This project embarks on a fascinating journey to uncover how various natural and human-related factors have influenced air pollution, specifically PM2.5 particle concentrations, across Europe over the past two decades. PM2.5 particles are microscopic pollutants in the air that can have serious health impacts, and understanding their sources and behavior is crucial for improving air quality. To achieve this, our research will employ modern machine learning techniques to analyze large amounts of data from 2003 to 2023. We will examine different factors, including weather conditions, land use types, emissions from various sources, and other environmental variables. By doing so, we aim to provide a comprehensive picture of how these elements have contributed to PM2.5 levels over time and across different locations on the continent. The main assumption of our study is that due to climate changes and the reduction of pollution levels in Europe over the years, the impact of certain natural and anthropogenic factors on air pollution may have changed.

One of the key tools we will use is a special method called Shapley value analysis (SHAP). This is an innovative approach in the field of interpretable machine learning. It can be compared to a detective that helps us to understand the role of each factor influencing the prediction of PM2.5 concentrations in the model. This method allows us to see which factors are most important in shaping pollution levels and how their influence changes over different time scales. For example, we might discover that some weather factors play a crucial role in winter, while others are more significant in summer, or that the role of these factors has changed over the years. Our research will utilize high-quality data from reputable sources, such as the Copernicus Atmosphere Monitoring Service (existing pollution models), ERA5 reanalysis (a detailed weather database), Corine Land Cover (land use database), WorldPop (population data), EMEP (emission data), and the Airbase air quality database. By integrating data from all these sources, we can create accurate models to predict PM2.5 levels.

To ensure our results are robust, we will compare different machine learning algorithms, including decision tree-based methods and neural networks. These algorithms represent different strategies or approaches to solving the problem. By comparing them, we can identify which ones provide the most reliable and consistent results for our study. Although this comparative analysis is not the main focus of our project, it will enhance the accuracy and credibility of our results. The core of our research lies in understanding the trends and patterns of the influence of various variables on PM2.5 pollution. We want to know how these patterns differ not only daily and seasonally but also over longer periods, such as decades. Additionally, we will examine how these trends appear in different parts of Europe, from bustling cities to quiet rural areas, also paying detailed attention to the climatic conditions in these regions.

By the end of this project, we expect to uncover hidden connections between climate changes, emission reductions, population growth, and air quality. The insights gained from this study will be valuable for policymakers, researchers, and the general public. By making our findings and collected data publicly accessible, we aim to support better policies and strategies to combat air pollution and protect public health across Europe. In summary, this project is an in-depth analysis of the complex network of factors influencing air pollution in Europe. Using advanced machine learning and high-quality data, we aim to reveal intricate patterns and trends for PM2.5 pollution over the past twenty years.