sampling, can improve the predictive accuracy of a multilayer perceptron. The model is trained on validated air quality data from the EEA and complies with the EU air monitoring policies.[5] Additionally, the data is included for training only if at least 85% of the annual measurements are available.[4] Since $PM_{2.5}$ concentration is the most documented measure in air quality, we focus on this data in our analysis.[34] The model with AL exhibits a better learning curve than the baseline with random sampling. However, due to the strongly imbalanced dataset, the model becomes biased towards the most represented labels and requires further tuning to improve its prediction accuracy across all labels.

Contents

Chapter 1

Introduction

Air pollution poses a major global health hazard.[11] Based on a wide range of scientific research, experts determine pollutant concentration levels that can be considered safe for human health over different exposure periods.[11][59][30] The World Health Organization (WHO) and the European Environment Agency (EEA) both state that respiratory and cardiovascular diseases are the main health risks from air pollution.[33][6] These health risks are broadly defined and life threatening, with one in nine deaths worldwide being attributed to air pollution.[39]

Particulate matter with a diameter of less than 10 and 2.5 microns (PM_{10} and $PM_{2.5}$) is the most documented and used measure in evaluating the health effects of air pollution, mainly due to the small size and health risks related with it.[34] According to EEA, 96% of the EU's urban population is exposed to unhealthy levels of $PM_{2.5}$.[6]

Understanding and controlling air quality is crucial to mitigate the effects of air pollution.[60][33] Despite this, there are data inconsistencies and gaps in current monitoring practices, most notably in low and middle income countries. This hinders the control of air pollution and contributes to health inequalities between populations around the world.[31][25]

Active Learning (AL) proves to be resource and cost efficient for training models on large amounts of unlabeled data, given that an oracle (e.g., a human annotator) is available to label these instances upon request. Through query strategies, AL identifies and requests unlabeled instances which are expected to improve and provide the most information to the model, once they get labeled by an oracle and added back to the training data.[56][52] Considering this, our goal in this thesis is to explore the role of AL in improving air quality monitoring in Europe. For our analysis, we first train a baseline model using EEA validated air quality data and then apply two AL approaches:

1. **Badge** which uses gradient embeddings and k-means++ to select the

most informative and representative instances for labeling.[9]

2. **TypiClust** which we combine with constrained clustering (COP-KMeans). The aim is to ensure that the instances selected for labeling are representative of their clusters that contain domain knowledge through the constraints.[57][44]

To analyze the performance of AL, we refer to the model's learning curve and compare its predictive accuracy across different query strategies.

Chapter 2

Background

2.1 Air Pollution

Air pollution happens when chemical, physical or biological substances contaminate the air to an extent that it impairs the natural properties of the atmosphere.[33] Indoor air pollution is a result of people using harmful polluting fuels and technologies at and around the home. Outdoor air pollution comes from sources such as industrial emissions and vehicle exhaust.[34]

The WHO states that air pollution poses a significant threat to public health and ecosystems around the world. In addition, the EEA identifies air pollution as the biggest environmental health risk in Europe.[58][6] According to the WHO, the main contaminants in the air significantly impacting human health, include: $PM_{2.5}$/PM_{10}, Nitrogen Dioxide (NO_2), Ozone (O_3), Carbon Monoxide (CO) and Sulfur Dioxide (SO_2).[34]

2.2 $PM_{2.5}$ and Its Health Impacts

Particulate Matter ($PM_{2.5}$) is the most widely used measure in evaluating the health effects of air pollution. It refers to particulate matters of chemical, physical or biological substances with a diameter equal or less than 2.5 micrometers. $PM_{2.5}$ originates from coal and other fossil fuel burning in power plants, vehicle emissions, industrial processes, residential heating, and various other human and natural sources. Regardless of their origin and chemical composition, these particles are grouped together in air quality guidelines based on their small size and related health risks.[34]

Given their small size, $PM_{2.5}$ has the ability to penetrate deep into the lungs and even enter the bloodstream, jeopardizing health when exposed for long or even short periods of time. In 2020, according to the EEA, long-

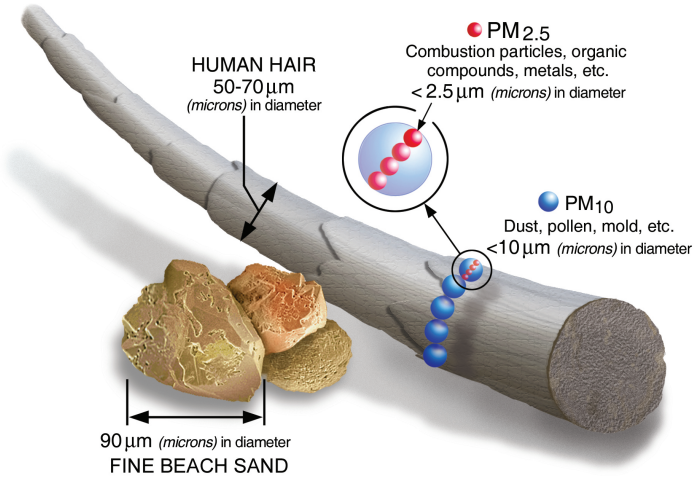

Figure 2.1: Size comparisons for PM particles.[8]

term exposure to $PM_{2.5}$ pollution was responsible for approximately 307,000 premature deaths in the European Union only.[2]

Short term exposure to high $PM_{2.5}$ levels can cause acute health effects, especially to people with pre-existing health conditions. These include irritation of the airways and increase of infection, heart attack, arrhythmia and even death risks.[28] In that regard, long term exposure has been associated with an increased risk of lung cancer, such as in 2013, $PM_{2.5}$ was classified as a cause of lung cancer by WHO's International Agency for Research on Cancer (IARC).[34]

Recent studies show that long term exposure is also being associated with an extended range of health risks that go beyond the known effects on the respiratory and cardiovascular systems. These include neurological disorders, renal dysfunction, gastrointestinal issues, reproductive health problems, metabolic disorders, immune system impairment, and potential escalation of diseases like COVID-19.[31][18]

2.2.1 WHO $PM_{2.5}$ Air Quality Targets

The WHO Global Air Quality Guidelines define targets for the reduction of PM2.5 pollution. Their annual $PM_{2.5}$ pollution targets and health hazard are defined as[59](P.78):

1. **Interim Target 1 (35 µg/m³):** Associated with significant health risks.

2. **Interim Target 2 (25 µg/m³):** Offers moderate reduction in health risks but remains above the recommended level for optimal health.

3. **Interim Target 3 (15 µg/m³):** Associated with further health improvements and reduced risks, particularly for cardiovascular and respiratory conditions.

4. **Interim Target 4 (10 µg/m³):** Represents an almost optimal level with reduced health risks.

5. **Air Quality Guideline (5 µg/m³):** The ideal target with the lowest risk of mortality and severe health effects.

2.3 Air Quality

Air quality refers to the air condition and is measured by the concentration of the harmful substances in it. To ensure safe air quality in global terms, the WHO guidelines set targets on six key air pollutants ($PM_{2.5}$, PM_{10}, NO_2, O_3, CO, SO_2).[60]

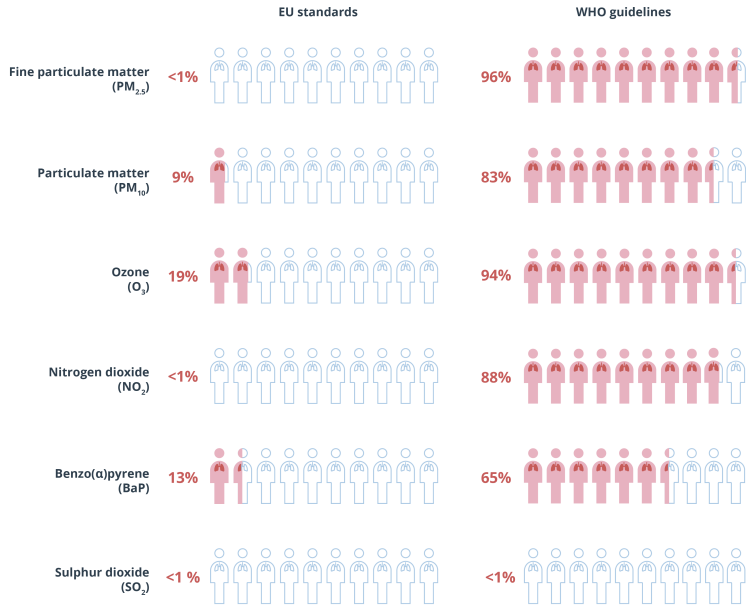

Figure 2.2: Percentage of the EU urban population exposed to air pollutant levels exceeding EU standards compared to WHO guidelines in 2022.[6]

Besides WHO, the European Union (EU), supported by the expertise of EEA, has also set its own air quality standards. They include the same six main pollutants as WHO and other substances. EU air quality standards are established for EU member states to work towards, by also keeping a balance between their environmental protection and economic development.[13] Moreover, the EU air quality standards are outlined in the Ambient Air Quality Directives, which makes them legally binding for all EU member states.[14]

In most cases, compared to the WHO guidelines, EEA defined higher tolerance levels for certain pollutants.[6] (Figure 2.2) From October 2024, EU Member States must implement Directive (EU) 2024/2881 to help reduce air pollution and support the EU's goal of achieving zero air pollution by 2050. The new air quality standards are further aligned with the WHO guidelines, including the reduction of annual limits for $PM_{2.5}$ by 2030.[13][38]

2.3.1 Challenges and Gaps in Monitoring

Air quality monitoring helps decision makers define and implement measures for a better environment and is therefore considered the key for improving air quality and public health. Currently, there is a global need for enhanced monitoring infrastructure and data transparency to effectively address air pollution challenges.[31][25]

The location of the city with the highest $PM_{2.5}$ levels can not be identified, since most countries have no $PM_{2.5}$ monitoring and the global mean population distance to a $PM_{2.5}$ monitor is 220 km by 2019.[25] The limited available data shows that countries with the strictest air quality guidelines generally have the lowest levels of particulate matter. In contrast, low- and middle-income

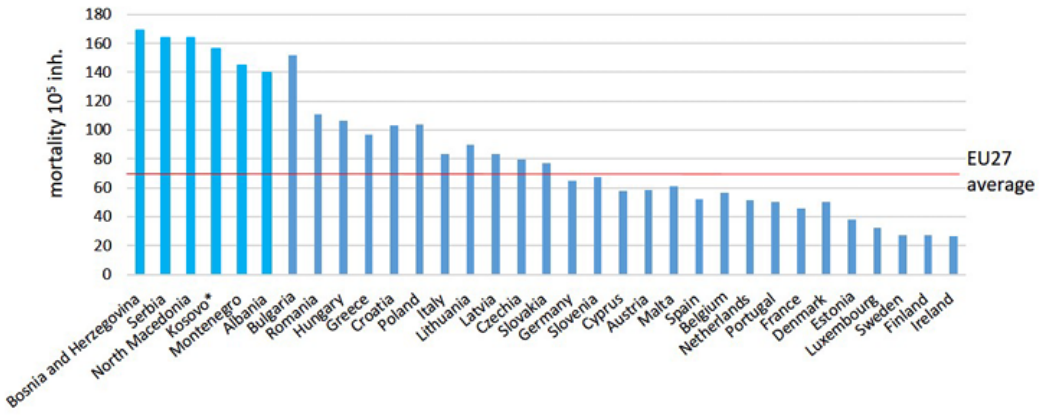

Figure 2.3: Annual mortality rates caused by $PM_{2.5}$ per 100,000 inhabitants.[12]

countries that significantly lack air quality monitoring are also most exposed to polluted air.[31][29]

The EEA's evaluations reveal that Balkan countries only meet the EU air quality monitoring standards to a limited extent.This is particularly concerning given the high levels of $PM_{2.5}$ pollution in the Western Balkans region.[12] (Figure: 2.3) For instance, $PM_{2.5}$ caused approximately 4,600 deaths in Albania alone in year 2021.[51]. Notably, Albania is also the only country being consistently grayed out in all European Union (EU) air quality publications.[1][3][6] This not only highlights a critical gap in data availability, but is also a hindrance to effective decision-making and policy integration of EU air quality frameworks.[51][16]

2.3.2 Monitoring Practices

Air quality measurement techniques include reference-grade monitors, low-cost sensors (LCS), remote sensing satellite instruments, passive diffusion samplers and personal direct reading instruments (PDRI).[32]

High cost reference systems typically provide highly accurate measurements of pollutants that rely on advanced calibration, hence, they are the most expensive technique to deploy. In contrast, LCS are more affordable and easier to operate, but their accuracy is often influenced by environmental conditions, which can make them unreliable. Remote sensing satellite instruments can assess "columns of air" and measure the total amount of particles from the ground to the upper atmosphere. However, these require complex models and high computational resources.

Passive diffusion samplers use chemical reagents and no electricity when collecting air quality data. Since they require careful analysis and validation from experts, they become impossible to implement for long-term or large-scaled monitoring. Likewise, PDRIs are not suitable for broader applications but rather for small-scaled applications.[32]

Advanced statistical approaches and machine learning enable researchers understand and predict air quality data together with their associated health risks. The paper "A review of machine learning for modeling air quality: Overlooked but important issues", highlights the following machine learning aspects that are widely used for air quality:[54]

- **Feature Engineering** which includes techniques such as treating outliers, missing values, feature normalization etc., as techniques to improve data quality and model accuracies.
- **Imbalanced data** can often lead to estimation biases, where models tend to predict the majority class. Techniques like resampling, sample

weighting are suggested to address this issue.

- **Validation Methods** including spatial and temporal strategies for evaluating models that can make reliable predictions on different points in time or geographic regions.
- **Regression Models** such as linear and multiple regressions that are used to quantify relationships between pollutant concentrations and health risks.
- **Random forests, Support Vector Machines and Neural Networks** are some of the machine learning algorithms used for classifying and predicting different pollution levels. In contrast to regression models, these algorithms also recognize non-linear patterns between multiple pollutant data.

2.4 Active Learning

In the field of machine learning (ML), the amount of unlabeled data grows exponentially, while obtaining high quality labeled data takes a lot of effort and resources. If human experts are essential in the labeling process, obtaining labeled data becomes even more expensive.[56][17]

Semi-Supervised Learning (SSL) is a domain in ML where a model is initially trained on a small amount of labeled data. Afterwards, it uses the distribution of the unlabeled data to further improve the model's performance. A typical SSL approach involves the model using its own predictions to label the data and retraining on them afterwards. The idea here is that the model can improve its prediction accuracy when trained on instances it is most confident about its predictions.[17]

Active Learning (AL) is also initially trained on a small amount of labeled data. However, AL proves to be efficient for training models on large amounts of unlabeled data, given that an oracle is available to label these instances upon request. Through query strategies, AL can identify and request the labels which are expected to improve the model, once they get labeled by an oracle and added back to the training data. By focusing on the most uncertain or diverse data points, AL models achieve comparable or even better performances than when trained on a fixed labeled dataset. In a nutshell, AL's goal is to train the model efficiently on queried instances, rather than on large, randomly labeled datasets. This makes AL advantageous in scenarios where enough unlabeled data is available, but the labeled data is scarce.[56][52]

The main difference between AL and traditional SSL lies in their assumptions. AL assumes that an oracle (e.g., a human annotator) is available to label

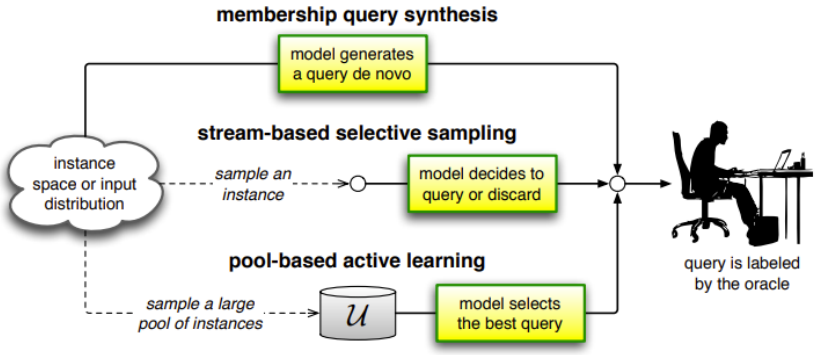

Figure 2.4: Three main AL scenarios.[52]

the data and SSL assumes the distribution and structure of the unlabeled data contains information that the model can effectively learn from.[56][17]

The surveys by Settles (2009) and Tharwat et al. (2023) outline three main scenarios for how active learning interacts with the data(Figure 2.4)[56][52]:

1. **Membership Query Synthesis** generates queries synthetically from scratch and does not select them from a real dataset. Since this approach does not include searching in a real dataset, it is considered a computationally efficient approach. However, this scenario can also generate unrealistic or impractical queries, especially when human experts are the oracle. For example in text data, membership query synthesis can make labeling challenging to the human, if unrealistic combinations of words that do not make linguistically sense are queried.

2. **Stream-Based Selective Sampling** queries the data points sequentially one after another. The learning algorithm evaluates each instance and depending on which query strategy, it decides whether to discard it or query a label for it. This scenario is computationally efficient and well suited for applications where processing power or memory space is scarce, such as mobile devices.

3. **Pool-Based Sampling** requires access to a large set or pool of unlabeled data. For the query, depending on the query strategy applied, the model evaluates all or a subset of the pool at once. This makes the approach computationally intensive but also highly effective for static datasets.

2.4.1 Query Strategies

In general, AL algorithms differ mainly in the query strategies they use to select the most informative or representative data samples for labeling.

These strategies are broadly grouped into[56]:

1. **Information-Based** query strategies calculate an informativeness score on each of the unlabeled instances. The higher the informativeness score, the greater the likelihood that labeling these instances would improve the model. The queries typically consist of a batch of selected instances in which the model is least certain about its predictions.

2. **Representation-Based** query strategies calculate a representativeness score on each unlabeled instance. This score quantifies how well an instance represents the overall structure or distribution of the unlabeled data. Querying instances with the highest representativeness scores helps AL train a stable and robust model in its early training stages.

3. **Meta-Active Learning** uses reinforcement or meta-learning techniques to initially learn a policy that determines the most effective query strategy. Based on the data and settings, it can dynamically select the best query strategy at different training stages and optimize model performance.

4. **Combined Approaches** integrate informative and representative-based query strategies to identify and query instances that are highly informative and representative in each batch.

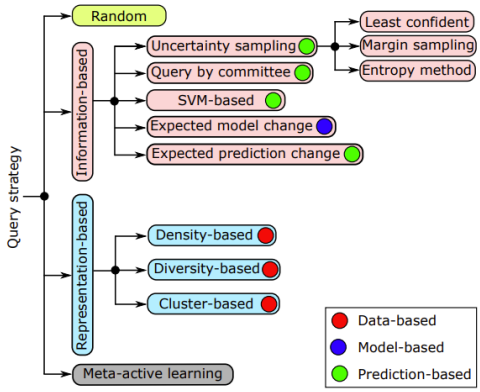

Figure 2.5: Taxonomy of query strategies for AL.[56]

2.4.2 Challenges

AL is highly dependent on its query strategy and can sometimes introduce redundant labels. Ash et al. (2020) mention in their paper on BADGE that implementing AL algorithm with information-based query strategies can result in the algorithm querying only instances the model is most uncertain about its predictions. Depending on the data and settings, the model may query very similar instances near its existent decision boundary, preventing the model learn from the broader data distribution and diverse instances. On the other hand, using representation-based query strategies may focus the algorithm query only diverse instances. Although these instances are as different from each other as possible, labeling them might not significantly improve the model if the model already predicts them correctly at earlier stages.[9] Moreover, methods that rely solely on information or representation in their query strategies, often fail to work consistently across different model architectures, batch sizes or datasets. For example, an AL approach might perform well on a residual neural network but not on a multilayer perceptron. Similarly, an approach can be ideal for large batch sizes but may under-perform on smaller batches.[9]

Hacohen et al. provide empirical evidence that querying strategies in deep AL should be selected depending on the budget and number of labeled examples available.(Figure 2.6) In low-budget regimes where the number of labeled examples is also small, the model is usually weak at the early training stages and its uncertainty estimates are not fully reliable. In such cases, deep AL models benefit more when using a representativeness-based query strategy. The model will select the most typical and diverse instances of the entire dataset, before its uncertainty even becomes reliable. In a high-budget regime, since available, the model is trained on a large amount of labeled data. Selecting uncertain and atypical instances with a information-based query strategy will potentially help the model learn new patterns and generalize better.[21]

AL also depends on the quality of the data it uses. For example, imbalanced data can make the model biased towards the most represented labels. As a result, it will query misleading instances and not perform as well as it could on a small and clean dataset. Moreover, it assumes the availability of initial labeled data, which is also not always the case in real life.[56] Hyperparameter tuning is also considered a challenge in the BADGE paper. The authors highlight that changing parameters can result in selecting entirely different instances for labeling under different configurations. This becomes an issue when labeling costs are already high, making AL a less efficient approach.[9] Besides, query budgeting should also be carefully considered for the cost efficiency.[56]

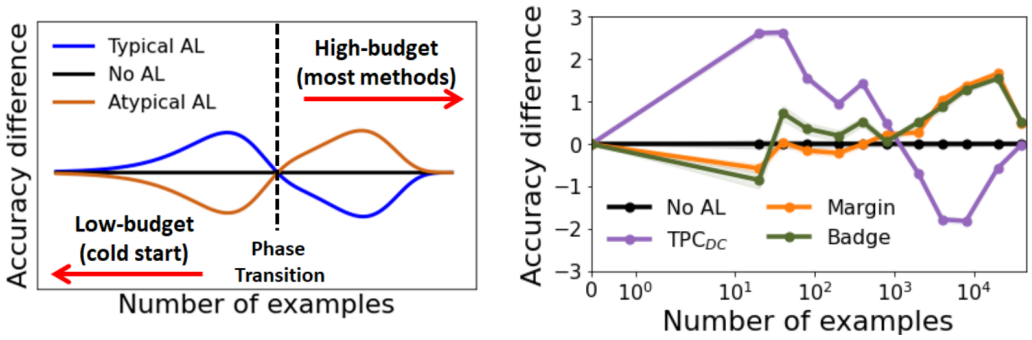

Figure 2.6: A comparison of AL query strategies, showing how accuracy changes with different budget and sample sizes.[21]

2.5 Clustering

Clustering is a widely applied technique and considered the most important question in unsupervised learning, where algorithms analyze unlabeled data to recognize patterns and structures, without any prior knowledge of the data.[61] [17] Clustering involves grouping a set of data points into subsets so that instances in the same cluster are more similar to each other than to those in other clusters. Additionally, instances in different clusters differ more from each other than from them in the same cluster.[23][61]

In most applications, clustering aims to group the data based on similarity or distance metrics.[23][61] Similarity and distance have an inverse relationship, meaning that the similarity between two instances typically increases, when their distance decreases and vice versa.[26] Similarity indicates the strength of the relationship between two instances and shows how similar they are with each other. In contrast, distance metrics show how far apart two instances are in the feature space and is often used to describe dissimilarity.[23][61]

When dealing with quantitative data, distance measures are often used to recognize relationships among instances. Examples of distance functions are Minkowski distance, Standardized Euclidean distance, Cosine distance, Pearson correlation distance and Mahalanobis distance. On the other hand, similarity measures are preferred more on qualitative data such as Jaccard similarity or Hamming similarity.[61] However, this distinction is not absolute and offers flexibility, as measures can be adapted or transformed depending on the context.[53]

Part of clustering algorithms is that they operate with different assumptions about the data. As a result, different clustering techniques often lead to different outputs on the same data.[23][61] K-Means, for example, assumes spherical clusters of similar size and variance, making it not practical for clus-

ters with more complex shapes. As shown in the scikit-learn examples, the K-Means algorithm leads to inaccurate results when applied to clusters with unequal variance or uneven sizes.[45][Figure 2.7][Figure 2.8]

2.5.1 Categories of Clustering Algorithms

The categorization of clustering algorithms provides insights into how they interact with various types of data and similarity measures. At a very basic level, traditional clustering algorithms are divided into three categories[23][61]:

1. **Partitional Clustering** main focus is to separate the input data into distinct spherical shaped clusters and typically requires a predefined number of clusters. Examples are K-Means, PAM, CLARA etc.

2. **Hierarchical Clustering** establishes clusters with a structure similar to a tree, using either a bottom-up or a top-down hierarchal approach, by adding the data points up or splitting them down accordingly. Examples are BIRCH, CURE, ROCK etc.

3. **Density-Based Clustering** groups data points that are dense and close together into the same cluster. Examples are DBSCAN, OPTICS etc.

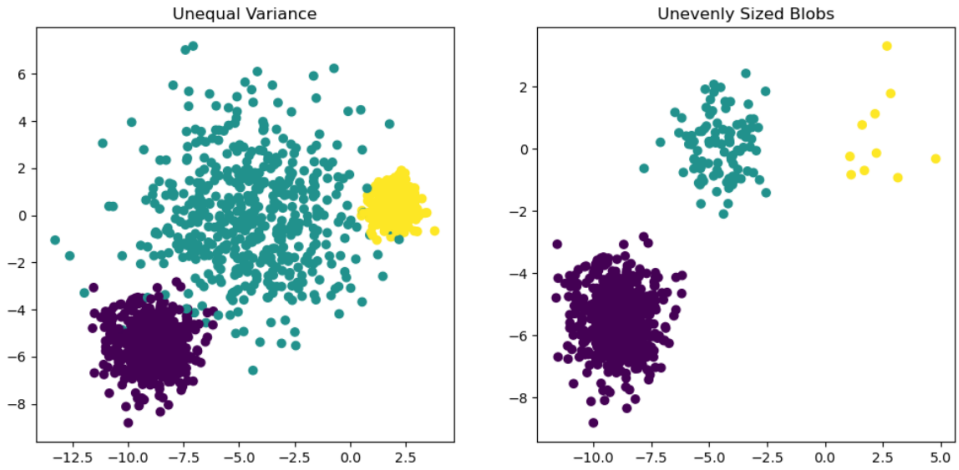

Figure 2.7: Ground truth clusters.[45]

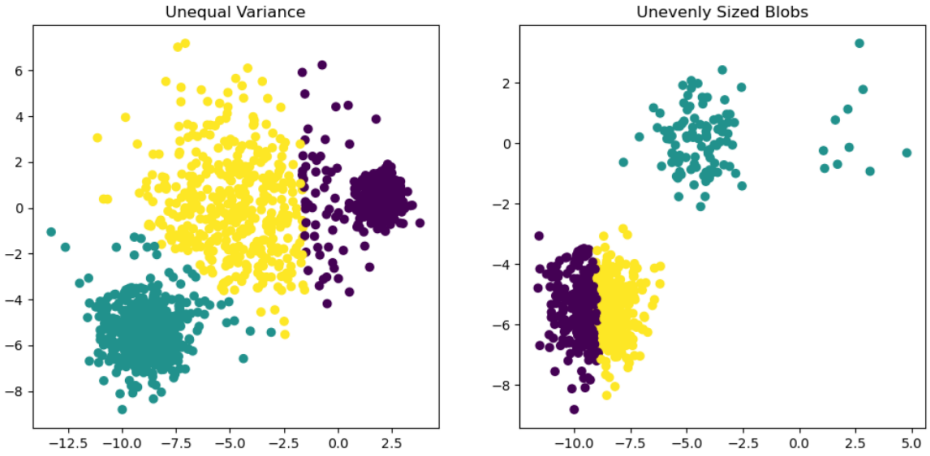

Figure 2.8: Unexpected KMeans clusters.[45]

Chapter 3

Related Work

3.1 Multilayer Perceptron

Multilayer Perceptron (MLP) is a neural network application composed of neurons distributed in 3 layers (input, hidden and output). Each neuron in the input layer corresponds to a feature in the training data. The hidden layer(s) consist of neurons that learn by optimizing their weights and biases using the back-propagation training algorithm. Initially, these weights are random. As data passes through the MLP, it makes predictions and updates the weights to minimize the prediction error. Moreover, the output layer of a MLP classfier contains the same amount of neurons as the target classes or labels.[40]

3.2 BADGE

Batch Active Learning by Diverse Gradient Embeddings (BADGE) is a method used in AL, designed for deep neural networks in a pool-based AL setup.[9] Its query strategy aims to combine informativeness and representativeness of the unlabeled instances, by utilizing gradient embeddings (vectors in the gradient space) and k-means++ clustering algorithm.

The model with BADGE is initially trained on a labeled dataset and then used to predict all labels of the unlabeled data pool. On each of these pseudo-labeled instances, the gradient embeddings are calculated to represent how much the parameters of the output layer in the deep neural network would change, given that the pseudo-label is also the true label. The lengths of these gradients capture how uncertain the model is about its predictions.

In order to select uncertain instances but also diverse, BADGE implements k-means++ initialization in its query strategy. The algorithm selects a "k"

number of gradient embeddings one by one. The first point is randomly selected and each new point is selected according to how far away it is from the already selected points. The further away, the higher the probability for a gradient embedding to be selected. As a result, the queried instances, chosen based on these diverse gradient embeddings, will provide information to the model by also representing the entire unlabeled dataset well.

3.3 Constrained K-Means

In unsupervised learning, traditional clustering algorithms perform without any prior knowledge, whereas constrained clustering incorporates domain knowledge when grouping the data, making it a SSL approach. An example of constrained clustering is COP-Kmeans, a variation of the standard k-means algorithm, that involves domain knowledge when setting rules on how the dataset should be clustered. These rules are typically called constraints, also defined as pairs of data points where: Must-link constraints specify that two points must be in the same cluster and cannot-link constraints ensure that two points are not assigned to the same cluster.[57]

A Python implementation of COP-KMeans is available in the GitHub repository by Behrouz Babaki.[10] It is a partitional clustering algorithm that uses the euclidean distance by default. Hence, it requires pre-defined constraints and the number of clusters k. Just like k-means, COP-Kmeans randomly initializes k cluster centers and assigns their nearest data points to each. Additionally, COP-Kmeans ensures that pairwise constraints are not violated during these assignments. In other words, if two data points are connected by a must-link constraint, they must be assigned to the same cluster, and if they are connected by a cannot-link constraint, they must be assigned to different clusters. If assigning a point to any of the clusters would violate a constraint, the algorithm leaves it out and does not assign it to any cluster. After the assignment process, the algorithm updates its cluster centers by calculating the average value of the points in each cluster and repeats the process until it converges or reaches the maximum number of iterations.

Wagstaff et al. (2001) highlight that COP-KMeans is sensitive to the order of the data points. This means that while the same constraints are respected, different data orders can lead to completely different clusters. Furthermore, the number of clusters is requested before-hand while the optimal number of k clusters is usually unknown in the real world. They also discuss on extending constrained clustering to hierarchical algorithms and soft constraints, so that not every single point is assigned strictly to one distinct cluster only.[57]

3.4 TypiClust

TypiClust is an AL query strategy designed to select the most typical or representative instances of a dataset. A model is initially trained using self-supervised learning or a set of labeled data. Then, it utilizes TypiClust, which uses a clustering algorithm to group the unlabeled data according to the task and the settings. Once the clustering is completed, TypiClust selects the most typical instances of each cluster by querying the most central instance of each cluster.

The authors of TypiClust tested the strategy on different deep learning tasks and found that it outperforms other AL query strategies in low-budget scenarios. Their results show that TypiClust models learn better in their early training stages, by selecting the most representative instances of the dataset, before the uncertain ones become reliable.[21]

Chapter 4

Contribution

4.1 Data Source

The data used in this thesis follows the EEA protocols and is sourced from the EEA's Table publisher. It is a publicly accessible tool with datasets on several air pollutants at different locations, environment types, activity levels and time intervals.[5]

Our focus is on 2018 annual mean $PM_{2.5}$ levels across 19 EU countries. The filters applied to ensure alignment with EU policies and improve data quality are[4]:

1. **E1a and AQ Report Data:** Ensures the dataset is based on validated data according to the European Commission standards.

2. **Data Coverage:** Includes annual statistics in the dataset, only if the data coverage is at least 85% over the year.

3. **Sampling Points (Data Flow C):** Ensures the data is collected from air quality monitoring stations that meet the EU standards of the air quality network.

4. **Single Sampling Point:** Ensures the data is collected from a single representative sensor at each site, preventing bias from multiple sensors and duplicate entries in the dataset.

5. **Hourly Observation Frequency:** Increases the dataset from ca. 500 in daily observation, to ca. 800 in hourly observations.

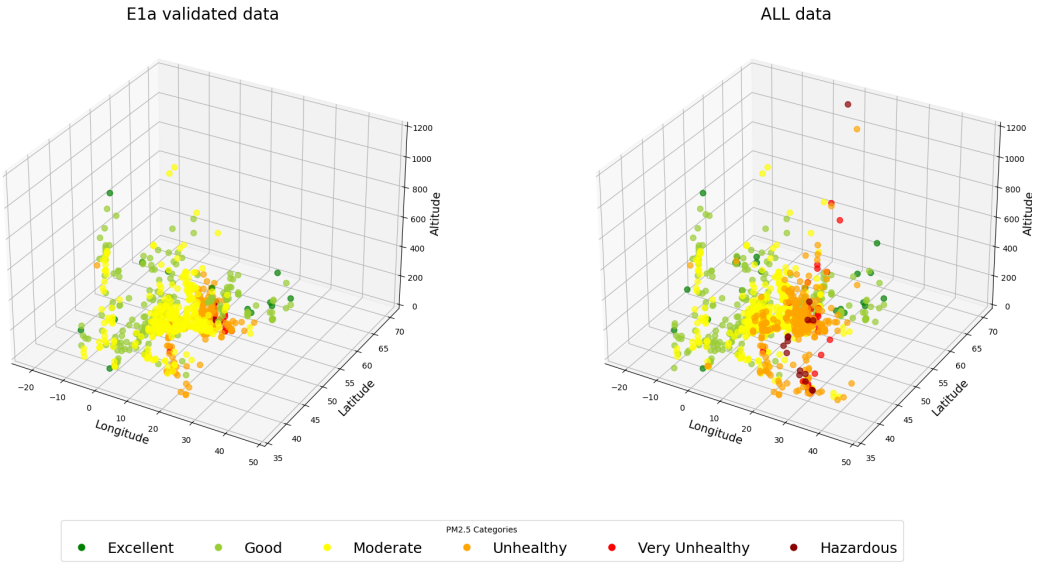

Figure 4.1: $PM_{2.5}$ Annual mean levels across Europe in 2018.

4.2 Data Preprocessing

The dataset is structured in a pandas `DataFrame` format.[35] Its rows represent annual mean records of single sampling points across EU, whereas the columns represent the features and target variable. To train a neural network model, we convert the dataset into arrays of real numbers with a mean of 0 and a standard deviation of 1, so that the model can understand and process it effectively.[47]

4.2.1 Handling Missing Values

377 missing entries are identified in the dataset, explicitly in the features "City" and "City Population". Replacing the missing values with zero for "City Population" do not show any improvement in model's prediction accuracy. Additionally, the correlation score with the target variable is 0.02 for "City Population" and -0.08 for "City" after frequency encoding.(Figure 4.2) Given the limited time and the effectiveness of AL on larger datasets, we remove these two features from the current analysis.[52]

4.2.2 Feature Selection

Following attributes with their respective ranges, relevant to the $PM_{2.5}$ levels, are selected for the analysis:

- **Categorical**:

1. Air Quality Station Area (Environment Type[7]: Urban, Suburban, Rural)
2. Air Quality Station Type (Activity Level[7]: Background, Industrial, Traffic)
3. Country (Austria, Belgium, Croatia, Czechia, Finland, France, Germany, Greece, Iceland, Italy, Lithuania, Netherlands, Norway, Poland, Portugal, Slovakia, Spain, Sweden, United Kingdom)
4. *City (285 cities)*

- **Numerical**:

1. Air Pollution Level (3.519 to 39.375)
2. Latitude (35 to 72),
3. Longitude (-25 to 60)
4. Altitude (-2.0 to 1225).
5. *City Population(48058.0 to 9845879.05)*

Note: The data is filtered to include only rows where latitude and longitude fall within specified bounds for EU.[27]

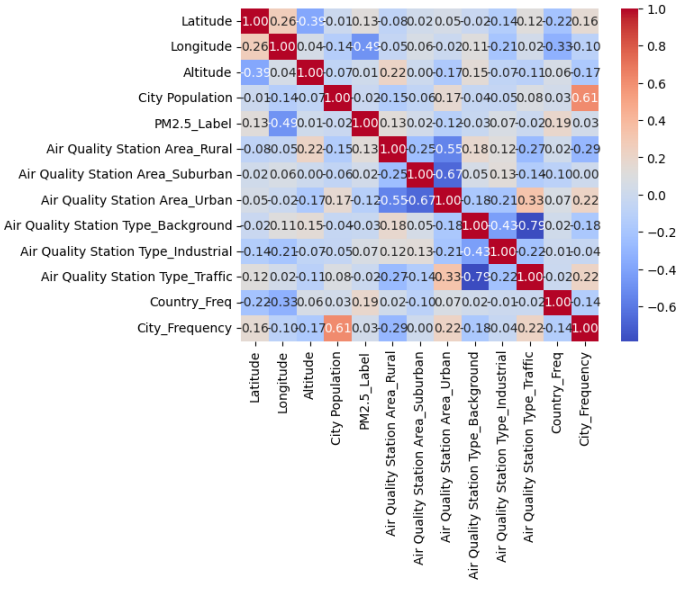

Figure 4.2: Pearson correlation matrix of the dataset features, calculated using the `pandas.DataFrame.corr()`.

4.2.3 Feature Transformation

Categorical features are encoded as numerical, since neural networks require vectors of real numbers as input to understand and learn effectively.[22]

- **Ordinal Encoding**

  We define "$PM_{2.5}$ Label" as our target variable. It represents the categorization of the "Air Pollution Level" into the intermediate targets from the WHO Air Quality Guidelines. (Section 2.2.1) We apply ordinal encoding, to ensure that the ordinal information is kept. This means, a numerical increase in the label corresponds to better air quality and lower health risks.[22] We use `OrdinalEncoder()` from the scikit-learn library.[48] Table 4.1 provides an overview of the $PM_{2.5}$ category labels and concentrations.

$PM_{2.5}$ Label	**$PM_{2.5}$ Category**	**Air Pollution Level ($\mu g/m^3$)**	**Count**
0	Hazardous	$PM_{2.5} > 35$	3
1	Very Unhealthy	$25 < PM_{2.5} \leq 35$	22
2	Unhealthy	$15 < PM_{2.5} \leq 25$	166
3	Moderate	$10 < PM_{2.5} \leq 15$	395
4	Good	$5 < PM_{2.5} \leq 10$	205
5	Excellent	$PM_{2.5} \leq 5$	14

Table 4.1: Ordinal encoding and distribution of $PM_{2.5}$ air pollution levels.

- **One-Hot Encoding**

  One-hot encoding is a basic encoding technique that does not assume any ordinal relationships between the categories. The categorical data is simply converted into vectors of 0's and 1's. Namely, each category is expressed as a vector of 0's and a single 1 corresponding to the category itself.[22]

  We apply one-hot encoding to "Air Quality Station Area" and "Air Quality Station Type" using the `get_dummies()` function from the pandas library.[37] The resulting encoded columns are illustrated in Table 4.2 and 4.3.

- **Frequeny Encoding**

  Frequency encoding tends to be suitable for processing categorical features including a large number of unique categories. The value assigned to each category is based on how often it occurs in the dataset.[19] We

apply this technique on the features "Country" and *"City"* using the `value_counts()` function from the pandas library.[36] (Table 4.4)

Air Quality Station Area	Rural	Urban	Suburban
Urban	0	1	0
Rural	1	0	0
Suburban	0	0	1
...	...	...	...

Table 4.2: One-hot encoding representation for "Air Quality Station Area".

Air Quality Station Type	Industrial	Background	Traffic
Industrial	1	0	0
Traffic	0	0	1
Background	0	1	0
...	...	...	...

Table 4.3: One-Hot encoding representation for "Air Quality Station Type".

Country	Country_Frequency
Belgium	58
Belgium	58
Finland	12
...	...

Table 4.4: Frequency encoding representation for "Country".

4.2.4 Data Skewness

The "$PM_{2.5}$ Label" distribution shows that label 3 ("Moderate") has the highest density of 395 instances, while label 0 ("Hazardous") is the most underrepresented with only 3 instances. (Table 4.1, Figure 4.1) Besides, the features "Air Quality Station" and "Air Quality Type" also do not have equally distributed categories, with their highest instance densities in the categories 'Urban' and 'Background'. (Figure 4.3)

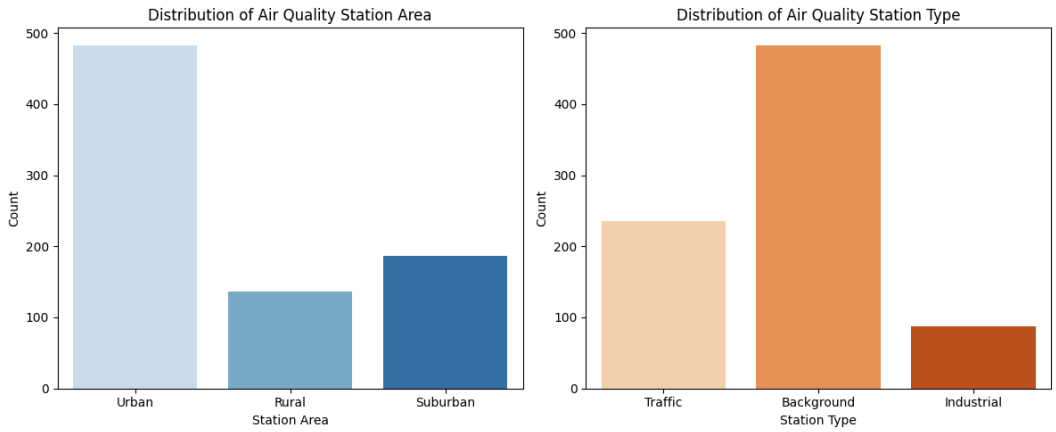

Figure 4.3: Distribution of "Air Quality Station Area" and "Air Quality Type".

4.2.5 Splitting Training and Test Data

The dataset contains 805 data points, where we split 80% (644 instances) for training and 20% (161 instances) for testing and evaluating the model. For this we use the `train_test_split()` function from the scikit-learn library.[50]

4.2.6 Data Scaling

The training data consists of 10 features with different scales and ranges, including both positive and negative values. This usually increases the difficulty of a model to learn effectively from the data.[15] Therefore, we standardize our training data to rescale the distribution of features such that the mean of all observed values is 0 and standard deviation is 1. To achieve this, we imported the `StandardScaler()` function from the scikit-learn library.[49]

4.3 Baseline Model

We configure a multilayer perceptron (`mlp`) as our baseline model using `MLP Classifier()` from scikit-learn.[47] After fitting it to the randomly shuffled training data, the model is configured with three neurons and one hidden layer, as this shows the best performance. We tested with 1 and 2 hidden layers, with up to 5 neurons per layer.

To evaluate the performance of the (`mlp`), we use k-fold cross-validation. This splits the training data into k-folds where the model is iteratively trained and validated in different subsets of each. The average accuracies of the test and validation sets are calculated and plotted in Figure 4.4, using the function `LearningCurveDisplay()` from scikit-learn.[46]

The model is also incrementally trained with its prediction accuracy plotted in Figure 4.5. In contrast to Figure 4.4, the model here is incrementally fitted on the training set and tested on the test set only. The red dashed line represents the prediction accuracy once the model is fitted on the full training data, particularly 68%.

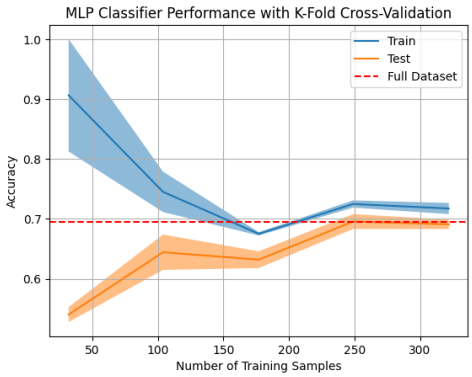

Figure 4.4: Learning curve for the MLP Classifier using k-fold cross-validation.(`layer=1; neuron=3; k=2; seed=42`)

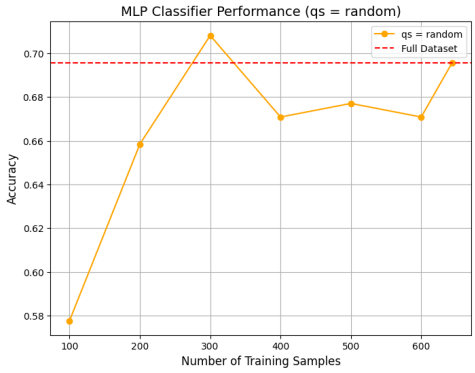

Figure 4.5: Learning curve for the MLP Classifier with random sampling.(`layer=1; neuron=3; seed=42`)

4.4 Active Learning

In our AL setup, we define following key parameters:

- `M`: number of labeled instances randomly selected to initially train the model.
- `X_target`: unlabeled data pool and the target number of labeled instances.
- `B`: batch size to be labeled in each AL iteration.
- `qs`: query strategy (assigned from scikit-activeml).[24]
- `cop_clusters`: number of clusters in `COPKmeans()`.
- `k`: number of clusters in `TypiClust()`.

At first, the `MLPClassifier()` is trained on `M` random instances. The remaining instances have their labels "removed" and set to `NaN` using `MISSING_LABEL` from skactiveml. The functions `labeled_indices()` and `unlabeled_indices()` are imported from skactiveml to track the instance indices depending on their label. Wrapping the MLP with `SklearnClassifier()` from skactiveml, enables the model to handle and differentiate from labeled and unlabeled instances abundantly during the AL.[43] As the AL iterates, based on the strategy `qs`, it queries batches of `B` instances for labeling. Since we do not have an oracle to provide the labels, we simulate this by using the original labels from the original dataset.[24] The AL iterates until `X_target` instances have been selected and labeled. In our case `X_target` corresponds to the original training set.

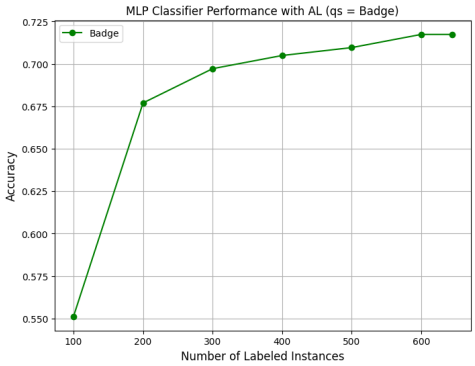

Figure 4.6: Learning curve for the MLP Classifier with Badge. (`M=100`; `B=100`; `X_target=644`; `seed=42`)

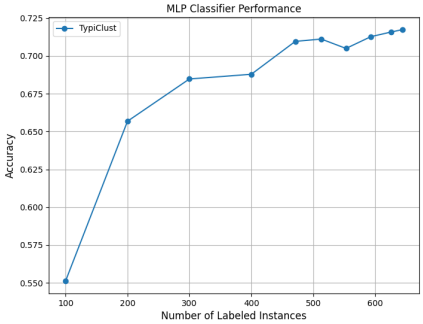

Figure 4.7: Learning curve for the MLP Classifier with COP-Kmeans and TypiClust. (`M=100`; `cop_clusters=5`; `k=1`; `B=20`; `X_target=644`; `seed=42`)

4.4.1 Badge Approach

We assign `qs=Badge()` from scikit-activeml.pool in our first AL approach.[42] After the MLP is initially trained on `M` randomly selected instances, the subsequent training iterations are performed with AL. The model receives the unlabeled data pool and calculates the informativeness score for each instance using gradient embeddings with respect to the MLP model's final layer parameters. With `k-means++` initialization, it aims to select representative instances of the unlabeled dataset, by querying instances based on their diverse gradient embeddings. The queried batches are labeled and then repeatedly added to the labeled training set for the next learning iteration, until the labeled dataset reaches `X_target`.

4.4.2 COP-KMeans and TypiClust Approach

Our second AL approach is a combination of semi-supervised learning with AL, where the data is first clustered with `COP-Kmeans()` from Behrouz Babaki and then queried for labeling using `TypiClust()` from scikit-activeml.pool.[10][44] This implementation of `TypiClust()` is designed to accept only clustering algorithms that are compatible with the scikit-activeml API.[44] Since `COP-Kmeans()` is not by default, we apply it separately on the dataset to first embed the domain knowledge and create the constrained clusters. Once the dataset is grouped into `cop_clusters` without violating any of the pre-defined constraints, we apply AL with `TypiClust()`.

`TypiClust()` uses `kmeans++` on each of the `cop_clusters`, to further group the instances into `k` "sub-clusters" and query `B` instances for labeling. Since `TypiClust()` queries instances nearest to cluster centers, we expect the queried instances to be the most representative of their respective `cop_clusters`. Moreover, because each AL iteration queries representative batches from all `cop_clusters` simultaneously, we also expect to gain informativeness through the domain knowledge embedded in the cluster constraints.

The `COP-Kmeans()` constraints are defined as pairs of instance indices that guide the algorithm to group instances together or not. Since the EEA distinguishes air quality stations by environment types and activity levels ("Air Quality Station Area" and "Air Quality Station Type"), we follow this categorization and define must-link and cannot-link constraints to capture pollution patterns across these different settings.[7]

- **Must-link:**
    1. Air quality stations in the same area are clustered together. (e.g. all urban stations together)
    2. Air quality stations of the same type are clustered together. (e.g. all traffic stations together)
- **Cannot-link:**
    1. Since their pollution sources and concentration levels differ significantly, urban and traffic stations should not be clustered with rural and background stations. (e.g. urban+traffic vs. rural+background)

4.4.2.1 Pseudo Code

This section offers a step by step overview of our AL approach with `COP-Kmeans()` from Behrouz Babaki and `TypiClust()` from scikit-activeml.pool[10][44]:

1. *Set parameters `M, B, cop_clusters`*
2. *Define `cannot_link` constraints as pairs of instance indices that should not be clustered together.*
3. *Define `must_link` constraints as pairs of instance indices that should be clustered together.*
4. *Apply `COP-Kmeans()` for grouping the dataset into `cop_clusters`.*
5. *Register the cluster labels and their corresponding instance indices in a dictionary.*
6. *Initialize `y` with instance labels set to `NaN`.*
7. *Assign the true labels to the `M` initially selected instance indices in `y`.*
8. *While the number of labeled instances is less than `X_target`:*
    - *Train the model on the available labeled data and record accuracy.*
    - *Query `B` unlabeled instance indices from `k` clusters by applying `TypiClust()` to each of the `cop_clusters`.*
    - *Assign the true labels to the queried instance indices in y.*

4.4.2.2 Learning Curve and Cluster Visualization

The `COP-Kmeans()` function from Behrouz Babaki does not contain a `random_state` parameter and will randomly initialize different cluster centers on each run. As a result, the clusters and learning curves with Typiclust variate on each run. (Figure 4.8)

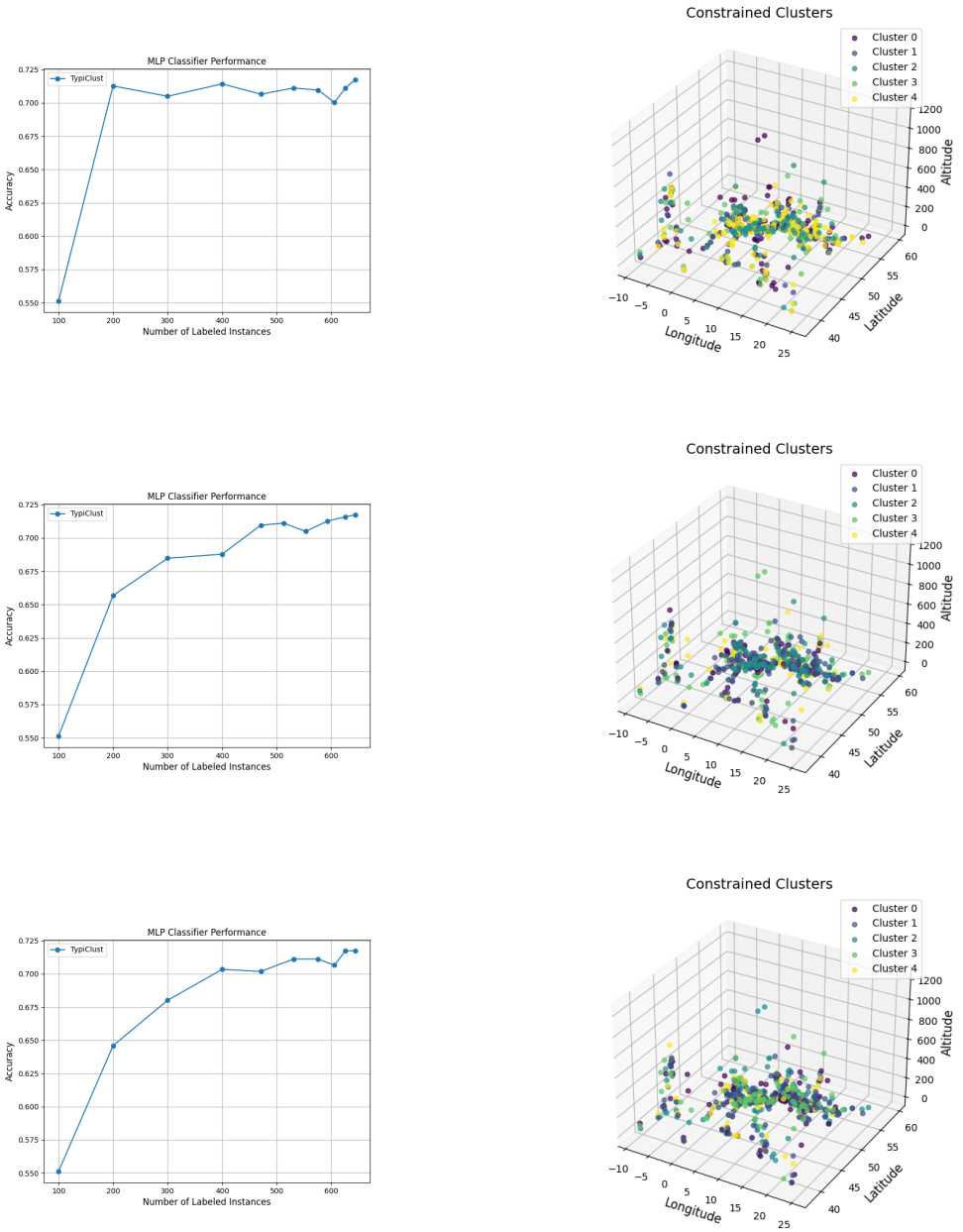

Figure 4.8: Learning curve and cluster visualization with COP-Kmeans. (`M=100`; `cop_clusters=5`; `k=1`; `B=20`; `X_target=644`; `seed=42`).

Chapter 5

Experiments

In this section we evaluate the `mlp` classifier's performance via iterative sampling and the query strategies `Random`, `Badge()` and `TypiClust()` with constrained clustering. Once the model finishes training on the entire training data, we compute the model's overall and label accuracy under different settings. Overall accuracy is computed as the percentage of correctly predicted instances across all labels whereas label accuracy is computed as the percentage of correctly predicted instances for each label. To achieve this, we run experiments using four random seeds (42, 123, 999, 1616) and compute the mean and standard deviation of the model's prediction accuracy values from these runs. In our experiments, the overall and label accuracy across all AL query strategies converge to the same value. Therefore, to reduce redundancy, we report results only for the baseline and AL.

We experiment with `cop_clusters` of sizes 7,5 and 3 and set the query size `"B"` to 100. If a cluster from `COP-Kmeans()` contains fewer instances than the query size, `"B"` dynamically adapts to the number of elements available within that cluster. `TypiClust()` from scikit-activeml operates with `k=5` clusters by default. To preserve the original clustering from `COP-Kmeans()`, we set `k=1`. Eventually, we experiment with `k=5` for `TypiClust()` and check whether this configuration leads to significant changes in accuracy.

Given that our dataset is highly imbalanced, we aim to determine if eliminating the underrepresented $PM_{2.5}$ labels (0: Hazardous, 1: Very Unhealthy, 5: Excellent) and focusing on the most represented labels (2: Unhealthy, 3: Moderate, 4: Good) leads to a more stable learning curve and improved prediction accuracy.

Furthermore, the training data features "Air Quality Station Area" and "Air Quality Station Type" are also imbalanced, with 'Background' type and 'Urban' area being the most dominant categories. After training our model on the 2018 validated dataset from the EEA, we set the model to query a batch

with `Badge()` from all available EEA data (E1a validated and unvalidated) from 2018 (Figure 4.1). We analyze the distribution of the real $PM_{2.5}$ labels of the queried instances, as well as the distribution by category of the features "Air Quality Station Area" and "Air Quality Station Type".

5.1 Comparison of the Baseline Model and Active Learning

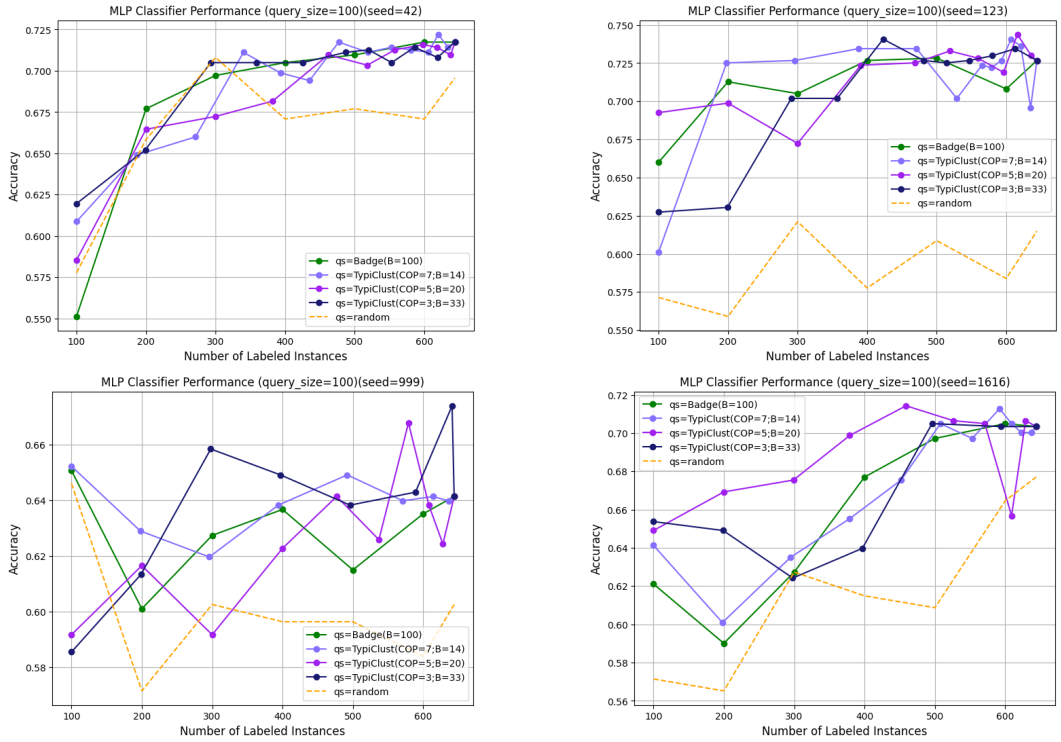

Figure 5.1: Learning curve with different query strategies across random seeds.(`M=100; X_target=644; k=1`)

	Baseline	AL
Label 0	0.0 ± 0.0	0.0 ± 0.0
Label 1	0.32 ± 0.32	0.0 ± 0.0
Label 2	0.79 ± 0.04	0.71 ± 0.04
Label 3	0.56 ± 0.21	0.83 ± 0.02
Label 4	0.31 ± 0.31	0.56 ± 0.09
Label 5	0.0 ± 0.0	0.0 ± 0.0
Overall	0.65 ± 0.4	0.70 ± 0.03

Table 5.1: Mean and standard deviation of model prediction accuracy for baseline and AL across different seeds.

5.1.1 TypiClust Cluster Size (k=5)

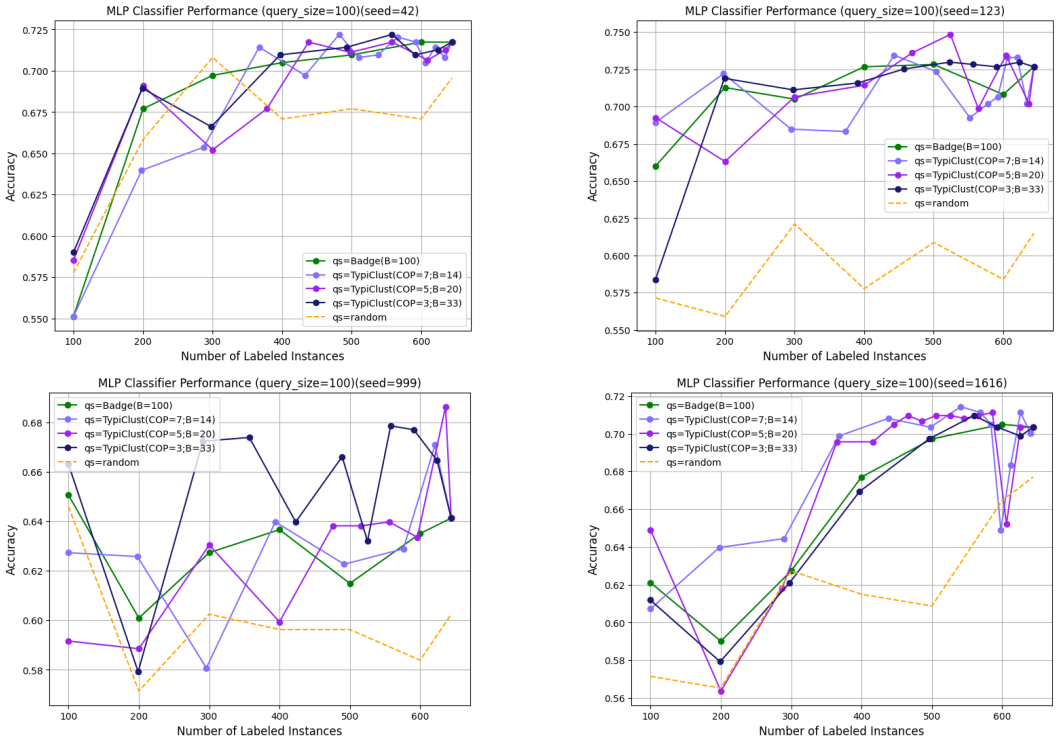

Figure 5.2: Learning curve with different query strategies across random seeds when using TypiClust with k=5. (`M=100; X_target=644; k=5`)

	Baseline	AL
Label 0	0.0 ± 0.0	0.0 ± 0.0
Label 1	0.32 ± 0.32	0.0 ± 0.0
Label 2	0.79 ± 0.04	0.71 ± 0.04
Label 3	0.56 ± 0.21	0.83 ± 0.02
Label 4	0.31 ± 0.31	0.56 ± 0.09
Label 5	0.0 ± 0.0	0.0 ± 0.0
Overall	0.65 ± 0.4	0.70 ± 0.03

Table 5.2: Mean and standard deviation of model prediction accuracy for baseline and AL across different seeds when using TypiClust with k=5.

5.1.2 Performance with Three Labels

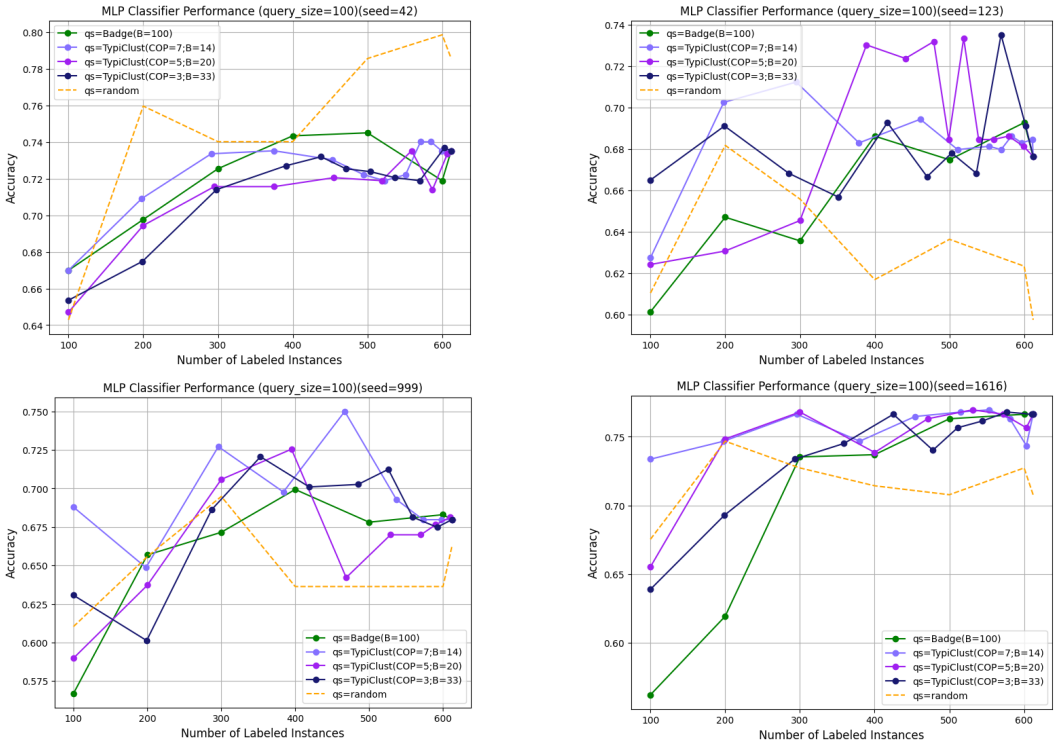

Figure 5.3: Learning curve with different query strategies across random seeds and a reduced label set. (`M=100; X_target=644; k=1`)

	Baseline	AL
Label 2	0.72 ± 0.08	0.71 ± 0.03
Label 3	0.80 ± 0.04	0.85 ± 0.03
Label 4	0.42 ± 0.18	0.47 ± 0.16
Overall	0.69 ± 0.07	0.71 ± 0.04

Table 5.3: Mean and standard deviation of model prediction accuracy for baseline and AL across different seeds and a reduced label set.

5.1.3 BADGE Query on All Data

- "All data" refers to the combination of E1a validated data and unvalidated data from 2018, as provided by the EEA. (4.1)
- The "Dataset" column refers to the total number of instances available for each label or category within the entire available dataset.
- Query size `B` = 100

$PM_{2.5}$ Label	**Description**	**Query**	**Dataset**
0	Hazardous	5	18
1	Very Unhealthy	4	45
2	Unhealthy	23	237
3	Moderate	37	457
4	Good	27	246
5	Excellent	4	27

Table 5.4: "$PM_{2.5}$ Label" query distribution on all data.

Air Quality Station Type	**Query**	**Dataset**
Background	52	633
Traffic	31	289
Industrial	17	108

Table 5.5: "Air Quality Station Type" query distribution on all data.

Air Quality Station Area	**Query**	**Dataset**
Urban	55	639
Suburban	27	227
Rural	18	164

Table 5.6: "Air Quality Station Area" query distribution on all data.

Air Quality Station (Type + Area)	Query	Dataset
Background + Urban	25	360
Background + Rural	13	132
Industrial + Urban	5	35
Industrial + Rural	4	30

Table 5.7: Query composition of two most represented (Background, Urban) and underrepresented (Industrial, Rural) categories.

5.2 Observations

The experiments show that multiple model initializations with different seeds demonstrate different learning curves, with `seed=42` displaying the least fluctuations and most stable learning curve. Likewise, across all seeds, AL with Badge() consistently demonstrates a steadier learning curve than other query strategies.

Throughout all experiments, AL maintains a smaller standard deviation in its predictions and achieves a slightly better overall prediction accuracy than the baseline model. This suggests that our AL model learns more efficiently by querying the most informative instances. In our case, this improvement is mainly due to AL performing better on well-represented $PM_{2.5}$ Labels (2, 3, 4), while struggling with underrepresented ones (0, 1, 5). The underrepresented labels are classified with an accuracy of 0.0% by AL, revealing that the model does not learn enough from the rare instances to successfully predict them. (Figure 5.1, Table 5.1)

When using `TypiClust()` with k=5 instead of k=1, we do not observe any changes in the prediction accuracy of the model. The learning curves with `TypiClust()` differ, but are also very similar and may be solely a result of randomness. Therefore, we cannot assume that setting k=5 in `TypiClust()` leads to a more diverse and representative selection of instances when working with more than one sub-cluster within the constrained clusters. (Figure 5.2, Table 5.2)

Reducing the $PM_{2.5}$ labels to 2, 3, and 4 improves the overall prediction accuracy of the baseline model from 65% to 69%, but shows almost no improvement with AL. Moreover, AL appears with a decrease in label 4 prediction accuracy, suggesting that the model learns most when all labels are included. (Table 5.3)

After assigning to `Badge()` the task of querying instances from the unlabeled pool of all available EEA data, 87% of the query's true labels were spread across the 3 most represented $PM_{2.5}$ labels, leaving only 13% to the other 3 underrepresented labels. (Table 5.4) Also 50% of the queried instances

belonged to the categories 'Urban' and 'Background', which were also the most frequent categories observed in the features "Air Quality Station Area" and "Air Quality Station Type" accordingly. (Table 5.5, Table 5.6) Moreover, 25% of the queried batch contained instances that had both categories 'Urban' and 'Background', while only 4% were selected from the least represented categories 'Rural' and 'Industrial'. (Table 5.7)

Chapter 6

Discussion

AL proves to be an effective ML approach when a large amount of unlabeled data is available, but labeling them is too difficult or demands too much resources.[56][55][52] However, as described in Section 2.4.2, AL comes with its own challenges. One major issue affecting AL, as well as deep learning in general, is the effective handling of imbalanced datasets.[55][20] This thesis serves as an analysis example facing this challenge, where a multilayer perceptron classifier learns by applying AL to the imbalanced EEA annual air quality data. After training, the model is completely unable to accurately predict three out of six $PM_{2.5}$ labels, which represent the $PM_{2.5}$ air pollution levels based on the WHO air quality guideline targets.

Even though BADGE has the most stable learning curve among all query strategies used in this thesis, the model remains biased towards the most represented labels. BADGE queries new instances from all EEA annual data for 2018, consisting mainly of instances belonging to the most represented labels and categories. This suggests that the model tends to reinforce existing biases rather than correct them.

If a model is trained mostly on instances belonging to the most represented labels, it learns fine-grained details about them. Consequently, the model becomes more uncertain about small changes in the labels on which it already learned a lot.[55] Since BADGE selects diverse instances the model is most uncertain about, the strategy tends to query instances of the most represented labels. This refines the model's prediction on these labels, but also reinforces bias and label imbalances in the queries.

Similarly to our AL approach with BADGE, constrained clustering with TypiClust improved the overall prediction accuracy, while still failing to correctly predict the underrepresented labels. The paper by Gilhuber et al. highlights examples of using SSL and AL on imbalanced data and how SSL can unintentionally hinder effective learning.[20] Motivated by their findings, we

hypothesize that a similar issue may be occurring in our case. Although we implement domain knowledge through the clustering process, it remains to be seen whether these constraints further increase the label imbalance by grouping minority labels into large clusters where the most represented labels dominate. This reduces the likelihood of minority labels representing their clusters and also being selected by TypiClust.

Real World Consequences

The inability of AL to correctly classify underrepresented PM2.5 labels has significant real-world implications. The model fails to identify extreme cases related with dangerously high PM2.5 concentrations, which in turn contributes to inequality in data monitoring.

Future Work

Our findings in this thesis demonstrate that, while AL enhances the model's learning curve and overall prediction accuracy, it becomes biased and prioritizes the most frequent labels in the dataset. To address this limitation, Tharwat et al. propose a novel AL approach [55] that focuses on identifying key regions in the data, to improve the representation of minority labels. It remains to be examined whether this approach is applicable and can reduce the biases in our AL framework.

Bibliography

[1] European Environment Agency. Air quality in europe 2021: Air quality status briefing 2021, 2021.

[2] European Environment Agency. Health impacts of air pollution in europe, 2021, 2021.

[3] European Environment Agency. Europe's air quality status 2023, 2023.

[4] European Environment Agency. Air quality annual statistics calculated by the eea - feature catalogue, 2024.

[5] European Environment Agency. Air quality e-reporting: Air quality annual statistics calculated by the eea, 2024.

[6] European Environment Agency. Europe's air quality status 2024, 2024.

[7] European Environment Agency. Monitoring station classifications and criteria for including them in the eea's assessment products, 2024.

[8] U.S. Environmental Protection Agency. Particulate matter (pm) basics, 2024.

[9] Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. *arXiv preprint arXiv:1906.03671*, 2020.

[10] Behrouz Babaki. Cop-kmeans version 1.5, July 2017.

[11] Hanna Boogaard, Katherine Walker, and Aaron J Cohen. Air pollution: the emergence of a major global health risk factor. *International Health*, 11(6):417–421, 10 2019.

[12] Belis C, Djatkov D, Lettieri T, Jones A, Wojda P, Banja M, Muntean M, Paunovic M, Niegowska MZ, Marinov D, Poznanović G, Pozzoli L,

Dobricic S, Zdruli P, and Vandyck T. Status of environment and climate in the western balkans. Scientific analysis or review KJ-NA-31077-EN-N (online),KJ-NA-31077-EN-C (print),KJ-NA-31-077-EN-E, Luxembourg (Luxembourg), 2022.

[13] European Commission. Air quality, 2024.

[14] European Commission. Eu air quality standards, 2024.

[15] Lucas B.V. de Amorim, George D.C. Cavalcanti, and Rafael M.O. Cruz. The choice of scaling technique matters for classification performance. *arXiv preprint arXiv:2212.12343*, 2022.

[16] Laureta Dibra. Air pollution related policies in albania and their implementation challenges, 2018.

[17] Jesper Engelen and Holger Hoos. A survey on semi-supervised learning. *Machine Learning*, 109, 02 2020.

[18] Amanda Garcia, Eduarda Santa-Helena, Anna De Falco, Joaquim de Paula Ribeiro, Adriana Gioda, and Carolina Rosa Gioda. Toxicological effects of fine particulate matter (pm2.5): Health risks and associated systemic injuries—systematic review. *Water, Air, & Soil Pollution*, 234(6):346, 2023.

[19] GeeksforGeeks. Categorical data encoding techniques in machine learning, 2025.

[20] Sandra Gilhuber, Rasmus Hvingelby, Mang Ling Ada Fok, and Thomas Seidl. How to overcome confirmation bias in semi-supervised image classification by active learning. *arXiv preprint arXiv:2308.08224v1*, 2023.

[21] Guy Hacohen, Avihu Dekel, and Daphna Weinshall. Active learning on a budget: Opposite strategies suit high and low budgets. *CoRR*, abs/2202.02794, 2022.

[22] John T. Hancock and Taghi M. Khoshgoftaar. Survey on categorical data for neural networks. *Journal of Big Data*, 7(1):28, 2020.

[23] Jasmine Irani, Nitin Pise, and Madhura Phatak. Clustering techniques and the similarity measures used in clustering: A survey. *International Journal of Computer Applications*, 134(7):9–14, 2016.

[24] Daniel Kottke, Marek Herde, Tuan Pham Minh, Alexander Benz, Pascal Mergard, Atal Roghman, Christoph Sandrock, and Bernhard Sick. scikit-activeml: A Library and Toolbox for Active Learning Algorithms. *Preprints*, 2021.

[25] Randall V. Martin, Michael Brauer, Aaron van Donkelaar, Gavin Shaddick, Urvashi Narain, and Sagnik Dey. No one knows which city has the highest concentration of fine particulate matter. *Atmospheric Environment: X*, 3:100040, 2019.

[26] Madhuri S. Mulekar and C. Scott Brown. *Distance and Similarity Measures*, pages 1–16. Springer New York, New York, NY, 2017.

[27] NCESC. What is the latitude and longitude bound of europe?, 2025.

[28] State of Global Air. Health effects of air pollution: Factsheet, 2020.

[29] State of Global Air. Particulate matter (pm) and health, 2024.

[30] World Health Organization. *Ambient air pollution: a global assessment of exposure and burden of disease*. World Health Organization, 2016.

[31] World Health Organization. Billions of people still breathe unhealthy air: new who data, 2022.

[32] World Health Organization. *Overview of methods to assess population exposure to ambient air pollution*. World Health Organization, Geneva, 2023.

[33] World Health Organization. Air pollution - world health organization (who), 2024.

[34] World Health Organization. Health impacts of air pollution: Types of pollutants, 2024.

[35] pandas. pd.dataframe, 2024.

[36] pandas. pandas.dataframe.value_counts, 2025.

[37] pandas. pandas.get_dummies, 2025.

[38] European Parliament and Council of the European Union. Directive (eu) 2024/2881 of the european parliament and of the council of 23 october 2024 on ambient air quality and cleaner air for europe (recast), 2024.

[39] UN Environment Programme. Air pollution note – data you need to know, 2023.

[40] Hassan Ramchoun, Mohammed Amine, Janati Idrissi, Youssef Ghanou, and Mohamed Ettaouil. Multilayer perceptron: Architecture optimization and training. *International Journal of Interactive Multimedia and Artificial Inteligence*, 4:26–30, 01 2016.

[41] J. Rentschler and N. Leonova. Global air pollution exposure and poverty. *Nature Communications*, 14:4432, 2023.

[42] scikit activeml. Badge, 2025.

[43] scikit activeml. Sklearnclassifier, 2025.

[44] scikit activeml. Typiclust, 2025.

[45] scikit learn. Demonstration of k-means assumptions, 2024.

[46] scikit learn. Learningcurvedisplay, 2025.

[47] scikit learn. Mlpclassifier, 2025.

[48] scikit learn. Ordinalencoder, 2025.

[49] scikit learn. Standardscaler, 2025.

[50] scikit learn. train_test_split, 2025.

[51] European External Action Service. Let's work together to address air pollution and save lives, 2024.

[52] Burr Settles. Active learning literature survey. Technical Report 1648, University of Wisconsin–Madison, 2009.

[53] Ali Seyed Shirkhorshidi, Saeed Aghabozorgi, and Teh Ying Wah. A comparison study on similarity and dissimilarity measures in clustering continuous data. *PLOS ONE*, 10:1–20, 12 2015.

[54] Die Tang, Yu Zhan, and Fumo Yang. A review of machine learning for modeling air quality: Overlooked but important issues. *Atmospheric Research*, 300:107261, 2024.

[55] Alaa Tharwat and Wolfram Schenck. Balancing exploration and exploitation: A novel active learner for imbalanced data. *Knowledge-Based Systems*, 210:106500, 2020.

[56] Alaa Tharwat and Wolfram Schenck. A survey on active learning: State-of-the-art, practical challenges and research directions. *Mathematics*, 11(4):820, 2023.

[57] Kiri Wagstaff, Claire Cardie, Seth Rogers, and Stefan Schrödl. Constrained k-means clustering with background knowledge. pages 577–584, 01 2001.

[58] World Health Organization. Climate impacts of air pollution, 2023.

[59] World Health Organization. Who global air quality guidelines, 2023.

[60] World Health Organization. Who global air quality guidelines - questions and answers, 2023.

[61] Dongkuan Xu and Yingjie Tian. A comprehensive survey of clustering algorithms. *Annals of Data Science*, 2(2):165–193, 2015.